ORIGINAL ARTICLE

Derivation and Validation of Automated Electronic Search Strategies to Identify Pertinent Risk Factors for Postoperative Acute Lung Injury

ANAS ALSARA, MD; DAVID O. WARNER, MD; GUANGXI LI, MD; VITALY HERASEVICH, MD, PhD; OGNJEN GAJIC, MD; AND DARYL J. KOR, MD

OBJECTIVE: **To develop and validate time-efficient automated electronic search strategies for identifying preoperative risk factors for postoperative acute lung injury.**

PATIENTS AND METHODS: **This secondary analysis of a prospective cohort study included 249 patients undergoing high-risk surgery between November 1, 2005, and August 31, 2006. Two independent data-extraction strategies were compared. The first strategy used a manual review of medical records and the second a Web-based query-building tool. Web-based searches were derived and refined in a derivation cohort of 83 patients and subsequently validated in an independent cohort of 166 patients. Agreement between the 2 search strategies was assessed with percent agreement and Cohen κ statistics.**

RESULTS: **Cohen κ statistics ranged from 0.34 (95% confidence interval, 0.00-0.86) for amiodarone to 0.85 for cirrhosis (95% confidence interval, 0.57-1.00). Agreement between manual and automated electronic data extraction was almost complete for 3 variables (diabetes mellitus, cirrhosis, H_2-receptor antagonists), substantial for 3 (chronic obstructive pulmonary disease, proton pump inhibitors, statins), moderate for gastroesophageal reflux disease, and fair for 2 variables (restrictive lung disease and amiodarone). Automated electronic queries outperformed manual data collection in terms of sensitivities (median, 100% [range, 77%-100%] vs median, 87% [range, 0%-100%]). The specificities were uniformly high (≥96%) for both search strategies.**

CONCLUSION: **Automated electronic query building is an iterative process that ultimately results in accurate, highly efficient data extraction. These strategies may be useful for both clinicians and researchers when determining the risk of time-sensitive conditions such as postoperative acute lung injury.**

Mayo Clin Proc. 2011;86(5):382-388

ALI = acute lung injury; COPD = chronic obstructive pulmonary disease; DDQB = Data Discovery and Query Builder; DM = diabetes mellitus; EMR = electronic medical record; GERD = gastroesophageal reflux disease; MCLSS = Mayo Clinic Life Sciences System; PPI = proton pump inhibitor

Acute lung injury (ALI) is a devastating postoperative respiratory complication and a leading cause of postoperative respiratory failure,[1-3] with a mortality rate of up to 45% in certain surgical populations.[4,5] Moreover, treatment options are limited once the condition is fully established. Earlier identification of at-risk populations may allow the implementation of effective ALI prevention strategies. Recognizing that numerous baseline factors can modify a patient's response to illness or injury and the likelihood of developing ALI, we recently developed an ALI risk prediction model for mixed medical and surgical populations.[6,7] This score assigns points both for conditions that predispose patients to ALI (eg, shock, aspiration, sepsis, pancreatitis, pneumonia, high-risk surgery, high-risk trauma) and

ALI-modifying factors (eg, sex, excess alcohol use, obesity, chemotherapy, diabetes mellitus [DM], smoking) at the time of hospital admission. We have shown the cumulative score to be a reliable predictor of the risk of developing ALI during hospitalization.

A key remaining limitation to the early identification of patients at high risk of postoperative ALI is the inability to identify risk factors in a timely manner. Currently, most investigators retrieve patient data manually from the medical record and reintroduce the data into research databases. This process, which is time-consuming and inefficient, risks inaccuracies due to errors in data entry.[8,9] Its scalability is also limited. Adoption of the electronic medical record (EMR) provides new opportunities for automated disease surveillance strategies. In particular, the availability of high-throughput information technology solutions for extracting clinical data provides an opportunity to improve the efficiency and conduct of clinical and translational research.[10-12] However, electronic data extraction is not without limitations.[13-15] Validation of data accuracy and completeness is necessary because data can be missing or incorrect as a result of the data-entry process or of corruption during data transmission, storage, or retrieval.[11]

This investigation aimed to determine the accuracy of innovative, Web-based automated electronic data-extraction strategies for identifying risk factors for postoperative ALI. To achieve this objective, we compared these novel data-collection strategies with an independent manual data collection. We hypothesized that Web-based data-extraction strategies would identify pertinent risk factors for postoperative ALI with an accuracy comparable to or exceeding that achieved with manual data collection.

For editorial comment, see page 373

PATIENTS AND METHODS

After obtaining approval by the Mayo Clinic Institutional Review Board, we performed a secondary analysis of a prospective cohort study comparing 2 independent strategies for identifying pertinent preoperative risk factors for postoperative ALI.

Our study population was obtained from a previous prospective cohort evaluation investigating the association between intraoperative ventilatory parameters and the development of early (<5 days) postoperative ALI.[3] Briefly, patients were included if mechanically ventilated for more than 3 hours during general anesthesia for intermediate- and high-risk surgical procedures. Patients were excluded if they denied permission to use their health information for research, they were younger than 18 years, they had prevalent major risk factors for lung injury or respiratory failure, or they had previously required mechanical ventilation. In a nested case-control design, 83 cases and 166 matched control patients (a total of 249 study patients) were identified from the full cohort from November 1, 2005, through August 31, 2006.

ALI RISK FACTORS

We analyzed data on preoperative clinical variables with evidence suggesting an association with postoperative ALI. Variables were categorized into 2 groups: preoperative comorbid conditions and preoperative medications. Preoperative comorbid conditions included DM,[3,16-18] chronic obstructive pulmonary disease (COPD),[3,19] restrictive lung disease,[20-22] cirrhosis, and gastroesophageal reflux disease (GERD). Classes of preoperative medications included proton pump inhibitors (PPIs), H_2-receptor antagonists, statins,[23] chemotherapeutic agents (within 6 months of the surgical procedure),[22,24] and immunosuppressants (within 6 months of the surgical procedure).[25-27] Amiodarone was also included as a medication of interest.[28,29] We interrogated the medical record for all medications contained within each class. Systemic corticosteroids were included in the list of medications considered for immunosuppressive therapy. Inhaled corticosteroids were not included.

It is important to note that a large portion of the included study population was referred from other health care facilities. As a result, documentation of the criteria leading to the diagnosis of a pertinent risk factor was often absent. Moreover, the retrospective nature of this investigation prevented prospective acquisition of this information. In light of these limitations, formal diagnostic criteria for the comorbid conditions of interest were not required. Rather, the comorbid conditions (eg, DM, COPD, restrictive lung disease, cirrhosis, GERD) were considered present if a physician documented the diagnosis in the EMR before the surgical procedure. Medications (eg, PPIs, H_2-receptor antagonists, statins, amiodarone) were considered present if they were being administered at the time of hospital admission. Because previous publications have suggested an increased risk of ALI with administration of chemotherapeutic and immunosuppressive medications up to 6 months before surgery, we increased the evaluation period for determining the presence of these medications to this interval (6 months before surgery).[24]

DATA-EXTRACTION STRATEGIES

Manual Data Extraction. The EMR of all study patients was interrogated by 1 of 2 trained study coordinators. Each was instructed to use standard operating procedures for the extraction of clinical data from the medical record. If a variable of interest was not identified in the EMR, it was assumed to be absent or negative. Similarly, if the variable was mentioned in the negative form (eg, "patient has no history of…" or "patient denies…") or if it was listed in the patient's family history but was not specifically assigned to the study patient, it was again assumed to be absent or negative. The research coordinators responsible for manual data extraction were not involved in the automated electronic data-extraction process.

Automated Electronic Data-Extraction Strategies—Mayo Clinic Life Sciences System and Data Discovery and Query Builder. Mayo Clinic has established a partnership with IBM to collaboratively develop a sophisticated data warehouse (Mayo Clinic Life Sciences System [MCLSS]), which contains a near real-time normalized replicate of Mayo Clinic's EMR. This warehouse is developed from multiple original clinical data sources, including highly annotated, full-text clinical notes, laboratory tests, diagnostic findings, demographics, and related clinical data from the year 2000 onward. Mayo Clinic's EMR data are extracted, transformed, and loaded into MCLSS using IBM's WebSphere Commerce Analyzer, creating DB/2 Universal Database structures of Mayo Clinic's normalized clinical data. Clinical patient data are mapped to standard medical terminologies using LexGrid (Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN) natural language processing technology.

The MCLSS also provides approved users with a query-building tool called the *Data Discovery and Query Builder* (DDQB). The DDQB is a Web-based application configured for query building that is intended to help physicians and researchers interrogate data files contained in the MCLSS. The DDQB allows users to identify administrative, demographic, laboratory, and diagnostic data of interest within the EMR. At the center of the DDQB software is the data abstraction model, an XML-based component that maps

end-user content expertise to physical representations. As the data within the MCLSS change, the data-abstraction model is modified while maintaining the validity of previously generated queries.

The DDQB allows users to build complex queries using Boolean logic, groupings, comparison operators, and relational joins without requiring programming knowledge. For new users, a 1-hour orientation is provided before use. The DDQB can then be used to query administrative data such as diagnostic and procedure-related codes. Complex query construction and free text searches of the EMR are also available to approved users. Because we wanted to develop near real-time data-collection strategies, we elected to pursue free text data searches rather than searching administrative codes such as *International Classification of Diseases, Ninth Revision* diagnostic codes; the latter are not applied in a time-efficient manner, and concerns have been raised about their accuracy in identifying comorbid conditions of interest.[14,30-32] Free text data queries were also used when interrogating the EMR for medications of interest. All MCLSS/DDQB searches were run by 1 of 2 physician investigators (A.A. or D.J.K.). Neither of these investigators participated in the manual extraction of clinical data.

For the automated extraction of medical comorbid conditions, MCLSS/DDQB was used to interrogate the EMR of each study patient during the 5-year period preceding surgery. Queries were restricted to the following sections of the clinical notes: history of the present illness, past medical/surgical history, and diagnosis. To optimize sensitivity, each query was designed to identify both the disease of interest and the common synonyms, acronyms, and abbreviations used to represent the disease. To improve specificity, we excluded negative forms of the condition of interest (eg, "patient *does not have* a history of…" or "there is *no* history of…"). We also excluded diagnoses referring to a family history of the condition. Negative forms such as *denies*, *doesn't*, and *no* were excluded as well. Query building was an iterative process in a derivation cohort of 83 study patients (ALI cases). Searches were performed and the results analyzed, and, on the basis of these results, more key terms, synonyms, acronyms, and abbreviations were added. Once finalized, the queries were validated in a second cohort of 166 patients (controls). A representative query is provided in the Figure. For the automated extraction of patient medications, free text searches were limited to the "current medication" section of the clinical notes. We limited the evaluation interval for medications of interest to the 3 months before surgery, extending it to 6 months for chemotherapy and immunosuppressive therapy. Clinical notes were interrogated for generic and trade names of all formulations within each specific pharma-
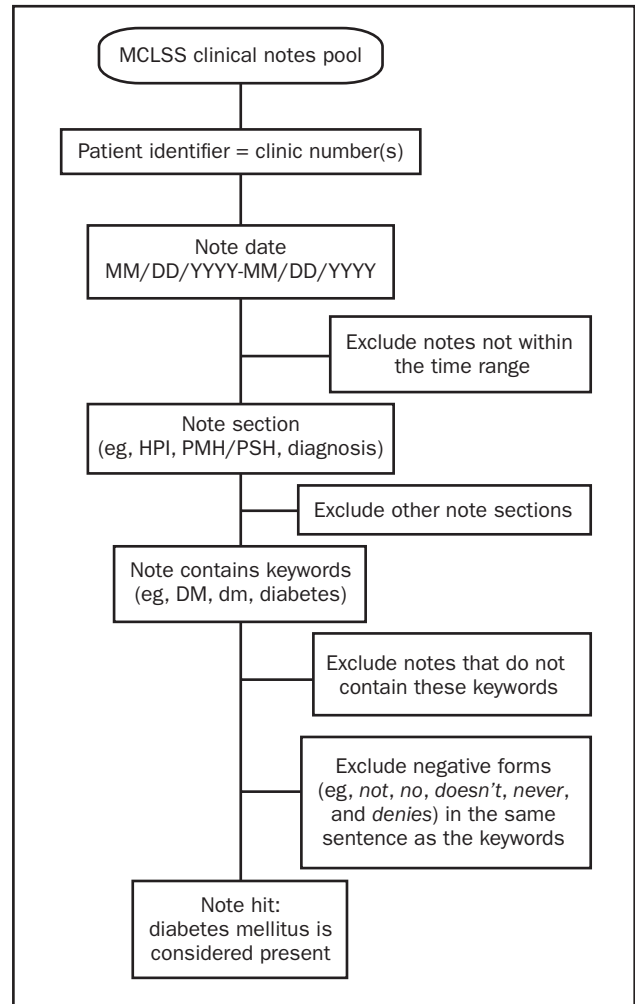

FIGURE. Electronic query for identifying a diagnosis of diabetes mellitus (DM) in the study population. HPI = history of present illness; MCLSS = Mayo Clinic Life Sciences System; PMH/PSH = past medical and surgical history.

ceutical class. A representative medication search query is shown in Table 1.

VALIDATION OF THE MANUAL AND ELECTRONIC DATA-EXTRACTION TECHNIQUES

After all data had been extracted, concordance between the 2 data-extraction techniques was assessed. All discordant and concordant negative results were investigated with an exhaustive manual review of the medical record (A.A. and D.J.K.). The specific procedures for this secondary review were dependent on the initial data-extraction result. All positive findings resulting from an automated Web-based query provide a direct link to the EMR identifying the exact location of the variable of interest in the patient's medical record. For all discordant results, where the condition was

TABLE 1. Electronic Query for Identifying the Presence of Statin Therapy at the Time of Hospital Admission in the Study Population

Section of clinical notes	Search terms	Additional criteria	Patient identifier
Current medications	rosuvastatin OR crestor OR pravachol OR pravastatin OR livalo OR pitavastatin OR altoprev OR mevacor OR lovastatin OR lescol OR fluvastatin OR zocor OR simvastatin OR lipitor OR atorvastatin	Note date on clinical notes is between MM/DD/YYYY and MM/DD/YYYY	Clinic number(s) (equals one of)

identified as being present with a Web-based query, this link was used to evaluate the accuracy of the initial positive result. For discordant results with a positive finding during manual review of the EMR, a detailed secondary review was performed to ensure its validity. For concordant negative results, 2 physician investigators (A.A. and D.J.K.) extensively reevaluated the EMR to confirm the absence of the variable of interest. In this final exhaustive review, interrogation was not limited to specific sections of the clinical notes (eg, "past medical history" or "active medications") but rather included all components. Concordant positive findings were considered true-positive results and were not further evaluated. Comorbid conditions (eg, DM, COPD, restrictive lung disease, cirrhosis, GERD) were considered present if documented in the EMR before the surgical procedure. Medications (eg, PPIs, H_2-receptor antagonists, statins, amiodarone) were considered present if they were being administered at the time of hospital admission. Medications identified as being present by either search strategy but determined during the final review of the EMR to have been discontinued before the surgical procedure were considered false-positive results. Chemotherapy and immunosuppressive therapy were considered present if administered at any time in the 6 months before the surgical procedure.

STATISTICAL ANALYSES

Because a true criterion standard for determining the presence of comorbid conditions and medications of interest is lacking, we report percent agreement and Cohen κ statistics as our primary statistical analysis comparing manual and automated electronic data-extraction procedures. As a secondary analysis, we have determined the sensitivity and specificity of both manual data extraction and automated electronic data extraction for each variable of interest. The final exhaustive review of the medical record was used as the "criterion standard" when determining the sensitivities and specificities of the 2 search strategies.

RESULTS

The prevalence of medical comorbid conditions and administration of medications of interest (as determined by the final review of the EMR) ranged from 1% for immu-

nosuppressive therapy to 50% for statin use. The percent agreement and associated Cohen κ statistics for each variable can be seen in Table 2. Due to a very low prevalence of preoperative chemotherapy and immunosuppressive therapy and the poor sensitivity for identifying these variables with manual review of the EMR, Cohen κ statistics are not reported for these variables. Of the comorbid conditions, DM and cirrhosis showed the greatest agreement and amiodarone therapy the least. Using the Cohen κ statistic magnitude guidelines described by Landis and Koch,[33] we determined that the agreement between manual and automated electronic data-extraction techniques was almost complete for 3 variables (DM, cirrhosis, H_2-receptor antagonists), substantial for 3 variables (COPD, PPIs, statins), moderate for a single variable (GERD), and fair for 2 variables (restrictive lung disease and amiodarone).

The rates of concordance and discordance are presented in Table 3. The sensitivity and specificity of each data-extraction strategy for all variables considered are presented in Table 4. Sensitivities for identifying variables of interest with manual data extraction ranged from a low of 0% for the infrequent variables chemotherapy and immunosuppressive therapy to a high of 100% for restrictive lung disease and cirrhosis (median sensitivity, 87%). Sensitivities for identifying variables of interest with MCLSS/

TABLE 2. Percent Agreement and Cohen κ Statistics Comparing Manual vs Automated Electronic Data-Extraction Strategies in the Validation Cohort (N=166)

Predictor	Percent agreement	Cohen κ statistic (95% CI)
Comorbid conditions		
Cirrhosis	99	0.85 (0.57-1.00)
COPD	95	0.79 (0.63-0.94)
Diabetes mellitus	94	0.81 (0.69-0.92)
Gastroesophageal reflux disease	88	0.53 (0.33-0.73)
Restrictive lung disease	99	0.40 (0.01-0.80)
Medications		
Amiodarone	92	0.34 (0.00-0.68)
Chemotherapy	97	NA
H_2-receptor antagonists	98	0.83 (0.64-1.00)
Immunosuppressive therapy	99	NA
Proton pump inhibitors	92	0.80 (0.69-0.90)
Statins	86	0.72 (0.62-0.83)

CI = confidence interval; COPD = chronic obstructive pulmonary disease; NA = not available.

TABLE 3. **Prevalence of Risk Factors for Postoperative Acute Lung Injury and the Concordance Between Manual and Automated Electronic Data-Extraction Strategies in the Validation Cohort (N=166)**[a]

Predictor	Prevalence (%)[b]	Concordant results		Discordant results	
		CP	CN	D(MP)	D(EP)
Comorbid conditions					
Cirrhosis	3 (2)	3	162	1	0
COPD	21 (13)	15	144	1	6
Diabetes mellitus	33 (20)	27	129	5	5
Gastroesophageal reflux disease	31 (19)	14	133	12	7
Restrictive lung disease	3 (2)	3	155	3	5
Medications					
Amiodarone	9 (5)	4	149	5	8
Chemotherapy	5 (3)	0	161	0	5
H$_2$-receptor antagonists	10 (5)	8	155	2	1
Immunosuppressive therapy	1 (1)	0	164	0	2
Proton pump inhibitors	43 (26)	37	116	9	4
Statins	83 (50)	65	78	7	16

[a] CN = concordant negative results; COPD = chronic obstructive lung disease; CP = concordant positive results; D(EP) = discordant results (electronic search positive, manual search negative); D(MP) = discordant results (electronic search negative, manual search positive).
[b] Determined by exhaustive review of the electronic medical record.

DDQB ranged from 77% for amiodarone therapy to 100% for COPD, restrictive lung disease, cirrhosis, H$_2$-receptor antagonists, PPIs, chemotherapy, and immunosuppressive therapy (median sensitivity, 100%). The sensitivities for identifying variables of interest with the Web-based electronic queries were higher than the corresponding sensitivities with manual review of the medical record for 7 of the 11 variables evaluated, identical for 3 variables, and lower for a single variable (statin therapy). The specificities were uniformly high (≥97%) for both manual and electronic data-extraction strategies.

DISCUSSION

This study compared 2 independent strategies for identifying pertinent risk factors for postoperative ALI. Our results suggest fair to excellent agreement between time-consuming manual review of the medical record and a highly efficient automated electronic search tool using data from a near real-time copy of the EMR. Comparison of the 2 strategies to an exhaustive review of the EMR revealed that the sensitivities of automated search strategies were consistently equal to or greater than that for the manual reviews of the medical record. The specificities of the searches were high and comparable for both data-collection strategies.

Early disease recognition with timely intervention has resulted in dramatic improvements in outcome for patients with acute coronary syndrome,[34,35] severe sepsis/septic shock,[36] and numerous other life-threatening conditions. Similarly effective strategies for the treatment of ALI have been suggested in preclinical studies.[37-41] Unfortunately, these promising preclinical results have not translated to the clinical setting, likely in part because strategies are applied too late in the progression of disease. Considering the limitations of clinical treatment trials and the disappointing results to date, prevention of ALI may prove a more effective approach than treating advanced disease. Working to this end, we have recently developed a lung injury prediction score for predicting a patient's risk of ALI at the time of hospital admission.[6,7] However, a remaining obstacle for the successful implementation of prediction models is the timely identification of variables known to portend risk. Historically, these variables have been identified with manual review of the medical record. This pro-

TABLE 4. **Calculated Sensitivities and Specificities for Manual and Automated Electronic Search Strategies in the Validation Cohort (N=166)**

Predictor	Manual review of the medical record		Automated electronic search	
	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Comorbid conditions				
Cirrhosis	1.00 (0.44-1.00)	1.00 (0.98-1.00)	1.00 (0.44-1.00)	0.99 (0.96-0.99)
COPD	0.81 (0.60-0.92)	0.98 (0.94-0.99)	1.00 (0.84-1.00)	0.99 (0.96-0.99)
Diabetes mellitus	0.91 (0.76-0.97)	0.99 (0.95-0.99)	0.91 (0.76-0.97)	0.99 (0.95-0.99)
Gastroesophageal reflux disease	0.61 (0.44-0.76)	0.98 (0.95-0.99)	0.84 (0.67-0.92)	1.00 (0.97-1.00)
Restrictive lung disease	1.00 (0.44 – 1.00)	1.00 (0.98-1.00)	1.00 (0.44-1.00)	0.99 (0.96-0.99)
Medications				
Amiodarone	0.66 (0.35-0.88)	0.98 (0.95-0.99)	0.78 (0.45-0.94)	0.97 (0.93-0.99)
Chemotherapy	0 (0-0.43)	1.00 (0.98-1.00)	1.00 (0.56-1.00)	1.00 (0.98-1.00)
H$_2$-receptor antagonists	0.90 (0.59-0.98)	0.99 (0.95-0.99)	1.00 (0.72-1.00)	1.00 (0.98-1.00)
Immunosuppressive therapy	0 (0-0.79)	0.99 (0.97-0.99)	1.00 (0.21-1.00)	1.00 (0.98-1.00)
Proton pump inhibitors	0.88 (0.75-0.95)	0.97 (0.92-0.99)	1.00 (0.92-1.00)	0.97 (0.93-0.99)
Statins	0.95 (0.88-0.98)	0.96 (0.90-0.99)	0.84 (0.75-0.91)	0.97 (0.91-0.99)

CI = confidence interval; COPD= chronic obstructive lung disease.

cess is time-consuming, inefficient, and frequently prone to error.[8,9]

The development and implementation of efficient automated electronic data-extraction strategies may provide an opportunity to identify pertinent risk factors and assess patient risk in near real-time. However, automated electronic search strategies have their own limitations. Pertinent data can be missing, errors can be made with data entry, and corruption can occur with data management, storage, and retrieval.[11] Valid results also depend on the accurate acquisition of patient information at the time of the initial patient encounter. Therefore, validation of new data-extraction techniques is mandatory.[8,9]

It should also be emphasized that, although well-designed validated queries can be highly efficient, the query-building process is typically iterative and can often be quite laborious. This is particularly true of free text queries, in which abbreviations, synonyms, and acronyms are commonplace. Additionally, key terms are often presented in the negative form and in a non–patient-specific format such as family history. We noted the identification and handling of the negative forms of key terms to be particularly important when optimizing search specificity.

Alternatives exist to the free text search strategies used in this study. The Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) is endorsed by health care standard organizations as the most comprehensive terminology for coding data in clinical information systems.[42] Previous investigations have noted generally excellent content coverage with SNOMED CT.[43,44] However, limitations exist with this system as well. For time-sensitive searches, the coding process can introduce an undesirable delay in the time from data entry into the EMR to SNOMED CT code allocation. SNOMED CT also appears to map poorly to some important critical care conditions (eg, severe sepsis[45]) as well as to numerous diagnoses included in common critical care severity of illness scoring systems, such as the APACHE (Acute Physiology and Chronic Health Evaluation) IV survey.[46] Similarly, SNOMED CT content coverage may not be as robust in the preoperative setting.[47] Finally, the reliability of SNOMED CT coding is imperfect and requires repeated training and testing.[48]

More advanced natural language processing techniques could also be used for identifying the textual information of interest. However, these techniques require extensive training and specific expertise as well as large training datasets.[49,50] Instead, we aimed to develop and validate portable automated strategies that could be used by clinician scientists who lack this specific expertise. It is unclear that more advanced natural language processing techniques would improve the accuracy of data extraction.[51,52]

Our study has several important limitations. First, we focused on 2 categories of data: patient comorbid conditions and medications. The completed queries were accurate, but considerable variability was noted in the early phases of the query-building process. The iterative nature of query building prevents us from generalizing our findings to other medical conditions or medications of interest because each would require independent validation. Similar validation would be required for other data elements such as demographics or laboratory results. A second limitation is our lack of strict definitions for the comorbid conditions of interest. The retrospective nature of this investigation and the substantial referral population prevented the use of strict diagnostic criteria. Rather, we chose to accept the documentation of a disease of interest in the EMR as sufficient for disease presence. It is possible that some of the patients in this investigation were inappropriately assigned comorbid conditions without meeting the strict diagnostic criteria. Although we attempted to identify active medication administration at the time of the surgical procedure, we cannot be certain that patients were taking the medications listed in their EMR. This limitation could only be effectively addressed with a prospective study design. Another limitation is the lack of a true criterion standard for determining the presence of our variables of interest. To allow a comparison of the 2 search strategies evaluated, we performed an exhaustive review of the medical record in an effort to create a criterion standard. This process has limitations and may have resulted in biased sensitivities and specificities for the queries performed. Moreover, the limited sample size results in rather imprecise estimates for the reported sensitivities and specificities, particularly among the lower frequency variables. A final important limitation is the single-center nature of this study and our study population. Although our specific results cannot be generalized to other institutions, the concepts presented have the potential for broad applicability.

CONCLUSION

Our findings validate the use of an automated electronic data-extraction tool for identifying pertinent preoperative risk factors for postoperative ALI. These strategies may be used to assist both clinicians and researchers when determining the risk of time-sensitive postoperative complications such as ALI.

REFERENCES
1. Hudson LD, Steinberg KP, Stapleton RD, Wang BM, Rubenfeld GD, Caldwell ES. Epidemiology of acute lung injury and ARDS. *Chest*. 1999;116:74S-82S.
2. Stapleton RD, Wang BM, Hudson LD, Rubenfeld GD, Caldwell ES, Steinberg KP. Causes and timing of death in patients with ARDS. *Chest*. 2005;128:525-532.
3. Fernandez-Perez ER, Sprung J, Afessa B, et al. Intraoperative ventilator settings and acute lung injury after elective surgery: a nested case control study. *Thorax*. 2009;64:121-127.

4. Kutlu CA, Williams EA, Evans TW, Pastorino U, Goldstraw P. Acute lung injury and acute respiratory distress syndrome after pulmonary resection. *Ann Thorac Surg*. 2000;69:376-380.

5. Ruffini E, Parola A, Papalia E, et al. Frequency and mortality of acute lung injury and acute respiratory distress syndrome after pulmonary resection for bronchogenic carcinoma. *Eur J Cardiothorac Surg*. 2001;20:30-36, discussion 36-37.

6. Trillo-Alvarez C, Cartin-Ceba R, Kor DJ, et al. Acute lung injury prediction score: derivation and validation in a population based sample [published online ahead of print June 18, 2010]. *Eur Respir J*. doi: 10.1183/09031936.00036810.

7. Gajic O, Dabbagh O, Park PK, et al; US Critical Illness and Injury Trials Group; Lung Injury Prevention Study Investigators (USCIITG-LIPS). Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. *Am J Respir Crit Care Med*. 2011;183(4):462-470.

8. Overhage JM, Suico J, McDonald CJ. Electronic laboratory reporting: barriers, solutions and findings. *J Public Health Manag Pract*. 2001;7:60-66.

9. Wurtz R, Cameron BJ. Electronic laboratory reporting for the infectious diseases physician and clinical microbiologist. *Clin Infect Dis*. 2005;40:1638-1643.

10. Brownstein JS, Murphy SN, Goldfine AB, et al. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. *Diabetes Care*. 2010;33(3):526-531.

11. Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers M, Weinstein RA. Development of a clinical data warehouse for hospital infection control. *J Am Med Inform Assoc*. 2003;10:454-462.

12. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology*. 2005;16:336-341.

13. Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. *J Am Med Inform Assoc*. 2000;7:42-54.

14. Baron JA, Weiderpass E. An introduction to epidemiological research with medical databases. *Ann Epidemiol*. 2000;10:200-204.

15. Baron RJ. Quality improvement with an electronic health record: achievable, but not automatic. *Ann Intern Med*. 2007;147:549-552.

16. Honiden S, Gong MN. Diabetes, insulin, and development of acute lung injury. *Crit Care Med*. 2009;37:2455-2464.

17. Gong MN, Thompson BT, Williams P, Pothier L, Boyce PD, Christiani DC. Clinical predictors of and mortality in acute respiratory distress syndrome: potential role of red cell transfusion. *Crit Care Med*. 2005;33:1191-1198.

18. Moss M, Guidot DM, Steinberg KP, et al. Diabetic patients have a decreased incidence of acute respiratory distress syndrome. *Crit Care Med*. 2000;28:2187-2192.

19. Algar FJ, Alvarez A, Salvatierra A, Baamonde C, Aranda JL, Lopez-Pujol FJ. Predicting pulmonary complications after pneumonectomy for lung cancer. *Eur J Cardiothorac Surg*. 2003;23:201-208.

20. Gajic O, Dara SI, Mendez JL, et al. Ventilator-associated lung injury in patients without acute lung injury at the onset of mechanical ventilation. *Crit Care Med*. 2004;32:1817-1824.

21. Kreider ME, Hansen-Flaschen J, Ahmad NN, et al. Complications of video-assisted thoracoscopic lung biopsy in patients with interstitial lung disease. *Ann Thorac Surg*. 2007;83:1140-1144.

22. Naito Y, Tsuchiya S, Ishihara S, et al. Impact of preexisting pulmonary fibrosis detected on chest radiograph and CT on the development of gefitinib-related interstitial lung disease. *Am J Clin Oncol*. 2008;31:340-344.

23. Shyamsundar M, McKeown STW, O'Kane CM, et al. Simvastatin decreases lipopolysaccharide induced pulmonary inflammation in healthy volunteers. *Am J Respir Crit Care Med*. 2009;179(12):1107-1114.

24. Iscimen R, Cartin-Ceba R, Yilmaz M, et al. Risk factors for the development of acute lung injury in patients with septic shock: an observational cohort study. *Crit Care Med*. 2008;36:1518-1522.

25. Malik SW, Myers JL, DeRemee RA, Specks U. Lung toxicity associated with cyclophosphamide use: two distinct patterns. *Am J Respir Crit Care Med*. 1996;154:1851-1856.

26. Pham PT, Pham PC, Danovitch GM, et al. Sirolimus-associated pulmonary toxicity. *Transplantation*. 2004;77:1215-1220.

27. Vlahakis NE, Rickman OB, Morgenthaler T. Sirolimus-associated diffuse alveolar hemorrhage. *Mayo Clin Proc*. 2004;79:541-545.

28. Wolkove N, Baltzan M. Amiodarone pulmonary toxicity. *Can Respir J*. 2009;16:43-48.

29. Saussine M, Colson P, Alauzen M, Mary H. Postoperative acute respiratory distress syndrome: a complication of amiodarone associated with 100 percent oxygen ventilation. *Chest*. 1992;102:980-981.

30. McCarthy EP, Iezzoni LI, Davis RB, et al. Does clinical evidence support ICD-9-CM diagnosis coding of complications? *Med Care*. 2000;38:868-876.

31. Nadkarni PM, Darer JA. Migrating existing clinical content from ICD-9 to SNOMED. *J Am Med Inform Assoc*. 2010;17(5):602-607.

32. Kern EF, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res*. 2006;41:564-580.

33. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.

34. Julian DG. Treatment of cardiac arrest in acute myocardial ischaemia and infarction. *Lancet*. 1961;2:840-844.

35. Reimer KA, Lowe JE, Rasmussen MM, Jennings RB. The wavefront phenomenon of ischemic cell death; 1: myocardial infarct size vs duration of coronary occlusion in dogs. *Circulation*. 1977;56:786-794.

36. Rivers E, Nguyen B, Havstad S, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med*. 2001;345:1368-1377.

37. ARDS Network. Ketoconazole for early treatment of acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. *JAMA*. 2000;283:1995-2002.

38. ARDS Clinical Trials Network; National Heart, Lung, and Blood Institute (NHLBI); National Institutes of Health (NIH). Randomized, placebo-controlled trial of lisofylline for early treatment of acute lung injury and acute respiratory distress syndrome. *Crit Care Med*. 2002;30:1-6.

39. Jepsen S, Herlevsen P, Knudsen P, Bud MI, Klausen NO. Antioxidant treatment with N-acetylcysteine during adult respiratory distress syndrome: a prospective, randomized, placebo-controlled study. *Crit Care Med*. 1992;20:918-923.

40. Meade MO, Jacka MJ, Cook DJ, Dodek P, Griffith L, Guyatt GH. Survey of interventions for the prevention and treatment of acute respiratory distress syndrome. *Crit Care Med*. 2004;32:946-954.

41. Zeiher BG, Artigas A, Vincent JL, et al. Neutrophil elastase inhibition in acute lung injury: results of the STRIVE study. *Crit Care Med*. 2004;32:1695-1702.

42. *Essentials of the US Hospital IT Market*. 4th ed. Chicago, IL: HIMSS Analytics; 2009. http://www.himssanalytics.org/docs/4thEditionEssentialsIntroductionFinal.pdf. Accessed February 23, 2011.

43. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clin Proc*. 2006;81:741-748.

44. Rosenbloom ST, Brown SH, Froehling D, et al. Using SNOMED CT to represent two interface terminologies. *J Am Med Inform Assoc*. 2009;16:81-88.

45. Shahpori R, Doig C. Systematized Nomenclature of Medicine-Clinical Terms direction and its implications on critical care. *J Crit Care*. 2010;25(2):364.e1-364.e9.

46. Bakhshi-Raiez F, Cornet R, de Keizer NF. Cross-mapping APACHE IV "reasons for intensive care admission" classification to SNOMED CT. *Stud Health Technol Inform*. 2008;136:779-784.

47. Ahmadian L, De Keizer NF, Cornet R. The use of SNOMED CT for representing concepts used in preoperative guidelines. *Stud Health Technol Inform*. 2009;150:658-662.

48. Chiang MF, Hwang JC, Yu AC, Casper DS, Cimino JJ, Starren JB. Reliability of SNOMED-CT coding by three physicians using two terminology browsers. *AMIA Annu Symp Proc*. 2006:131-135.

49. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. 2004;11:392-402.

50. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. *J Am Med Inform Assoc*. 2009;16:596-600.

51. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform*. 2006;39:589-599.

52. Divita G, Tse T, Roth L. Failure analysis of MetaMap Transfer (MMTx). *Medinfo*. 2004;107(pt 2):763-767.