Hindawi Publishing Corporation
Journal of Cancer Epidemiology
Volume 2011, Article ID 983271, 5 pages
doi:10.1155/2011/983271

*Research Article*

# Evaluation of the Reliability of Electronic Medical Record Data in Identifying Comorbid Conditions among Patients with Advanced Non-Small Cell Lung Cancer

**Catherine E. Muehlenbein,[1] J. Russell Hoverman,[2] Stephen K. Gruschkus,[3] Michael Forsyth,[3] Clara Chen,[3] William Lopez,[3] Anthony Lawson,[1] Heather J. Hartnett,[3] and Gerhardt Pohl[1]**

[1] *Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN 46285, USA*
[2] *Texas Oncology, Dallas, TX 75251, USA*
[3] *Healthcare Informatics, US Oncology, The Woodlands, TX 77380, USA*

Correspondence should be addressed to Catherine E. Muehlenbein, cemuehlenbein@lilly.com

Received 30 September 2010; Revised 10 February 2011; Accepted 28 February 2011

Academic Editor: S. Dubinett



*Background*. Traditional methods for identifying comorbidity data in EMRs have relied primarily on costly and time-consuming manual chart review. The purpose of this study was to validate a strategy of electronically searching EMR data to identify comorbidities among cancer patients. *Methods*. Advanced stage NSCLC patients ($N = 2,513$) who received chemotherapy from 7/1/2006 to 6/30/2008 were identified using iKnowMed, US Oncology's proprietary oncology-specific EMR system. EMR data were searched for documentation of comorbidities common to advanced stage cancer patients. The search was conducted by a series of programmatic queries on standardized information including concomitant illnesses, patient history, review of systems, and diagnoses other than cancer. The validity of the comorbidity information that we derived from the EMR search was compared to the chart review gold standard in a random sample of 450 patients for whom the EMR search yielded no indication of comorbidities. Negative predictive values were calculated. *Results*. The overall prevalence of comorbidities of 22%. Overall negative predictive value was 0.92 in the 450 patients randomly sampled patients (36 of 450 were found to have evidence of comorbidities on chart review). *Conclusion*. Results of this study suggest that efficient queries/text searches of EMR data may provide reliable data on comorbid conditions among cancer patients.

## 1. Background

Use of electronic medical records (EMRs) has grown dramatically over the last decade. Traditional methods for identifying comorbidity data in EMRs for use in clinical research have relied primarily on comprehensive manual chart review. While highly sensitive for extracting data from EMR systems, they are very costly and time consuming as they need to be performed by trained chart abstractors [1].

The objective of the current study was to develop and validate an efficient and reliable algorithm to identify comorbidities among advanced non-small cell lung cancer (NSCLC) patients using relatively simple queries and text field searches of EMR data.

In 2010, cancer of the lung or bronchus resulted in approximately 222,000 new cases and 167,000 deaths within the United States [2]. Approximately 85% of lung cancer patients will have NSCLC; most will have advanced stage disease at the time of diagnosis, and 5-year survival is less than 10% in this patient population [3, 4].

The median age of NSCLC patients is now slightly less than 70 years [5], and 60% of cases occur in patients ≥65 years [6]. As a result, patients often present with several concurrent comorbid conditions that may impair organ function [7, 8] and result in reduced therapeutic benefit from chemotherapy [9]. In addition, common risk factors for NSCLC, such as smoking or excessive alcohol

use [10], also place patients at greater risk for other serious medical conditions such as heart disease, chronic obstructive pulmonary disease (COPD), other cancers, and stroke.

The presence of NSCLC together with a serious comorbid condition has the potential to impact treatment decision making and clinical outcomes [11]. Retrospective observational research on NSCLC outcomes using EMR data should thus include information on patients' comorbid conditions, making the accurate and efficient identification of comorbidity data vital. Information on adapting a clinical comorbidity index for use with ICD-9-CM administrative databases is available in the medical literature [1]. Lacking in the literature is clarity on whether efficient electronic searches and queries of oncology EMR data can yield similar comorbidity information in the absence of claims data and without costly and lengthy manual chart abstraction.

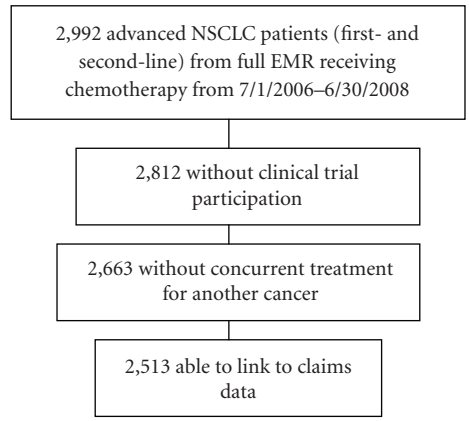


Figure 1: Sample selection flow chart. Abbreviations: EMR: electronic medical record. NSCLC: non-small cell lung cancer.

## 2. Methods

*2.1. Data Sources.* This retrospective study utilized data from US Oncology's EMR system iKnowMed (iKM). iKM is an oncology-specific EMR system that captures outpatient practice encounter history for patients under care including (but not limited to) laboratory, diagnosis, therapy administration, line of therapy, staging, comorbidities, and performance status information. This product is being implemented across the US Oncology Comprehensive Strategic Alliance (CSA) network. iKM currently is used by more than 900 community-based oncology providers. For the time-frame of this study, iKM was active in approximately 82% of the CSA network.

Due to new initiatives to utilize an electronic medical record system, US Oncology began to implement this electronic record early on in US Oncology's history. Initially, the EMR was designed to capture the information that was previously collected on the paper record. Currently, US Oncology researchers have the ability to conduct retrospective studies and gain access to the electronic data by submitting a request to the US Oncology Institutional Review Board (IRB). This current study was submitted to and approved by the US Oncology IRB. There was minimal patient risk related to this retrospective study, and patient confidentiality was rigorously maintained per US Oncology policies. Only US Oncology employees on the study team had access to the data and were involved in data manipulation and analysis. Waiving informed consent did not adversely affect patient rights as they were unlikely to directly benefit from this study. Only the minimum number of variables necessary to accomplish the goals of this project were extracted. Only data that fell within the study time period were used and all results are reported in the aggregate. Electronic chart reviews were conducted only for the purpose of validating data that were abstracted through programmatic queries of the iKnowMed database.

*2.2. Patient Selection Criteria and Characteristics.* As shown in Figure 1, this study included advanced NSCLC patients who initiated a first- or second-line chemotherapy regimen in the 30-month period from July 1, 2006 through June 30, 2008. Patients in the 1st line cohort were not included in the 2nd line cohort, even if they initiated 2nd line therapy during the 30-month period. All patients received care at a US Oncology Network practice utilizing iKM capabilities. NSCLC disease was required to be Stage IIIB or IV or recurrent (defined as disease that was initially diagnosed as early stage that progressed to metastatic disease). Patients who were enrolled on formal IRB approved clinical trials were excluded from this analysis as were any patients receiving care for other primary cancers during the study time period.

*Phase 1: Algorithm for Identification of Comorbidities .* To efficiently identify specific comorbid conditions, programmatic queries were applied to standardized fields within the EMR (i.e., concomitant illness tables, specific diagnosis tables, review of symptoms, physical examination tables, and past medical history table). To supplement the programmatic queries, keyword searches (see the appendix) were conducted of available text-based fields (e.g., dictation notes or nurse progress notes). There were no time restrictions placed on the queries or text field searches meaning comorbidities found at any time in the EMR were eligible.

*Phase 2: Validation of Algorithm .* We evaluated the reliability of the EMR query/text field search by comparing comorbidity status of patients as identified via the query/text field search algorithm to the comorbidity status as identified via the comprehensive chart review (the gold standard). To do this, the algorithm was used to identify the patients in the study sample without comorbidities. Seven groups of patients were created by randomly sampling from among these patients: one group ($N = 150$) had no evidence of any comorbidities while 6 groups ($N = 50$ each) did not have evidence of a single prespecified comorbidity. For all patients in these 7 groups, comprehensive chart reviews (i.e., the gold standard) were conducted to verify that they were truly negative for comorbidities. Reliability (as demonstrated

by negative predictive value) of the algorithm was calculated as follows:

$$\text{Negative predictive value} = \frac{\text{True negatives}}{(\text{True negatives} + \text{False negatives})}. \quad (1)$$

Negative predictive values greater than 80% were considered to be acceptable. Negative predictive values greater than 90% were considered to be good quality.

*2.3. Statistics.* Negative predictive values between specific groups were compared using an exact version of Pearson's chi-square test. 95% confidence intervals were calculated using an exact test for proportions. All analyses were run in SAS version 8.2 (Cary, North Carolina).

## 3. Results

*3.1. Patient Disposition.* Figure 1 summarizes the study population selection. During the study period, 2,992 patients with advanced NSCLC received chemotherapy at a US Oncology Network site that had iKM EMR capabilities. Of these, 2,513 patients (84%) were included in this study. Patient demographic and clinical characteristics as well as comorbidity status (as documented via the EMR search algorithm) are summarized in Table 1.

*3.2. Concordance between Comprehensive Electronic Chart Review and EMR Search Algorithm.* The measured reliability of the search algorithm is summarized in Table 2. Of the 150 patients randomly chosen who were negative for all selected comorbidities via search algorithm, 16 were false negatives (via "gold standard" comprehensive chart review), for a negative predictive value of 89%. Within the 6 groups of randomly selected patients ($N = 50$ each) who were negative for specific comorbidities via search algorithm, negative predictive values ranged from 82% for diabetes (9 of 50 patients were found to have diabetes on chart review) to 100% for cerebrovascular disease and were found to differ significantly in the six disease-specific samples ($P = .007$). Considering the 450 patients from all groups, the overall negative predictive value was 92% (36 of 450 were found to have evidence of comorbidities on chart review).

Table 1: Patient characteristics of study sample ($N = 2,513$).

| | $N$ (%) |
|---|---|
| Line of therapy | |
| First-line | 2004 (80%) |
| Second-line | 509 (20%) |
| Age | |
| Mean (SD) | 69.7 |
| Median (range) | 70.7 (30–92) |
| Gender | |
| Male | 1378 (55%) |
| Female | 1135 (45%) |
| Stage at diagnosis | |
| I-IIIA | 538 (24%) |
| IIIB-IV | 1664 (76%) |
| Missing | 311 |
| Performance status | |
| 0 | 1045 (46%) |
| 1 | 1032 (46%) |
| 2+ | 175 (8%) |
| Missing | 261 |
| Comorbidity status | |
| Any comorbidity (1+ comorbidities) | 553 (22%) |
| Specific comorbidities | |
| Moderate/severe renal disease | 37 (1.5%) |
| Congestive heart failure | 106 (4.2%) |
| Chronic obstructive pulmonary disorder | 218 (9%) |
| Cerebrovascular disease | 33 (1.3%) |
| Diabetes | 162 (6.4%) |
| Peripheral vascular disease | 113 (4.5%) |
| Myocardial infarction | 21 (0.8%) |
| Liver disease | 16 (0.6%) |

## 4. Conclusions

Results of the current study suggest that efficient programmatic queries and text field searches of EMR data may provide reliable data on comorbid conditions among cancer patients. Among this population of advanced NSCLC patients, the rate of comorbidities was 22%. The most prevalent comorbidity, as expected, was COPD. This is consistent with research in operable patients with NSCLC, which has found rates of COPD as high as 50% [12].

Traditional methods for identifying comorbidity data in EMRs for use in clinical research have relied primarily on comprehensive manual chart review. While this study relied on chart reviews to be a highly sensitive gold standard for extracting data from EMR systems, we also characterized them as costly and time consuming as they were conducted by nurses trained in chart abstraction over the course of several weeks. As EMR data sources become more widely utilized in health services and outcomes research, it will be important to develop methods of electronically searching the data in ways that are as clinically accurate but provide significant savings in both time and money compared to manual chart reviews.

Using an algorithm that incorporated programmatic queries and keyword searches of US Oncology's cancer-specific EMR system (iKM), we were able to reliably and accurately identify comorbidities. In a validation study using comprehensive electronic chart reviews, the algorithm yielded an acceptable false negative rate (8%).

The benefits of search algorithms of electronic records compared with comprehensive manual chart reviews have been previously reported [1, 13–15]. These include the reduction in time and expense for data extraction as well as the identification of data in free-text chart notes that otherwise would be difficult to locate. While previous papers

Table 2: Measured reliability of algorithm (combination of programmatic queries and key word searches).

| Electronic chart review group | True negatives (via chart review) | False negatives (via chart review) | Negative predictive value | 95% confidence interval |
|---|---|---|---|---|
| $N = 450$ patients identified as negative via algorithm | 414 | 36 | 0.92 | 0.89–0.94 |
| Subgroup analyses of patients negative for comorbidities via algorithm | | | | |
| Group 1: Negative for all comorbidities ($N = 150$) | 134 | 16 | 0.89 | 0.83–0.94 |
| Group 2: Negative for diabetes ($N = 50$) | 41 | 9 | 0.82 | 0.69–0.91 |
| Group 3: Negative for cardiovascular disease ($N = 50$) | 47 | 3 | 0.94 | 0.83–0.99 |
| Group 4: Negative for cerebrovascular disease ($N = 50$) | 50 | 0 | 1.00 | 0.93–1.00 |
| Group 5: Negative for moderate/severe renal disease ($N = 50$) | 47 | 3 | 0.94 | 0.83–0.99 |
| Group 6: Negative for liver disease ($N = 50$) | 49 | 1 | 0.95 | 0.89–0.99 |
| Group 7: Negative for COPD ($N = 50$) | 46 | 4 | 0.92 | 0.81–0.98 |

COPD: chronic obstructive pulmonary disorder.

have focused on the experience of single institutions, our experience suggests that implementation within a large oncology network is feasible. Refinement of search algorithms will continue to grow in importance as the use of EMRs by the broader oncology community expands. The lack of standard reporting requirements has been an important limitation of attempts to use EMRs for clinical research purposes [16]. The validity of the data extracted in the current study reflects the strength of the iKM system in terms of data standardization. Efforts to improve the search algorithm include ongoing collaborations by US Oncology with their physicians to enhance and refine the capture of relevant free-text data and also to ensure the algorithm reflects advances in the field of oncology. US Oncology is also working to improve iKM by defining additional fields their physicians need to capture and document the patients' clinical information, as well as continuing to train their physicians to enhance documentation and data capture rates.

While the use of oncology-specific EMR data and outpatient claims data provides a very detailed and comprehensive view of the medical oncology setting, a limitation of the study is the likelihood of missing data on comorbidity-specific resource utilization that may occur outside the US Oncology network, particularly those associated with noncancer-related events. Another limitation is that we studied only the negative predictive value of the new algorithm. To ascertain sensitivity would require a different and potentially arduous study design in which one captured sufficiently large numbers of patients with each of the comorbidities via random selection from the population using the gold standard manual chart review process. Our assumption is that given the serious nature of the comorbidities of interest, they were likely to have been reported to the attending physician. It remains for future study whether the comorbidities were subsequently properly captured in iKM.

Comorbidities as defined in the current study relate to physical conditions that are not a part of lung cancer, but that may have an impact on treatment choices, safety, and outcomes. The current search strategy using efficient queries of standardized EMR data may be widely applicable to other oncology settings and research questions. The current report suggests it is possible to extract research data in a reliable and timely manner.

## Appendix

### iKM Search Terms Used in This Study

DM: Diabetes mellitus
Congestive heart failure
Peripheral vascular disease
Aortic aneurysm
Aortic dissection
Thoracic aneurysm
Abdominal aneurysm
Venous thromboembolism
Myocardial infarction
Coronary artery disease
CAD
HTN
Cereb
Hemorrhage
Cranial
Occlusion
Embolism
Obstruction
Stroke
Thrombosis
Narrowing

Transient
TCI
CVA
TIA
Chronic glomerulonephritis
Nephritis
Nephropathy
Kidney
Renal failure
CRI
Cirrhosis
Hepatitis
Portal
Liver
Hepatic
Hepatorenal
Varices
CPD
COPD
Chronic and pulmonary
Bronchitis.

## Acknowledgments

The authors wish to thank Albert Alva and Joseph Darragh for their guidance in utilizing iKnowMed data for this project. Research funding was supported by Eli Lilly and Company. This research was previously presented at the International Society For Pharmacoeconomics and Outcomes Research (ISPOR) 15th Annual International Meeting; Atlanta, Georgia; May 15–19, 2010.

## References

[1] L. Seyfried, D. A. Hanauer, D. Nease, R. Albeiruti, J. Kavanagh, and H. C. Kales, "Enhanced identification of eligibility for depression research using an electronic medical record search engine," *International Journal of Medical Informatics*, vol. 78, no. 12, pp. e13–e18, 2009.

[2] A. Jemal, R. Siegel, J. Xu, and E. Ward, "Cancer statistics, 2010," *Ca: A Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 277–300, 2010.

[3] F. A. Shepherd, "Screening, diagnosis, and staging of lung cancer," *Current Opinion in Oncology*, vol. 5, no. 2, pp. 310–322, 1993.

[4] J. Walling, "Chemotherapy for advanced non-small-cell lung cancer," *Respiratory Medicine*, vol. 88, no. 9, pp. 649–657, 1994.

[5] R. J. Havlik, R. Yancik, S. Long, L. Ries, and B. Edwards, "The National Institute on Aging and the National Cancer Institute SEER collaborative study on comorbidity and early diagnosis of cancer in the elderly," *Cancer*, vol. 74, no. 7, pp. 2101–2106, 1994.

[6] B. K. Edwards, H. L. Howe, L. A. G. Ries et al., "Annual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on U.S. cancer burden," *Cancer*, vol. 94, no. 10, pp. 2766–2792, 2002.

[7] M. Vercelli, A. Quaglia, C. Casella et al., "Cancer patient survival in the elderly in Italy. ITACARE Working Group," *Tumori*, vol. 83, no. 1, pp. 490–496, 1997.

[8] T. Wasil and S. M. Lichtman, "Clinical pharmacology issues relevant to the dosing and toxicity of chemotherapy drugs in the elderly," *Oncologist*, vol. 10, no. 8, pp. 602–612, 2005.

[9] H. Wildiers, M. S. Highley, E. A. de Bruijn, and A. T. van Oosterom, "Pharmacology of anticancer drugs in the elderly population," *Clinical Pharmacokinetics*, vol. 42, no. 14, pp. 1213–1242, 2003.

[10] J. R. Molina, P. Yang, S. D. Cassivi, S. E. Schild, and A. A. Adjei, "Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship," *Mayo Clinic Proceedings*, vol. 83, no. 5, pp. 584–594, 2008.

[11] J. Ngeow, S. S. Leong, F. Gao et al., "Impact of comorbidities on clinical outcomes in non-small cell lung cancer patients who are elderly and/or have poor performance status," *Critical Reviews in Oncology/Hematology*, vol. 76, no. 1, pp. 53–60, 2009.

[12] A. López-Encuentra, "Comorbidity in operable lung cancer: a multicenter descriptive study on 2992 patients," *Lung Cancer*, vol. 35, no. 3, pp. 263–269, 2002.

[13] J. M. Fisk, P. Mutalik, F. W. Levin, J. Erdos, C. Taylor, and P. Nadkarni, "Integrating query of relational and textual data in clinical databases: a case study," *Journal of the American Medical Informatics Association*, vol. 10, no. 1, pp. 21–38, 2003.

[14] W. Gregg, J. Jirjis, N. M. Lorenzi, and D. Giuse, "StarTracker: an integrated, web-based clinical search engine," *AMIA Annual Symposium Proceedings*, vol. 2003, p. 855, 2003.

[15] F. Malamateniou, G. Vassilacopoulos, and J. Mantas, "A search engine for virtual patient records," *International Journal of Medical Informatics*, vol. 55, no. 2, pp. 103–115, 1999.

[16] K. J. Kristianson, H. Ljunggren, and L. L. Gustafsson, "Data extraction from a semi-structured electronic medical record system for outpatients: a model to facilitate the access and use of data for quality control and research," *Health Informatics Journal*, vol. 15, no. 4, pp. 305–319, 2009.