Original Article

Linkage of Indiana State Cancer Registry and Indiana Network for Patient Care Data

Laura P. Ruppert, MHA[a]; Jinghua He, PhD, MPH[b]; Joel Martin[c]; George Eckert[d,e,f]; Fangqian Ouyang[d]; Abby Church, MPH[c]; Paul Dexter, MD[c,d,g]; Siu Hui, PhD[c,d]; David Haggstrom, MD[c,d,h,i]

Abstract: Background: Large automated electronic health records (EHRs), if brought together in a federated data model, have the potential to serve as valuable population-based tools in studying the patterns and effectiveness of treatment. The Indiana Network for Patient Care (INPC) is a unique federated EHR data repository that contains data collected from a large population across various health care settings throughout the state of Indiana. The INPC clinical data environment allows quick access and extraction of information from medical charts. The purpose of this project was to evaluate 2 different methods of record linkage between the Indiana State Cancer Registry (ISCR) and INPC, determine the match rate for linkage between the ISCR and INPC data for patients diagnosed with cancer, and to assess the completeness of the ISCR based on additional validated cancer cases identified in the INPC EHRs. Methods: Deterministic and probabilistic algorithms were applied to link ISCR cases to the INPC. The linkage results were validated by manual review and the accuracy assessed with positive predictive value (PPV). Medical charts of melanoma and lung cancer cases identified in INPC but not linked to ISCR were manually reviewed to identify true incidence cancers missed by the ISCR, from which the completeness of the ISCR was estimated for each cancer. Results: Both deterministic and probabilistic approaches to linking ISCR and INPC had extremely high PPV (>99%) for identifying true matches for the overall cohort and each subcohort. The combined match rate for melanoma and lung cancer cases identified in the ISCR that matched to any patient occurrence in INPC (not by disease) was 85.5% for the complete cohort, 94.4% for melanoma, and 84.4% for lung cancer. The estimated completeness of capture by the ISCR was 84% for melanoma and 98% for lung cancer. Conclusion: Cancer registries can be successfully linked to patients' EHR data from institutions participating in a regional health information organization (RHIO) with a high match rate. A pragmatic approach to data linkage may apply both deterministic and probabilistic approaches together for the diverse purposes of cancer control research. The RHIO has the potential to add value to the state cancer registry through the identification of additional true incident cases, but more advanced approaches, such as natural language processing, are needed.

Key words: *electronic health records, record linkage*

Introduction

With the passage of the Indiana General Assembly's Public Law 174-1985 in 1985, the Indiana State Cancer Registry (ISCR) was established "for the purpose of recording all cases of malignant disease and other tumors and precancerous diseases required to be reported by federal law or federal regulation or the National Program of Cancer Registries that are diagnosed or treated in Indiana, and compiling necessary and appropriate information concerning those cases, as determined by the state department, in order to conduct epidemiologic surveys of cancer and to apply appropriate preventive and control measures." Reporting for both providers and hospitals began on January 1, 1987.[1]

Population-based cancer registries have been widely used to study the epidemiology of various cancers, including incidence across geographical locations and time. Cancers captured in registries are well characterized around the time of diagnosis, including type, stage, and initial treatment. These case characteristics are seldom captured elsewhere, but they are invaluable for researchers in conducting in-depth epidemiological studies such as geographical variations of surveillance[2] and time trends of treatment patterns for specific cancer types and stages, eg, stage IV oral cavity and pharyngeal cancers.[3] The availability of mortality data associated with cancer registries also enables studies of factors on survival.[4] However, cancer registries are usually limited in other follow-up information such as subsequent adjuvant or chronic treatments, clinical course and patient outcomes or adverse events. Linking cancer registry patients' records to an individual's electronic health records (EHR) can create a resource for asking

[a]Indiana State Cancer Registry, Indiana State Department of Health, Indianapolis, Indiana. [b]Merck & Co., Inc. [c]Regenstrief Institute, Inc, Indianapolis, Indiana. [d]Indiana University School of Medicine, Indianapolis, Indiana. [e]Richard M. Fairbanks School of Public Health, Indianapolis, Indiana. [f]Indiana CTSI (Clinical and Translational Sciences Institute), Indianapolis, Indiana. [g]Eskenazi Health, Indianapolis, Indiana. [h]VA Health Services Research and Development Center for Health Information and Communication, Richard L. Roudebush VA Medical Center, Indianapolis, Indiana. [i]Indiana University Center for Health Services and Outcomes Research, Indianapolis, Indiana. Address correspondence to Laura P. Ruppert, MHA, Indiana State Department of Health. Email: lruppert@isdh.in.gov.

This research was funded by a grant under the Merck–Regenstrief Program in Personalized Health Care Research and Innovation, a collaboration between Merck, Sharp & Dohme and the Regenstrief Institute.

This journal article was supported by DP003884 funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

more complex questions in longitudinal, population-based studies, especially regarding patterns of follow-up care.

In a working example of such a resource, the ISCR data were linked to the Indiana Network for Patient Care (INPC). The INPC is a unique federated EHR data repository that contains data collected from a large population across various health care settings throughout the state of Indiana. The INPC includes clinical data from 103 Indiana hospitals, 41 core hospital systems, and 60 community clinics, as well as state and local public health departments. Each participating institution provides common data elements, which can include inpatient admission/discharge information; outpatient visit information; laboratory values; microbiology, pathology, radiology, and cardiology reports; and clinical notes that can be analyzed via natural language processing. The INPC was originally developed by the Regenstrief Institute, which developed an accompanying clinical data environment to allow quick access and extraction of information from medical charts. The purpose of this project was to develop and validate linkage algorithms to match the cancer cases in ISCR to medical records in the INPC. The linked records were used to assess the completeness of the ISCR in capturing specific cancers in Indiana. The findings have implications for the value and design of future longitudinal studies that make use of linked cancer registry and EHR data.

Methods

Cohort Selection

Three cohorts were selected from the ISCR for this study: 1 overall cohort encompassing all cancer patients and 2 subcohorts consisting of melanoma and lung cancer cases.

Complete Cohort. The complete cohort of cancer patients was selected from all entries with a primary date of diagnosis from January 1, 2005 to December 31, 2013 in the ISCR. To allow for the most complete diagnosis and first round treatment information, cancer cases are reported to the ISCR within 6 months of diagnosis or first round treatment. This expanded range of dates ensured that the cancer registry capture was complete for cases in the 2005–2012 time frame. Population-based data collection approaches have undergone progressive changes in the 3 decades since the ISCR was first established. Starting in 2003, the ISDH Cancer Registry implemented the Facility Oncology Registry Data Standards (FORDS) coding standard, developed by the Commission on Cancer (CoC). Consequently, we assembled our cohort over a 10-year time frame from 2005–2013 to both encompass data collected after the FORDS standard was fully implemented, as well as to allow for a sufficient cohort of cancer cases for analysis.

To maximize the likelihood that ISCR cancer cases would be identified in the INPC, these cases were further restricted to those submitted to the ISCR by 61 health care institutions that send EHRs to INPC. Of these 61 institutions, 42.6% were accredited by the CoC. For this estimate, institutions include both hospitals, as well as integrated delivery care systems that may encompass more than a single hospital.

Selection of Melanoma and Lung Cancer Cases from the ISCR. These specific cancers were chosen from the complete cohort because experience at the ISCR suggested that the capture rate was relatively higher for lung cancer, but lower for melanoma compared to other cancers. Specific cancer cases were selected from the ISCR cohort by histology code. The Surveillance, Epidemiology, and End Results (SEER) Program and International Classification of Diseases for Oncology, 3rd edition (ICD-O-3) list categorizes melanoma as 8700–8799 (**http://seer.cancer.gov/icd-o-3/**). Lung cancer cases were selected from the complete cohort by including all SEER ICD-O-3 codes C340–C349.

The mean ages (and standard deviations) of the melanoma and lung cancer cases were 58.5 (16.6) and 66.0 (11.4) years, respectively. Of the melanoma cases, 45% were female and 98% were white (1% or less other races); of the lung cancer cases, 48% were female, 88% were white, and 11% were African American.

Linkage Algorithms Applied between ISCR and INPC

An attempt to match all eligible cancer cases from the complete ISCR cohort to the INPC was made using 2 different linkage approaches.

Deterministic Linkage. The Regenstrief Global Linkage Algorithm, which is run daily on the INPC production database to link newly generated clinical data to existing patient records in the INPC master file. The Global Linkage Algorithm is considered a conservative deterministic algorithm.[6] Deterministic algorithms assess whether record pairs agree or disagree on a given set of identifiers, where agreement is assessed as a binary ("all-or-nothing") outcome.[7] For this study's purposes, Global Linkage made use of name, date of birth, gender, ZIP code, telephone number, and Social Security number, whenever these data elements were available.

Probabilistic Linkage. A majority of patients in the ISCR had a value representing the medical record number (MRN) of the submitting institution, which should have very high specificity if matched to the MRN in the INPC. Therefore, separate probabilistic linkage processes were run, based upon whether the institution and MRN matched between ISCR and INPC among all possible pairs from the 2 data sources. Probabilistic algorithms assign different weights for each record field based upon the probability that agreement on this field increases or decreases the probability that the 2 records refer to the same person. Probabilistic linkages allow imperfect matches due to partially inaccurate or missing data. The specific probabilistic linkage algorithm used for each linkage process is named RecMatch, a Regenstrief-developed probabilistic matching program based on the Felligi-Sunter model.[5] To limit the number of pairs being considered, RecMatch functions by first selecting blocking variables. Each eligible match pair within a block must exactly match on the blocking variables. Other data fields are then evaluated for similarity and a score is generated based on their likelihood of being a true match. All eligible matches scoring above that cut-off score are considered true matches. Multiple blocks based on different blocking variables were used, and pairs identified

as matches from any block were considered to be matches. For the complete cohort, the institution ID was required to be one of the blocking variables within each block in order to keep the number of potential pairs of matches within feasible computational parameters. For the 2 subcohorts, no such requirement was necessary.

Validation of Matches for Testing Optimal Linkage Method and Match Rate between ISCR and INPC

Pairs of identifiers from the ISCR and INPC that were declared as matches by both the Global and MRN/probabilistic algorithms were considered true matches. Pairs declared as matches by 1 algorithm, but not another, were manually reviewed by 2 reviewers to determine the "true" match status. Medical record review was used as the reference standard for evaluating the performance of the linkage algorithms.

Evaluation of Linkage Algorithms for Linkage Method and Match Rate between ISCR and INPC

Within each "zone" where pairs of identifiers are declared matches by only 1 algorithm, the proportion of true matches was estimated based on the validation results. To arrive at an estimate of the positive predictive value (PPV) of each linkage algorithm, the estimate of each "zone" was combined with the presumed 100% accuracy in the zone in which all pairs were declared matches by both algorithms.[8]

Estimating completeness of ISCR's capture of melanoma and lung cancer cases

A subset of patients identified as having cancer in INPC, but who were not identified as having cancer through linkage to the ISCR, were sampled to estimate the completeness of the ISCR. Patients were selected from the INPC if they had a first occurrence of an ICD-9 diagnostic code of lung cancer (162.2-162.9) or melanoma (172.X) between the dates January 1, 2005 to December 31, 2012. A subset of 200 charts of each cancer type was randomly selected for manual review by 2 reviewers to determine if each case was a true incident case of the specific cancer within the time period. The estimated number of true incident cases in the INPC that were not found in the ISCR was used to estimate the completeness of the ISCR.

Results

Evaluation of the Performance of Linkage Algorithms

Complete Cohort. From 2005–2013, a total of 202,153 cases were submitted to the ISCR from institutions also reporting data to the INPC. Application of the deterministic algorithm to these 202,153 cases in the ISCR resulted in 132,893 cases being validated as matches to patients in the INPC.

For 126,779 cases, ISCR MRN matched to a corresponding MRN in the INPC; in this instance, the probabilistic algorithm did not converge because manual validation of a random sample of pairs of identifiers in this group showed a 99% accuracy by MRN alone. For 75,374 cases, the ISCR MRN did not match to any MRN in the INPC; in this scenario, the probabilistic algorithm declared 22,804 as matched to patients in INPC. The stratified MRN/probabilistic algorithm approach declared a total of 149,583 ISCR cases as matched to patients in the INPC.

Overall, a total of 172,895 ISCR cases could be matched to the INPC using either of the 2 algorithms, resulting in an overall match rate of 85.5%. These results are summarized in Table 1. From each cell, a random sample of pairs of patient identifiers was manually reviewed (2 independent reviewers) to determine whether each pair came from the same patient. The sample sizes of the manual reviews are shown in parentheses in Table 1. The 2 independent reviewers had high agreement on whether a pair was truly the same patient (interrater κ = 0.988). The estimated PPV was 99.96% (s.e. = 0.04%) for the deterministic algorithm and 99.39% (0.19%) for the stratified MRN/probabilistic algorithm.

Table 1. Numbers of Indiana State Cancer Registry Cases Declared Matched to Indiana Network for Patient Care (Numbers Sampled for Manual Review) for the Complete Cohort

Probabilistic Algorithm		*Deterministic Algorithm*	
		Match	**No Match**
MRN Match		94,134	32,645 (400)
MRN No Match	Probabilistic Match	15,447	7,357 (400)
	No Probabilistic Match	23,312 (400)	29,258 (200)

MRN, medical record number.

Melanoma and Lung Cancer Cohorts. After all eligibility criteria were met, the total number of cases in the melanoma (n = 6,853) and lung cancer (n = 31,565) cohorts were determined over the study period. For melanoma, 6,471 of the original 6,853 ISCR cases could be linked to INPC using any of the algorithms, a match rate of 94.4%. For lung cancer, 26,662 of the 31,565 ISCR patients were linked to INPC, a match rate of 84.4%. For each of these 2 cohorts, the cases linked by each algorithm, and in combination, are shown in Table 2. For melanoma, the estimated PPV was 99.9% for the probabilistic algorithm and 100% for the deterministic algorithm if cases identified by both algorithms were assumed true matches. For lung cancer, the respective PPV estimates were 99.8% and 100%.

The probabilistic algorithm has a lower PPV than the deterministic algorithm for both cohorts. Although the sensitivity of each of the 2 algorithms cannot be estimated without reviewing some cases missed by both, their sensitivity can be compared using McNemar's test, which is based on only the counts in the discrepant cells (ie, matched by 1 algorithm but not the other). The McNemar's test was highly significant for both cancer cohorts because there were many more cases found in 1 cell (identified by deterministic, but not probabilistic algorithm) than found in

Table 2. Numbers of Melanoma and Lung Cancer Cases in Indiana State Cancer Registry Declared Matched to Indiana Network for Patient Care (Numbers Sampled for Manual Review)

Probabilistic Algorithm		*Deterministic Algorithm*	
		Match	**No Match**
Melanoma	MRN/ Probabilistic Match	5,894 (0)	94 (72)
	MRN/ No Probabilistic Match	23,312 (200)	29,258 (0)
Lung Cancer	MRN/ Probabilistic Match	22,198 (0)	292(165)
	MRN/ No Probabilistic Match	3,944 (200)	4,903 (0)

MRN, medical record number.

the other (identified by probabilistic, but not deterministic algorithm).

Completeness of the ISCR

A search for melanoma administrative codes in INPC with a first diagnosis date between January 1, 2005 and December 31, 2012 yielded 9,043 cases, 3,083 (34.1%) of which were found in the ISCR. Among the 5,960 cases that did not link to the ISCR, a chart review of INPC data from a random sample of 199 patients with any text report data was undertaken; this chart review was intended to determine whether the INPC was identifying patients who should have been found in the cancer registry, or if the patients were incorrectly identified as having cancer by the INPC. Of the 199 patients, 44 (22%) were confirmed as true incidence cases in the time period. Therefore, the estimated capture rate of melanoma by the ISCR was 84%.

A search for lung cancer administrative codes in INPC over the same time period yielded 21,259 lung cancer cases, 13,593 (63.1%) of which were found in the cancer registry. Of 200 charts reviewed from the patients not identified in the ISCR, only 15 (7.5%) were confirmed as true incident cases, leading to an estimate of 98% completeness of the ISCR.

To further investigate true cancer cases not captured by the ISCR, 78 unique melanoma cases were delivered to the ISCR for manual review, 39 of which were truly not captured in the ISCR, rather than a failure of the linkage algorithm. When 74 validated INPC lung cancer cases not linked to ISCR were investigated by the ISCR, only 14 were not found to be independent patients in the ISCR, and 3 cases represented disagreements (2 were coded as the incorrect cancer and 1 was a lung metastasis all on the INPC side).

Discussion

The state population-based data linkage described in this project represents an uncommon linkage of state cancer registry (ISCR) cases with federated EHR data from the INPC, an regional health information organization (RHIO). Prior population-based cancer registry linkages have commonly involved the use of insurance claims data, either public[9] or private.[10] Compared to linkages with Medicare claims focused upon older populations, the linkage of a state cancer registry with EHR data leverages longitudinal, electronic data which documents care delivered to all of the general population served by several community-based health care institutions. Therefore, EHR data linkages hold the promise of generating knowledge about cancers more common in younger populations, eg, testicular cancer, thyroid cancer, lymphoma, and leukemia. Compared to administrative claims, EHR data also has the potential to provide more clinically detailed information, such as the results of lab or imaging tests, than the event-based billing information available in insurance claims. For this reason, it has been proposed that quality measures should preferably be based upon clinical data from EHRs, rather than administrative claims.[11]

The overall match rate of 88.5% discovered here is encouraging, suggesting that information about longitudinal, follow-up care may be ascertained among a significant proportion of cancer patients shared between the ISCR and INPC. Based upon these findings, this Indiana state-based partnership will continue moving forward to explore how this data resource can best be implemented to meet the cancer control, policy, and health services research needs of the state's population. A growing number of RHIOs exist throughout the United States, [12] and we recommend that other state departments of health and cancer registry programs explore the possibilities for collaboration with local partners. While both the population reach, and clinical functionality, of Health Information Exchanges (HIEs) will vary geographically[13]; currently, the opportunities made possible by cancer registry-EHR/HIE linkages in the field of cancer control are numerous.

Cancer control covers the continuum of care from prevention to end-of-life care. Given the complementary nature of cancer registry and EHR data, merging these 2 data repositories has the potential to create a unique resource for many types of epidemiologic studies and clinical research topics. The EHR data, again, offers a longitudinal perspective that enables the ascertainment of services before, during, and after cancer diagnosis. Clinical data before diagnosis can be used to measure functional status and comorbidities[14] that might influence treatment decisions, while after initial treatment, the EHR data can be used to evaluate adjuvant or chronic treatment, surveillance procedures, and long-term outcomes, such as anticipated and unanticipated late effects of cancer treatment.[15]

Trade-offs existed in the choice between the deterministic and probabilistic algorithms. While the probabilistic algorithm identified more matches than the deterministic algorithm across the complete cancer cohort, the deterministic algorithm had a higher PPV than the probabilistic algorithm. One contributing factor to the difference may be a higher duplicate rate associated with probabilistic approaches.[16] Ultimately, the pragmatic decision was made to implement both deterministic and probabilistic

algorithms together, as the PPV associated with both was quite high. Even this small degree of error may be unacceptable for clinical uses; but for the purpose of longitudinal, epidemiologic cancer control studies, this threshold is still determined to be reasonable.

Reporting to the ISCR had a higher completion rate for lung cancer compared to melanoma, based upon the additional cases identified in the INPC. Prior study has reported that timeliness of reporting varies by cancer type.[17] Such variation may be explained by the fact that different types of cancers are more likely to be diagnosed in different health care settings. Similar to timeliness, completeness may also be influenced by the reporting institution. Cancers diagnosed in hospitals may more often be reported to the state cancer registry than those diagnosed in non-hospital settings. Specifically, lung cancer is more likely to be not only diagnosed, but treated, in hospitals than melanoma,[17] and thus, may have more complete reporting than melanoma which is more likely to be diagnosed in independent laboratories or physicians' clinics.

These findings confirm that administrative data alone (in this case, from the RHIO) has a limited ability for cancer case identification due to high false positive rates, reinforcing that ICD-9 data should not be used as a standalone approach.[18] In fact, the INPC does not have access to comprehensive administrative data from any single insurance source, further limiting its potential for case identification. Among the INPC cases that could not be linked to the ISCR, only a small proportion (22% for melanoma and 7.5% for lung cancer) could be validated as true cancer incident cases during the study period. The state cancer registry data still serves a vital function in the identification of incident cases (which is not possible from claims data) with detailed site and staging information. Data from EHRs are unlikely to further enrich the state cancer registry with information about previously unrecognized incident cancer cases without the addition of natural language processing abilities across an adequate supply of clinical documents.[19]

Conclusion

In summary, it is concluded that by linking the ISCR with the INPC, the ISCR is able to identify missing cancer cases. Although the accuracy of the ISCR is high, identification of any missing cases adds value to the overall accuracy of the ISCR, and it ensures proper incidence and mortality can be assessed and targeted approaches for cancer control can be implemented across the state. One can also ascertain that for epidemiological studies based on large databases such as a HIEs and EHRs, case identification using cancer registries that can be linked to EHRs will provide definitively diagnosed cancer cases with the added advantage of rich data on treatment, disease progression, and outcomes.

Most, but not all, patients with specific cancers identified by ICD-9 codes in the INPC could be linked to the ISCR. Among those who could not be linked, about half were found to be false negatives from the registry perspective, ie, a cancer was present based on manual review of their EHRs in INPC. The public health importance of this approach is significant. The potential of a HIE to capture cancer cases in real time, especially cases that are not otherwise identified by the state cancer registry, suggests future models for disease surveillance using EHR data.

References

1. Indiana Code § 16-38-2-1 (1985): Cancer registry; establishment.
2. Davis F, Nagamuthu C, Ross J, Megyesi J. Current status of brain tumor surveillance in Canada and why it matters. *J Registry Manage.* 2015;42:139-145.
3. White-Gilbertson S, Nelson S, Zhan K, Xiao C, Cope L, Day T. Analysis of the National Cancer Database to describe treatment trends in Stage IV oral cavity and pharyngeal cancers in the United States. *J Registry Manage.* 2015;42:146-151.
4. Sandar M, Hsiang LG, Yew CK, Guat LB. Use of population-based cancer registry data to determine the effect of timely treatment on the survival of colorectal cancer patients. *J Registry Manage.* 2015;42:130-138.
5. Fellegi IP, Sunter SB. A theory for record linkage. *J Am Stat Assoc*. 1969;40:1183-1210.
6. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp*. 2002;305-309.
7. Dusetzina SB, Tyree S, Meyer AM, et al. *Linking Data for Health Services Research: A Framework and Instructional Guide*. Rockville, MD: Agency for Healthcare Research and Quality; 2014.
8. Rosenman M, He J, Martin J, et al. Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory. *J Am Informatics Assoc*. 2014;21:345-352.
9. Engels EA, Pfeiffer RM, Ricker W, Wheeler W, Parsons R, Warren JL. Use of surveillance, epidemiology, and end results–medicare data to conduct case-control studies of cancer among the US elderly. *Am J Epidemiol.* 2011;174:860-870.
10. Doebbeling BN, Wyant DK, McCoy KD, et al. Linked insurance-tumor registry database for health services research. *Med Care*. 1999;37:1105-1115.
11. Tang PC, Ralston M, Arrigotti MF, Qureshi L, Graham J. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Informatics Assoc*. 2007;14:10-15.
12. Adler-Milstein J, Bates DW, Jha AK. Operational health information exchanges show substantial growth, but long-term funding remains a concern. *Health Aff (Millwood)*. 2013;32:1486-1492.
13. Adler-Milstein J, Bates DW, Jha AK. A survey of health information exchange organizations in the United States: implications for meaningful use. *Ann Intern Med.* 2011;154:666-671.
14. Klabunde CN, Legler JM, Warren JL, Baldwin LM, Schrag D. A refined comorbidity measurement algorithm for claims-based studies of breast, prostate, colorectal, and lung cancer patients. *Ann Epidemiol.* 2007;17:584-590.
15. Doyle JJ, Neugut AI, Jacobson JS, Grann VR, Hershman DL. Chemotherapy and cardiotoxicity in older breast cancer patients: a population-based study. *J Clin Oncol.* 2005;23:8597-8605.
16. Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. *J Am Med Informatics Assoc.* 2014;21:97-104.
17. Smith-Gagen J, Cress RD, Drake CM, Felter MC, Beaumont JJ. Factors associated with time to availability for cases reported to population-based cancer registries. *Cancer Causes Control.* 2005;16:449-454.
18. Baldi I, Vicari P, Di Cuonzo D, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. *J Clin Epidemiol.* 2008;61:373-379.
19. D'Avolio LW, Nguyen TM, Farwell WR, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Informatics Assoc.* 2010;17:375-382.