Automated Extraction of Date of Cancer Diagnosis from EMR Data Sources

Jeremy L. Warner, M.D., M.S.[1,2], Lucy Wang B.S.[3], Ravi Atreya B.S.[2], Pam Carney R.N., M.S.N.[3], Joe Burden, B.S.[3], Mia A. Levy, M.D., Ph.D.[1-3]

[1]Department of Medicine, Division of Hematology & Oncology, Vanderbilt University, Nashville, TN; [2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; [3]Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN

Abstract

A diagnosis of cancer is a feared event with life-changing repercussions. However, there are many different types of cancer and the emergent phenotype for any one cancer type can be vastly different between patients. The basis for describing cancer phenotypes is outcome, whether measured as survival or a surrogate parameter such as progression-free survival. To date, the estimation of survival for large unselected groups of patients has been hampered by the fact that simple information such as the date of diagnosis is not recorded in a structured format in most electronic medical record (EMR) systems. We describe here an approach to estimate date of cancer diagnosis through the use of structured and unstructured components of a comprehensive EMR, and compare those results to a gold standard for date of cancer diagnosis. Our preliminary results are encouraging and indicate that this algorithm could be utilized to estimate outcome measures on large populations of cancer patients.

Background and Significance

Cancer is an umbrella term for a heterogeneous set of disease phenotypes with widely varying prognoses and treatments. Many national and international efforts, the most well-known of which is the American Join Committee on Cancer Tumor-Node-Metastasis (TNM) Staging System (1), have enabled the discretization of individual prognosis into well-defined subgroups that can be tracked in a prospective fashion. In order to accurately evaluate any particular cancer population for clinical, research, financial, operational, and/or quality reporting purposes, it is necessary to capture detailed information about an individual patient's "cancer journey" (2). This journey, which includes prevention, screening, diagnosis, treatment, survivorship, and end-of-life care, has no defined beginning. However, by convention, the date of diagnosis is considered a key event for the reporting of outcomes, whether these be quality measures such as time to first surgery or outcome measures such as progression-free and overall survival. Currently, much of the relevant information, including the date of diagnosis, is locked within the narrative portions of the electronic medical record (EMR).

While date of diagnosis is one of the key elements that tumor registrars identify, not all cancer patients are analytic cases, which are defined in Tennessee for example as "cases diagnosed at the accessioning facility and/or administration of any of the first course of treatment after the [tumor] registry's reference date" (3). Similar definitions apply across the North American Association of Central Cancer Registries (NAACCR) sites. At tertiary care centers, much cancer care is delivered at the time of relapse or progression, and there is generally no legal obligation on the part of tumor registrars to capture detailed information on these patients. Additionally, the process of manual abstraction of date of diagnosis can be quite tedious and may not be entirely accurate. In prior work, we have shown that inter-annotator agreement (IAA) between two blinded clinical experts is only 61% for date of diagnosis, with discrepancies ranging from -100 days to +365 days on a small sample (N=49) of patients (4).

EMRs may offer an attractive solution for the automated extraction of date of cancer diagnosis. EMRs aggregate a wide variety of structured and unstructured data sources which may include the needed information. These include diagnosis codes, procedure codes, clinical documents, and the metadata associated with these documents. Here, we describe and evaluate a heuristic algorithm for determining date of cancer diagnosis that utilizes a variety of these data sources at our institution, the Vanderbilt University Medical Center (VUMC).

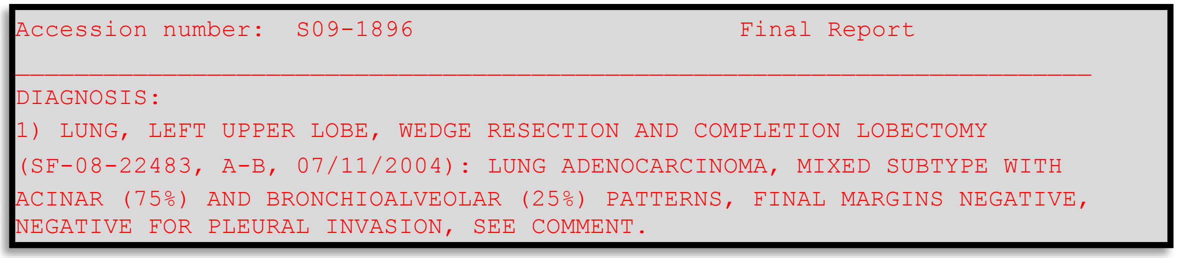
Methods

Algorithm Description: The date of diagnosis algorithm was developed through an iterative process on a population of approximately 3000 cancer patients who underwent molecular profiling as part of their clinical care at VUMC, as of April 30th 2014. This population was selected because the molecular profile tests are very specific to cancer subtypes and we could thus reliably assume that the patients of interest definitely had a cancer diagnosis. The data source used for this analysis was the VUMC Research Derivative (RD), an identifiable database of clinical and related data derived from VUMC's clinical information systems and restructured for research and quality programs (5). The RD contains diagnosis, treatment, demographic, and outcome data that are recorded in structured, semi-structured, or free text fields, including data provided by the VUMC Tumor Registry (TR). While the dates and titles of documents scanned from outside institutions, e.g. outside pathology reports, are retained, the scanned images are not included in the RD. This research was determined to be exempt by the Vanderbilt University Institutional Review Board (IRB #131613 and #140697); all authors with access to data had the appropriate HIPAA training. The algorithm is described below in pseudocode, and thereafter in the text:

```
Pseudocode for Date of Diagnosis Algorithm

Definitions
1. First VUMC Encounter Date (First_VUMC_Enc_Date) = least of the following dates:
    • First CPT code date
    • First completed outpatient or inpatient encounter date
    • First ICD-9-CM code (any) date
2. First Cancer Diagnosis Code Date (First_Cancer_ICD_Date):
    • First ICD-9-CM date of a candidate ICD-9-CM code (140-209 inclusive)
3. First Pathology Report Date (First_Path_Report_Entry_Date):
    • First Pathology Report entry date that has at least one other note within +-14 days
4. Outside Pathology Review Date (First_Clin_Notes_Accession_Date)
    • First "Accession number and dates" found in Clinical Lab notes

Algorithm

STEP 1: Patients diagnosed at VUMC
If First_Cancer_ICD_Date is at least 45 days later than First_VUMC_Enc_Date, then the patient is considered
to have been diagnosed at VUMC. In this case, First_Cancer_ICD_Date is used as date of cancer diagnosis
(Date_of_Diagnosis). If First_Cancer_ICD_Date does not exist, return Date_of_Diagnosis as NULL.

STEP 2: Patients diagnosed elsewhere
If First_Cancer_ICD_Date is within 45 days of First_VUMC_Enc_Date, then the patient is considered to have
been referred to VUMC after a primary diagnosis elsewhere.
1.  If both First_Path_Report_Entry_Date and First_Clin_Notes_Accession_Date exist, then use earlier of two
    as Date_of_Diagnosis.
2.  If only First_Path_Report_Entry_Date exists, then use First_Path_Report_Entry_Date as
    Date_of_Diagnosis.
3.  If only First_Clin_Notes_Accession_Date exists, then use First_Clin_Notes_Accession_Date as
    Date_of_Diagnosis.
4.  If neither of these two dates are available, then use First_Cancer_ICD_Date as Date_of_Diagnosis.
5.  If First_Cancer_ICD_Date does not exist, return Date_of_Diagnosis as NULL.
```

The first task of the algorithm was to determine whether the cancer was diagnosed at VUMC or elsewhere. This was done by comparing the date of first encounter at VUMC to the date when a candidate International Classification of Diseases, Clinical Modification (ICD-9-CM) code was recorded. The candidate codes are from ICD-9-CM Chapter Two: Neoplasms (140-239), excluding codes for benign neoplasms, carcinoma *in situ*, neoplasms of uncertain behavior, or neoplasms of unspecified nature (210-239). The date of first encounter was the first occurrence of any Current Procedural Terminology (CPT®) code, or the date of the first completed encounter as recorded in the outpatient or inpatient scheduling systems (not all clinical visits are billed). Patients with at least 45 days elapsed between the first encounter date and the first candidate ICD-9-CM code date were considered to have established care at VUMC and to thus have been diagnosed at VUMC.

From this point, the algorithm branched down two paths. For the "VUMC-diagnosed" patients, the date of the first candidate ICD-9-CM code was used as a proxy for date of diagnosis. For the others, a more complicated evaluation was required. This included analyzing the metadata of outside pathology reports, which are scanned into the chart and thus not themselves accessible. If an outside pathology report was dated within +/- 14 days of other reports (e.g. radiology, clinical documents) then it was considered a candidate report; otherwise, it may have been a report that was included in a large document transfer (e.g. pathology from a benign colonoscopy is included in a lung cancer patient's file because "all pathology reports" are requested at the time of referral to VUMC). It is also customary for the outside pathology to be reviewed by VUMC pathologists, at which point a pathology consultation report is generated. These reports, when present, are dated to the date that they are reviewed, but usually contain a stereotyped reference to the original accession number and date, which are amenable to regular expression extraction. An example is shown in **Figure 1** with details of this sub-algorithm described in the figure caption.



```
Accession number:  S09-1896                      Final Report
_________________________________________________________________________
DIAGNOSIS:
1) LUNG, LEFT UPPER LOBE, WEDGE RESECTION AND COMPLETION LOBECTOMY
(SF-08-22483, A-B, 07/11/2004): LUNG ADENOCARCINOMA, MIXED SUBTYPE WITH
ACINAR (75%) AND BRONCHIOALVEOLAR (25%) PATTERNS, FINAL MARGINS NEGATIVE,
NEGATIVE FOR PLEURAL INVASION, SEE COMMENT.
```

Figure 1: An example of a consultative pathology report with original date of diagnosis (07/11/2004). This date is extractable with a regular expression. Accession numbers and dates have been altered to preserve anonymity.

Evaluation Methodology: In order to evaluate the algorithm, we undertook two independent validations. In the first, approximately 2.5% of identified charts in the training cohort were randomly selected for quality assurance (QA) review. All reviewed charts underwent manual abstraction by at least two abstractors with clinical subject matter expertise, and binary IAA (agree vs. disagree) was calculated by Cohen's kappa (κ) (6). Any disagreements in manual abstraction were adjudicated by a third abstractor, with persistent discrepancies resolved by discussion between the three abstractors. In the second validation, the manually abstracted diagnosis dates of 1500 randomly selected patients with a single entry in the VUMC Tumor Registry NAACCR table, between the years 2009 and 2013, were compared to the calculated diagnosis dates. The date range was chosen to cover a five-year period that ended more than six months ago, as the TR is permitted to lag in manual abstraction by up to six months. The patients were restricted to those with a single entry so as to exclude patients with metachronous cancers (multiple primaries occurring at intervals). Tumor site and histology information was also extracted from the NAACCR table.

Results

The algorithm was manually evaluated on 75 randomly selected charts during its iterative development. This process led to several improvements in the algorithm. IAA of the chart abstractors for date of cancer diagnosis was κ=0·79; all but one discrepancy were resolved by adjudication. In the one unresolved case, all abstractors could agree on a date of diagnosis within a 14-day interval. The median absolute discrepancy between the manually and algorithmically determined dates of diagnosis was two days (IQR, zero to 260 days).

The algorithm was then evaluated on the much larger sample from the TR. The randomly selected patients represented a variety of tumor sites and histologies, as shown in **Table 2**. All patients had at least one ICD-9-CM code (any) and 1399 (93%) had at least one ICD-9-CM code in the candidate list for cancer codes. All but one patient had a completed encounter, and all patients had at least one CPT® code – such that the *First VUMC Encounter Date* could be calculated for all eligible patients. We were thus unable to calculate a date of cancer diagnosis for 101 patients (7%); manual review of a portion of these indicated that they had benign cerebral neoplasms or *in situ* neoplasms, which are reportable by the TR but not considered malignant cancers. 867/1399 patients (62%) had a calculable *First Pathology Report Date*, and 610 patients (44%) had a calculable *Outside Pathology Review Date*.

Table 2. Five most common primary tumor sites and tumor histologies in the TR sample. No one tumor site or tumor histology dominates this cohort, but common sites and histologies are well-represented.

Primary Tumor Site	**Patients, N (%)**	**Tumor Histology**	**Patients, N (%)**
C619: Prostate	220 (16)	814: adenocarcinoma	359 (26)
C421: Lymphoma	86 (6)	807: squamous cell carcinoma	156 (11)
C341: Lung	79 (6)	801: carcinoma, NOS	70 (5)
C649: Kidney	67 (5)	831: clear cell adenocarcinoma	53 (4)
C504: Breast	55 (4)	850: duct carcinoma	50 (4)
Other (171 sites)	892 (64)	Other (100 histologies)	711 (51)
Total	**1399 (100)**	**Total**	**1399 (100)**

The calculated date of cancer diagnosis agreed with the TR-designated date of cancer diagnosis with a relative discrepancy of median 0 days (IQR 0 days to +14 days). 579 patients (41%) were exact matches. There were a number of outliers, with 15.5%, 11%, and 8% of the population having an absolute discrepancy of more than 90 days, 180 days, and one year, respectively. A histogram of the date discrepancies within +/- one year of the referent TR-designated date of diagnosis is shown in **Figure 2**.

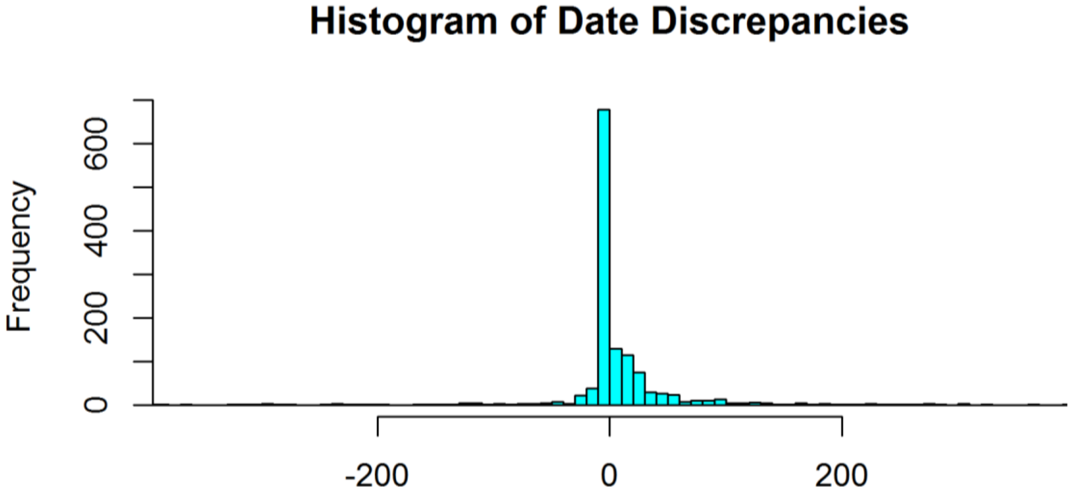


Figure 2: Histogram of date discrepancies, in days, for calculated date of cancer diagnosis, as compared to tumor registry date of diagnosis. Positive deflections represent a calculated date that is later than the tumor registry date. Outliers beyond +/- 365 days (one year) are not shown.

We manually reviewed some of the 63 extreme outliers (defined as a greater than 1000-day discrepancy). The results are summarized in **Table 3**. Four of the eight were identified by *First Cancer Diagnosis Code Date* and four by *First Pathology Report Date*. For the four identified by ICD-9-CM, two were the result of incorrect usage of ICD-9-CM codes, and two were due to skin cancer codes (non-melanoma skin cancers are not recorded by the TR although technically these patients therefore had more than one malignancy). For the four identified by pathology, one (patient #1) had a benign diagnosis, one (patient #5) had a second cancer which was not captured as an analytic case due to its diagnosis and treatment elsewhere, and two had matching tumor sites & histologies with differing dates. For these two cases, the TR recorded the recurrence as a new analytic case per the NCI SEER 2007 Multiple Primary and Histology Coding Rules, Rule #7 (personal communication, Judith Roberts).

Table 3. A sample (N=8) of extreme outliers.

Patient	**TR Year of Diagnosis**	**TR Diagnosis**	**Algorithm Year of Diagnosis**	**Algorithm Trigger & Evidence (Pathology or ICD-9-CM Code)**
1	2011	Prostate cancer	2001	*First Pathology Report Date*: benign lipoma
2	2012	Pilocytic astrocytoma	2005	*First Pathology Report Date*: pilocytic astrocytoma

3	2010	Gastrointestinal carcinoid	2003	*First Cancer Diagnosis Code Date*: 173.1 other malignant neoplasm of skin of eyelid, including canthus
4	2009	Myxofibrosarcoma	2006	*First Cancer Diagnosis Code Date*: 189.0 renal cell carcinoma (used clinically as a rule-out code)
5	2009	Hepatocellular carcinoma	2003	*First Pathology Report Date*: lung cancer
6	2009	Pituitary adenoma (benign)	2012	*First Cancer Diagnosis Code Date*: 194.3 malignant neoplasm of pituitary gland (used in error)
7	2013	Melanoma	2007	*First Cancer Diagnosis Code Date*: 173.3 other malignant neoplasm of skin of other and unspecified parts of face
8	2012	Lung cancer	2007	*First Pathology Report Date*: lung cancer

Conclusion and Discussion

We have demonstrated a generic approach to the determination of date of cancer diagnosis that performs well, as compared to the gold standard of tumor registrar manual abstraction, on a variety of cancer types and histologies. Our approach, when coupled with other methods of EMR data extraction such as the determination of co-morbidities (7, 8), exposure to chemotherapy regimens (9, 10) and other drugs (11), and exploration of genomic factors (12) can form the basis of a rapid-learning system for oncology (13). Tools such as the one described here may also aid quality reporting systems such as the Quality Oncology Practice Initiative (QOPI) as they transition from manual to electronic capture systems (14).

There are several limitations to our current approach. Our requirement that there elapse at least 45 days between the first encounter date and the first cancer ICD-9-CM code as a definition for an “established” patient does not account for the possibility of a patient presenting to establish care and being diagnosed simultaneously (e.g. an otherwise well individual who does not ordinarily participate in medical care presenting to the emergency room with acute leukemia). It is likely that future iterations of the algorithm could refine the identification of a patient who has established care at VUMC. Of note, there is no specific definition for the establishment of care, which can occur anytime during a continuum of illness. In fact, we have shown that tumor registrars regularly suffer a very poor positive predictive value in case identification, which leads to a large time effort merely to identify analytic vs. non-analytic cases (15).

Another limitation is the challenge of handling patients with metachronous cancer and/or recurrences. As illustrated by the manual review of extreme outliers, 25% had recurrences and 50% had multiple neoplasms (despite our efforts to exclude this possibility by restricting to patients with a single TR entry). Recurrence poses a particular challenge if the original tumor tissue is not available for review, in which case the TR will consider the presentation as a new analytic case with a new date of diagnosis (16). When cancer subtype-specific anchors such as targeted molecular testing or biomarkers are available, the algorithm can be modified to handle patients with multiple cancer diagnoses with relative ease; when this information is not available, the algorithm would have to be further modified to search for specific diagnoses if a cohort with multiple primaries were considered. Whether this algorithm could be successfully extended to other institutions would depend on the particular nature of metadata captured on outside records, as well as how pathology referrals are handled and documented. Efforts by ourselves (17, 18) and others (19, 20) to normalize and standardize clinical EMR data may ease the generalizability of the algorithm.

It is also worth noting that much of the later stages of refinement of the algorithm, as well as future work, was focused on the handling of outlying “edge” cases. These edge cases are identified through manual QA, which is by far the rate-limiting step in algorithm development. The danger of diminishing returns on investment, as well as the risk of over-fitting to the training data, are always evident. The results reported here are likely adequate for certain explorations of data but continue to require verification based on the specific task of interest. The TR-derived date of diagnosis remains the gold standard and should continue to be used, when available.

In conclusion, our algorithm for the automated extraction of date of cancer diagnosis performs reasonably well, with the exception of several significant outliers. When tumor registry date of cancer diagnosis is not available, this algorithm offers a good alternative for the estimation of diagnosis date. As a first pass, its utilization may also relieve some of the burden of manual data abstraction currently placed on tumor registrars. When used in combination with other dated endpoints, e.g. date of death or date of progression, the data extract from this algorithm can form the basis for survival analyses, which are essential for the elucidation of cancer phenotypes.

References

1. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17(6):1471-4.
2. Inman AW. The Cancer Journey. Biofeedback. 2010;38(1):24-7.
3. Tennessee Cancer Registry: Tennessee Department of Health; [cited 2014 August 23rd]. Available from: http://health.state.tn.us/TCR/.
4. Warner JL, Anick P, Drews RE. Physician inter-annotator agreement in the Quality Oncology Practice Initiative manual abstraction task. J Oncol Pract. 2013;9(3):e96-102.
5. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: The Vanderbilt approach. J Biomed Inform. 2014.
6. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296-8.
7. Cano C, Blanco A, Peshkin L. Automated identification of diagnosis and co-morbidity in clinical records. Methods Inf Med. 2009;48(6):546-51.
8. Hall WH, Ramachandran R, Narayan S, Jani AB, Vijayakumar S. An electronic application for rapidly calculating Charlson comorbidity score. BMC Cancer. 2004;4:94.
9. Levy MA, Giuse DA, Eck C, Holder G, Lippard G, Cartwright J, et al. Integrated information systems for electronic chemotherapy medication administration. J Oncol Pract. 2011;7(4):226-30.
10. Bhatia H, Levy M. Automated plan-recognition of chemotherapy protocols. AMIA Annu Symp Proc. 2011;2011:108-14.
11. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc. 2014.
12. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417-28.
13. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-learning system for cancer care. J Clin Oncol. 2010;28(27):4268-74.
14. Campion FX, Larson LR, Kadlubek PJ, Earle CC, Neuss MN. Advancing performance measurement in oncology: quality oncology practice initiative participation and quality outcomes. J Oncol Pract. 2011;7(3 Suppl):31s-5s.
15. Naser R, Roberts J, Salter T, Warner JL, Levy M. An informatics-enabled approach for detection of new tumor registry cases. J Registry Manag. 2014;41(1):19-23.
16. Johnson C, Peace S, Adamo P, Fritz A, Percy-Laurry A, Edwards B. The 2007 multiple primary and histology coding rules. Bethesda, MD: Surveillance, Epidemiology and End Results Program, National Cancer Institute. 2007.
17. Yu P, Artz D, Warner J. Electronic health records (EHRs): supporting ASCO's vision of cancer care. Am Soc Clin Oncol Educ Book. 2014;34:225-31.
18. Warner J, Hughes KS, Krauss JC, Maddux S, Yu PP, Shulman LN et al. The Clinical Oncology Treatment Plan and Summary implementation guide: an interoperable HL7® document standard to improve the quality of cancer care. American Society of Clinical Oncology Annual Meeting 2014: Abstract 6603.
19. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. Health Aff (Millwood). 2014;33(7):1229-35.
20. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181-5.