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

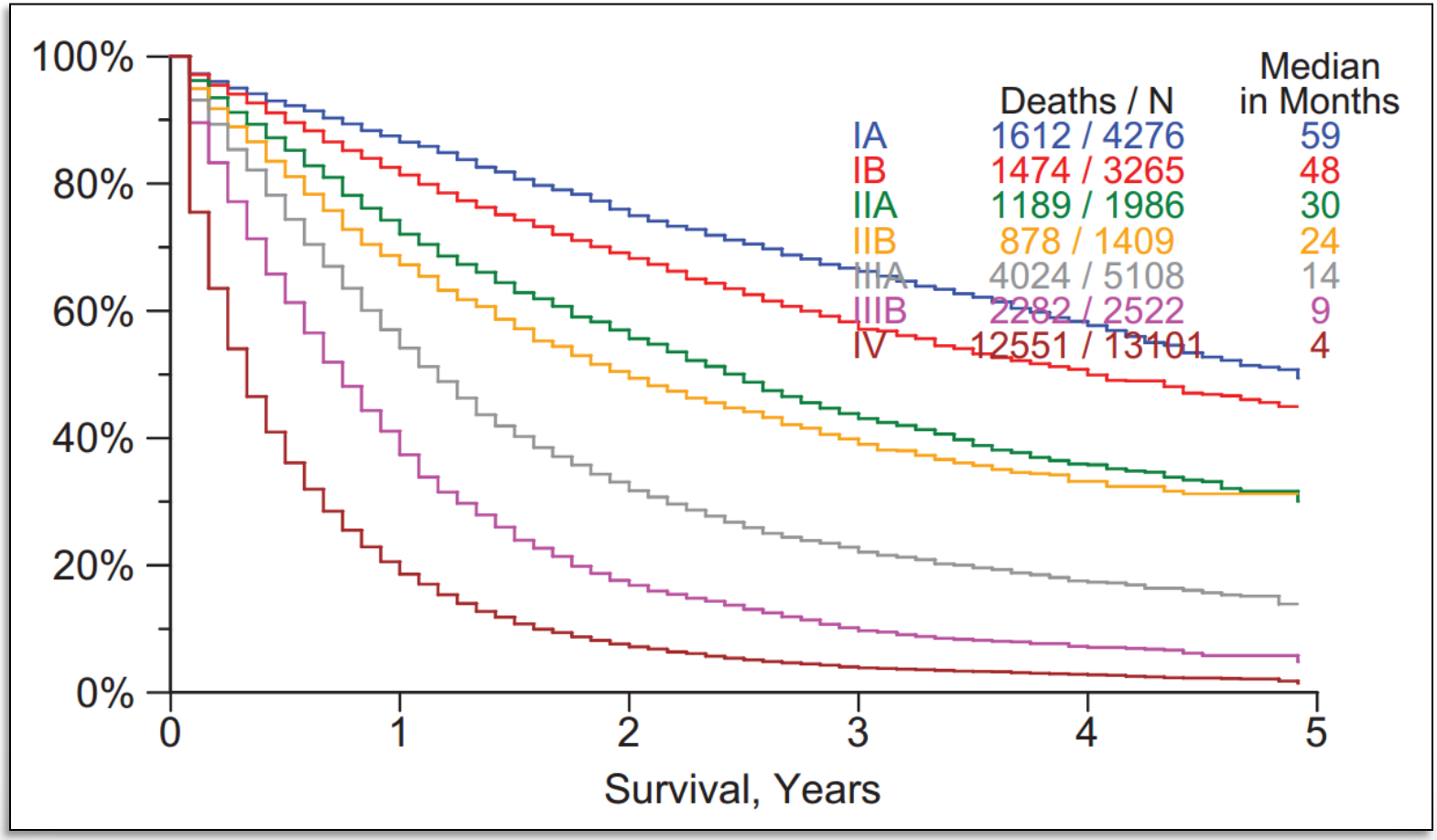
Outline

- Background
- Approach
- Results
- Future Directions
- Conclusion

Background: Why does Date of Cancer Diagnosis Matter?

- **Prognosis** from cancer diagnosis depends on:
  - Histology (type of cancer)
  - Stage at diagnosis (extent of spread)
  - Performance status at diagnosis
  - Comorbidities at diagnosis
- In all cases, prognosis is measured *relative to the date of cancer diagnosis*

Example: Lung Cancer Survival by Stage (AJCC 7th Ed.)



Groome et al. JTO 2007

Background cont.

- **Efficacy** of antineoplastic therapies is measured by:
  - Overall survival (OS)
  - Progression free survival (PFS)
  - Time to next treatment/therapy (TTNT)
  - Disease/relapse free intervals (DFS/EFS/DFI etc.)
  - Response rate (RR)
- OS and first-line measures (other than RR) are *relative to the date of cancer diagnosis*

Date of Cancer Diagnosis: Challenges

- Definition is not unambiguous
  - Do symptoms count? Abnormal imaging results?
- From the NCIthesaurus
  - New Diagnosis (Code C54731)
  - **Definition:** The unprecedent (sp) recognition of the presence of a disease, condition, or injury based on investigation, analysis or tests for signs and symptoms.
- As a result, this data element is usually not recorded in structured format in EMRs.

Challenges, cont.

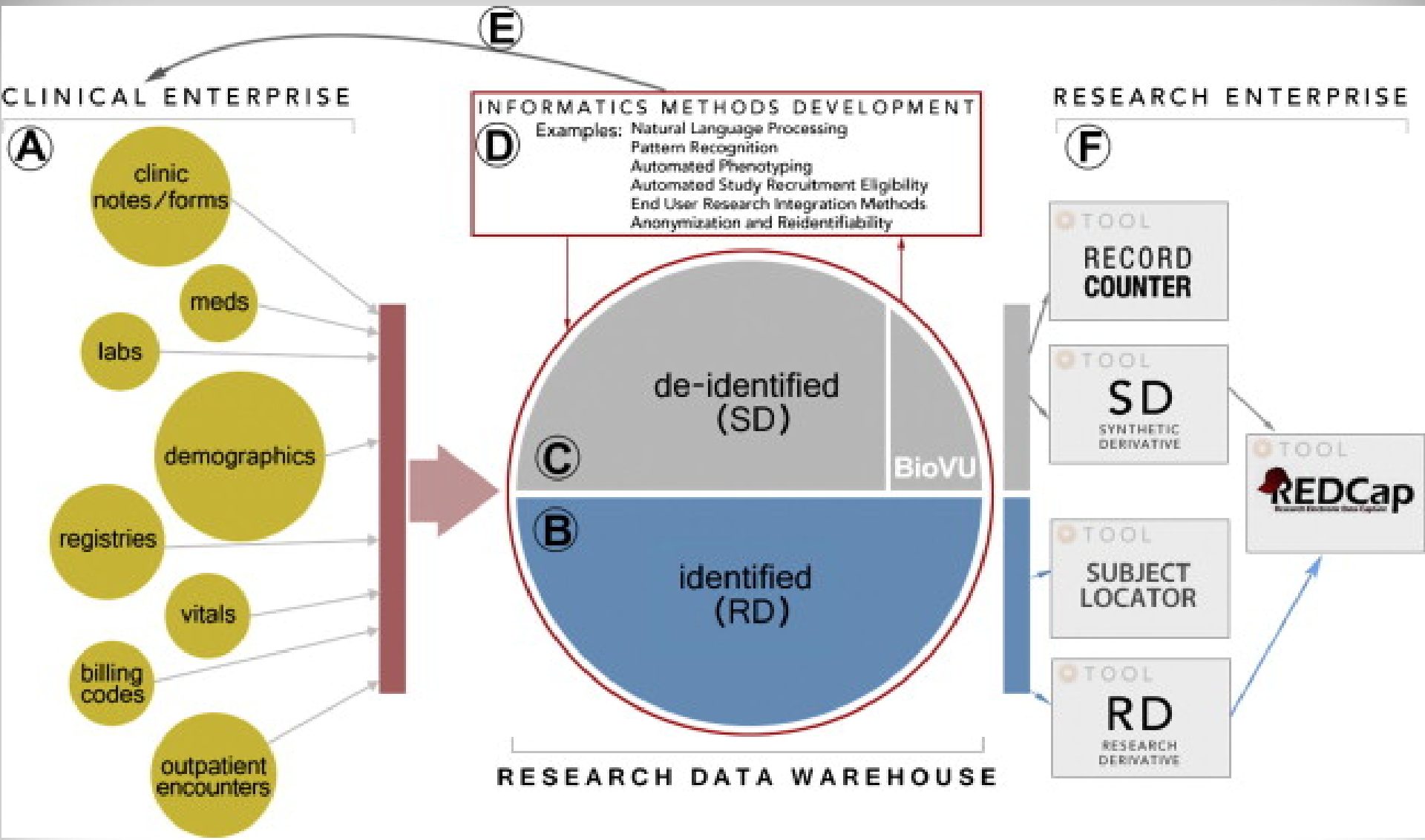
- Cancer registries record date of cancer diagnosis on a *subset of patients*, based on manual abstraction.
  - May be limited to "analytic cases" only; patients diagnosed elsewhere or long ago may not be included in the registry
  - Subject to a 6-month delay
  - Possibly prone to errors

Approach

- **Our definition:** Date of cancer diagnosis is the date that a biopsy is obtained that establishes an *unequivocal histologic diagnosis of cancer*.
- **Goal:** Accurately determine date of cancer diagnosis for all cancer patients, automatically and in real time.
- **Evaluation:** Two stage validation compares to:
  - Manual abstraction, during development
  - Vanderbilt tumor registry, after finalization

Approach, cont.

- Initial algorithm built iteratively, with a series of evaluations and improvements.
  - Data source: 3000 cancer patients (5 subtypes) who had molecular testing at Vanderbilt
  - Approx. 2.5% (75 patients) manually reviewed
- Finalized algorithm evaluated
  - Data source: 1500 randomly selected patients with a single cancer diagnosis (any subtype) recorded in the Vanderbilt tumor registry


E
CLINICAL ENTERPRISE
A
clinic notes/forms
meds
labs
demographics
registries
vitals
billing codes
outpatient encounters
INFORMATICS METHODS DEVELOPMENT
D
Examples: Natural Language Processing
Pattern Recognition
Automated Phenotyping
Automated Study Recruitment Eligibility
End User Research Integration Methods
Anonymization and Reidentifiability
de-identified (SD)
C
BioVU
B
identified (RD)
RESEARCH DATA WAREHOUSE
RESEARCH ENTERPRISE
F
TOOL
RECORD COUNTER
TOOL
SD
SYNTHETIC DERIVATIVE
TOOL
REDCap
TOOL
SUBJECT LOCATOR
TOOL
RD
RESEARCH DERIVATIVE

Data Element #1: Clinical encounter

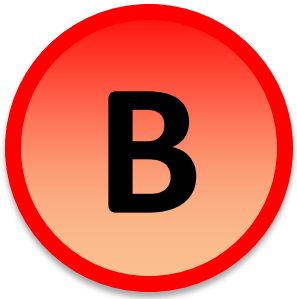
- Current Procedural Terminology (CPT®), 4th edition codes
  - 99201-99255, 99356-99357
- First completed outpatient or inpatient encounter date (from billing systems)
- First ICD-9-CM code (any)

Data Element #2: Cancer ICD-9 Code

- First ICD-9-CM code in the Neoplasms chapter (140-239), excluding
  - 210-229: Benign Neoplasms
  - 230-234: Carcinoma *In Situ*
  - 235-238: Neoplasms Of Uncertain Behavior
  - 239-239: Neoplasms Of Unspecified Nature

Problem: (at least) Two Distinct Patient Populations

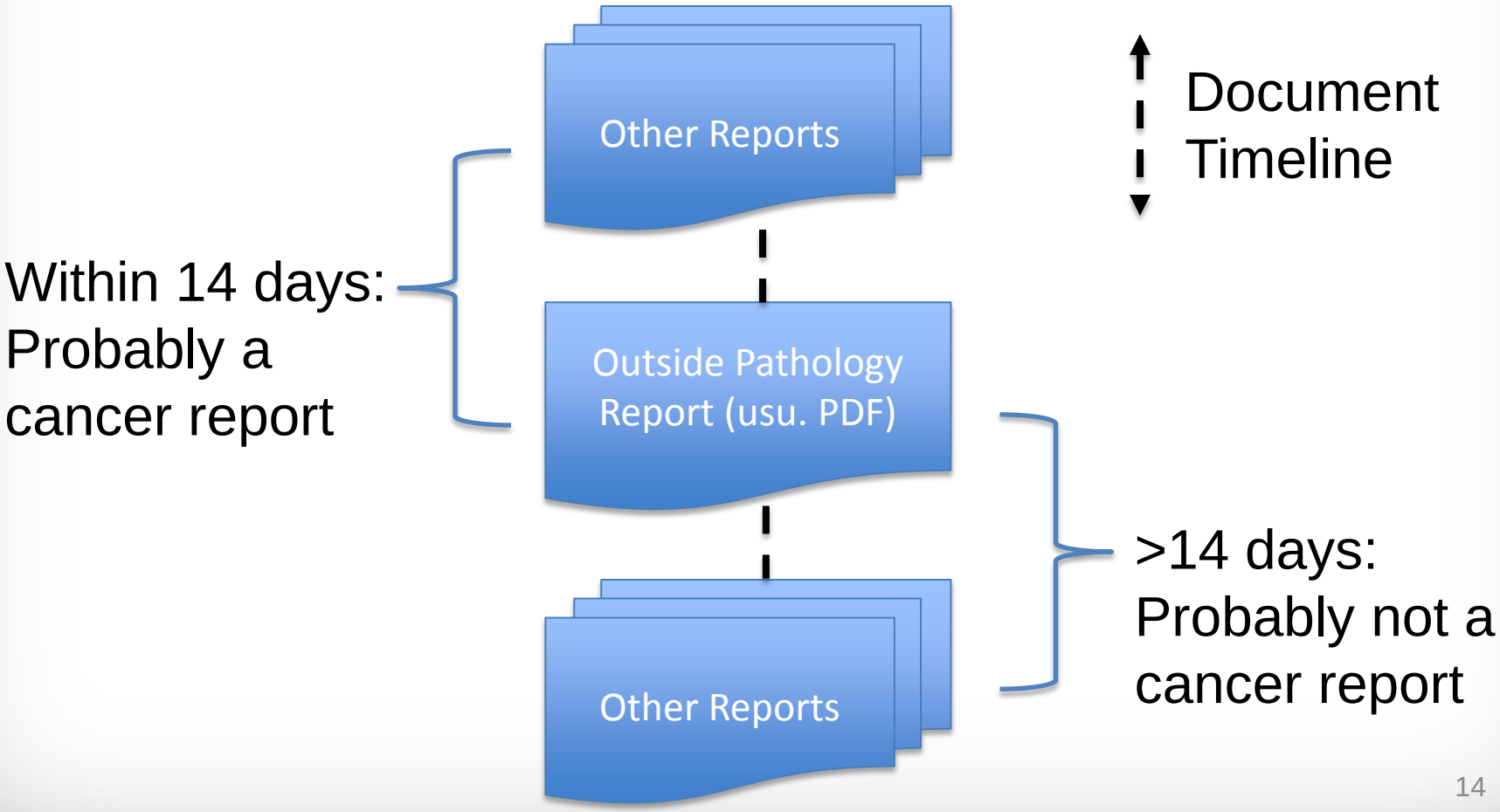
Patients diagnosed and treated at Vanderbilt (“established”).



Patients diagnosed elsewhere and treated at Vanderbilt (“referred”).

Heuristic: If first cancer ICD-9 code occurs at least 45 days after first encounter, assign to Group A.

Data Element #3: First Pathology Note Date



Data Element #4: Outside Pathology Review Date

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.

Pulling this information out of consultation reports could significantly improve results.

Improving the Custom Algorithm (N=36)

	Custom Algorithm without DE #4	Custom Algorithm with DE #4
Total Absolute Deviation	10,602 days	1230 days
Median	0	0
10th percentile	0	0
90th percentile	256 days	38 days

Finalized algorithm accuracy (N=75): Median absolute discrepancy two days (IQR, zero to 260 days)

Validation of Finalized Algorithm

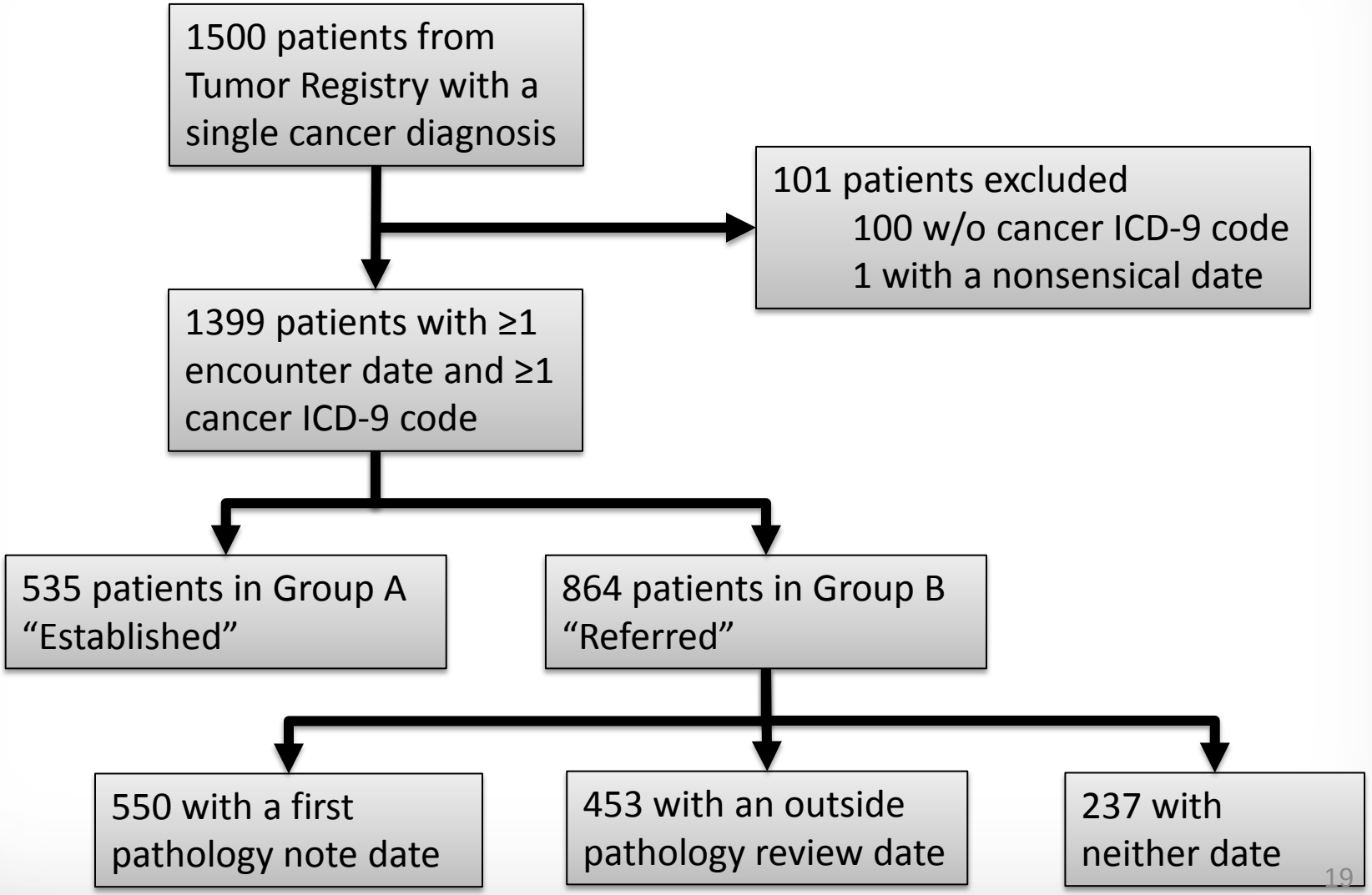
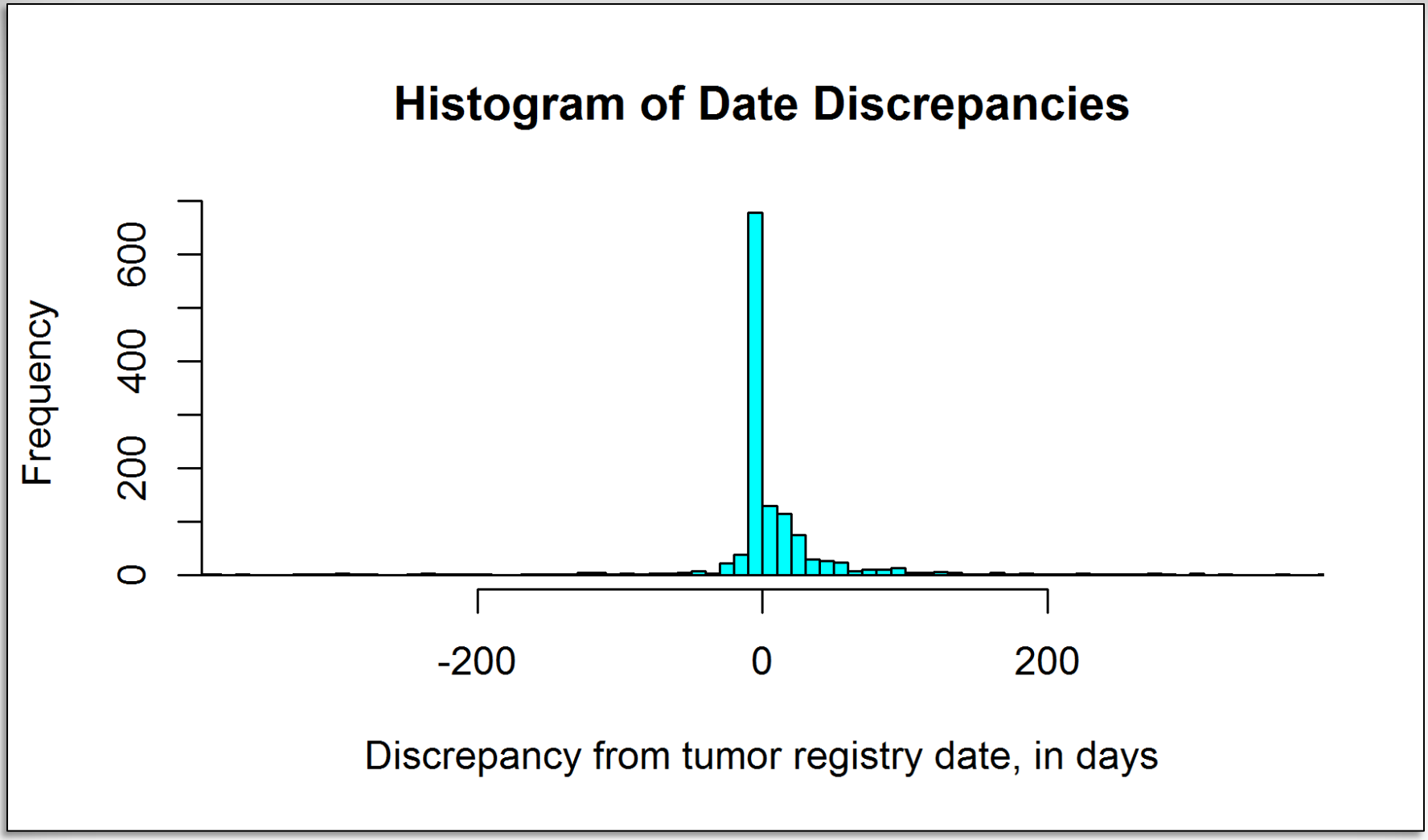


Table 2. Five most common primary tumor sites and tumor histologies in the TR sample. No one tumor site or tumor histology dominates this cohort, but common sites and histologies are well-represented.

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)**



Relative median actual discrepancy of 0 days (IQR 0 days to +14 days). 41% were exact matches.

Some Extreme Outliers (N=8)

TR Year	TR Diagnosis	Algorithm Year	Algorithm Trigger & Evidence (Pathology or ICD-9-CM Code)
2011	Prostate cancer	2001	First Pathology Report Date: **benign lipoma**
2009	Pituitary adenoma	2012	First Cancer Diagnosis Code Date: 194.3 malignant neoplasm of pituitary gland (**used in error**)
2013	Melanoma	2007	First Cancer Diagnosis Code Date: 173.3 other **malignant neoplasm of skin** of other and unspecified parts of face
2012	Lung cancer	**2007**	First Pathology Report Date: **lung cancer**

Types of Errors

- Second malignancy or recurrence present (5 of 8)
- ICD-9-CM code used incorrectly (2 of 8)
  - As a rule-out
  - In error (malignant code for benign condition)
- Algorithm captured a benign pathology report (1 of 8)

Limitations

- Heuristic does not account for a patient presenting to establish care and being diagnosed quickly (e.g. presenting to the emergency room with acute leukemia)
- Multiple cancers and recurrences remain challenging (also for registrars!)
- Chasing edge cases may lead to overfitting, limiting generalizability

Future Directions

- Secondary research on EHR cohorts of cancer patients can benefit from this algorithm.
- Along with date of death algorithm, enables **survival analysis**.
- Algorithm could assist tumor registrars in case identification and speed their abstraction process.

Conclusion

- We have demonstrated an algorithm for the automated determination of date of cancer diagnosis from EHR data.
- Performance on a large unselected cohort is very good, with median error of 0 days.
- Extreme outliers do exist, mostly due to second malignancies or recurrences.