

October 29, 2018

Data Considerations

Critical Dataset Features

Amy Abernethy, MD, PhD

Chief Medical Officer / Chief Scientific Officer & SVP - Oncology, Flatiron Health *(a member of the Roche Group)*

Adjunct Professor of Medicine, Duke University School of Medicine

@dramyabernethy ♦♦ amy@flatiron.com

Disclosures

Amy Abernethy is Chief Medical Officer, Chief Scientific Officer and SVP, Oncology at Flatiron Health, a member of the Roche Group, and has stock ownership in Roche

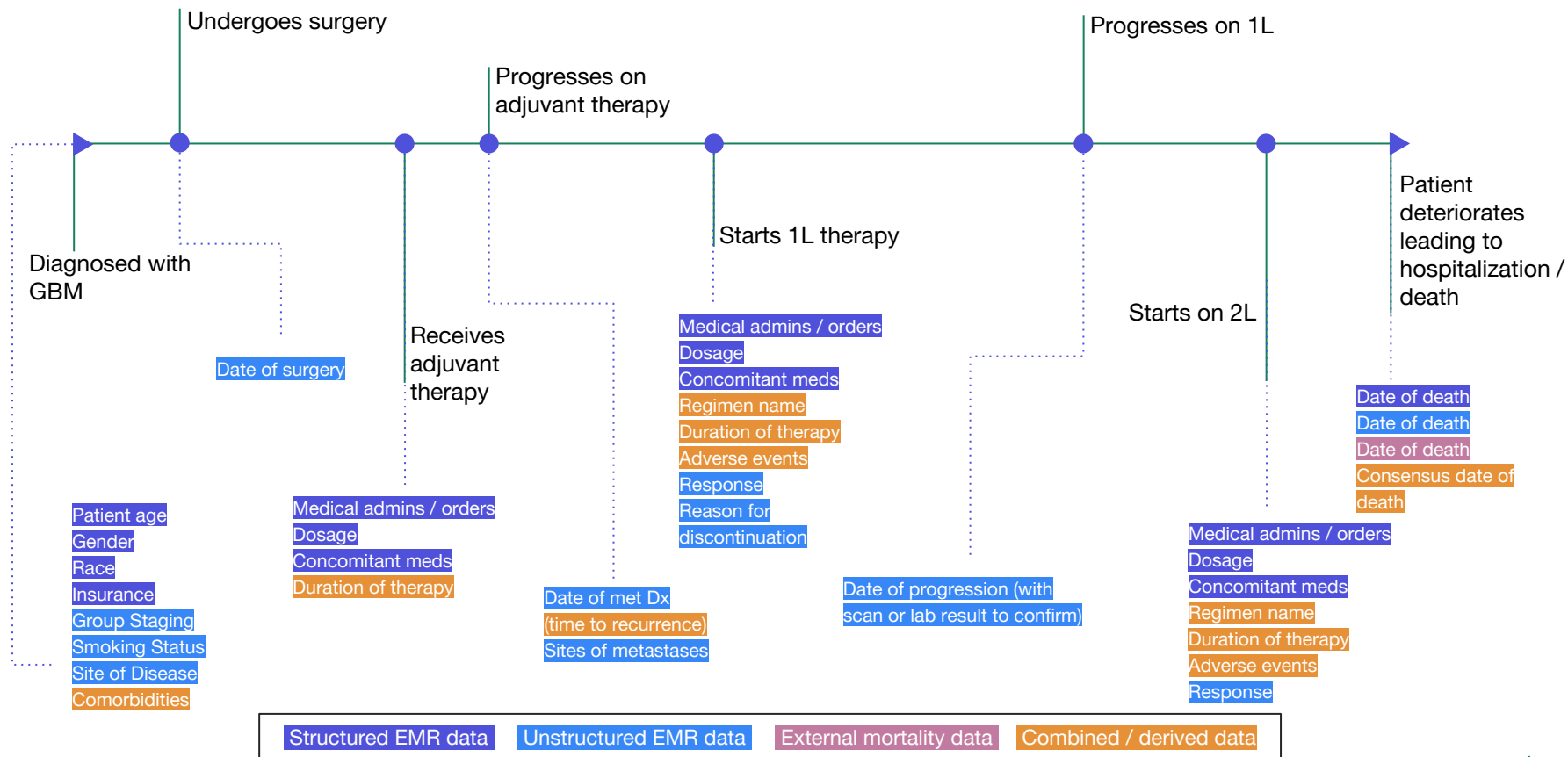
Other disclosures include:

- Athenahealth (NASDAQ: ATHN), Board of Directors, Stock ownership
- CareDx (NASDAQ: CDNA), Board of Directors, Stock ownership
- Orange Leaf Associates, LLC, Owner
- Highlander Partners, Senior Advisor (<http://highlander-partners.com/>)
- SignalPath Research, Advisor (<http://signalpath.com/>)
- The One Health Company, Special Advisor (<http://www.theonehealthcompany.com/>)
- RobinCare, Advisor (<https://getrobinicare.com/>)
- KelaHealth, Inc. - peri-surgical risk prediction, Advisor (kelahealth.com)
- Patent Pending: Current patent pending for a technology that facilitates the extraction of unstructured information from medical records.

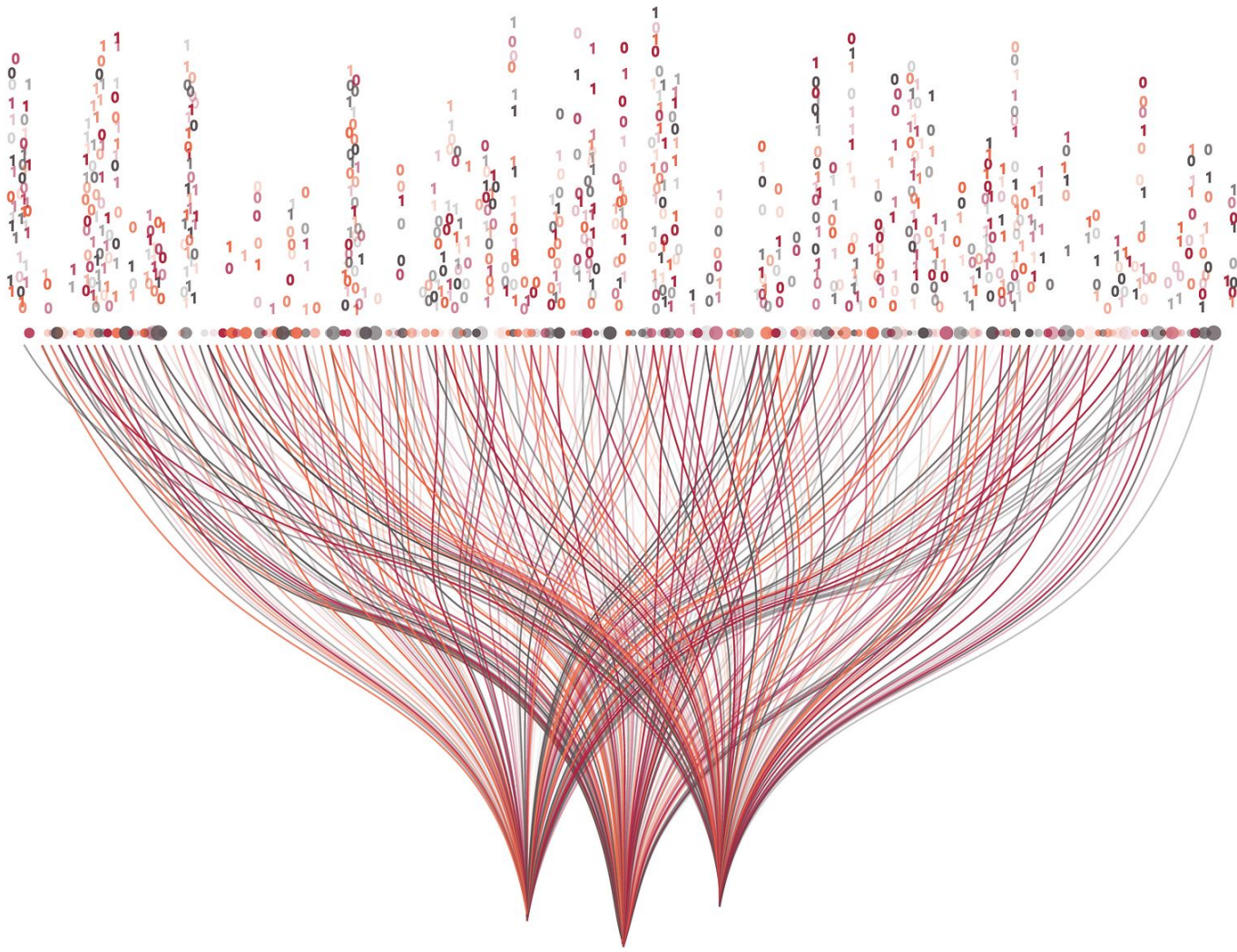
1. Importance of having all of the aspects of the data you need to create algorithms, etc

Reliable data produce reliable algorithms and reliable results

A comprehensive view of the patient journey



*Relative timing not exact



Evaluating relevancy and quality

Data Relevancy

- Availability of key data elements
 - Exposure
 - Outcome
 - Covariate
 - Patient-level linking (if applicable)
- Representativeness
- Sufficient subjects
- Longitudinality



Data Quality

- Accuracy
 - Validity
 - Conformance
 - Plausibility
 - Consistency
- Completeness
- Provenance
- Transparency of data processing

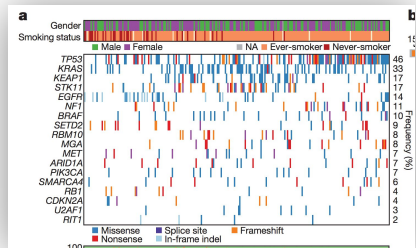


Fit-for-Purpose Data

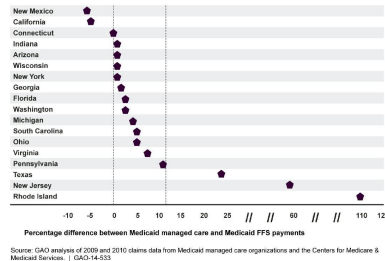
Within the given clinical and regulatory context, the real-world dataset is of sufficient quality, as well as relevant, robust, and representative.

2. Different data types have different technical, contextual and governance features which influence reliability

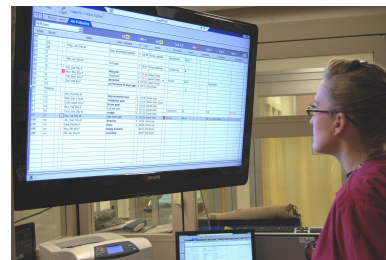
Different data types have different technical, contextual and governance features



Instrumentation data
(e.g., genomics,
imaging, immune
profiling)



Administrative data
(e.g., claims, mortality)



Longitudinal clinical
data (e.g, electronic
health records,
registries)



Patient generated data
(e.g., PROs,
biosensors)

Example: Imaging Data



- Raw images are best if available, but require massive storage capabilities and are hard to de-identify
 - Pixelated data is not the same as an interpretation of findings
 - Information needed for subsequent analyses and algorithms is the interpretation
- An alternative is to focus on radiology reports in the EHR
 - Reports must be curated into an analytical dataset
 - Quality of the interpretation varies depending on reviewer
 - Findings in routine clinical practice vary from third party independent review for clinical trials
- Adequate interpretation of scans requires information about context
- Standardized systems for proxy data from scans (e.g., RECIST)
 - Rely on a series of features that may not be available outside of clinical trials (e.g., target lesions, prior scans)
 - May miss clinical features - must understand the pitfalls of the approach

Example: Imaging Data



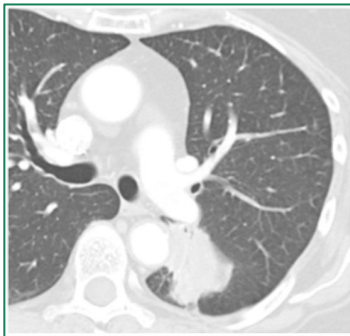
- Raw images are best if available, but require massive storage capabilities and are hard to de-identify
 - Pixelated data is not the same as an interpretation of findings
 - Information needed for subsequent analyses and algorithms is the interpretation
- An alternative is to focus on radiology reports in the EHR
 - Reports must be curated into an analytical dataset
 - Quality of the interpretation varies depending on reviewer
 - Findings in routine clinical practice vary from third party independent review for clinical trials
- Adequate interpretation of scans requires information about context
- Standardized systems for proxy data from scans (e.g., RECIST)
 - Rely on a series of features that may not be available outside of clinical trials (e.g., target lesions, prior scans)
 - May miss clinical features - must understand the pitfalls of the approach

Example: Imaging Data



- Raw images are best if available, but require massive storage capabilities and are hard to de-identify
 - Pixelated data is not the same as an interpretation of findings
 - Information needed for subsequent analyses and algorithms is the interpretation
- An alternative is to focus on radiology reports in the EHR
 - Reports must be curated into an analytical dataset
 - Quality of the interpretation varies depending on reviewer
 - Findings in routine clinical practice vary from third party independent review for clinical trials
- Adequate interpretation of scans requires information about context
- Standardized systems for proxy data from scans (e.g., RECIST)
 - Rely on a series of features that may not be available outside of clinical trials (e.g., target lesions, prior scans)
 - May miss clinical features - must understand the pitfalls of the approach

Curation of Unstructured Data



Gross Description

The specimen is received in formalin labeled with the patient's name. It consists of a 1 x 0.3 x 0.1 cm aggregate of pink-tan to red-pink soft tissue cores and fragments entirely submitted in one block.

Dictated by: GREGORY W SMITH PA

Entered: 02/06/12 - 1544 JAM

Microscopic Description

The specimen consists of a well differentiated adenocarcinoma, favor lung primary. CK7 and TTF are positive. CK20 is negative. A colleague agrees with this malignant diagnosis.

Dictated by: THOMAS J GRIFONE MD

Entered: 02/07/12 - 1423 SML

Diagnosis

SPECIMEN SUBMITTED AS TRUCUT BIOPSY LEFT LUNG NODULE:
- WELL-DIFFERENTIATED ADENOCARCINOMA, FAVOR LUNG PRIMARY.
- SEE ABOVE.

Impression:

1. Large left upper lobe bronchogenic carcinoma extending to the left hilum markedly increased in size since prior study
2. Mediastinal adenopathy increased in size particularly subcarinal space. The adenopathy in the AP window has undergone partial necrosis since previous exam.
3. Stable right apical nodularity possibly scar
4. Emphysema

6-2 → 4.2 cm
JS PET
121

Currently restaging study showed no soft tissue disease. Bone scan showed stable bony metastasis, with the exception of 1 new lesion in the left superior pubic rami. The nature of this new left superior pubic rami lesion is not clear even though it is possible this is new metastatic lesion; however, it is unusual to have an isolated progression yet rest of the bony metastasis are stable. The patient is completely asymptomatic. His tumor markers are stable; therefore, we decided to continue the current management and we will get followup studies in 3-4 months.

Example: Imaging Data



- Raw images are best if available, but require massive storage capabilities and are hard to de-identify
 - Pixelated data is not the same as an interpretation of findings
 - Information needed for subsequent analyses and algorithms is the interpretation
- An alternative is to focus on radiology reports in the EHR
 - Reports must be curated into an analytical dataset
 - Quality of the interpretation varies depending on reviewer
 - Findings in routine clinical practice vary from third party independent review for clinical trials
- Adequate interpretation of scans requires information about context
- Standardized systems for proxy data from scans (e.g., RECIST)
 - Rely on a series of features that may not be available outside of clinical trials (e.g., target lesions, prior scans)
 - May miss clinical features - must understand the pitfalls of the approach

Context: Accounting for Changing Interpretations Over Time

Impression:

1. Large left upper lobe bronchogenic carcinoma extending to the left hilum markedly increased in size since prior study
2. Mediastinal adenopathy increased in size particularly subcarinal space. The adenopathy in the AP window has undergone partial necrosis since previous exam.
3. Stable right apical nodularity possibly scar
4. Emphysema

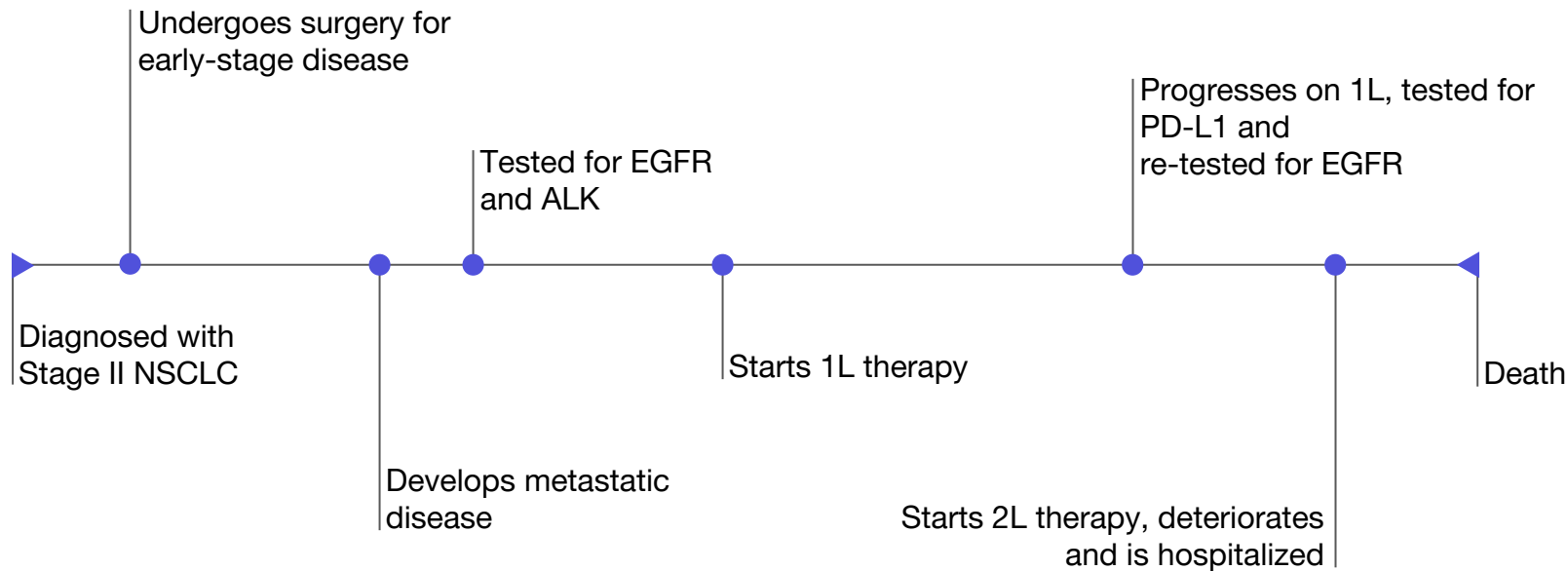
6-2
JS
4.2 cm
PE
121

Currently restaging study showed no soft tissue disease. Bone scan showed stable bony metastasis, with the exception of 1 new lesion in the left superior pubic rami. The nature of this new left superior pubic rami lesion is not clear even though it is possible this is new metastatic lesion; however, it is unusual to have an isolated progression yet rest of the bony metastasis are stable. The patient is completely asymptomatic. His tumor markers are stable; therefore, we decided to continue the current management and we will get followup studies in 3-4 months.

Example: Imaging Data



- Raw images are best if available, but require massive storage capabilities and are hard to de-identify
 - Pixelated data is not the same as an interpretation of findings
 - Information needed for subsequent analyses and algorithms is the interpretation
- An alternative is to focus on radiology reports in the EHR
 - Reports must be curated into an analytical dataset
 - Quality of the interpretation varies depending on reviewer
 - Findings in routine clinical practice vary from third party independent review for clinical trials
- **Adequate interpretation of scans requires information about context**
- Standardized systems for proxy data from scans (e.g., RECIST)
 - Rely on a series of features that may not be available outside of clinical trials (e.g., target lesions, prior scans)
 - May miss clinical features - must understand the pitfalls of the approach



Clinical events are a combination of clinical, pathological, radiological, & biomarker data - *in context*

Undergoes surgery for
early-stage disease

Tested for EGFR
and ALK

Progresses on 1L, tested for
PD-L1 and
re-tested for EGFR

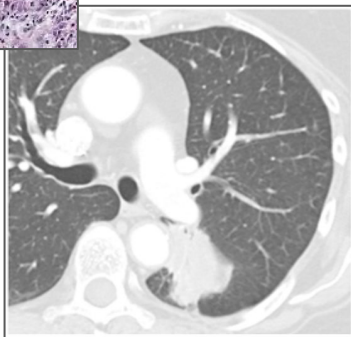
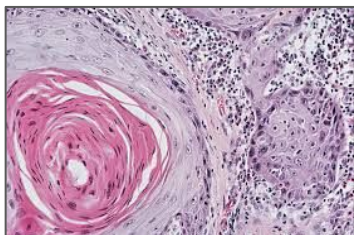
Death

Diagnosed with
Stage II NSCLC

Starts 1L therapy

Develops metastatic
disease

Starts 2L therapy, deteriorates
and is hospitalized



Gross Description

The specimen is received in formalin labeled with the patient's name. It consists of a 1 x 0.3 x 0.1 cm aggregate of pink-tan to red-pink soft tissue cores and fragments entirely submitted in one block.

Dictated by: GREGORY W SMITH PA
Entered: 02/06/12 - 1544 JAM

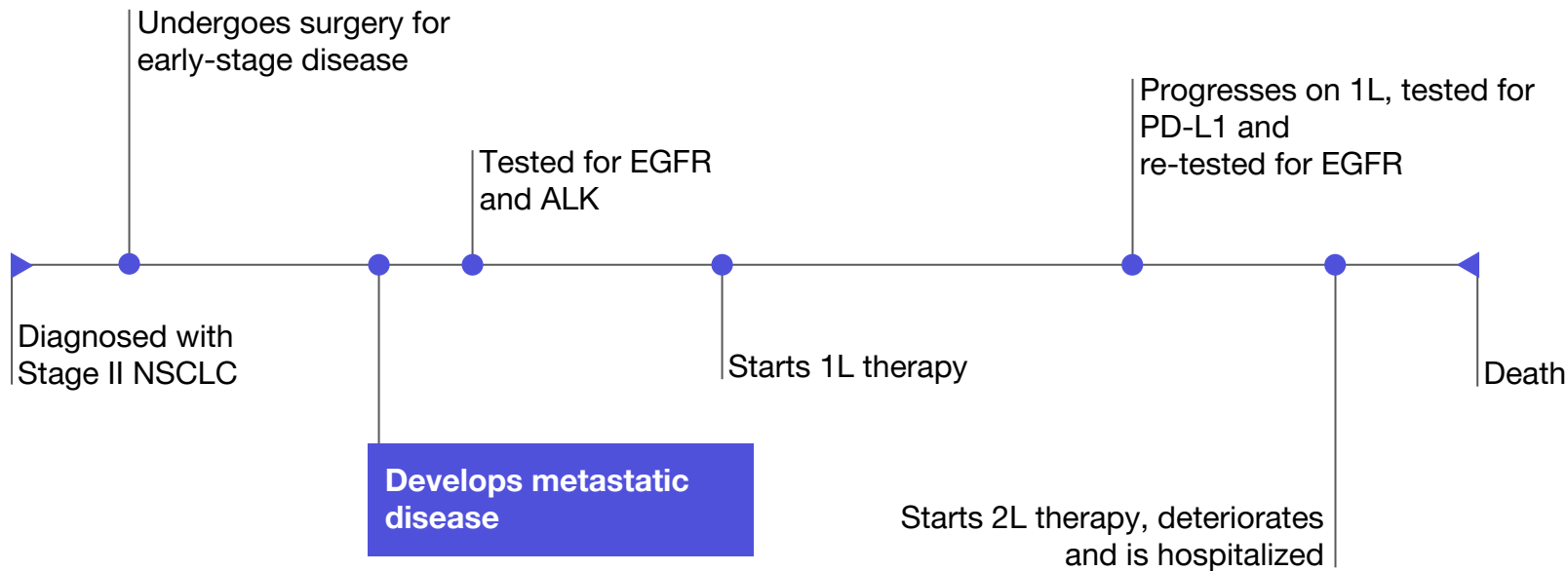
Microscopic Description

The specimen consists of a well differentiated adenocarcinoma, favor lung primary. CK7 and TTF are positive. CK20 is negative. A colleague agrees with this malignant diagnosis.

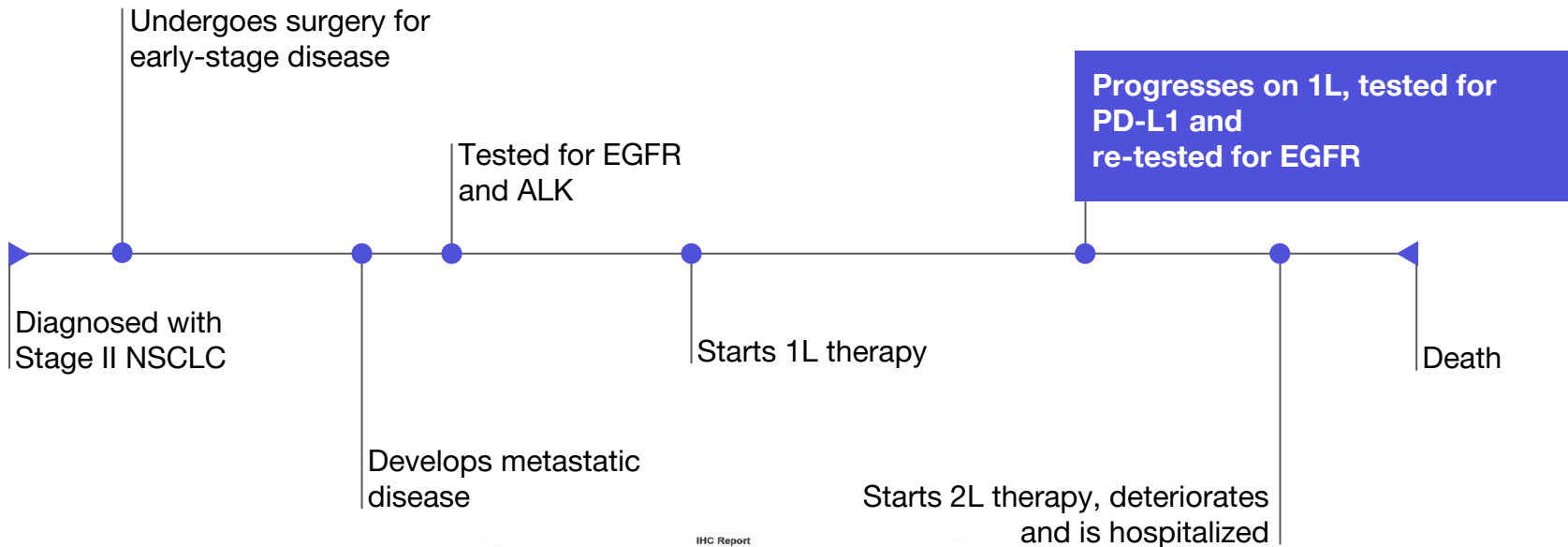
Dictated by: THOMAS J GRIFONE MD
Entered: 02/07/12 - 1423 SML

Diagnosis

SPECIMEN SUBMITTED AS TRUCUT BIOPSY LEFT LUNG NODULE:
- WELL-DIFFERENTIATED ADENOCARCINOMA, FAVOR LUNG PRIMARY.
- SEE ABOVE.



Path?



Time to progression is dependent on when patient is evaluated

IHC Report

Lung, Right Upper Lobe Tissue

Review: Manual Assay Type: NEGATIVE

Tumor Stained: 0 Reference Range: NEGATIVE < 50% POSITIVE >= 50%

PD-L1, 22C3

Review: Manual Assay Type: NEGATIVE

Tumor Stained: 0 Reference Range: NEGATIVE < 1% POSITIVE >= 1%

Results: NEGATIVE, ELIGIBLE FOR OPDIVO®

PD-L1, 28-8

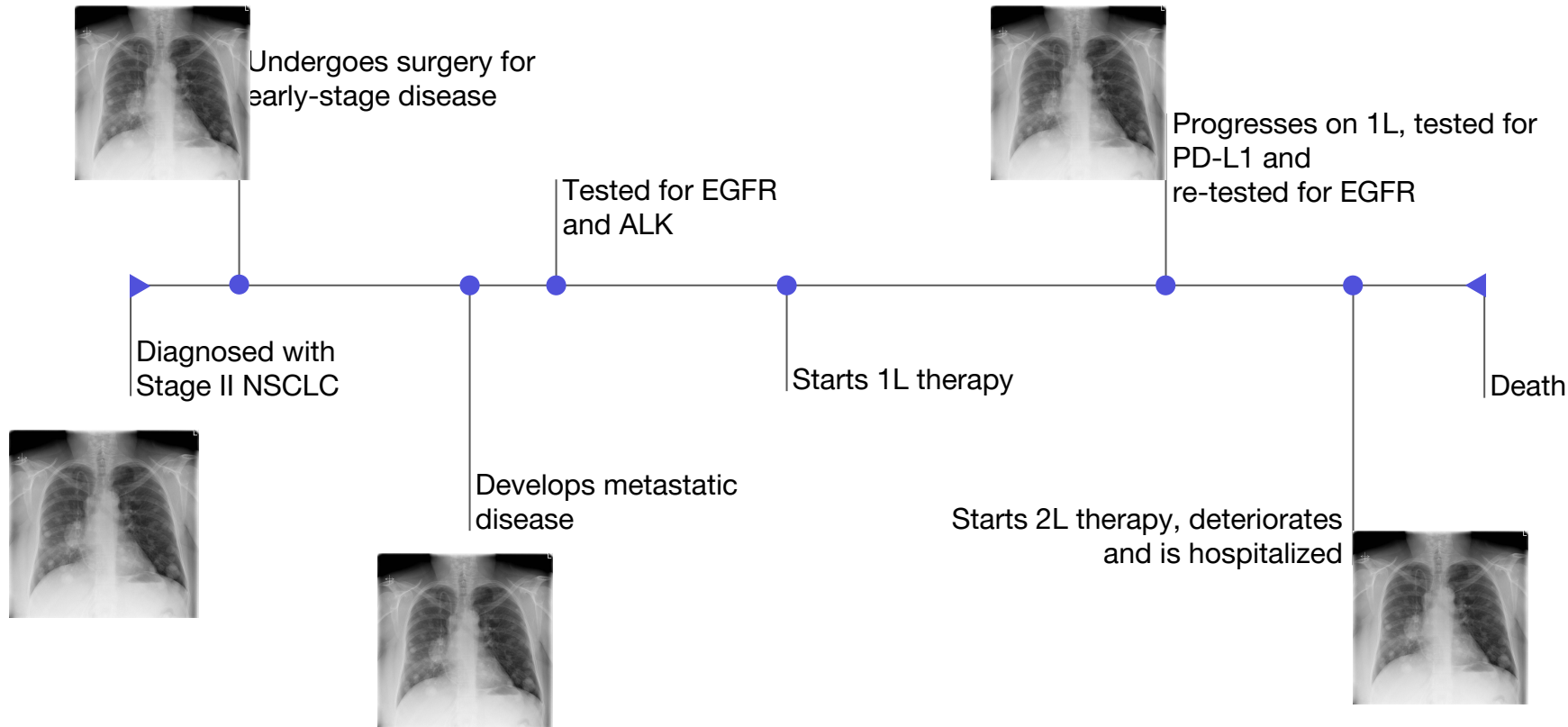
Comments: All non-small cell lung cancer patients are eligible for OPDIVO® (pembrolizumab) regardless of their PD-L1 status. The professional interpretation was performed at Flatiron, Inc., 6455 Mission Court, West Bloomfield, MI, 48324, CLIA: 2302013964

Impression:

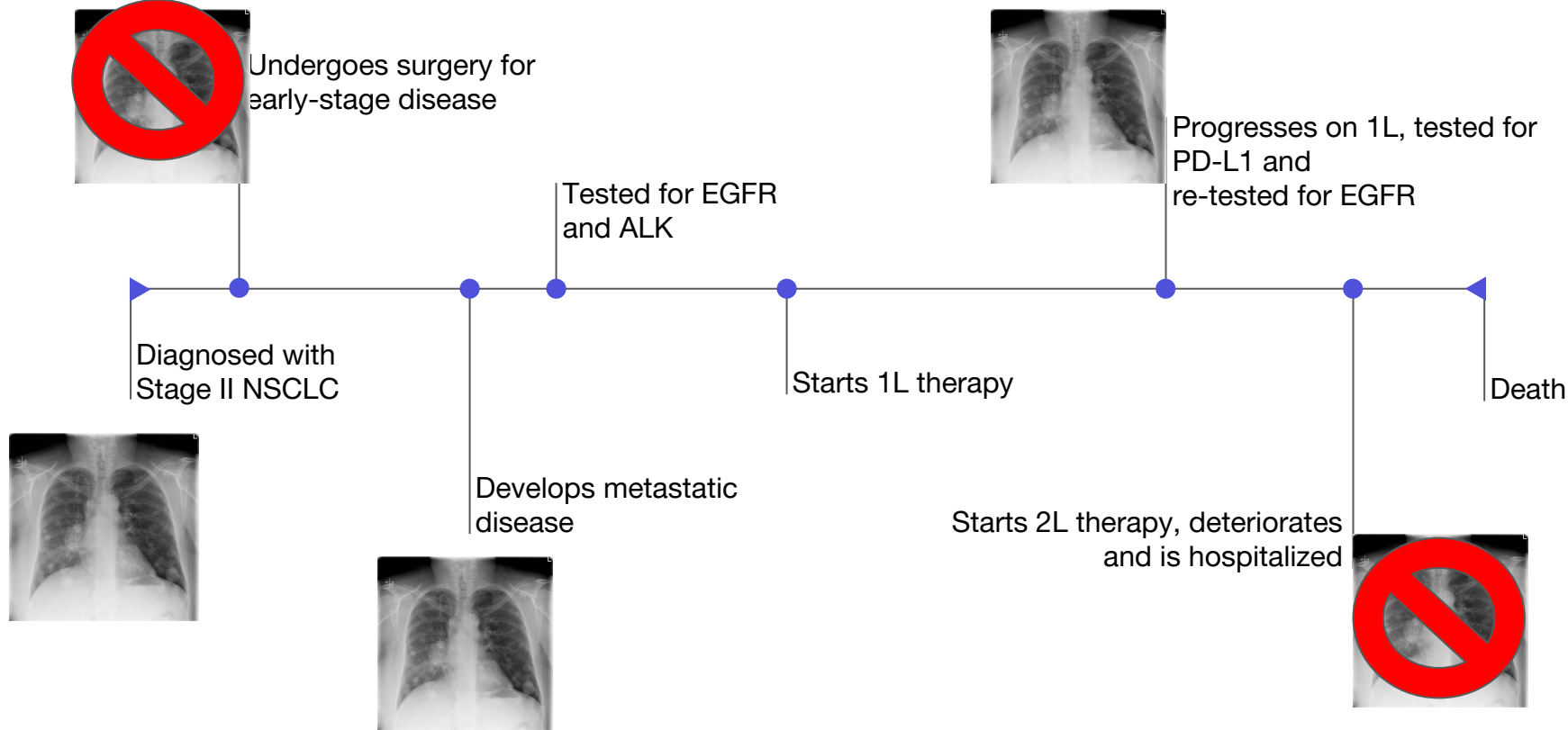
1. Large left upper lobe bronchogenic carcinoma extending to the left hilum, markedly increased in size since prior study
2. Mediastinal adenopathy increased in size particularly subcarinal space. The adenopathy in the AP window has undergone partial necrosis since previous exam.
3. Stable right apical nodularity possibly scar
4. Emphysema

6-2 JS 7-4-20 12/

Currently restaging study showed no soft tissue disease. Bone scan showed stable bony metastasis, with the exception of 1 new lesion in the left superior pubic rami. The nature of this new left superior pubic rami lesion is not clear even though it is possible this is new metastatic lesion; however, it is unusual to have an isolated progression yet rest of the bony metastasis are stable. The patient is completely asymptomatic. His tumor markers are stable; therefore, we decided to continue the current management and we will get followup studies in 3-4 months.



Complete datasets increase reliability



Complete datasets increase reliability

Example: Imaging Data

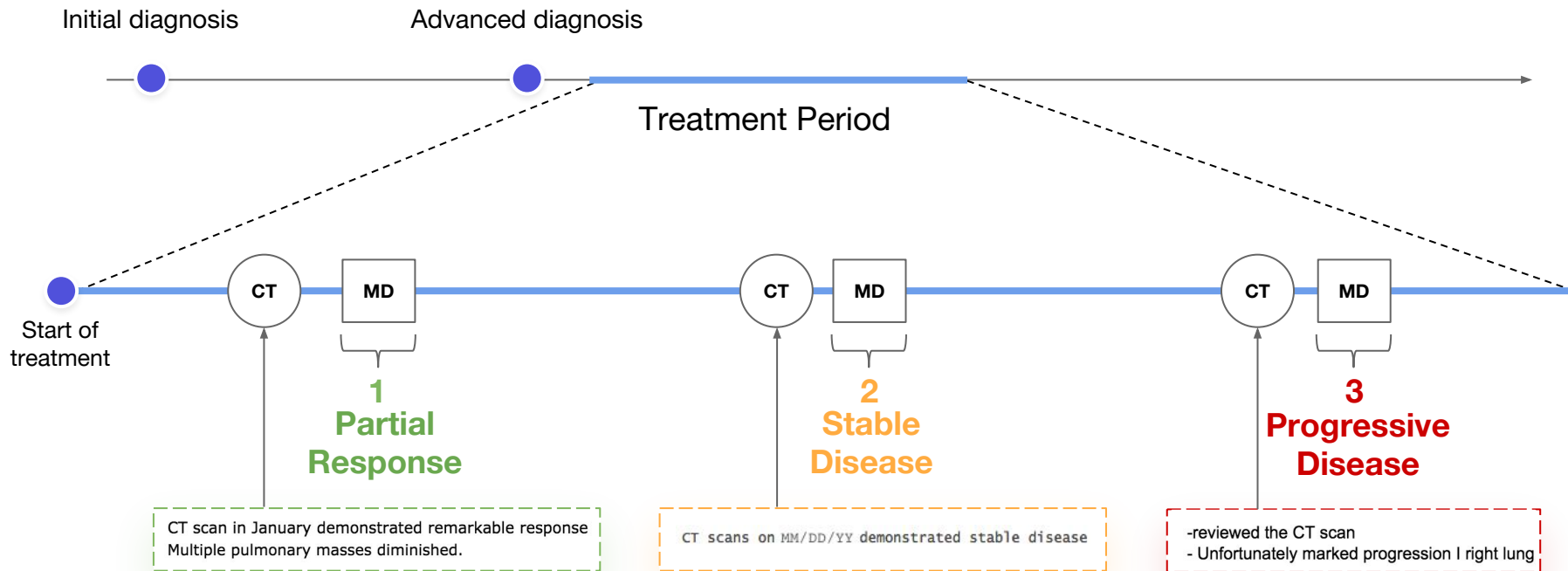


- Raw images are best if available, but require massive storage capabilities and are hard to de-identify
 - Pixelated data is not the same as an interpretation of findings
 - Information needed for subsequent analyses and algorithms is the interpretation
- An alternative is to focus on radiology reports in the EHR
 - Reports must be curated into an analytical dataset
 - Quality of the interpretation varies depending on reviewer
 - Findings in routine clinical practice vary from third party independent review for clinical trials
- Adequate interpretation of scans requires information about context
- Standardized systems for proxy data from scans (e.g., RECIST)
 - Rely on a series of features that may not be available outside of clinical trials (e.g., target lesions, prior scans)
 - May miss clinical features - must understand the pitfalls of the approach

rwTR Assessment Categories

Category	Description
Complete Response (CR)	Complete resolution of all visible disease.
Partial Response (PR)	Partial reduction in size of visible disease in some or all areas without any areas of increase in visible disease. Captures a decrease in disease volume even though disease is still present.
Stable Disease (SD)	No change in overall size of visible disease. Also includes cases where some lesions increased in size and some lesions decreased in size.
Progressive Disease (PD)	Increase in visible disease and/or presence of any new lesions. Includes cases where the clinician indicates progressive disease, PD, or POD as the overall assessment.
Pseudoprogession	Clinician indicates pseudoprogession or related terminology (e.g., tumor flare).
Indeterminate response	Clinician specifically indicates that the response is “indeterminate” or “uncertain,” or if the clinician’s interpretation of the scan(s) cannot be mapped to one of the above assessment categories.
Not documented	Clinician’s note references this response imaging (e.g., “Patient had recent scan”) but does not mention any assessment of tumor response.

Response is tied to exposure to a therapy



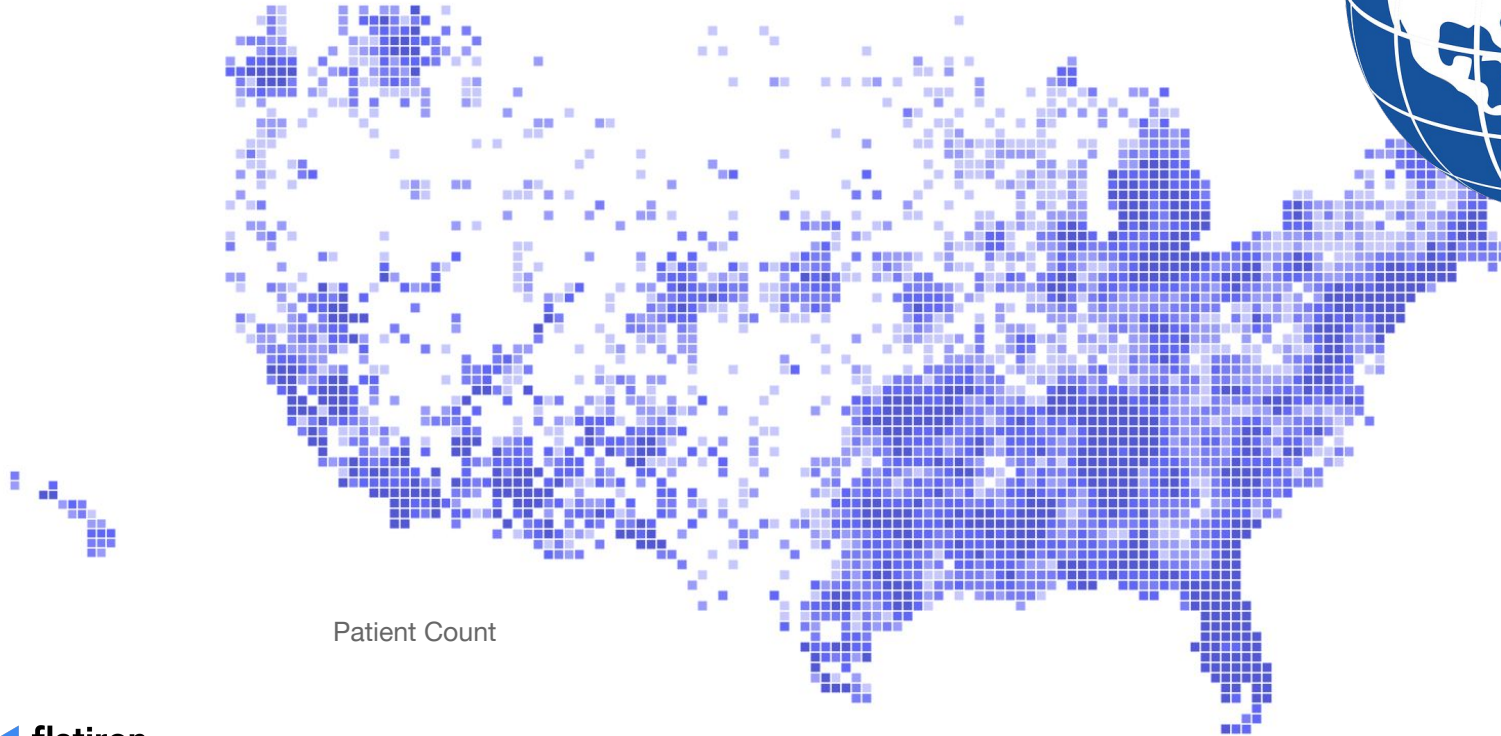
Governance and De-ID



- Who approves access?
- What kind of access?
- How is access QA'ed?

All ultrasound machines store the patient's name and MRN as intrinsic parts of the image and are displayed whenever the image is displayed (Fig. 3). The DICOM headers can be deidentified, but the PHI in the image remains. In addition to the patient's name and MRN, some formats display the birth date, scan date and time, or facility name. <https://www.ajronline.org/doi/full/10.2214/AJR.13.11789>

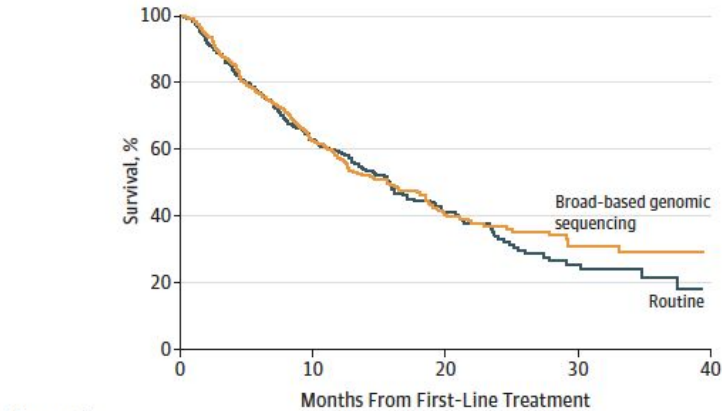
3. Representativeness of datasets
(genomic, geographic); risk of bias in
the underlying data



Patient Count

Does genomic testing improve survival for lung cancer patients?

Figure 2. Kaplan-Meier Estimates of Patients With Broad-Based Genomic Sequencing vs Routine Testing Propensity Score-Matched Sample (n = 1038)



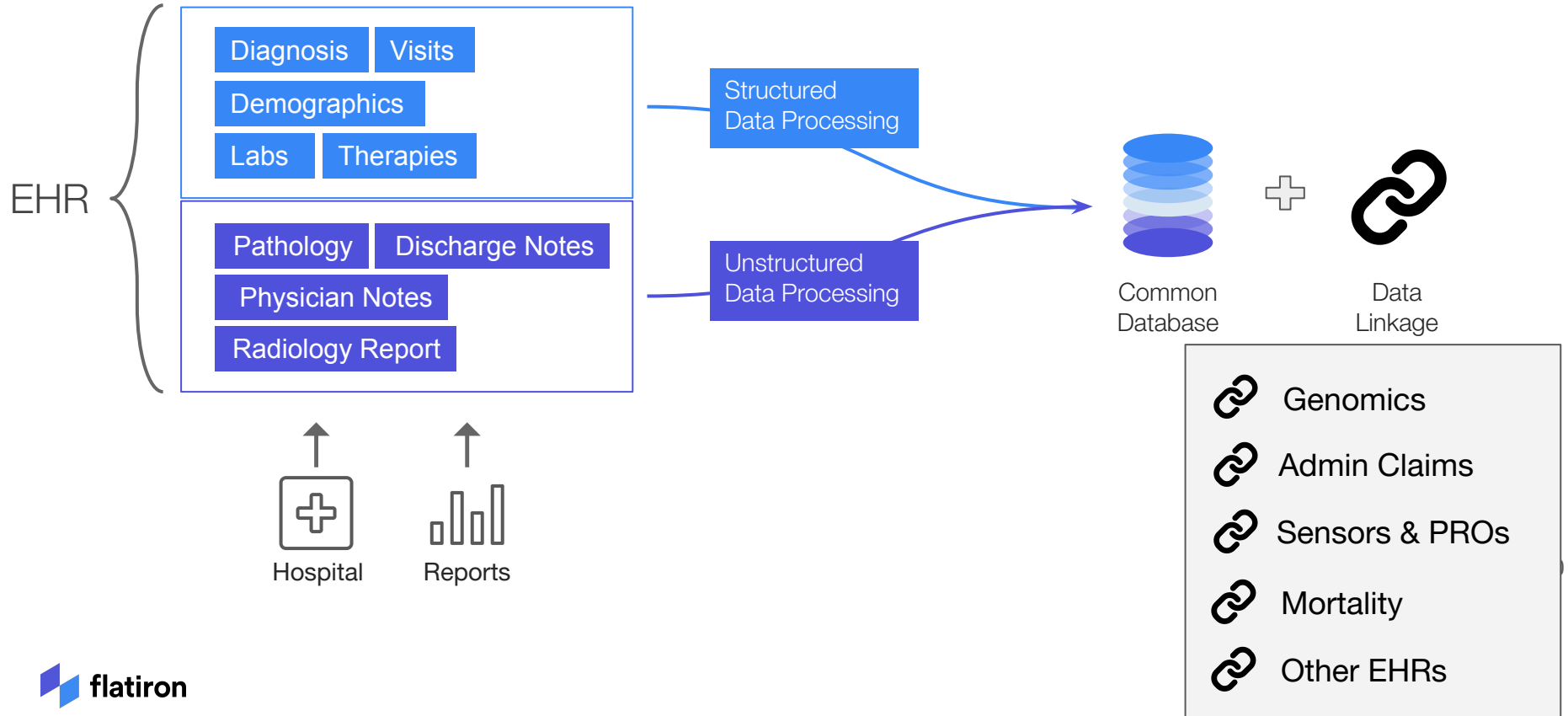
No. at risk	Testing method	515	195	64	23	8
	Broad-based genomic sequencing					
	Routine	513	192	66	18	<5

	No. (%)		P Value ^a
	Broad-Based Genomic Sequencing (n=875)	Routine Testing (n=4813)	
Total			
Age at diagnosis, y			
≤45	41 (4.7)	89 (1.9)	
46-55	106 (12.1)	588 (12.2)	
56-65	277 (31.7)	1367 (28.4)	<.001
66-75	307 (35.1)	1642 (34.1)	
76-85	144 (16.5)	1127 (23.4)	
Sex			
Male	407 (46.5)	2247 (46.7)	.93
Female	468 (53.5)	2566 (53.3)	
Race/ethnicity			
Non-Hispanic white	594 (67.9)	3023 (62.8)	
Non-Hispanic black	49 (5.6)	380 (7.9)	
Hispanic or Latino	28 (3.2)	166 (3.5)	<.001
Asian	39 (4.5)	172 (3.6)	
Other	98 (11.2)	377 (7.8)	
Unknown	67 (7.7)	695 (14.4)	

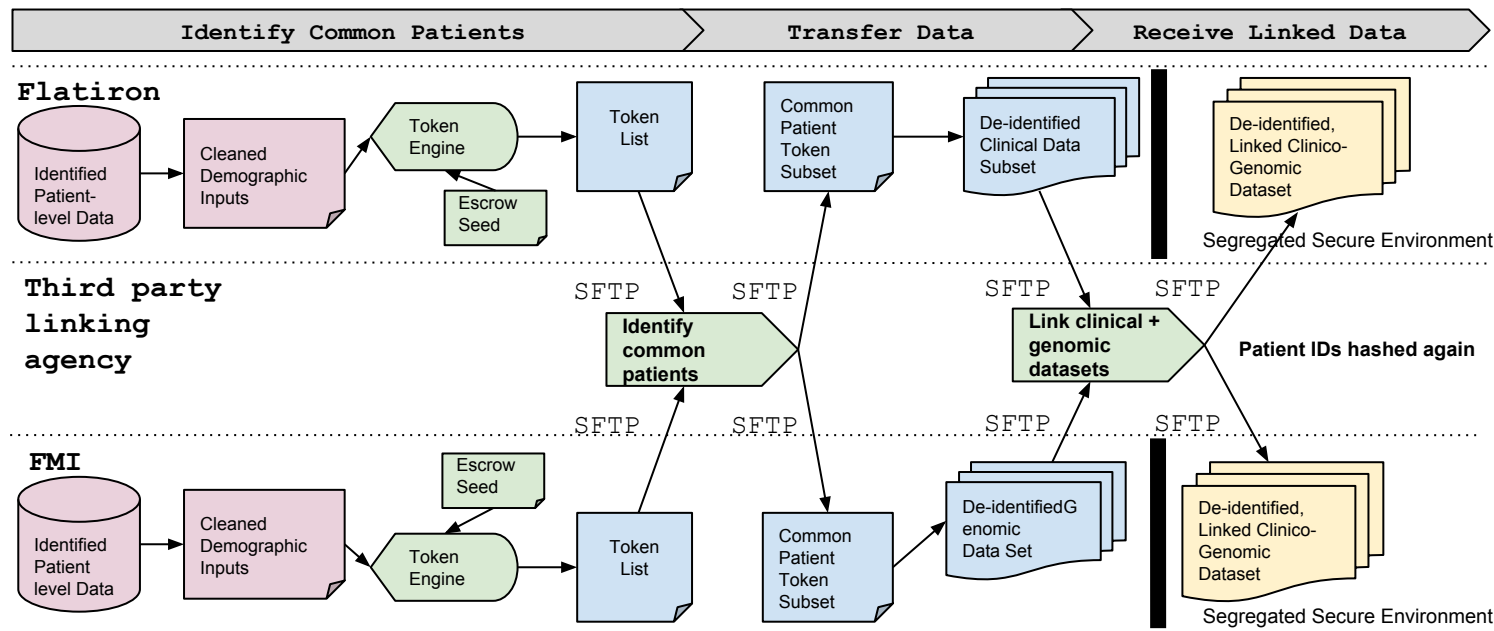
Presley et al. *JAMA* 2018

4. Methods to resolve data gaps

Data linkage to fill gaps



Data linkage solutions



Red - Identified Data

Blue - De-identified Data

Yellow - De-identified Linked Data

Green - Third party linking agency's technology

This entire process is repeated quarterly at FH-FMI for longitudinal refresh of the data (e.g., new chart abstraction)

Final linked clinico-genomic dataset cannot be re-identified or linked to other FH or FMI datasets (unique patient ID)

Evaluate data against a reference standard

E.g., gold standard = National Death Index

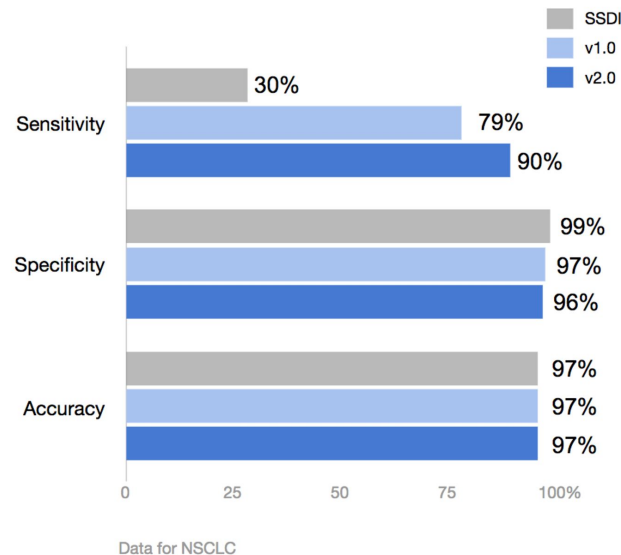
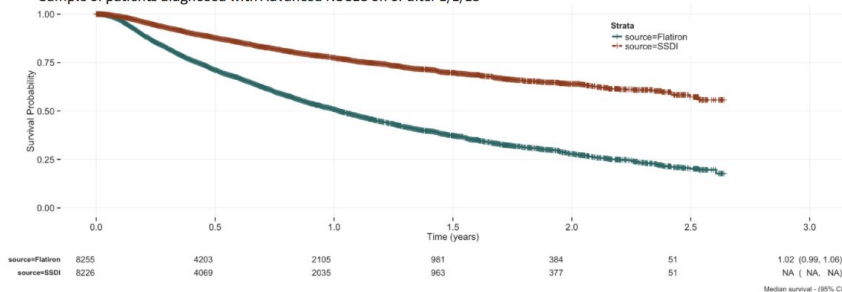


Dedicated field for Patient Date of Death (DoD)

Unstructured documents (e.g., death certificates, condolence cards)

Data vendors selected on basis of data coverage and recency

OS: Advanced NSCLC diagnosis to death
Sample of patients diagnosed with Advanced NSCLC on or after 1/1/13



Results:

- Capture of CCI by ICD codes was poor. The sensitivity of CCI ≥ 1 as measured by ICD codes as compared to abstracted data (gold standard) was 9% (95% CI: 5-14%)
- Using ICD codes to identify patients without any comorbidities works well most of the time. This is reflected in the high observed specificity and NPV (99%, CI: 98-100% and 72%, CI: 69-75%, respectively)
- Abstraction captured more comorbidities than ICD codes. However, comorbidity data abstracted from an Oncology EHR may itself be incomplete and not an ideal gold standard. Oncologists are unlikely to document a comorbidity unless it will affect cancer treatment decision-making

Background

Background

- The Charlson Comorbidity Index (CCI) is used to characterize risk in populations, and is typically calculated using ICD code algorithms from non-world data (RWG) [1]. However, nearly all electronic data are derived from the United States.
- Electronic health record (EHR) data are becoming a more common source of RWG for research, and are themselves a source of ICD codes.
- Using data captured in a specialty Oncology EHR, this study compared the accuracy of the CCI calculated using ICD codes versus data abstracted from the medical record. The abstracted data were considered to be the gold standard.

References

Study Population

- This study included all cancer non-small cell lung cancer (NSCLC) patients from the Flatiron Health Database (Figure 1).
- The Flatiron Health database is a longitudinal, demographic and geographically diverse database derived from EHR data. It includes data from > 200 cancer clinics and > 2 million active cancer patients in the United States.
- Patient-level data include structured and unstructured data. Structured data are pulled directly from the EHR. To extract data from unstructured data sources (such as clinic notes), Flatiron has developed a "technology-enabled" chart abstraction methodology.

Figure 1: Cohort selection



Table 2: Calculation of diagnostic accuracy of ICD codes for OAs as compared to the gold standard of abstraction.

for CDS as compared to the gold standard of abstraction

		Abstraction		
		CDS = 1	CDS = 0	
CDS Codes	CDS = 1	True Positive (A)	False Positive (B)	$PPV = A/(A+B)$ $NPV = C/(C+D)$
	CDS = 0	False Negative (C)	True Negative (D)	

↓ **Sensitivity**
↓ **Specificity**

^aQuin H. et al. Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment: Hospital Discharge Abstracts Using Data From 8 Countries. *Am J Epidemiol* 2015; 172:674-682. ^bCancer and metastases were not included in the CCI calculation for this study.

Results

Findings

- In 715 patients, capture of CGI by ICD codes was poor. The sensitivity of CGI as measured by ICD codes as compared to abstracted data (gold standard) was 50% (95% CI = 44-57) (Table 1).
- In 715 patients, ICD codes to identify patients without any comorbidities works well most of the time. This is reflected in the high standard specificity and NPV (95% CI = 90-100% and 73% - 99-100%, respectively) (Table 1).
- A similar pattern was observed across individual comorbidities: sensitivity = 30%, specificity = 90-100%, and NPV = 75-100%. Estimates of PPV had poor precision because of the low number of comorbidities identified via ICD codes (Table 1).
- Abstraction captured more comorbidities than ICD codes. Among patients whose CGI score differed depending on whether it was calculated using ICD codes or abstracted data, 60% of the patients whose CGI score was higher when calculated using ICD codes had at least one comorbidity entered on patients being classified as having no comorbidities from ICD codes to CGI = 0, but at least one comorbidity from abstraction to CGI = 1.

Table 4: OOI reclassification in 715 advanced NSCLC patients as measured to structured versus unstructured data

[illegible]

Limitations

Limitations

Notwithstanding, few comorbidities were identified overall, whether by ICD code or abstraction, which resulted in low precision of our results. The Patient Health Database is derived from an Oncology EHR, and data from medical records may be incomplete. For example, primary care physicians rarely enter data in the EHR, and the database does not include the incentive to enter ICD codes for comorbidities because it is usually used to enter other data. The database also does not include data to describe comorbidities in the medical record unless it will influence treatment decisions. We are therefore wary of underestimating the presence of comorbidities based on information from the database.

Conclusions/Discussion

Conclusions/Discussion

Using data from an Oncology EHR, ECO codes were not sufficient for identification of patients who were diagnosed with metastatic disease. ECO codes may not be sufficient to capture ECO-relevant comorbidities that are documented elsewhere in the EHR. Currently, data abstracted from an Oncology EHR may lack the incomplete and/or inconsistent data that may be present in a comprehensive source. It will affect cancer treatment decision-making.

Additional data from an Oncology EHR data source combined with additional data sources to increase the completeness of data for use in ECO calculation.

Additional analysis may include stratification by stage of diagnosis (diagnosed with early stage disease vs. diagnosed with advanced stage disease) compared to groups who are diagnosed with metastatic disease. Those who are diagnosed with early stage disease may have more complete capture of comorbidities in either structured or unstructured data because there are

Statistical Analysis

- We identified comorbidities used in the calculation of the modified CGI, not including cancer as a comorbidity, using statistics, using the ICD-9-CM ICD-9-CM codes and unstructured data from the Oncology 5350 (Table 1).
- We assessed severity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of presence of any comorbidity (CGI) as a risk factor for death, considering the structured data to be the gold standard for purposes of calculation (Table 2).
- We calculated the percent of patients whose CGI changed after treatment using ICD-9-CM codes versus structured data.
- Comorbidities were only considered if they occurred prior to the advanced diagnostic date.

© 2004 Blackwell Publishing Ltd *Journal of Internal Medicine* 255: 103–110

Comorbidity	Scores
Conjunctive heart failure	0
Dementia	0
Chronic pulmonary disease	1
Rheumatologic disease	1
Mild liver disease	0
Moderate or severe liver disease	4
Diabetes with chronic complications	1
Hemiplegia or paraplegia	0
Renal disease	1
AIDS/HIV	4
Any malignancy*	2
Malignant solid tumor*	0

Table 3: Comparison of structured and unstructured data for
estimation of contributions to TAE, whereas MOC is not used

	Association with depression (OR)	Association with anxiety (OR)	Association with PTSD (OR)	Association with any disorder (OR)
Any comorbidity (N = 5)	0.90 (0.86-0.94)	0.92 (0.88-0.96)	0.92 (0.88-0.96)	0.92 (0.87-0.96)
Post-traumatic stress	0.92 (0.88-0.96)	0.93 (0.89-0.97)	1.00 (0.96-1.04)	0.98 (0.93-1.03)
Manic/depressive	0.86 (0.81-0.91)	0.87 (0.82-0.92)	NA ^a	0.86 (0.81-0.92)
Chronic post-traumatic stress	0.93 (0.89-0.97)	0.93 (0.89-0.97)	0.79 (0.75-0.83)	0.79 (0.75-0.83)
Has a psychiatric disorder	0.76 (0.72-0.80)	0.76 (0.72-0.80)	0.79 (0.75-0.83)	0.87 (0.83-0.91)
Used any alcohol	0.93 (0.89-0.97)	0.93 (0.89-0.97)	0.69 (0.65-0.73)	0.78 (0.74-0.82)
Used drugs or smoke (not alcohol)	0.92 (0.88-0.96)	0.93 (0.89-0.97)	NA ^a	0.91 (0.87-0.95)
Diagnosed with chronic pain	0.93 (0.89-0.97)	0.93 (0.89-0.97)	0.79 (0.75-0.83)	0.79 (0.75-0.83)
Has a pain condition (not chronic pain)	0.93 (0.89-0.97)	0.93 (0.89-0.97)	0.79 (0.75-0.83)	0.79 (0.75-0.83)
Has a chronic pain condition	0.93 (0.89-0.97)	0.93 (0.89-0.97)	1.00 (0.96-1.04)	0.98 (0.93-1.03)
Has any	0.93 (0.89-0.97)	0.93 (0.89-0.97)	0.79 (0.75-0.83)	0.79 (0.75-0.83)

Presented at ISPOR, May 19-23, 2018 in Baltimore, MD.
For additional information, contact Christina Parnello at cparnello@salix.com

All authors are employees of Flatiron Health, a cancer-focused health technology company in New York, New York, USA and the source of study data.

5. Issues in Precision Medicine: Available biologic/biomarker data, small cohorts

Deep Phenotype is Needed

In an era of abundant data, merging biological information with deep clinical phenotype is more important than ever



The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data

Nucleic Acids Research, 2013, 1–9
doi:10.1093/nar/gkt1026

DEEP PHENOTYPING

The details of disease

Precision medicine demands precise matching of deep genomic and phenotypic models — and the deeper you go, the more you know.

S14 | NATURE | VOL 527 | 5 NOVEMBER 2015

Warner et al. *Genome Medicine* (2016) 8:113
DOI 10.1186/s13073-016-0371-3

Genome Medicine

REVIEW

Open Access



Integrating cancer genomic data into electronic health records

Jeremy L. Warner^{1,2,3*}, Sandeep K. Jain^{2,4} and Mia A. Levy^{1,2,3}

Precision medicine often yields small cohorts

THE NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations

David M. Hyman, M.D., Igor Puzanov, M.D., Vivek Subbiah, M.D., Jason E. Faris, M.D., Ian Chau, M.D., Jean-Yves Blay, M.D., Ph.D., Jürgen Wolf, M.D., Ph.D., Noopur S. Raje, M.D., Eli L. Diamond, M.D., Antoine Hollebecque, M.D., Radj Gervais, M.D., Maria Elena Elez-Fernandez, M.D., Antoine Italiano, M.D., Ph.D., Ralf-Dieter Hofheinz, M.D., Manuel Hidalgo, M.D., Ph.D., Emily Chan, M.D., Ph.D., Martin Schuler, M.D., Susan Frances Lasserre, M.Sc., Martina Makrutzki, M.D., Florin Sirzen, M.D., Ph.D., Maria Luisa Veronese, M.D., Josep Tabernero, M.D., Ph.D., and José Baselga, M.D., Ph.D.

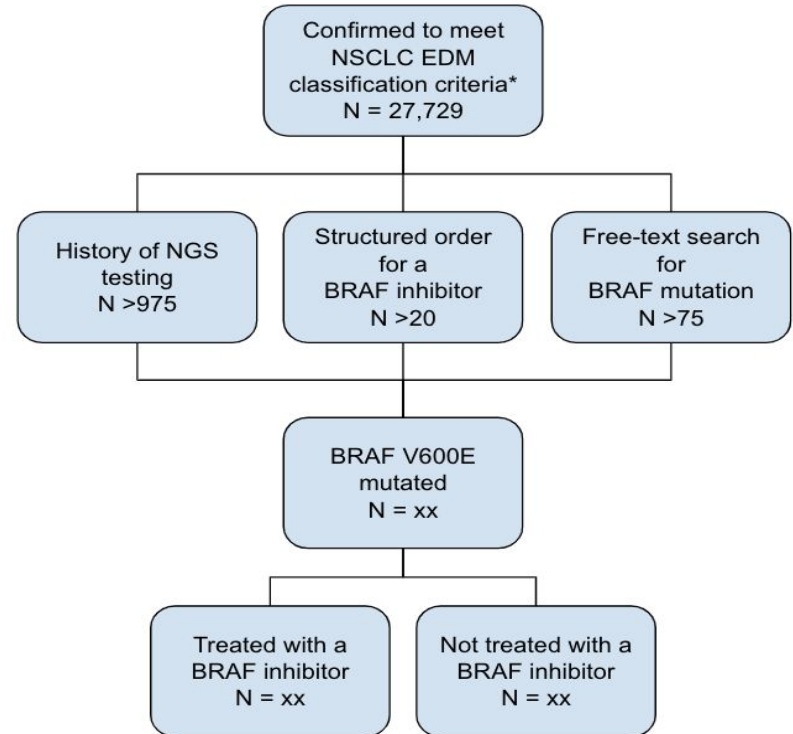
ABSTRACT

BACKGROUND

BRAF V600 mutations occur in various nonmelanoma cancers. We undertook a histology-independent phase 2 “basket” study of vemurafenib in BRAF V600 mutation-positive nonmelanoma cancers.

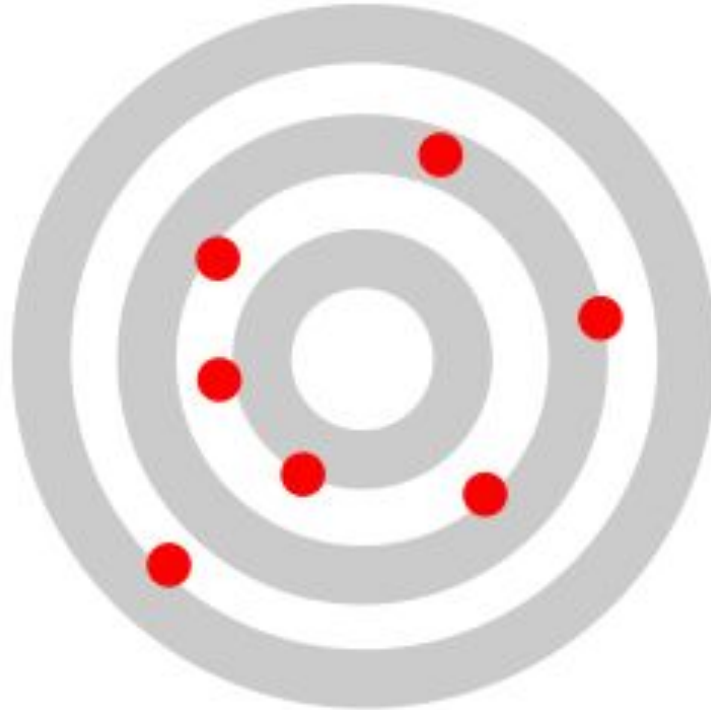
METHODS

We enrolled patients in six prespecified cancer cohorts: patients with all other

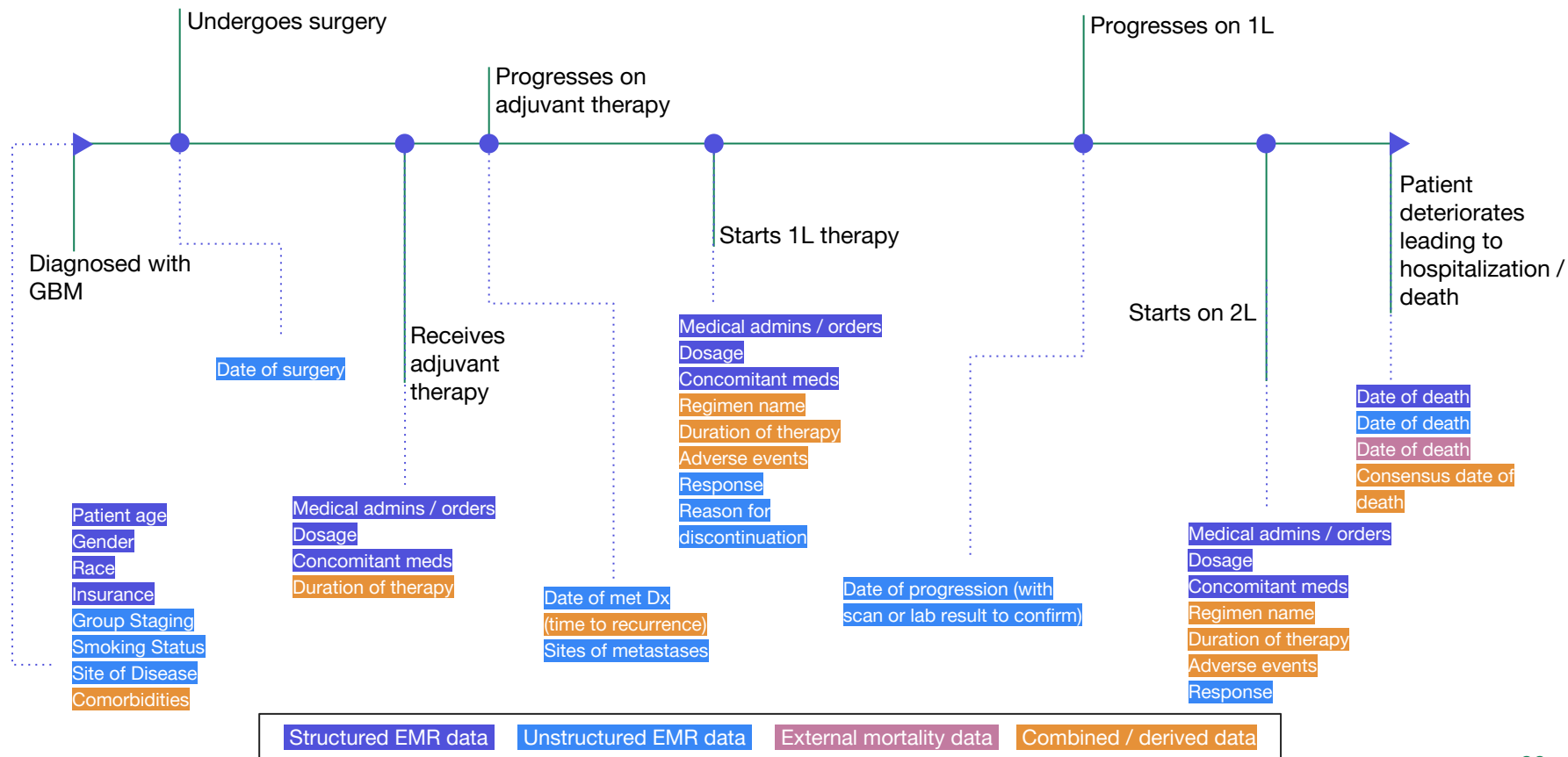


6. Data reliability and quality

***Reliable data produce reliable
algorithms and reliable results***



Documentation of source, quality and provenance



*Relative timing not exact

Need a consistent approach to documenting completeness and quality

Appendix B: Flatiron Health PD-L1: Inter-rater agreement and kappas on abstracted variable

Project:

FDA

PD-1 inhibitors in aNSCLC

Note: For questions where a high percentage of patients have a common answer (e.g., PD-L1 testing status), kappa may be significantly lower than inter-rater agreement. In these cases, it may be more accurate to use inter-rater agreement to measure reliability of the data.

Kappas scale

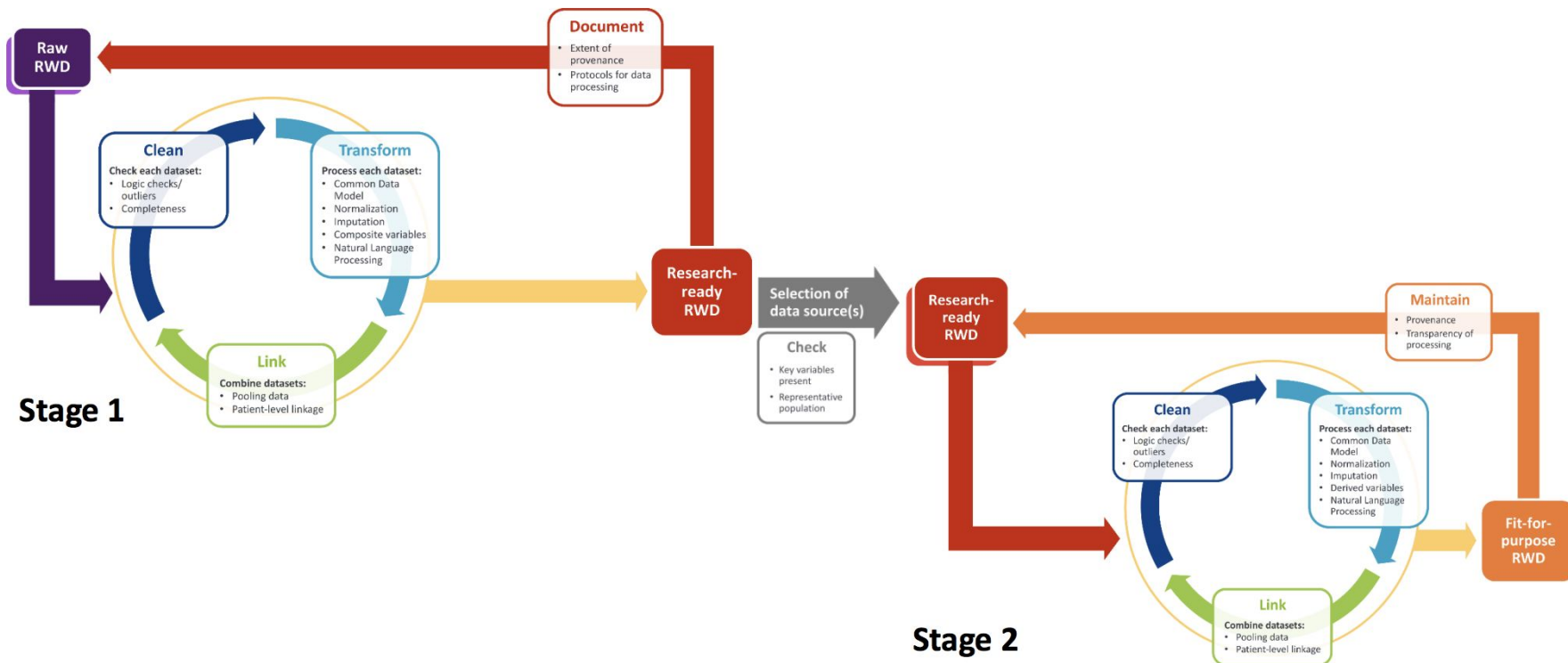
Almost perfect	0.8 to 1.0
Substantial	0.6 to 0.8
Moderate	0.4 to 0.6
Fair	0.2 to 0.4
Slight	0 to 0.2

Table: Enhanced_AdvancedNSCLC

Summary of variable inter-rater agreement and kappas

Variable	Description of variable	Corresponding question(s) on abstraction form	Question type	Inter-rater agreement (exact day for dates)	Kappa (exact agreement)	Kappa (30-day window for dates)
DiagnosisDate	Date of initial diagnosis	Enter the date of initial diagnosis	date	0.795	0.794	0.902
AdvancedDiagnosisDate	Date of diagnosis of advanced disease: first recurrence or metastasis	Enter the date of the first diagnosis of metastatic or advanced NSCLC	date	0.695	0.695	0.796
MetastaticDiagnosisDate	Date of diagnosis of metastatic disease	Enter the date of initial diagnosis [for ~55% of patients in the cohort who are diagnosed metastatic]	date	0.795	0.794	0.902
		Enter the date of distant metastatic diagnosis [for ~45% of patients in the cohort who are diagnosed non-metastatic]	date	0.527	0.476	0.557
Histology	Histology	Select the histology	drop down	0.947	0.894	
GroupStage	Group stage at time of initial diagnosis	Select the group stage	drop down	0.848	0.768	
SmokingStatus	Documented history of smoking	Smoking status	drop down	0.934	0.695	
EgfrTested	Indicator of whether the tumor was tested for a EGFR mutation	Was the tumor tested for a EGFR mutation?	boolean	0.927	0.84	
AlkTested	Indicator of whether the tumor was tested for an ALK rearrangement	Was the tumor tested for an ALK rearrangement?	boolean	0.901	0.791	
PdL1Tested	Indicator of whether the tumor was tested for PD-L1 expression	Was the tumor tested for PD-L1 expression?	boolean	0.901	0.547	
KrasTested	Indicator of whether the tumor was tested for a KRAS mutation	Was the tumor tested for a KRAS mutation?	boolean	0.894	0.728	
Ros1Tested	Indicator of whether the tumor was tested for a ROS-1 rearrangement	Was the tumor tested for a ROS-1 rearrangement?	boolean	0.881	0.725	

Research-ready databases



7. Recency and availability



8. Privacy and security



9. Policy Considerations

***Reliable data produce reliable
algorithms and reliable results***

Policy Considerations

- Require accurate complete datasets that incorporate many data types
- Details about the data matter
- Biased and unreliable data lead to biased and unreliable algorithms
- Accessible data at point of care
- While maintaining patient privacy
- Solving this requires input of many types of experts - clinical, analytic, software, hardware, privacy, etc
- Stimulate tech innovation that solves all of these problems - do not skimp
- Expect transparency about data reliability and quality - as well as accuracy of output from algorithms
- Create standards for documenting reliability, quality and accuracy at the data source, dataset and algorithm level
- Develop a the workforce