Database & Methods Cyberseminar Series

Session #5. *Phenotyping physiologic measurement of lung function in the VA electronic health record using automated tools*

VIReC
RESEARCHERS' GUIDE TO VA DATA

At the end of this cyberseminar, participants will be able to:

- Describe challenges to extracting pulmonary function test (PFT) data from VA electronic health record (EHR)-based data
- Understand the benefits and limits of automated tools that extract PFT numeric values from the EHR
- Name at least two potential implications of automated tools for extracting PFTs

Session roadmap

- PFTs, chronic obstructive pulmonary disease (COPD) clinical importance and COPD research
- Existing data sources to identify PFTs values
- Methods to extract data from existing sources
- Structured query language (SQL) tools for extracting PFT data
- Clinical example of use of extracted PFT data

Poll Question #1

I am interested in VA data primarily due to my role as:

a. Principal investigator/Co-PI

b. Research staff (Project coordinator, data manager, programmer)

c. Clinical Staff

d. Operations Staff

e. Other—Please describe via the Q & A function

Session roadmap

- **PFTs, chronic obstructive pulmonary disease (COPD) clinical importance and COPD research**
- Existing data sources to identify PFTs values
- Methods to extract data from existing sources
- Structured query language (SQL) tools for extracting PFT data
- Clinical example of use of extracted PFT data

Poll Question #2

How would you rate your knowledge of methods to ascertain pulmonary function test (PFT) data for Veterans in the VA?

a. 1 (No knowledge)

b. 2

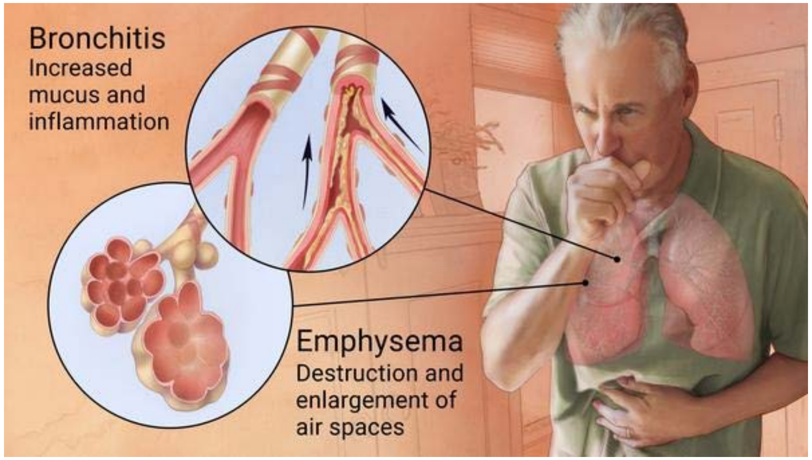
c. 3

d. 4

e. 5 (Expert)

Chronic obstructive pulmonary disease

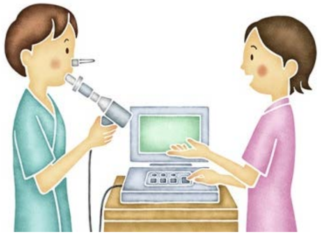
- 7-12% of adults worldwide
- 3rd leading cause of death in the United States
- Identified clinically and with pulmonary function tests (PFTs)


Bronchitis
Increased mucus and inflammation
Emphysema
Destruction and enlargement of air spaces

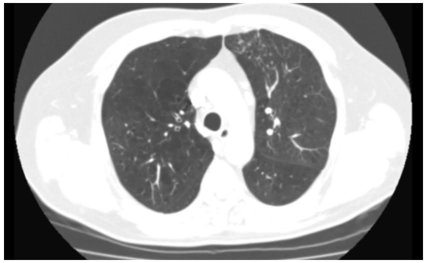
Chronic obstructive pulmonary disease (COPD)

- Identified using PFTs in the setting of symptoms

- Forced expiratory volume in one second (FEV1; in liters)
  - ✓Commonly used, highly reproducible
  - ✓Used to grade severity (GOLD stages 1-4)



- Also radiographic emphysema (Chest CT)



Research in COPD

- Accurate identification of COPD and its severity important for epidemiologic and clinical research
  - Determine guideline-concordant treatment and management
  - Inform estimates for prognosis
  - Identify potential participants for clinical trials

PFTs are critical to accurately identifying COPD

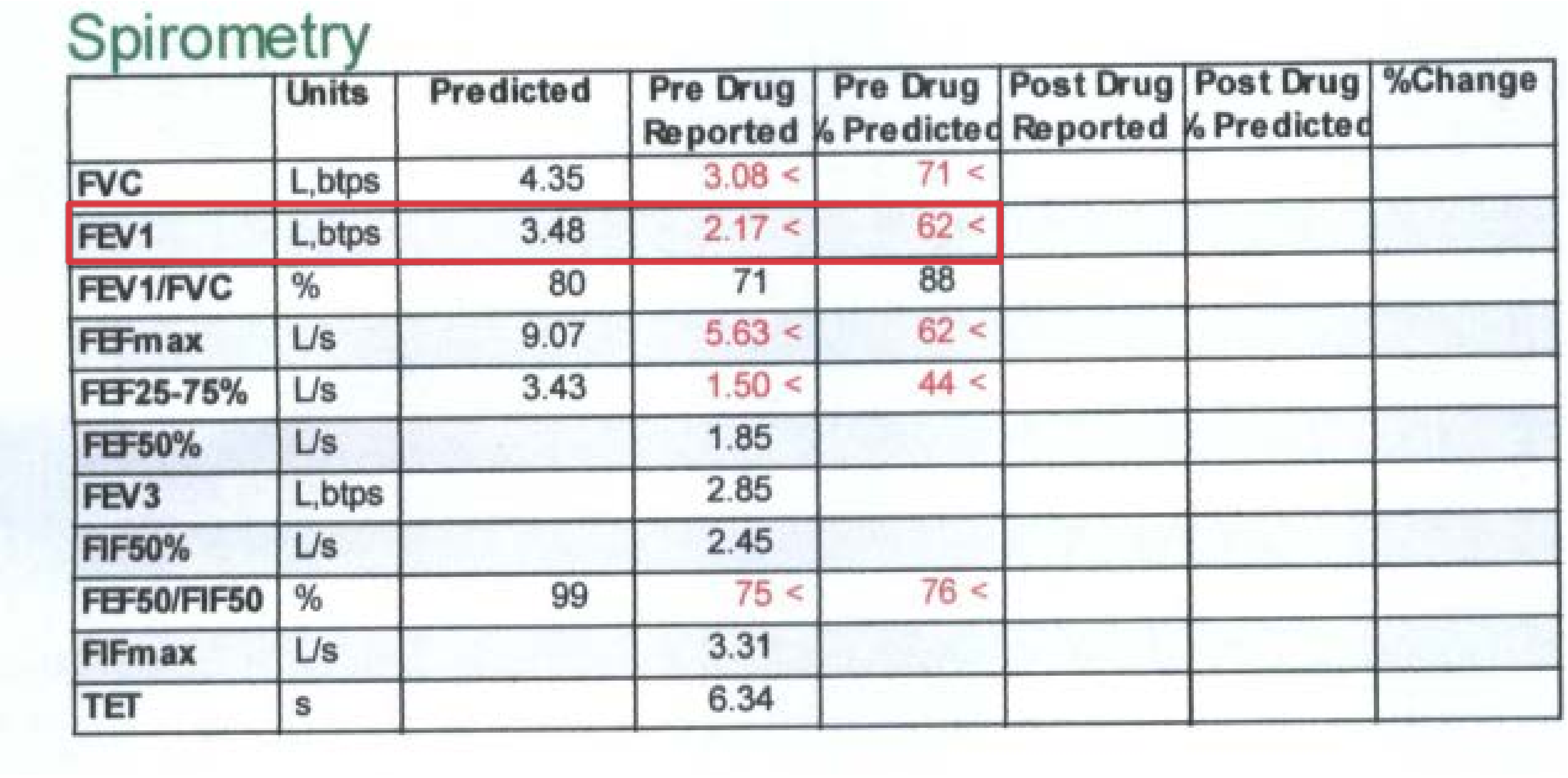
- PFTs assess:
  - Spirometry measures airway obstruction
  - Total lung capacity
  - Diffusion capacity across pulmonary/vascular interface

PFTs are critical to accurately identifying COPD

- PFTs assess:
  - **Spirometry measures airway obstruction**
  - Total lung capacity
  - Diffusion capacity across pulmonary/vascular interface

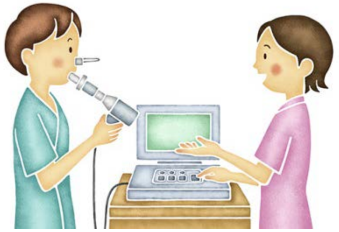
- Spirometry
  - Forced exhale
  - Maneuver that determines airway limitation
  - Absolute value and percent of predicted based on patients age, race, gender and height

PFT data is the standard for measuring COPD and its severity



Spirometry

	Units	Predicted	Pre Drug Reported	Pre Drug % Predicted	Post Drug Reported	Post Drug % Predicted	%Change
FVC	L,btps	4.35	3.08 <	71 <			
FEV1	L,btps	3.48	2.17 <	62 <			
FEV1/FVC	%	80	71	88			
FEFmax	L/s	9.07	5.63 <	62 <			
FEF25-75%	L/s	3.43	1.50 <	44 <			
FEF50%	L/s		1.85				
FEV3	L,btps		2.85				
FIF50%	L/s		2.45				
FEF50/FIF50	%	99	75 <	76 <			
FIFmax	L/s		3.31				
TET	s		6.34				



Session roadmap

- PFTs, chronic obstructive pulmonary disease (COPD) clinical importance and COPD research
- **Existing data sources to identify PFTs values**
- Methods to extract data from existing sources
- Structured query language (SQL) tools for extracting PFT data
- Clinical example of use of extracted PFT data

Poll Question #3

Have you ever used the VHA Corporate Data Warehouse (CDW) to extract numeric values?

a. Yes, both the Master File and Mini File

b. Yes, only the Mini File

c. Yes, only the Master File

d. No

COPD codes available but of limited utility

- ICD-9 diagnosis codes for identifying COPD rarely include severity

- Improved performance of ICD-9 codes when adding pharmacy data for treatment of maintenance/exacerbation

- Will not include objective measures for severity of airflow limitation, such as FEV_1, which is obtained from spirometry

Crothers et al. Pharmacoepidemiol Drug Saf, *Epub ahead of print* 2018.

COPD codes available but of limited utility

- PFT data (FEV$_1$) is in the EHR but in heterogeneous forms
  - Easily accessible fields in CDW
  - Numeric values in free text, semi-structured data in unstructured fields
- VA EHR illustrates these points
- Developing tools to extract FEV$_1$ would advance our understanding of COPD for hundreds of thousands of patients

Data source

- Comprehensive EHR data with structured and free text data
- Corporate Data Warehouse (CDW)
  - Structured data, Common Procedural Terminology (CPT) codes
  - Also unstructured text from progress notes, radiology reports, etc.
- PFTs identified by CPT codes (CPT-4: 94010, 94060, 94070, 94375, 95375, 95070, or 94664) but FEV$_1$ values not always entered into CDW

Session roadmap

- PFTs, chronic obstructive pulmonary disease (COPD) clinical importance and COPD research
- Existing data sources to identify PFTs values
- **Methods to extract data from existing sources**
- Structured query language (SQL) tools for extracting PFT data
- Clinical example of use of extracted PFT data

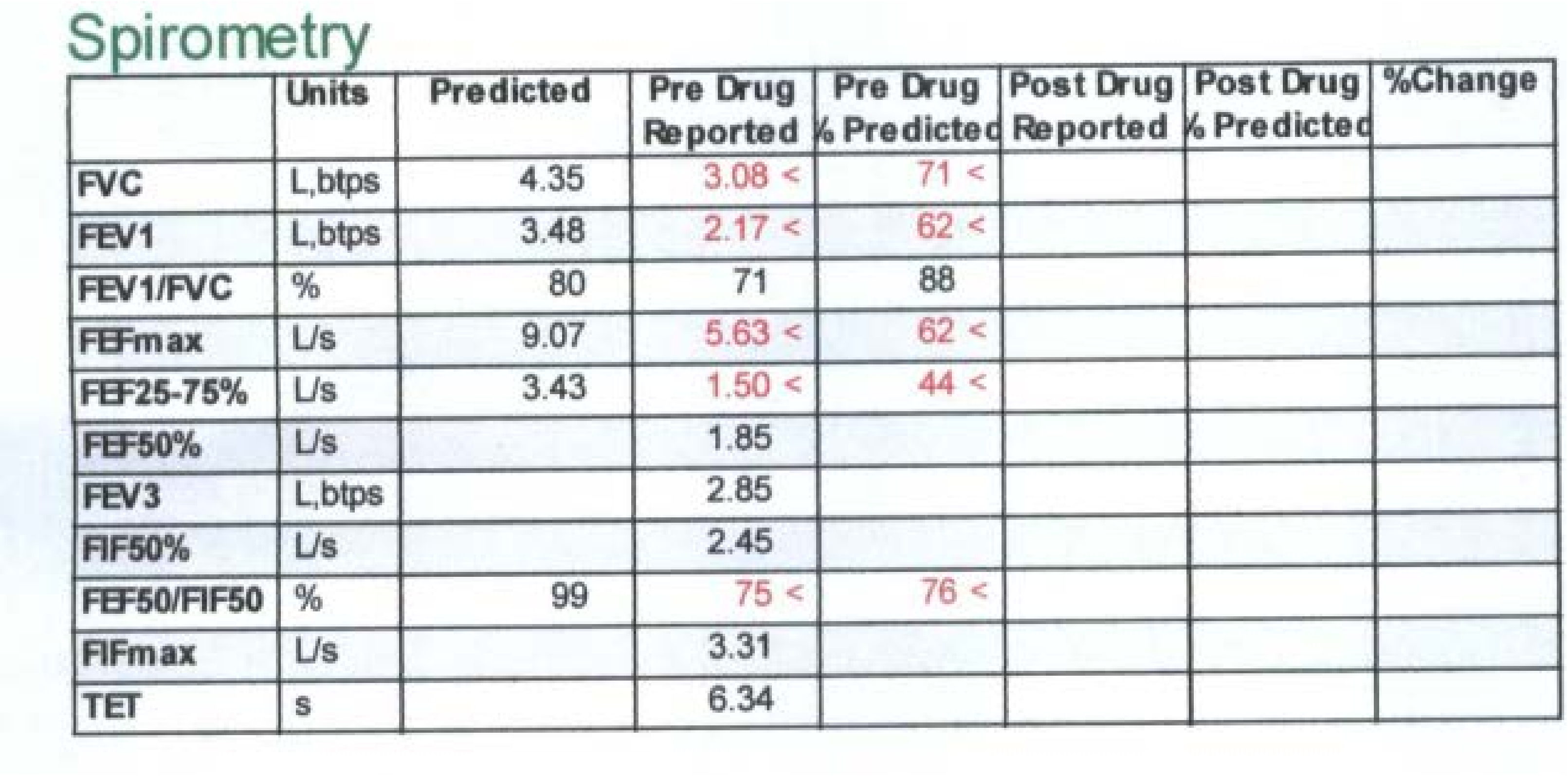
Methods used for two data sources

- Structured PFT data from CDW PFT domain reviewed and require minimal cleaning
- Text Integration Utilities (TIU) free-text data used to create Structured Query Language (SQL)-based tool
- SQL tool for full text keyword search in SQL algorithm
  - ‘FEV1’ and its variants
  - Negation for ‘fever’

Session roadmap

- PFTs, chronic obstructive pulmonary disease (COPD) clinical importance and COPD research
- Existing data sources to identify PFTs values
- Methods to extract data from existing sources
- **Structured query language (SQL) tools for extracting PFT data**
- Clinical example of use of extracted PFT data

How does PFT output appear for clinicians?



Spirometry

	Units	Predicted	Pre Drug Reported	Pre Drug % Predicted	Post Drug Reported	Post Drug % Predicted	%Change
FVC	L,btps	4.35	3.08 <	71 <			
FEV1	L,btps	3.48	2.17 <	62 <			
FEV1/FVC	%	80	71	88			
FEFmax	L/s	9.07	5.63 <	62 <			
FEF25-75%	L/s	3.43	1.50 <	44 <			
FEF50%	L/s		1.85				
FEV3	L,btps		2.85				
FIF50%	L/s		2.45				
FEF50/FIF50	%	99	75 <	76 <			
FIFmax	L/s		3.31				
TET	s		6.34				

How does PFT data appear in free text?

- TIUDocumentSID: 999999999
- Reference date: 01/01/1900
- SCRSSN: 9999999999
- PFT Results Pre % Pred Post %Predicted FVC 2.62L 50 3.02L 57 FEV-1 0.93L 22 1.10L 26 FEV-1/FVC 35 Very severe airflow obstruction with significant bronchodilator response. In July of 1999, the pre-bronchodilator FEV1 was 1.17L. Lung volumes were in the normal range at that time. In 1995, the FEV1 was 2.54L. The flow volume loop is truncated in both phases of respiration, and is consistent with the given history of a fixed extrathoracic obstruction.

How does PFT data appear in free text?

- TIUDocumentSID: 999999999
- SCRSSN: 999999999
- Reference date: 11/11/1111
- MEDICAL HISTORY: This xx-year-old right-handed white male served for 12 months in Vietnam. The veteran indicates that since the late 1960's he has had some dyspnea that has gradually worsened. He currently smokes a pack of cigarettes a day but in the past has smoked 3 to 4 packs per day. A chest x-ray on 00/00/00 showed chronic bronchial irritation. Pulmonary function tests on 11/11/11 showed an FVC of 86.9% predicted pre-bronchodilator and 95.5% post-bronchodilator. FEV-1 pre-bronchodilator was 66.4% of predicted. Post-bronchodilator it was 72.5%. The FEV-1/FVC ratio was 0.60 both pre-bronchodilator and post-bronchodilator.

FEV1 extraction algorithm

- For the document corpus, identify a set of documents (D) containing the keyword “FEV” (using full-text search)

- For each document (d_i) in D, extract 20-character snippets (each of which begin with “FEV”)

- Create a subset of snippets (S_{di}) such that each snippet begins with one of the following substrings: "FEV=", "FEV -1", "FEV- 1" and "FEV- 1."

FEV1 extraction algorithm

S_{di}

- Snippets with substrings: FEV=, FEV 1, FEV -1, FEV-1

- For each snippet s_j in S_{di} , extract the numeric value v_j

- Create a subset S'_{di} such that each snippet's numeric value (v_k) satisfies the following: $0.5 \leq v_k \leq 5.5$

We implemented the algorithm using SQL with fulltext search feature supported by MS SQLServer

Evaluating SQL tool performance

- Chart review by pulmonologist as reference standard
  - CDW FEV_1 value
  - SQL FEV_1 value extracted

Results

- 5,958 unique patients with 18,183 documents including FEV1

SQL tool increased FEV1 yield by 3849 (21%) compared with CDW alone

Results

Comparing SQL tool with chart review for FEV1 extraction (N=128 documents for 117 unique patients)	
Positive predictive value for identifying FEV1	89%
Kappa for correctly identifying “FEV1” entity	0.66
Spearman’s correlation (among quantifiable FEV1)	0.99

Session roadmap

- PFTs, chronic obstructive pulmonary disease (COPD) clinical importance and COPD research
- Existing data sources to identify PFTs values
- Methods to extract data from existing sources
- Structured query language (SQL) tools for extracting PFT data
- **Clinical example of use of extracted PFT data**

Lung cancer risk factors in general population

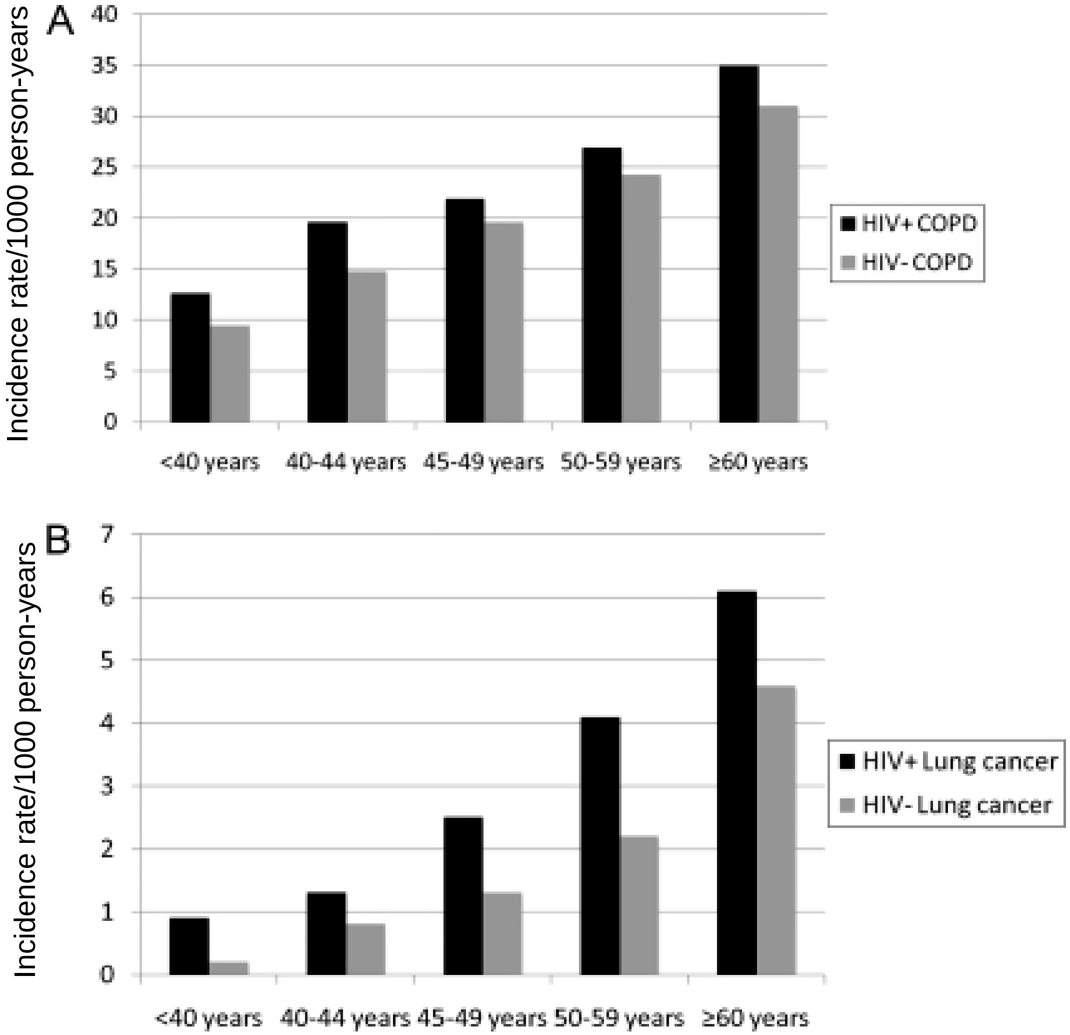
- Established lung cancer risk factors
  - Smoking
  - Age
  - COPD
  - Occupational factors

- Inconsistent associations between COPD severity of airflow limitation and lung cancer risk

COPD and lung cancer more common in HIV+

COPD

Lung cancer



Crothers et al. AJRCCM. 2011

COPD severity, lung cancer and HIV

- We asked how severity of airflow limitation impacted lung cancer risk in COPD patients
  - Does HIV status affect these associations?

- Used the Veterans Aging Cohort Study (VACS)

Data source: VACS

- HIV-infected (HIV+) and uninfected Veterans
    - 47,700 HIV+
    - 98,500 Uninfected
    - Matched by age, race, gender, VA site of care
    - Ongoing data collection but includes 1996-2015

Initial development in VACS

- COPD diagnosis based on ICD-9 codes (491.2x, 493.2x, 496) between 2000-2015

- Primary outcome: incident lung cancer

- Severity determined by extracted FEV_1 values ≥ 6 months prior to cancer dx from structured and unstructured data
  - GOLD stages 1-4 based on severity (3 & 4 collapsed due to small n)

Results

- 8612 (27% HIV+, 73% uninfected) with COPD and FEV1

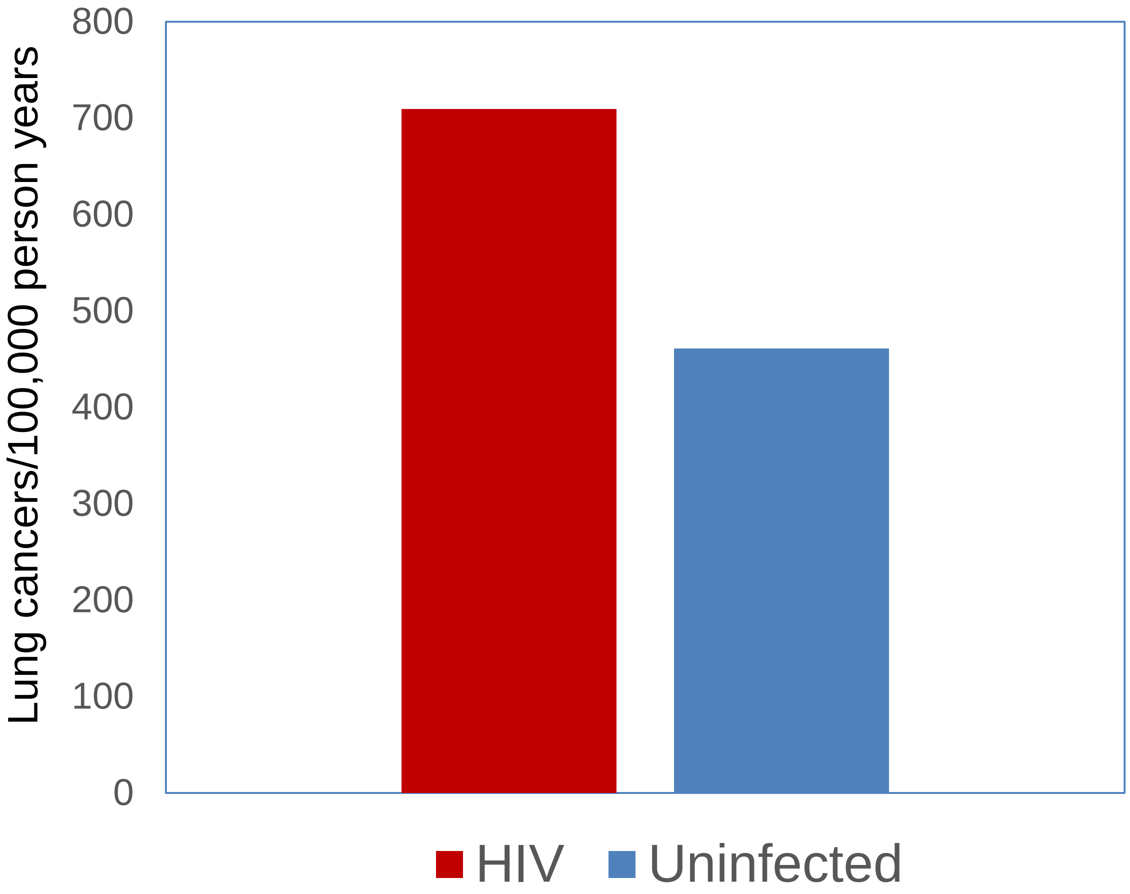
- Age and race similar between HIV+ and uninfected

- HIV+ more likely current smokers (70% vs. 64%; p < 0.001)

- GOLD stages similar between HIV+ and uninfected

GOLD 1	36%
GOLD 2	41%
GOLD 3&4	22%

Incidence of lung cancer higher in HIV+



Akgün et al. *Accepted, ATS*. 2018

Multivariable model for lung cancer risk

	Reference	Incidence rate ratio (IRR)	95% CI
HIV	Uninfected	1.34	(1.08, 1.70)
Age	Per 10 years	2.00	(1.79, 2.22)
Race/Ethnicity	White		
Black		0.97	(0.78, 1.19)
Hispanic		0.54	(0.31, 0.96)
Other		1.42	(0.67, 3.01)
Smoking status	Never		
Current		4.80	(3.04, 7.60)
Former		2.38	(1.44, 3.94)
COPD severity	GOLD Stage 1		
GOLD 2		1.30	(1.02, 1.64)
GOLD 3/4		1.45	(1.10, 1.92)

Akgün et al. *Accepted ATS*. 2018

Multivariable model for lung cancer risk

	Reference	Incidence rate ratio (IRR)	95% CI
HIV	Uninfected	1.34	(1.08, 1.70)
Age	Per 10 years	2.00	(1.79, 2.22)
Race/Ethnicity	White		
Black		0.97	(0.78, 1.19)
Hispanic		0.54	(0.31, 0.96)
Other		1.42	(0.67, 3.01)
Smoking status	Never		
Current		4.80	(3.04, 7.60)
Former		2.38	(1.44, 3.94)
COPD severity	GOLD Stage 1		
GOLD 2		1.30	(1.02, 1.64)
GOLD 3/4		1.45	(1.10, 1.92)

No interaction between COPD severity and HIV for lung cancer

Akgün et al. *Accepted ATS*. 2018

Conclusions

- Difficult to ascertain COPD severity
  - FEV_1 values can be extracted using SQL queries
  - Excellent ascertainment and good accuracy
  - Increases the yield for complete FEV_1 values in VA data

- Clinical questions addressed:
  - COPD severity is associated with lung cancer risk in HIV+ and uninfected
  - No interaction between HIV and COPD severity

Next steps

Develop COPD phenotype in VA populations incorporating FEV_1 for disease severity

- Evaluate whether SQL tool values have similar predictive values for outcomes

Apply rules to other electronic health record systems (e.g., Epic)

Acknowledgement & thanks to co-authors

- Keith Sigel, MD, MS
- Kei Cheung, PhD
- Farah Kidwai-Khan, MS
- Alex Bryant, MAS
- Cindy Brandt, MD, MPH
- Amy Justice, MD, PhD
- Kristina Crothers, MD

The work is supported by the following Yale Cancer Center and VACS grants:

- NIH/NCI Supplement P30CA016359-37S4
- U10 AA013566-completed
- U24 AA020794
- U01 AA020790
- Department of Veterans Affairs, Veterans Health Administration, and Office of Research and Development

Additional Resources

VIReC Options for Specific Questions

- Community knowledge sharing
- ~1,300 VA data users
- Researchers, operations, data stewards, managers
- Subscribe by visiting http://vaww.virec.research.va.gov/Support/HSRData-L.htm (VA Intranet)

HelpDesk

- Individualized support

virec@va.gov

(708) 202-2413

Quick links for VA data resources

Quick Guide: Resources for Using VA Data
http://vaww.virec.research.va.gov/Toolkit/QG-Resources-for-Using-VA-Data.pdf (VA Intranet)

VIReC: http://vaww.virec.research.va.gov/Index.htm (VA Intranet)

VIReC Cyberseminars: http://www.virec.research.va.gov/Resources/Cyberseminars.asp

VHA Data Portal: http://vaww.vhadataportal.med.va.gov/Home.aspx (VA Intranet)

VINCI: http://vaww.vinci.med.va.gov/vincicentral/ (VA Intranet)

Health Economics Resource Center (HERC): http://vaww.herc.research.va.gov (VA Intranet)

CDW: https://vaww.cdw.va.gov/Pages/CDWHome.aspx (VA Intranet)

Archived cyberseminar: What can the HSR&D Resource Centers do for you?
http://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=101

Contact information

VA Information Resource Center

Hines VA Hospital

virec@va.gov

708-202-2413

Kathleen Akgun
Kathleen.Akgun@va.gov

Next session:
September 10th @ 1 pm Eastern

Database & Methods Cyberseminar Series

Assessing Comorbidity

Denise M. Hynes, PhD
VA Information Resource Center