

Overview of Clinical Data Science

6.871/HST.956: Machine Learning for Healthcare

February 8, 2022

Dr. Madhur Nayan

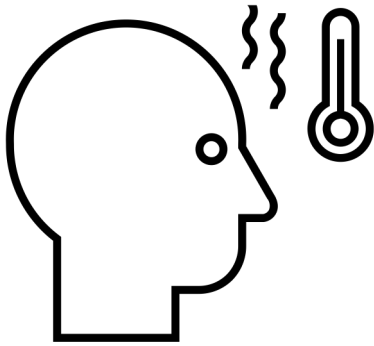


ML for Health Conferences

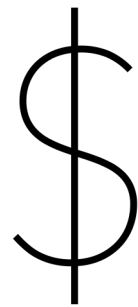
- Machine Learning for Health (ML4H)
 - Previously a NeurIPS workshop, separate symposium as of 2021
 - Last year, submissions due September
- Machine Learning for Healthcare (MLHC)
 - Submission deadline: April 14, 2022
 - Duke University, August 5-6th, 2022
- Symposium on Artificial Intelligence for Learning Health Systems (SAIL)
 - Submission deadline: TBD
 - Bermuda, May 23-25, 2022
- Conference on Health, Inference, and Learning (CHIL)
 - Submission deadline: January 14, 2022
- And more (NeurIPS, ICML, AAAI, etc.)

Stakeholders in Healthcare

Providers



Patient



Payer

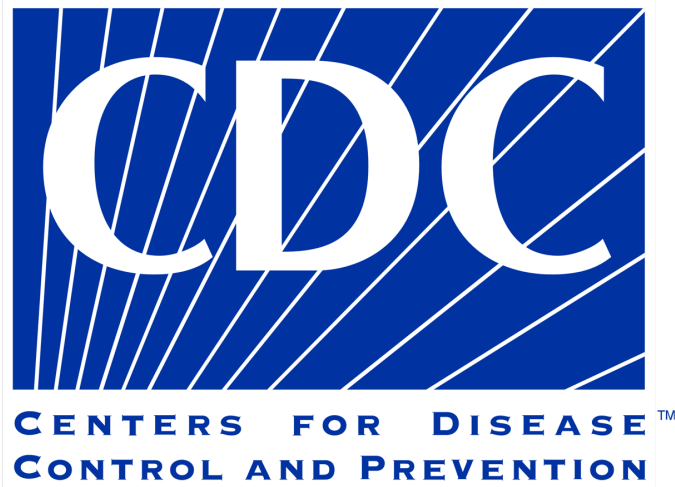


Policymaker

“The Four Ps” of healthcare

Stakeholders have different goals and expectations from the healthcare system

Policymaker



If You were EXPOSED to COVID-19 & Unvaccinated OR Vaccinated >6 mo. ago with Pfizer or Moderna vaccine or >2 mo. ago with J&J vaccine

Day 1-5 → **Day 5** → **Day 6-10**

Stay home Test if Possible Continue to wear a mask around others

If you can't quarantine you must wear a mask for 10 days.

If you develop symptoms get a test and stay home.

cdc.gov/coronavirus

If You Test POSITIVE for COVID-19 (regardless of vaccination status)

Day 1-5 → **Day 6-10**

Stay home If you have **no symptoms** or your **symptoms are resolving**, you can leave your house—continue to wear a mask around others.

If you have a fever, continue to stay home until your fever resolves.

cdc.gov/coronavirus

If You were EXPOSED to COVID-19 & Boosted

Day 1 → **Day 5** → **Day 10**

Wear a mask around others for 10 days. Test on day 5, if possible

If you develop symptoms get a test and stay home.

cdc.gov/coronavirus

If You were EXPOSED to COVID-19 & Unvaccinated OR Vaccinated >6 mo. ago with Pfizer or Moderna vaccine or >2 mo. ago with J&J vaccine

Day 1-5 → **Day 5** → **Day 6-10**

Stay home Test if Possible Continue to wear a mask around others

If you can't quarantine you must wear a mask for 10 days.

If you develop symptoms get a test and stay home.

cdc.gov/coronavirus

https://en.wikipedia.org/wiki/Centers_for_Disease_Control_and_Prevention

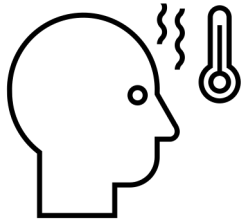
<https://www.stetson.edu/other/safer-stetson/isolation.php>

Overview of Clinical Data Science

- Topics of Discussion
 - Goals of Clinical Data Science
 - Sources of Clinical Data
 - Exploring Clinical Data
 - Challenges of Working with Clinical Data
 - Applying Clinical Data Science

Goals of Clinical Data Science

- Overall goal is to improve population health in a resource-effective manner
- Stakeholders have **different immediate goals** with clinical data



Patient



Providers



Payer



Policymaker

Mrs. Patel

- 65 year old female
- Presents to the ER with abdominal pain
- CT scan
 - <https://radiopaedia.org/cases/renal-cell-carcinoma-9>
- She is discharged from the ER and outpatient follow-up is arranged



Case courtesy of Dr Roberto Schubert,
Radiopaedia.org, rID: 14439

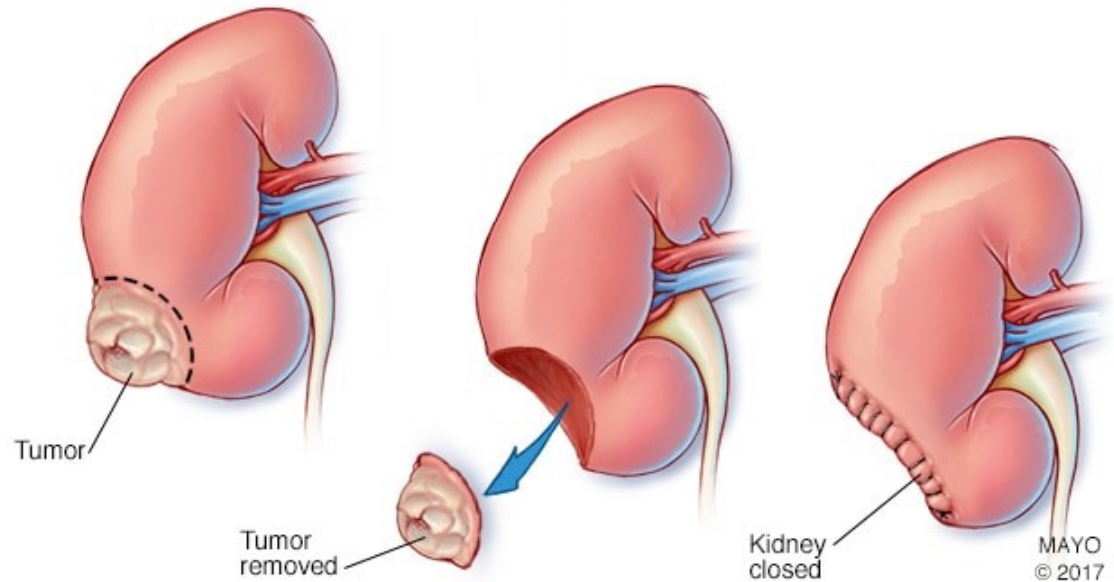
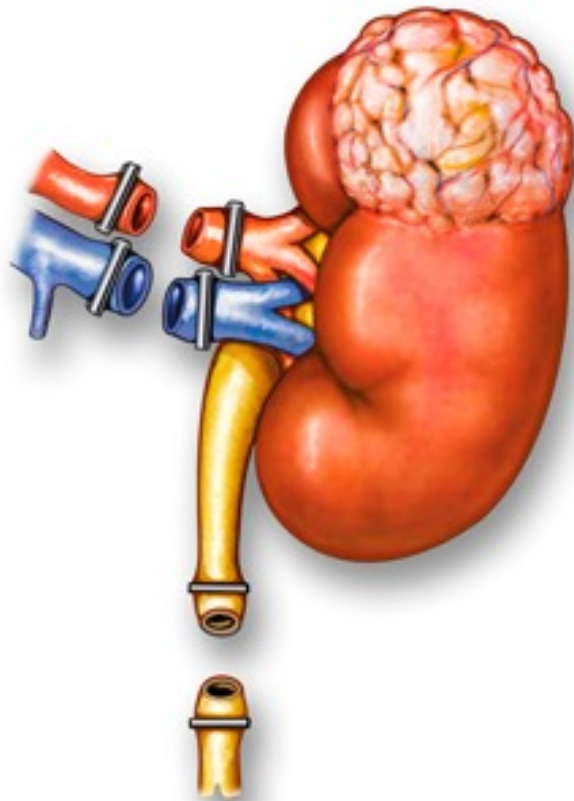
Patient/Provider Goals of Clinical Data Science

- Mrs. Patel is a 65 year old who was recently diagnosed with kidney cancer. She presents to your office. You discuss the diagnosis and treatment options. She has some questions.
 - After treatment, **what is the risk** of my cancer coming back before the Ultimate World Cruise (December 2023)?
 - **Will the risk** of my cancer coming back **change** if I get a partial nephrectomy instead of a radical nephrectomy?

Radical Nephrectomy

VS

Partial Nephrectomy



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

MAYO
© 2017

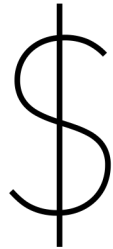
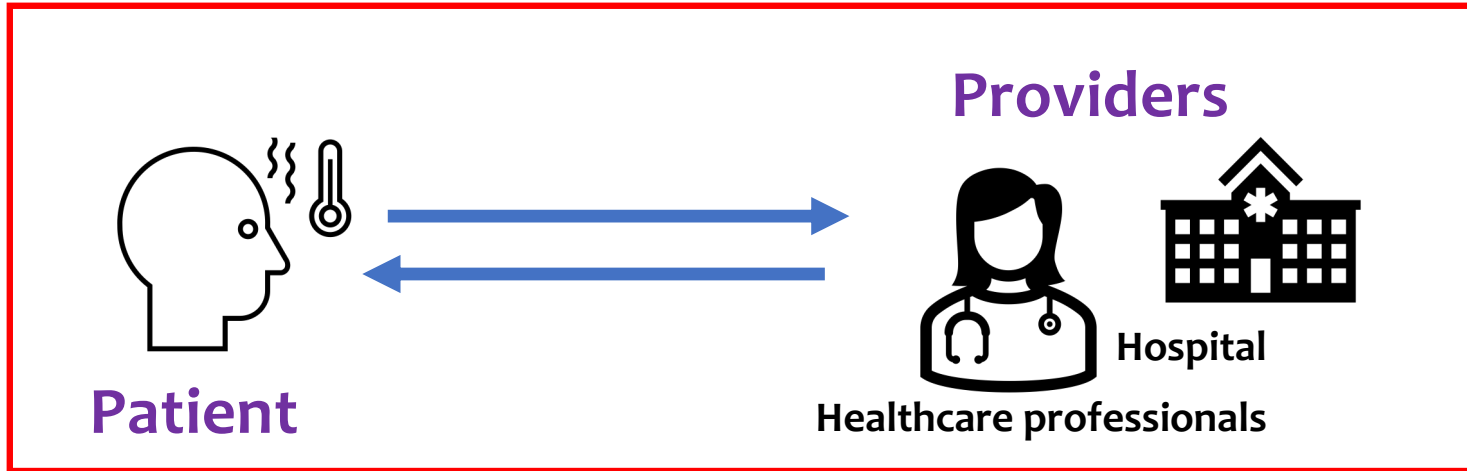
<https://www.fairbanksurology.com/robotic-radical-nephrectomy>
<https://www.mayoclinic.org/tests-procedures/nephrectomy/multimedia/img-20332175>

Patient/Provider Goals of Clinical Data Science

- Mrs. Patel is a 65 year old who was recently diagnosed with kidney cancer. She presents to your office. You discuss the diagnosis and treatment options. She has some questions.
 - After treatment, **what is the risk** of my cancer coming back before the Ultimate World Cruise (December 2023)?
 - **Will the risk** of my cancer coming back **change** if I get a partial nephrectomy instead of a radical nephrectomy?

How would you answer these questions using clinical data science?

Sources of Clinical Data



Payer



Policymaker

Provider Derived Data

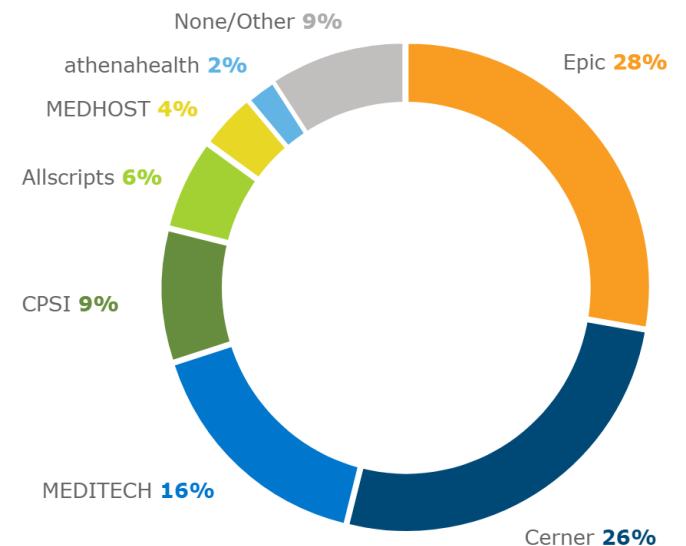
- Previously, paper charts were used for clinical documentation.
 - What are problems with paper charts?
- Electronic health records (EHR) are a digital version of the patient's paper chart.
 - Providers are reimbursed based on the EHR
- Examples of EHR databases: MIMIC, Mass General Brigham Research Patient Data Registry (RPDR)

Electronic Health Records in the US

- Different hospitals use different EHR systems
 - Largest EHR systems in US
 1. EPIC
 2. Cerner
 3. Meditech
 - To efficiently and accurately share clinical information, EHRs must be interoperable
 - Current EHRs are not interoperable

2018 US Acute Care Hospital Market Share

(n=5,447 acute care hospitals)



EPIC

The screenshot displays the EPIC Hyperspace interface. At the top, there are several notification banners: "Hyperspace", "2 : Chart Completion", "2 : My Incomplete Notes", "1 : Research ADT Event Notific...", and "0". Below these are navigation tabs: "Learning", "Home", "Schedule", "Patient Lists", "In Basket", "Queries", "Patient Station", "Status Board", "Mark Patients For Merge", and "Remind Me". A search bar with the name "Stork" is visible on the right. The main navigation bar includes buttons for "SnapShot", "Chart Review" (highlighted with a red box), "CRISP", "Room...", "Plan", "Wra...", "HM", "Results", "Synopsis", "Medic...", and "Immun...". Below this, the "Chart Review" section is active, with sub-tabs for "Encounters", "SnapShot", "Notes", "Labs", and "Pathology" (highlighted with a red box). The "Pathology" sub-tab is selected, showing options like "Preview", "Refresh (2:06 PM)", "Select All", "Deselect All", "Review Selected", "Route", "Lab Flowsheet", and "Add to Bookmarks". There are also filter options: "Filters", "Results Only", "Completed/Resulted", and "Hide Canceled". At the bottom, a table header is visible with columns: "Date/Time", "Specimen ID", "Test Type", "Status", "Collected by", and "Enco".

<http://apps.pathology.jhu.edu/team-path-md/pathology-for-core-clinical-clerkships/how-to-find-pathology-results-and-reports-on-epic/>

EMERGENCY MEDICINE EVALUATION NOTE

History of Present Illness

Chief Complaint: @EDCC@

HPI: @NAME@ is a @AGE@ @SEX@ ***

ROS: A complete 11 system ROS was performed (constitutional, eyes, ENMT, cardiovascular, respiratory, gastrointestinal, genitourinary, musculoskeletal, skin, neurological, psychiatric) and was negative aside from the pertinent positives and negatives noted in the HPI.

Previous History

@PMH@
 @PSH@
 @SOCH@
 @FAMHX@
 @ALLERGY@
 @MEDSCONDENSED@

Physical Exam

@VSHOSP@

Results

@EDLABS@
 @EDRADIOLOGY@
 The laboratory results, imaging results and other diagnostic exam results were reviewed in the EMR.

ED Course & Medical Decision Making

@EDMEDS@
 @EDCOURSE@

Procedures

@PROCDOC@

Diagnosis

@DIAGX@

Disposition

***Discharged
 @EDDISCHARGERX@

Sections of note

- History of Present Illness
- Previous History
- Physical Exam
- Results
- ED Course & Medical Decision Making
- Procedures
- Diagnosis
- Disposition

<https://www.acep.org/administration/quality/health-information-technology/epic-articles/things-you-can-do-on-your-own-epic/>

Star icon | **B** | abc | Undo | Redo | ? | + | Insert SmartText | Left Arrow | Right Arrow | SmartList | Menu Icon

1 2 3 4 5 6 7 8

EMERGENCY MEDICINE EVALUATION NOTE

History of Present Illness

Chief Complaint: @EDCC@

HPI: @NAME@ is a @AGE@ @SEX@ ***

ROS: A complete 11 system ROS was performed (constitutional, eyes, ENMT, cardiovascular, respiratory, gastrointestinal, genitourinary, musculoskeletal, skin, neurological, psychiatric) and was negative aside from the pertinent positives and negatives noted in the HPI.

<https://www.acep.org/administration/quality/health-information-technology/epic-articles/things-you-can-do-on-your-own-epic/>

☆ **B** + abc ↶ ↷ ? + Insert SmartText ↶ ↷ ➡ Insert SmartList ☰

⌂ 1 2 3 4 5 6 7 8

EMERGENCY MEDICINE EVALUATION NOTE

History of Present Illness

Chief Complaint: @EDCC@

HPI: @NAME@ is a @AGE@ @SEX@ ***

ROS: A complete 11 system ROS was performed (constitutional, eyes, ENMT, cardiovascular, respiratory, gastrointestinal, genitourinary, musculoskeletal, skin, neurological, psychiatric) and was negative aside from the pertinent positives and negatives noted in the HPI.

Previous History

@PMH@
 @PSH@
 @SOCH@
 @FAMHX@
 @ALLERGY@
 @MEDSCONDENSED@

Physical Exam

@VSHOSP@

Results

@EDLABS@
 @EDRADIOLOGY@
 The laboratory results, imaging results and other diagnostic exam results were reviewed in the EMR.

ED Course & Medical Decision Making

@EDMEDS@
 @EDCOURSE@

Procedures

@PROCDOC@

Diagnosis

@DIAGX@

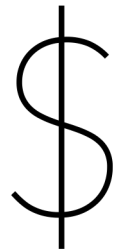
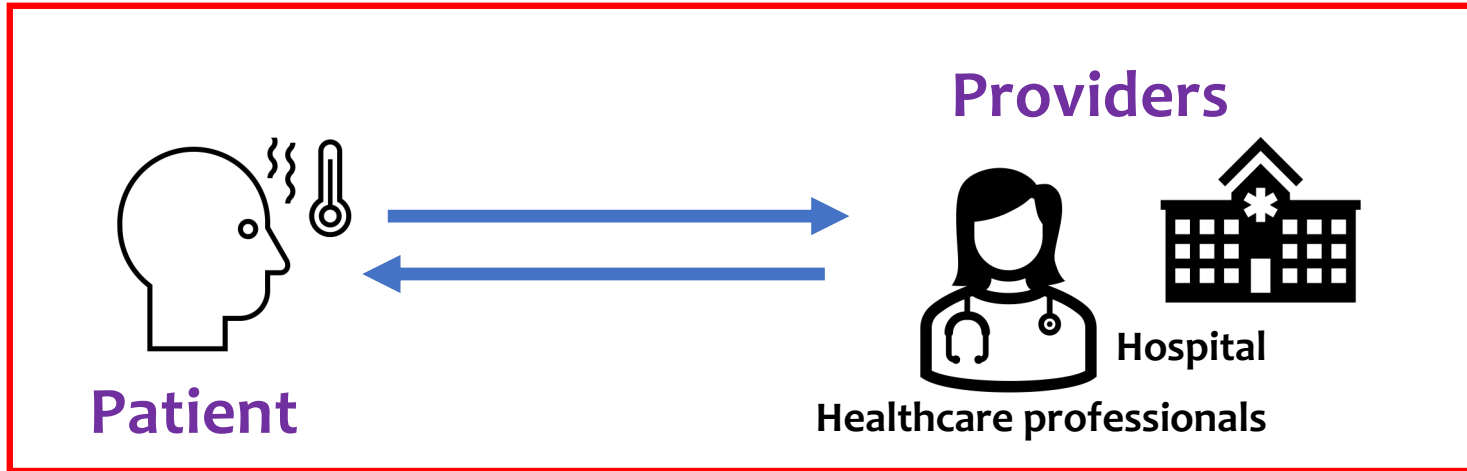
Disposition

***Discharged
 @EDDISCHARGERX@

What are potential problems with template notes?

<https://www.acep.org/administration/quality/health-information-technology/epic-articles/things-you-can-do-on-your-own-epic/>

Sources of Clinical Data



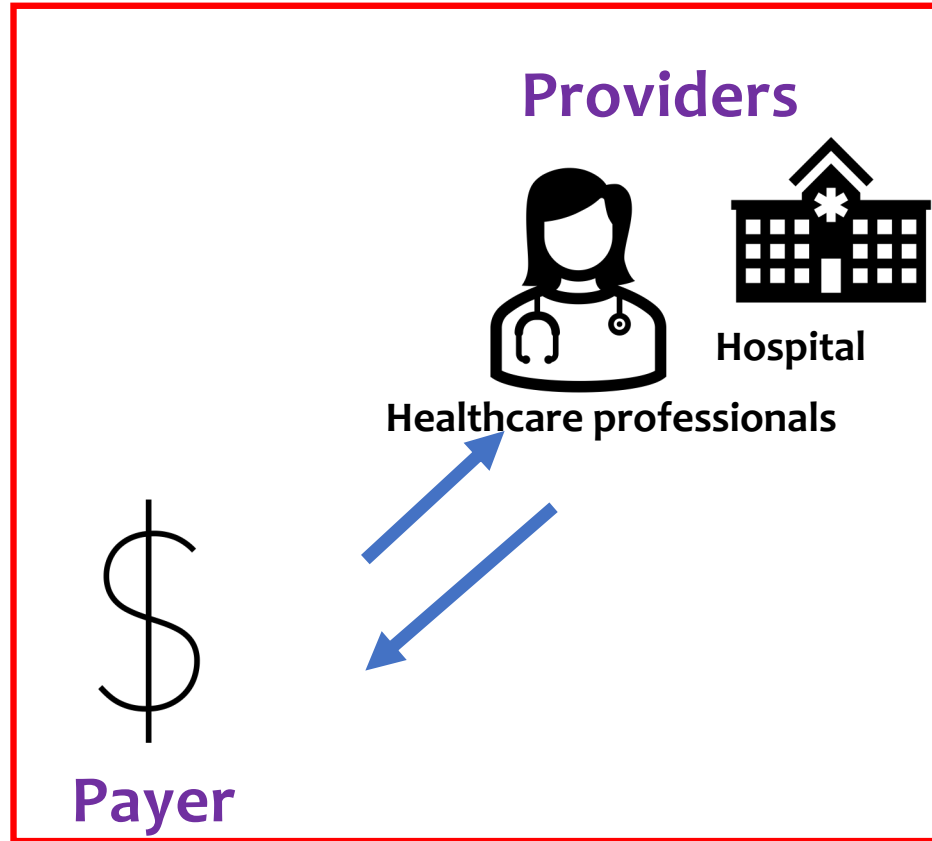
Payer



Policymaker

What are potential limitations of using EHR data only?

Sources of Clinical Data



Payer Derived Data

- Claims data
 - Consists of the **billing codes** that **providers** (physicians, hospitals, pharmacies, and other health care providers) submit to **payers**
 - Examples: IQVIA, IBM MarketScan, Optum, Medicare

<https://www.optum.com/content/dam/optum/resources/whitePapers/Benefits-of-using-both-claims-and-EMR-data-in-HC-analysis-WhitePaper-ACS.pdf>


Payer Derived Data

- Medicare Claims Data
 - Medicare
 - Federal health insurance program
 - Covers
 - Age \geq 65
 - Certain people under 65 with disabilities
 - People of any age with End Stage Renal Disease or amyotrophic lateral sclerosis

<https://www.sgim.org/communities/research/dataset-compendium/medicare-claims-data>

Payer Derived Data

CMS1500

		ABC Insurance Company Suite 600 567 Insurance Lane Big City IL 80605		CARRIER
HEALTH INSURANCE CLAIM FORM APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE (NUCC) 02/12				
PICA <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>		PICA <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>		
1. MEDICARE <input type="checkbox"/> (Medicare#)		MEDICAID <input type="checkbox"/> (Medicaid#)		
TRICARE <input type="checkbox"/> (ID#DoDM)		CHAMPVA <input type="checkbox"/> (Member ID#)		
GROUP HEALTH PLAN <input checked="" type="checkbox"/> (ID#)		FECA Bx/LUNG <input type="checkbox"/> (ID#)		
OTHER <input type="checkbox"/> (ID#)		1a. INSURED'S I.D. NUMBER (For Program in Item 1) X0123456789		
2. PATIENT'S NAME (Last Name, First Name, Middle Initial) Doe Jr, John, J		3. PATIENT'S BIRTH DATE MM DD YY 01 01 1987		
5. PATIENT'S ADDRESS (No., Street) 123 Main Street		6. PATIENT RELATIONSHIP TO INSURED Self <input type="checkbox"/> Spouse <input type="checkbox"/> Child <input checked="" type="checkbox"/> Other <input type="checkbox"/>		
CITY Anytown		7. INSURED'S ADDRESS (No., Street) 123 Main Street		
STATE IL		8. RESERVED FOR NUCC USE		
ZIP CODE 60610		CITY Anytown		
TELEPHONE (Include Area Code) (312) 5551212		STATE IL		
9. OTHER INSURED'S NAME (Last Name, First Name, Middle Initial) Doe, Mary, A		10. IS PATIENT'S CONDITION RELATED TO: a. EMPLOYMENT? (Current or Previous) <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO		
a. OTHER INSURED'S POLICY OR GROUP NUMBER X9876543210		b. AUTO ACCIDENT? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO		
b. RESERVED FOR NUCC USE		c. OTHER ACCIDENT? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO		
c. RESERVED FOR NUCC USE		11. INSURED'S POLICY GROUP OR FECA NUMBER A1234		
d. INSURANCE PLAN NAME OR PROGRAM NAME XYZ Insurance Company		a. INSURED'S DATE OF BIRTH MM DD YY 01 01 1958		
12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below. SIGNED Signature on File		b. OTHER CLAIM ID (Designated by NUCC) Y4 112233445566		
DATE 09/30/12		c. INSURANCE PLAN NAME OR PROGRAM NAME ABC Insurance Company		
13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below. SIGNED SOF		d. IS THERE ANOTHER HEALTH BENEFIT PLAN? <input checked="" type="checkbox"/> YES <input type="checkbox"/> NO # yes, complete items 9, 9a, and 9d.		
READ BACK OF FORM BEFORE COMPLETING & SIGNING THIS FORM.		PATIENT AND INSURED INFORMATION		

<https://fiachraforms.com/shop/1500-02-12-standard-paper-claim-form/>

Payer Derived Data

Diagnosis codes

CMS1500

Procedure codes

14. DATE OF CURRENT ILLNESS, INJURY, or PREGNANCY (LMP) MM DO YY QIAL 09 30 2012 431				15. OTHER DATE QUAL 454 MM DO YY 09 25 2012				16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION FROM MM DO YY TO MM DO YY 09 25 2012 TO 10 28 2012											
17. NAME OF REFERRING PROVIDER OR OTHER SOURCE DN Jane A Smith MD				17a. G2 ABC1234567890 17b. NPI 0123456789				18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES FROM MM DO YY TO MM DO YY 09 25 2012 TO 09 28 2012											
19. ADDITIONAL CLAIM INFORMATION (Designated by NUCC)				20. OUTSIDE LAB? <input checked="" type="checkbox"/> YES <input type="checkbox"/> NO \$ CHARGES 112500 00				22. REG/EMISSION CODE 7 ORIGINAL REF. NO. ABC12334567890											
21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY (Relate A-L to service line below (24E)) A. 998.59 B. 780.6 C. V18.0 D. E878.8 ICD Ind 9				23. PRIOR AUTHORIZATION NUMBER															
24. A. DATE(S) OF SERVICE		B. PLACE OF SERVICE		C. EMG		D. PROCEDURES, SERVICES, OR SUPPLIES		E. DIAGNOSIS		F. \$ CHARGES		G. DAYS OR UNITS		H. ICD-9-CM		I. RENDERING PROVIDER ID #			
From MM DO YY To MM DO YY		MM DO YY		Y N		CPT/HCPCS MODIFIER		POSTER						NPI		Z5678901234			
1 09 30 12 09 30 12 11 Y		11 Y		99241 25		ABCD		50 00		2 Y		NPI 9876543210		12345678901					
2 10 01 11 01 01 11 11 N		N		A6410 P2		ABSS		45 00		2 N		NPI 0123456789							
3												NPI							
4												NPI							
5												NPI							
6												NPI							
25. FEDERAL TAX I.D. NUMBER				26. PATIENT'S ACCOUNT NO. 12341234				27. ACCEPT ASSIGNMENT? <input checked="" type="checkbox"/> YES <input type="checkbox"/> NO				28. TOTAL CHARGE \$ 190 00				29. AMOUNT PAID \$			
31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS (I certify that the statements on the reverse apply to this bill and are made a part thereof.) Joe Smith MD Signed 09/30/12 DATE				32. SERVICE FACILITY LOCATION INFORMATION General Hospital 9876 Hospital Street Anytown IL 60610-9876 a. 567891234 b. G2A1234567890				33. BILLING PROVIDER INFO & PH# (312) 5552222 Physician Practice Inc 1234 Healthcare Street Anytown IL 60610-1234 a. 9876543210 b. G2Z5678901234				30. Rcvd for NUCC Use							

NUCC Instruction Manual available at: www.nucc.org

PLEASE PRINT OR TYPE

APPROVED OMB-0938-1197 FORM 1500 (02-12)

<https://fiachraforms.com/shop/1500-02-12-standard-paper-claim-form/>

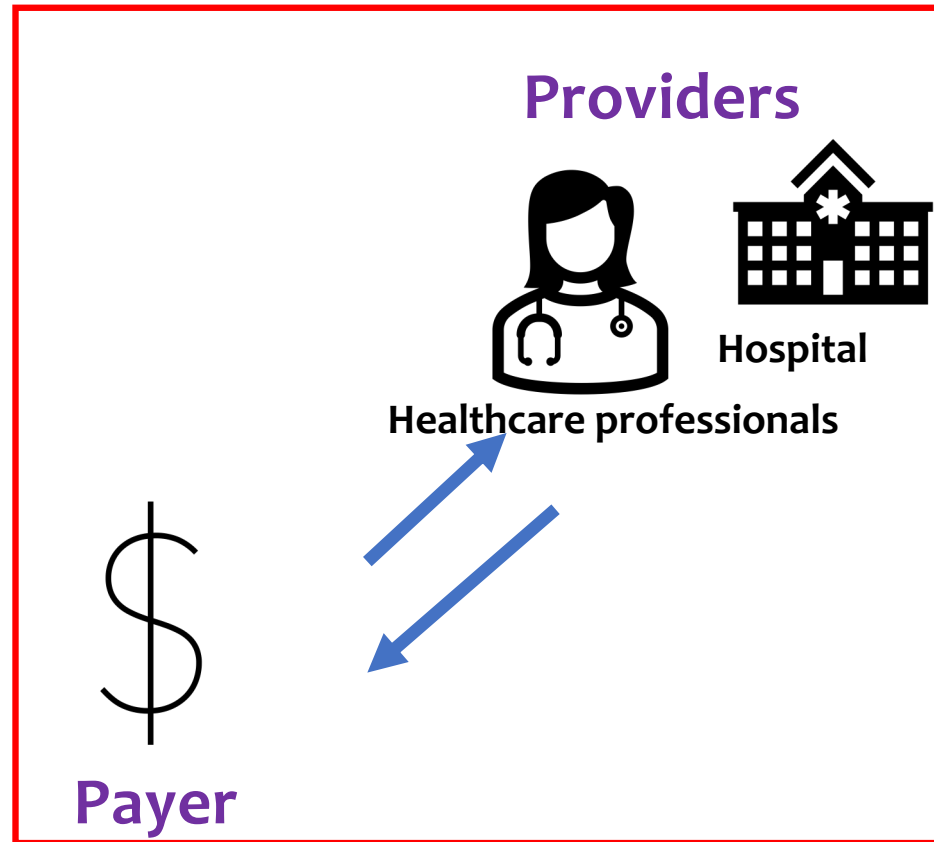
Payer Derived Data

- All-payer claims databases
 - Large State databases that include claims from private and public payers
 - Massachusetts All-Payer Claims Database
 - Releases data extracts to government agencies, payers, providers, provider organizations, and researchers
 - All applications to access the data are reviewed for conformity with legal requirements

<https://www.ahrq.gov/data/apcd/index.html>

<https://www.chiamass.gov/ma-apcd/>

Sources of Clinical Data

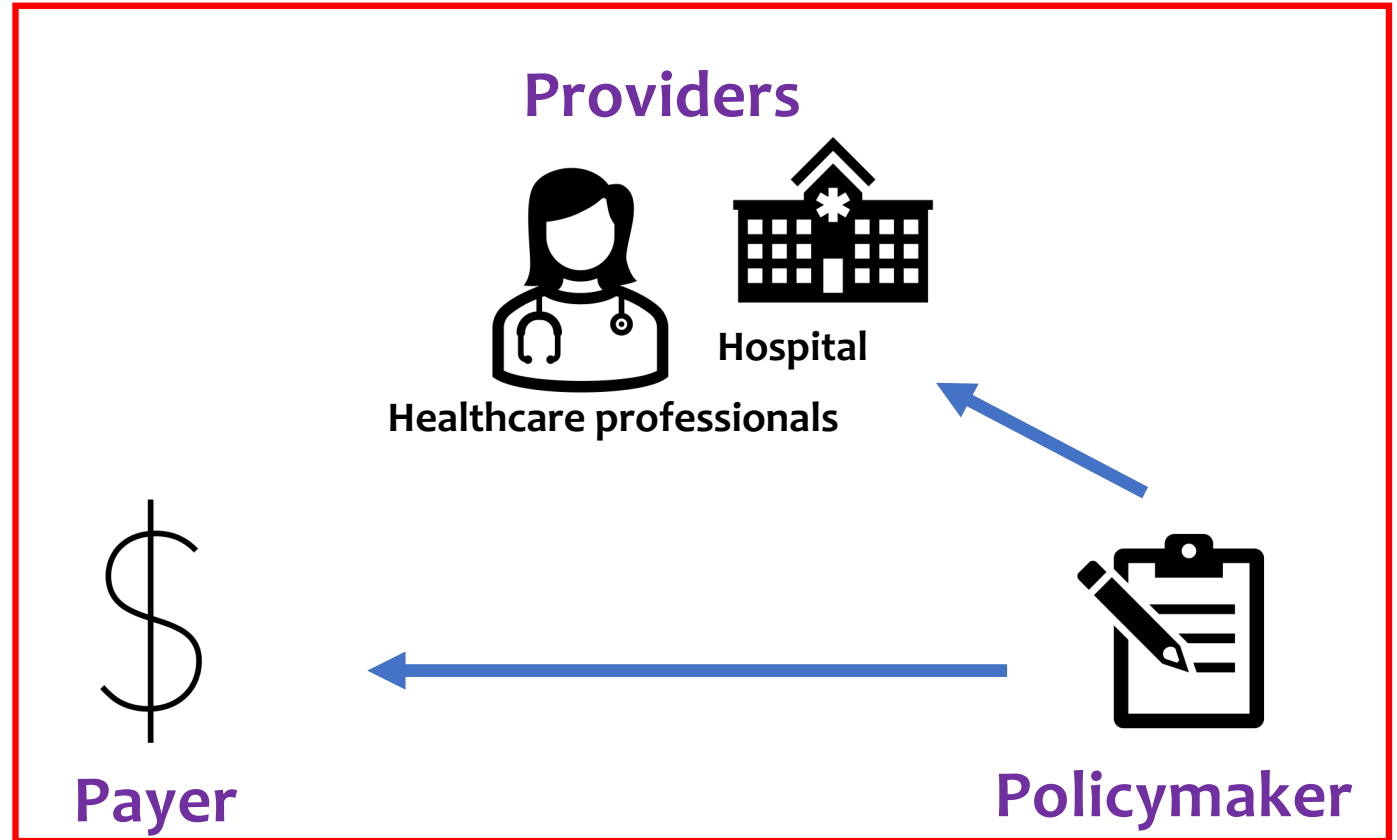


What are potential limitations of using payer data only?



Policymaker

Sources of Clinical Data



Policymaker Derived Data

- **National Cancer Database (NCDB)**

- Hospital registry data from **Commission on Cancer (CoC)**-accredited facilities
 - What is COC?
 - A program, from the American College of Surgeons, that **recognizes cancer care programs** for providing comprehensive, high-quality, and multidisciplinary patient centered care.
 - CoC accreditation
 - Granted to facilities that demonstrate compliance with the **CoC standards**

CoC Standards

1 Institutional Administrative Commitment	1
1.1 Administrative Commitment	3
2 Program Scope and Governance	5
2.1 Cancer Committee	7
2.2 Cancer Liaison Physician	9
2.3 Cancer Committee Meetings	10
2.4 Cancer Committee Attendance	11
2.5 Multidisciplinary Cancer Case Conference	12
3 Facilities and Equipment Resources	15
3.1 Facility Accreditation	17
3.2 Evaluation and Treatment Services	18
4 Personnel and Services Resources	21
4.1 Physician Credentials	23
4.2 Oncology Nursing Credentials	24
4.3 Cancer Registry Staff Credentials	26
4.4 Genetic Counseling and Risk Assessment	28
4.5 Palliative Care Services	31
4.6 Rehabilitation Care Services	33
4.7 Oncology Nutrition Services	34
4.8 Survivorship Program	36

5 Patient Care: Expectations and Protocols	39
5.1 College of American Pathologists Synoptic Reporting	41
5.2 Psychosocial Distress Screening	43
5.3 Sentinel Node Biopsy for Breast Cancer	45
5.4 Axillary Lymph Node Dissection for Breast Cancer	47
5.5 Wide Local Excision for Primary Cutaneous Melanoma	49
5.6 Colon Resection	50
5.7 Total Mesorectal Excision	52
5.8 Pulmonary Resection	53
6 Data Surveillance and Systems	55
6.1 Cancer Registry Quality Control	57
6.2 Data Submission (<i>Retired in 2021</i>)	59
6.3 Data Accuracy (<i>Retired in 2021</i>)	60
6.4 Rapid Cancer Reporting System: Data Submission	61
6.5 Follow-Up of Patients	62
7 Quality Improvement	65
7.1 Accountability and Quality Improvement Measures	67
7.2 Monitoring Concordance with Evidence-Based Guidelines	68
7.3 Quality Improvement Initiative	70
7.4 Cancer Program Goal	72

8 Education: Professional and Community Outreach	75
8.1 Addressing Barriers to Care	77
8.2 Cancer Prevention Event	78
8.3 Cancer Screening Event	80
9 Research	83
9.1 Clinical Research Accrual	85
9.2 Commission on Cancer Special Studies	87

https://www.facs.org/-/media/files/quality-programs/cancer/coc/optimal_resources_for_cancer_care_2020_standards.ashx

CoC Standards

- **5.1 College of American Pathologists Synoptic Reporting**
- **Definition and Requirements: 90% of the eligible cancer pathology reports are structured using synoptic reporting format as defined by the College of American Pathologists (CAP) cancer protocols, including containing all core data elements within the synoptic format.**

https://www.facs.org/-/media/files/quality-programs/cancer/coc/optimal_resources_for_cancer_care_2020_standards.ashx

CoC Standards

1 Institutional Administrative Commitment	1
1.1 Administrative Commitment	3
2 Program Scope and Governance	5
2.1 Cancer Committee	7
2.2 Cancer Liaison Physician	9
2.3 Cancer Committee Meetings	
2.4 Cancer Committee Attendance	
2.5 Multidisciplinary Cancer Case Conference	
3 Facilities and Equipment Resources	15
3.1 Facility Accreditation	17
3.2 Evaluation and Treatment Services	18
4 Personnel and Services Resources	21
4.1 Physician Credentials	23
4.2 Oncology Nursing Credentials	
4.3 Cancer Registry Staff Credentials	
4.4 Genetic Counseling and Risk Assessment	
4.5 Palliative Care Services	
4.6 Rehabilitation Care Services	33
4.7 Oncology Nutrition Services	34
4.8 Survivorship Program	36

5 Patient Care: Expectations and Protocols	39
5.1 College of American Pathologists Synoptic Reporting	41
5.2 Psychosocial Distress Screening	43
5.3 Sentinel Node Biopsy for Breast Cancer	45
5.4 Axillary Lymph Node Dissection for Breast Cancer	47
5.5 Wide Local Excision for Primary Cutaneous Melanoma	49
5.6 Colon Resection	50
5.7 Total Mesorectal Excision	52

Data reporting to NCDDB is required for accreditation

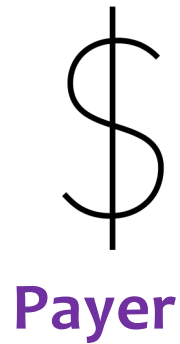
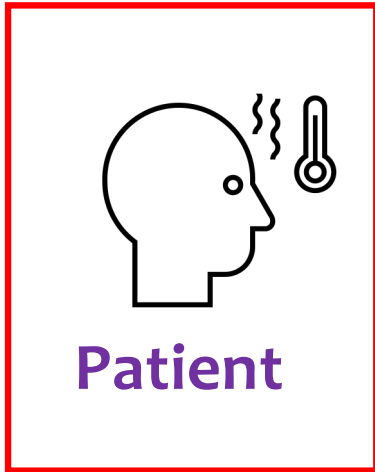
6.2 Data Submission (<i>Retired in 2021</i>)	59
6.3 Data Accuracy (<i>Retired in 2021</i>)	60
6.4 Rapid Cancer Reporting System: Data Submission	61
6.5 Follow-Up of Patients	62

What are potential limitations of using the NCDDB as a data source?

8.2 Education: Professional and Community Outreach	75
8.3 Addressing Barriers to Care	77
8.4 Cancer Prevention Event	78
8.5 Cancer Screening Event	80
9 Research	83
9.1 Clinical Research Accrual	85
9.2 Commission on Cancer Special Studies	87

https://www.facs.org/-/media/files/quality-programs/cancer/coc/optimal_resources_for_cancer_care_2020_standards.ashx

Sources of Clinical Data



Patient Derived Data

- PatientsLikeMe
 - Online community that allows members to find other patients like them, share and track their health data over time, and contribute to scientific research
 - Launched in 2006 for patients with amyotrophic lateral sclerosis
 - For-profit company
 - More than 600,000 registered members across more than 2900 conditions (as of February 2018)
 - Survey of members in 2016-2017
 - 67% furthered their understanding of how their condition could affect them
 - 63% on how to live better with their condition

Wicks, Paul, et al. "Scaling PatientsLikeMe via a "generalized platform" for members with chronic illness: web-based survey study of benefits arising." *Journal of medical Internet research* 20.5 (2018): e9909.

Patient Derived Data

Renal cell carcinoma



HOME

CONDITIONS

TREATMENTS

SYMPTOMS

SEARCH



Sign in

Join now

Members are tracking more than **2,800** conditions on PatientsLikeMe. See what they're saying about yours...

Cancer

Breast , Lung , Liver , Testicular , Prostate , Pancreatic , CLL (Chronic Lymphocytic Leukemia) , Non-Hodgkin's Lymphoma , Thyroid

Developmental and Chromosomal

Tay-Sachs , Autism Spectrum , Down Syndrome

Digestive and Intestinal

Crohn's Disease , IBS , Ulcerative Colitis

Endocrine

Diabetes: Type I , Type II , Hypothyroidism , Hyperthyroidism

Eye, Ear, Nose and Throat

Hearing Loss , Glaucoma , Macular Degeneration

Heart, Blood and Circulatory

Coronary Artery Disease , Hypertension , Iron Deficiency Anemia , Raynaud's Syndrome , Congestive Heart Failure , Cardiomyopathy , Aplastic Anemia

<https://www.patientslikeme.com/conditions/>

Patient Derived Data

Common symptoms reported by people with renal cell cancer

Common symptoms	How bad it is	What people are taking for it
Pain		Pregabalin, Gabapentin, Oxycodone
Fatigue	21 renal cell cancer patients report severe pain (17%) 	Amphetamine, Armodafinil, Motorized scooter/chair
Stress		Aromatherapy
Anxious mood		Clonazepam, Escitalopram, Acupuncture
Depressed mood		Venlafaxine, Sertraline, Aripiprazole

Reports may be affected by other conditions and/or medication side effects. We ask about general symptoms (anxious mood, depressed mood, fatigue, pain, and stress) regardless of condition.

Last updated: February 7, 2022

<https://www.patientslikeme.com/conditions/renal-cell-ca>


What is Venlafaxine?


- An antidepressant in a group of drugs called selective serotonin and norepinephrine reuptake inhibitors (SSNRIs).
- Affects chemicals in the brain that may become unbalanced and cause depression.
- Used to treat major depressive disorder, anxiety, and panic disorder.


Patient Derived Data

117 patient evaluations for Venlafaxine

Sep 3, 2012 (Started Oct 10, 2006)

Effectiveness  Moderate (for major depressive disorder)

Effectiveness  Moderate (for depressed mood)

Side effects  Mild (for Overall) (sexual dysfunction)

Adherence  Always

Burden  Not at all hard to take

Dosage: 100 mg Daily

Advice & Tips: Slight sexual dysfunction. As long as I take it several hours before sexual activity it is no problem. A big benefit is the leveling out of emotions.

Cost: < \$25 monthly

<https://www.patientslikeme.com/treatment/duloxetine>

Patient Derived Data

Research & Care

ResearchKit CareKit Developers Resources Publications Submit your App

Active Tasks

Use active tasks to capture sensor information.

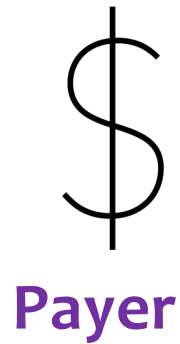
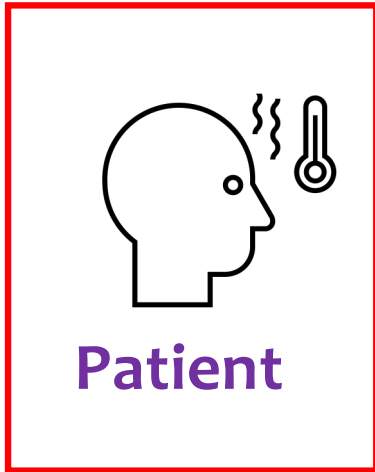
ResearchKit uses Apple device sensors to measure a variety of tasks. Select any of the below to view a sample.

- motor skills
- fitness
- cognition
- speech
- hearing

Tapping Speed
This activity measures your tapping speed.

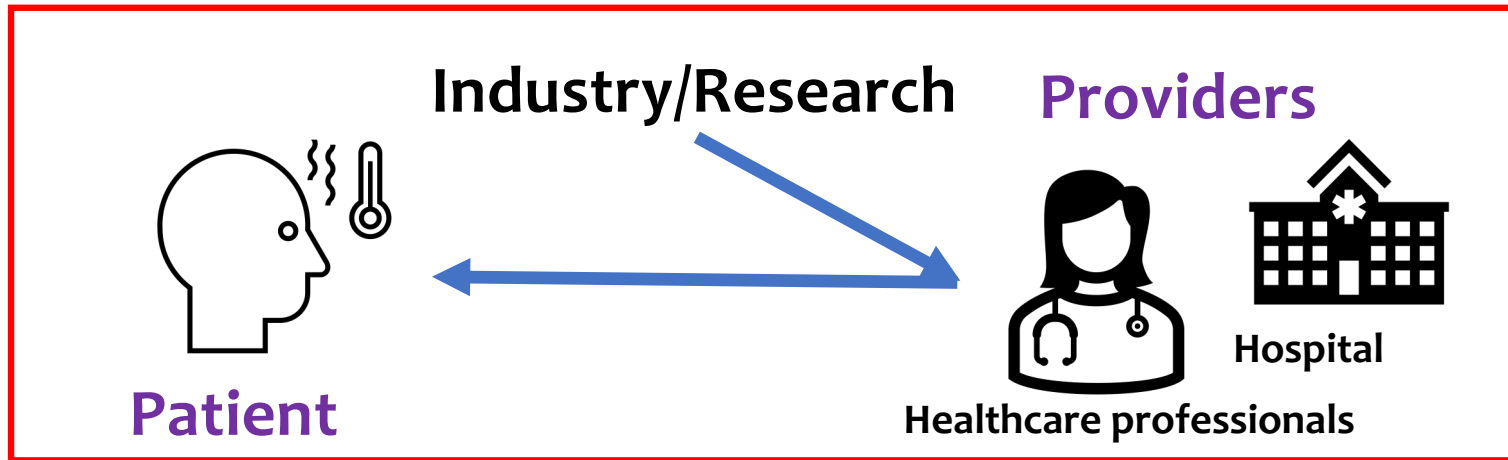
<https://www.researchandcare.org/researchkit/>

Sources of Clinical Data

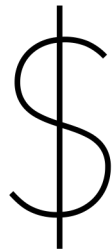


What are potential benefits of using patient derived data?

Sources of Clinical Data



- Study evaluating a new test or treatment
- May also be used for secondary analyses



Payer



Policymaker

Where does clinical data come from?

- Patient
- Providers
- Payer
- Policy-maker
- Industry
- Research

Caution

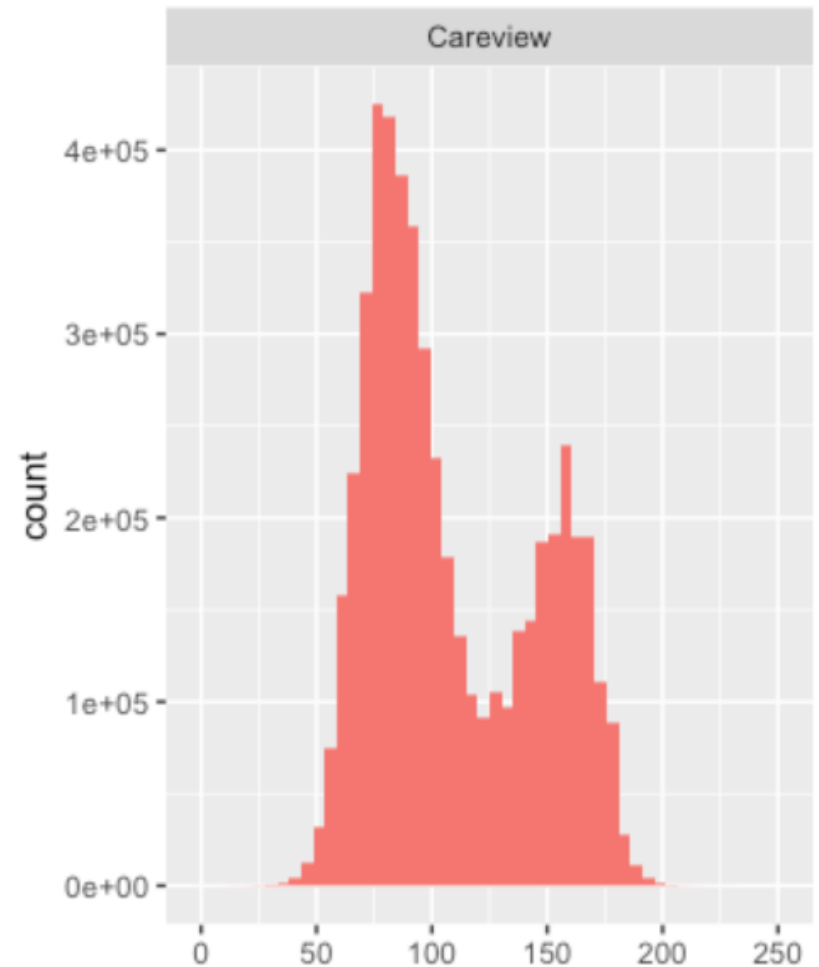
- None has complete data for
 - Individual patient
 - Population
- Usually not designed for research

Types of data in healthcare

- History
 - Symptoms and their details, past medical/surgical history, medications, allergies, family history, etc.
- Physical exam
 - Height, weight, BMI, vital signs (temperature, blood pressure, heart rate, etc), tenderness, erythema (redness), etc.
- Labs
 - Complete blood count, serum electrolytes, urine culture, blood culture, etc.
- Imaging
 - Chest x-ray, CT scan, bone scan, MRI, ultrasound, etc.
- Pathology
 - Biopsy, surgical pathology
- Genetics
 - Germline testing, etc.

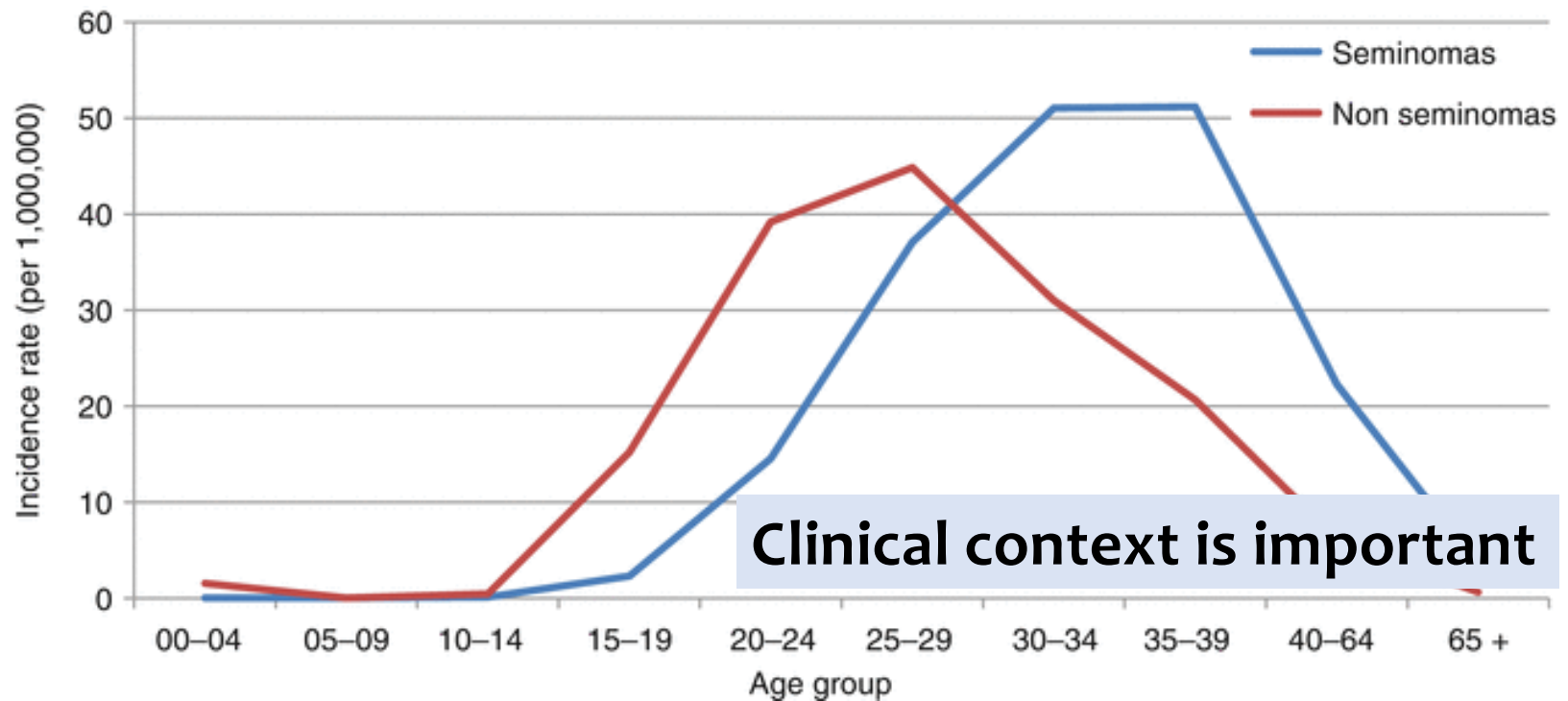
Exploring Clinical Data

- Medical Information Mart for Intensive Care (MIMIC)-III
 - Public de-identified dataset
 - Critical care data for over 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012
- Distribution of heart rates in the MIMIC-III chart (as recorded in Carevue)



Adapted from Dr. Szolovitz

Bimodal distribution in clinical care



Clinical context is important

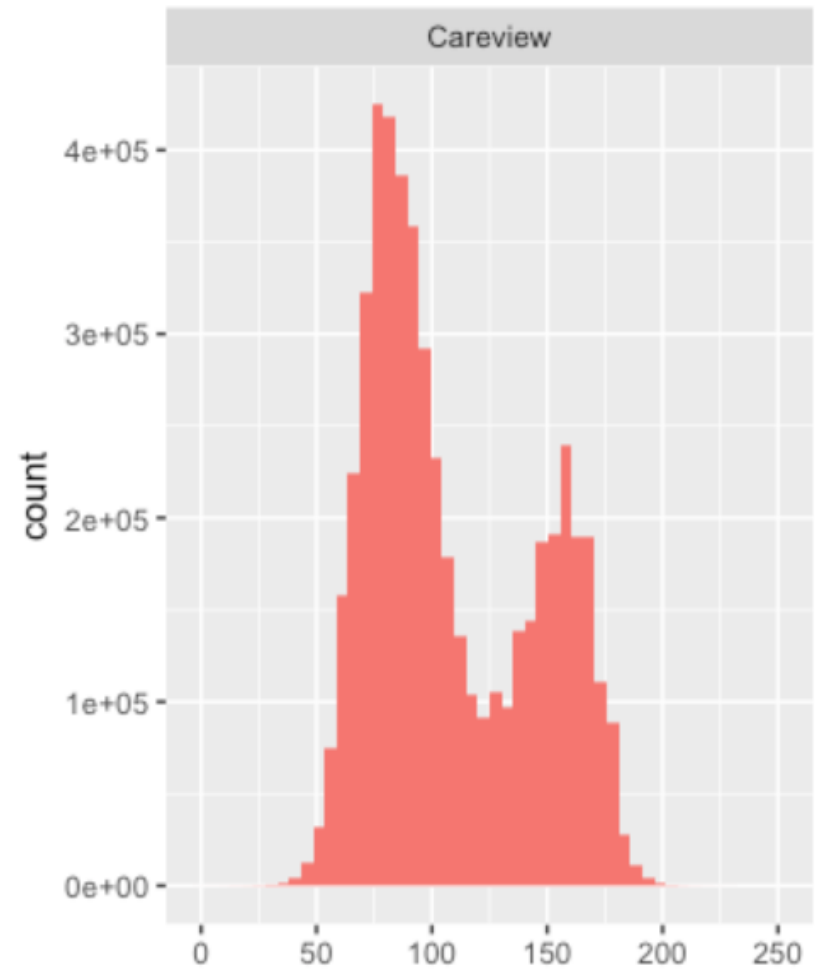
Gatta, Gemma, and Annalisa Trama. "Epidemiology of testicular Cancer." *Pathology of Testicular and Penile Neoplasms*. Springer, Cham, 2016. 3-18.

Hanna N, Timmerman R, Foster RS, et al. Epidemiology. In: Kufe DW, Pollock RE, Weichselbaum RR, et al., editors. *Holland-Frei Cancer Medicine*. 6th edition. Hamilton (ON): BC Decker; 2003. Available from:

<https://www.ncbi.nlm.nih.gov/books/NBK12708/>

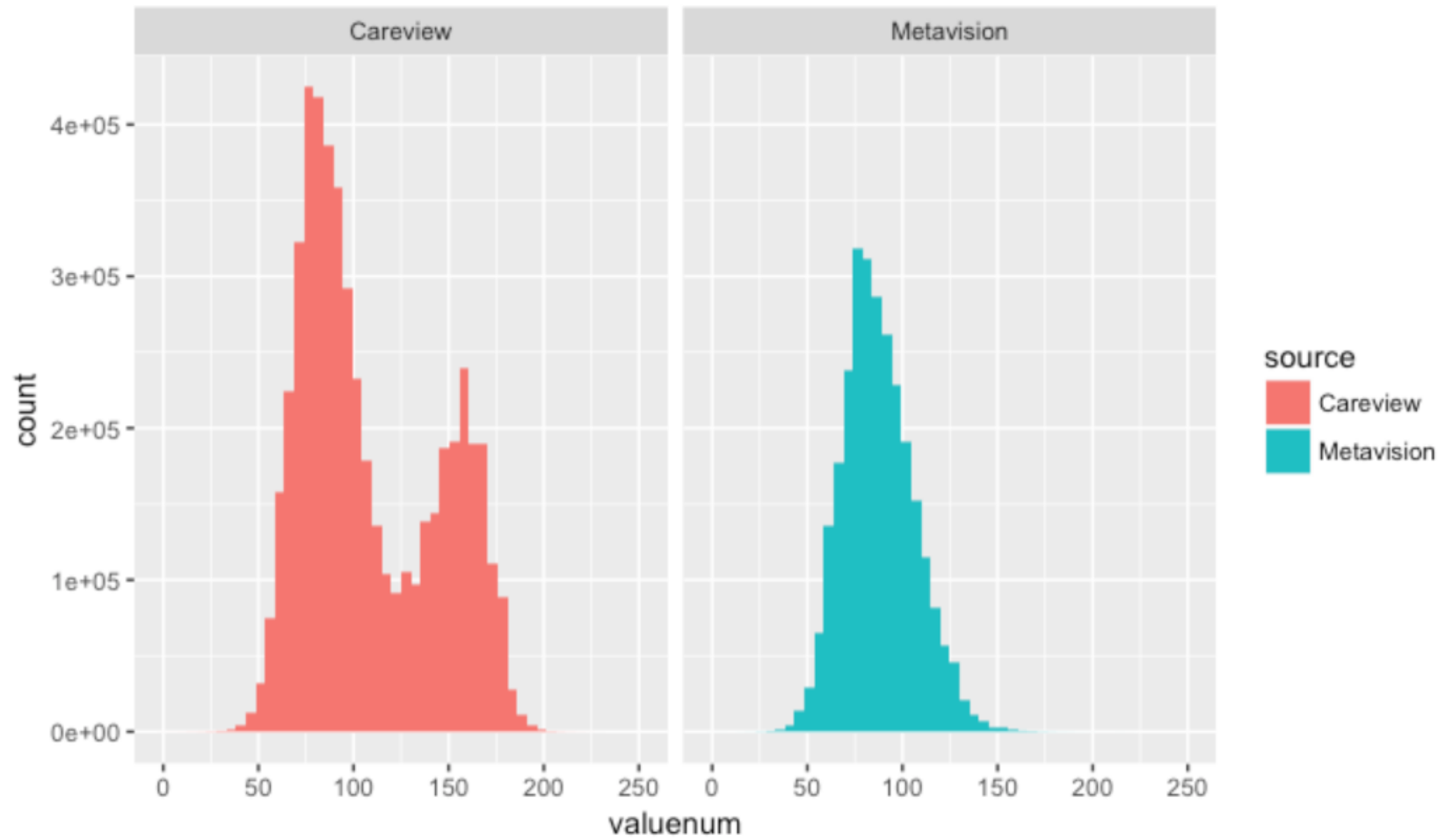
Exploring Clinical Data

- Distribution of heart rates in the MIMIC-III chart (as recorded in Carevue)



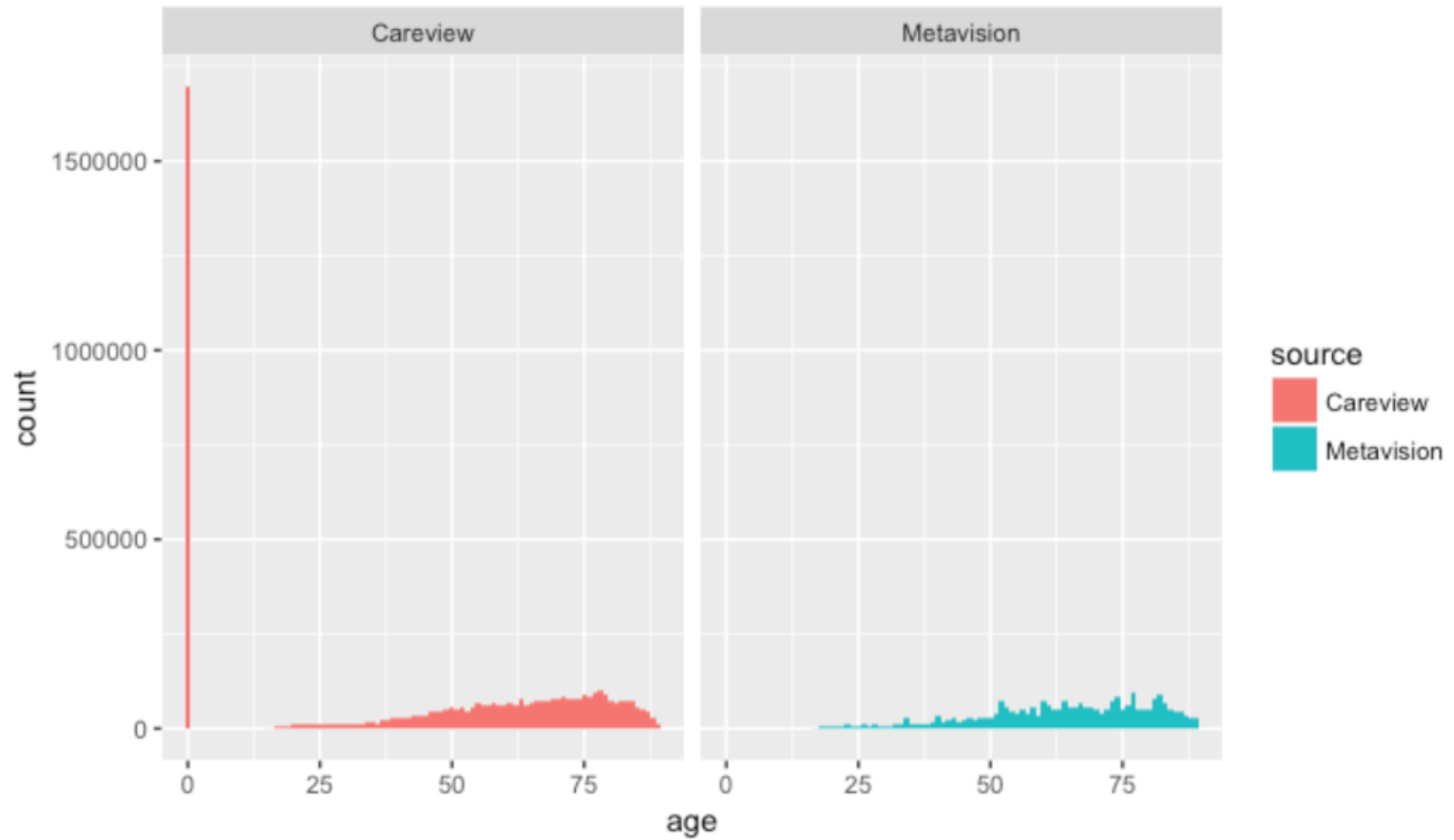
Adapted from Dr. Szolovitz

Comparison of Careview and Metavision heart rates, outliers removed



Adapted from Dr. Szolovitz

Age distribution of patients with recorded heart rates, age ≥ 90 suppressed

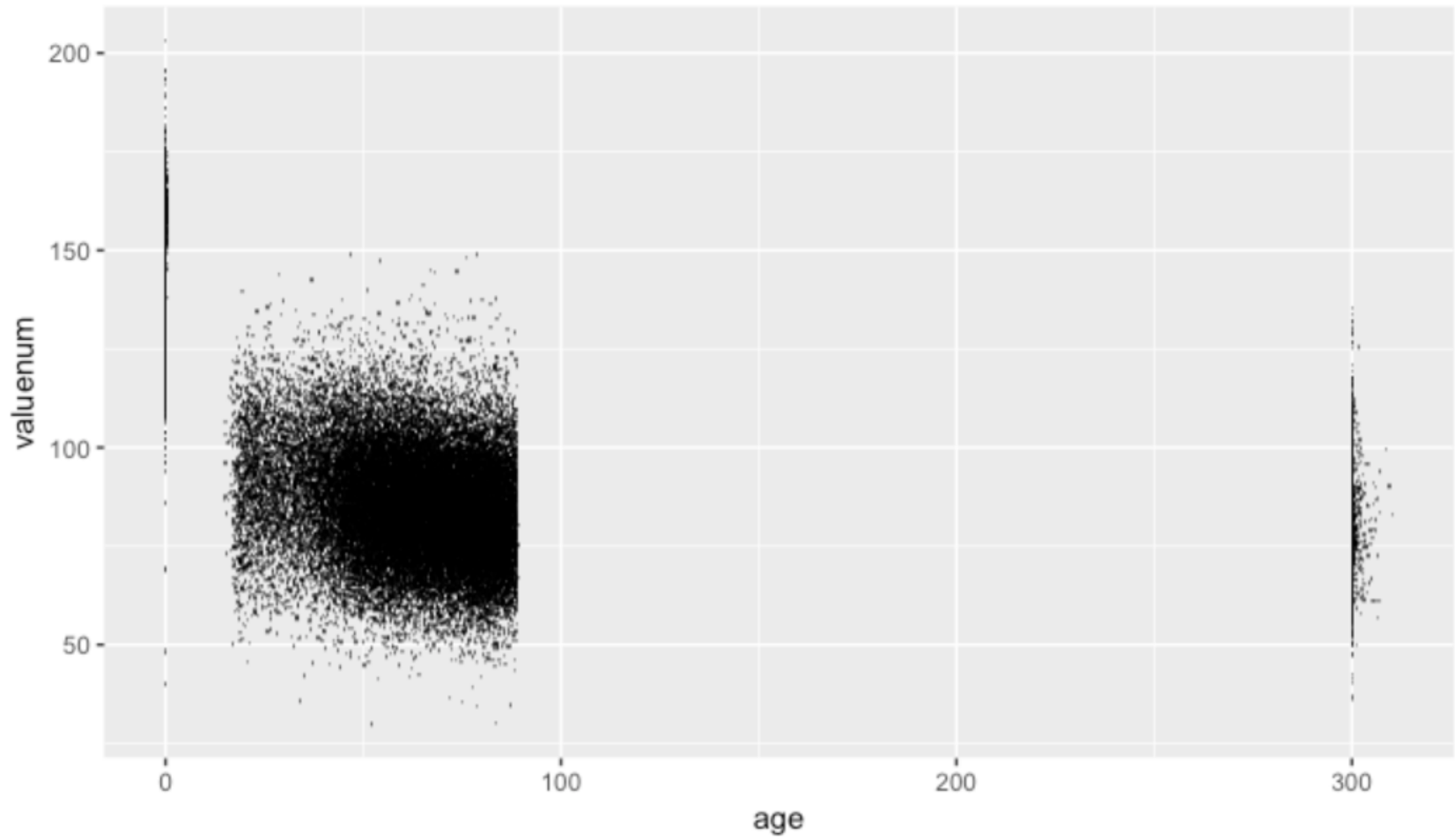


Adapted from Dr. Szolovitz



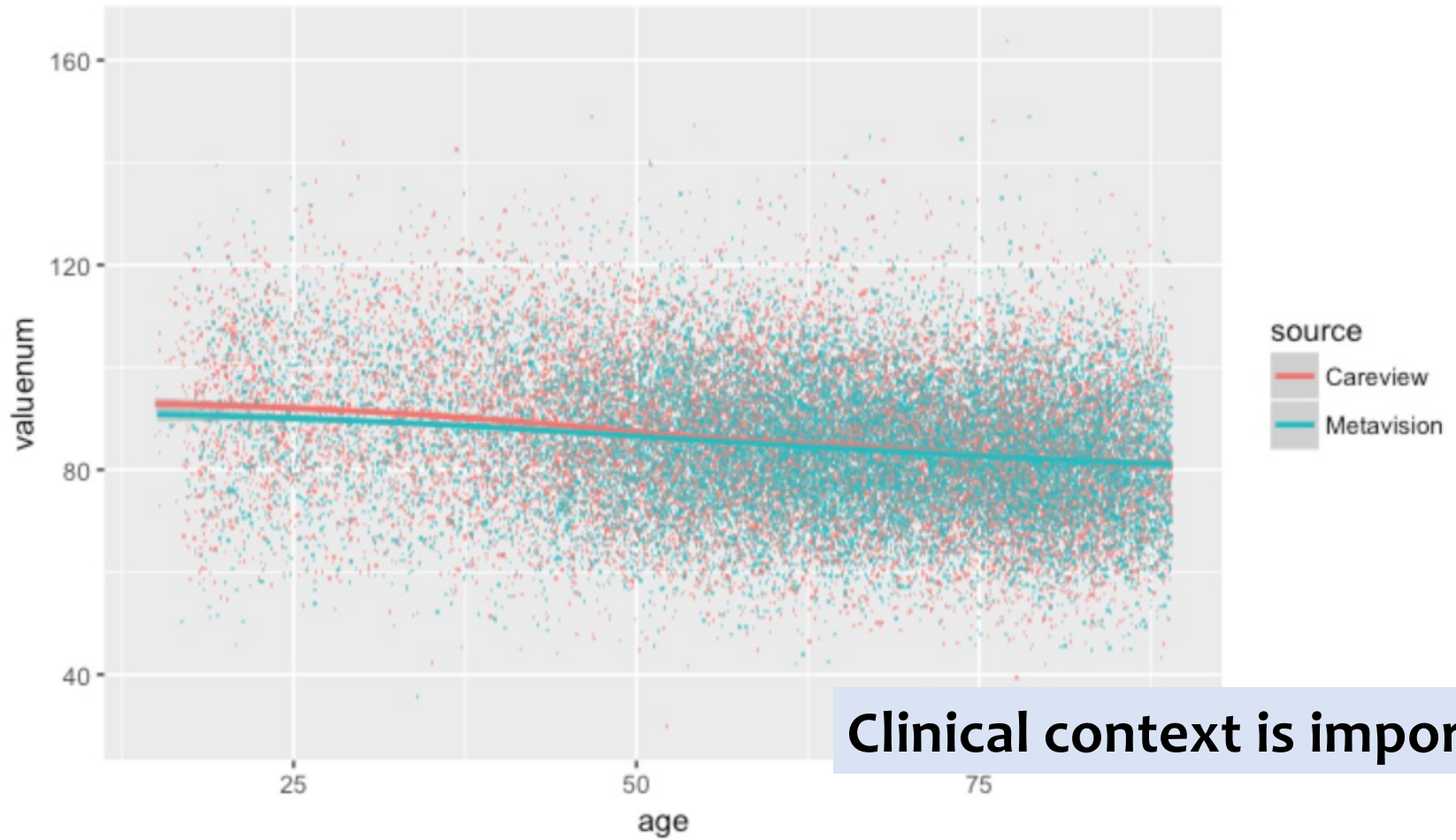
Adapted from Dr. Szolovitz

HR vs. Age in entire population



Adapted from Dr. Szolovitz

HR vs. Age in adults, smoothed



Clinical context is important

Adapted from Dr. Szolovitz

Back to Mrs. Patel

Practical Application of Clinical Data Science

- Formulate question to apply clinical data science
- Identify features and labels needed to answer clinical question
 - Review literature, discussion with experts, etc.
- Identify appropriate data source: NCDB (for demonstration)
 - Not for recurrence, but suppose death from any cause
- Obtain appropriate ethics approval
 - Review data dictionary (NCDB Participant User Data File), if available
- Exploratory data analysis
- Clean data... then clean data
- Analysis
- Report results

NCDB Data Dictionary

Table of Contents

- Layout of Data Dictionary Items..... 10
- Facility and Patient Demographics..... 12
 - Case Key 13
 - Facility Key 14
 - Facility Type 15**
 - Facility Location 16
- Patient Treated in More than One CoC Facility Flag..... 17
- Reference Date Flag..... 18
- Age at Diagnosis 19
- Sex 20
- Race 21
- Spanish/Hispanic Origin 24
- Primary Payor at Diagnosis 26

https://www.facs.org/-/media/files/quality-programs/cancer/ncdb/puf_data_dictionary.ashx

Facility Type

Data Dictionary Category: Facility and Patient Demographics

PUF Data Item Name: FACILITY_TYPE_CD

NAACCR Item #: Not applicable

Diagnosis Years Available: 2004 - 2018

Length: 1

Allowable Values: 1 - 4, blank

Description:

Each facility reporting cases to the NCDB is assigned a category classification by the Commission on Cancer Accreditation program. This item provides a general classification of the structural characteristics of each reporting facility.

Code	Definition
1	Community Cancer Program
2	Comprehensive Community Cancer Program
3	Academic/Research Program (includes NCI-designated comprehensive cancer centers)
4	Integrated Network Cancer Program
blank	Not available

https://www.facs.org/-/media/files/quality-programs/cancer/ncdb/puf_data_dictionary.ashx

Appendix A: Site-Specific Surgery Codes	333
Oral Cavity.....	334
Parotid and Other Unspecified Glands.....	336
Pharynx	338
Esophagus	340
Stomach	342
Kidney, Renal Pelvis, and Ureter.....	376

RX_SUMM_SURG_PRIM_SITE

30 = PN

50 = RN

Exploratory Data Analysis

PUF_CASE_ID
PUF_FACILITY_ID
FACILITY_TYPE_CD
FACILITY_LOCATION_CD
AGE
SEX
RACE
SPANISH_HISPANIC_ORIGIN
INSURANCE_STATUS
MED_INC_QUAR_00
NO_HSD_QUAR_00
UR_CD_03
MED_INC_QUAR_12
NO_HSD_QUAR_12
UR_CD_13
CROWFLY
CDCC_TOTAL_BEST
SEQUENCE_NUMBER
CLASS_OF_CASE
YEAR_OF_DIAGNOSIS
PRIMARY_SITE
LATERALITY
HISTOLOGY
BEHAVIOR
GRADE

	counts	%
White	435222	0.845290
Black	58260	0.113153
Unknown	5547	0.010773
Other	4048	0.007862
American Indian, Aleutian, or Eskimo	2432	0.004723
Other Asian, including Asian, NOS and Oriental, NOS	2119	0.004116
Chinese	1565	0.003040
Filipino	1400	0.002719
Asian Indian or Pakistani, NOS	932	0.001810
Japanese	913	0.001773
Korean	583	0.001132
Vietnamese	486	0.000944
Asian Indian	413	0.000802

TUMOR_SIZE

Describes the largest dimension of the diameter of the primary tumor in millimeters (mm).

	counts	%
99	135	0.000415
1	133	0.000409
128	130	0.000400
Tumor involvement of specified primaries	128	0.000393
111	126	0.000387
> 1 cm, < 2 cm	125	0.000384
123	115	0.000354

RX_SUMM_SURG_PRIM_SITE
30 = PN
50 = RN

Challenges of Working with Clinical Data

- Access
- Heterogeneity
- Noisy data
- Missing data

Access

- **Healthcare data is**
 - **Sensitive**
 - Details about an individual that they may want to keep private
 - **Protected** by the Health Insurance Portability and Accountability Act (HIPAA)
 - Federal law passed in 1996
 - Created standards to protect sensitive protected health information (PHI)
 - PHI is any information
 - Created, used, or disclosed in the course of providing a health care service, such as diagnosis or treatment
 - That can be used to identify an individual

<https://www.cdc.gov/phlp/publications/topic/hipaa.html>
<https://cphs.berkeley.edu/hipaa/hipaa18.html>

Personal health information

• Examples of PHI identifiers (18)

1. Names
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if*
3. All elements of dates (except year) for dates directly
4. Medical record numbers
5. Health plan beneficiary numbers
6. Account numbers
7. Certificate/license numbers
8. Vehicle identifiers and serial numbers, including license plate numbers

If the HIPAA Privacy Rule applies to your research, you must obtain an Authorization to use/disclose PHI or a Waiver of Authorization from the Institutional Review Board

9. Such age, except that such ages and elements may be aggregated into a single category of age 90 or older
10. Phone numbers
11. Fax numbers
12. Electronic mail addresses
13. Social Security numbers
14. Biometric identifiers, including finger and voice prints
15. Full face photographic images and any comparable images
16. Any other unique identifying number, characteristic, or code

<https://cphs.berkeley.edu/hipaa/hipaa18.html>

De-identified health information

- No restrictions on the use or disclosure of de-identified health information
- Methods to de-identify information
 - Formal determination by a qualified statistician
 - **Removal of specified identifiers** of the individual and of the individual's relatives, household members, and employers is required, and is adequate only if the covered entity **has no actual knowledge that the remaining information could be used to identify the individual**

Access

- De-identifying data is challenging
 - Removing ID may not be enough
- Even when data are de-identified, **fear of litigation and breach of privacy** discourages providers from sharing patient health data



Schwarz, Christopher G., et al. "Identification of anonymous MRI research participants with face-recognition software." *New England Journal of Medicine* 381.17 (2019): 1684-1686.

Heterogeneity

- Healthcare data and systems that used that integrate that data were not designed for research purposes
 - Various terms can be used to describe bladder cancer
 - Carcinoma of the bladder
 - Bladder malignancy
 - Malignant neoplasm of bladder
 - Etc.
- Moving toward standardization
 - International Classification of Diseases (ICD) codes

ICD codes

- Most widely used method of disease classification
- 11th edition released 2018, in effect from January 2022
 - 1st edition released 1990
 - 10th edition released 1994

Title	Reason for addition
Chapter 3: Diseases of the blood or blood-forming organs	These two chapters were split from a single chapter in ICD-10, recognizing differences in etiology, manifestations, and care
Chapter 4: Diseases of the immune system	
Chapter 7: Sleep–wake disorders	This topic has become more prominent since the 10th revision. The chapter mostly includes new concepts with some concepts moved from other chapters in ICD-10
Chapter 17: Conditions related to sexual health	This topic has become more prominent since the 10th revision. The chapter mostly includes concepts moved from other chapters in ICD-10, combined with some new concepts
Chapter 26: Traditional medicine conditions	This entirely new supplementary chapter in ICD-11 enables coding in terms of traditional medicine concepts, where required

Harrison, James E., et al. "ICD-11: an international classification of diseases for the twenty-first century." *BMC medical informatics and decision making* 21.6 (2021): 1-10.

ICD-11 Chapters

1.1A00–1H0Z Certain infectious or parasitic diseases

2.2A00–2F9Z Neoplasms

3.3A00-Diseases of the blood or blood-forming organsorgans

4.4A00–4B4Z Diseases of the immune system

5.5A00–5D46 Endocrine, nutritional or metabolic diseases

6.6A00–6E8Z Mental, behavioural or neurodevelopmental disorders

7.7A00–7B2Z Sleep-wake disorders

8.8A00–8E7Z Diseases of the nervous system

9.9A00–9E1Z Diseases of the visual system

10.AA00–AC0Z Diseases of the ear or mastoid process

11.BA00–BE2Z Diseases of the circulatory system

12.CA00–CB7Z Diseases of the respiratory system

13.DA00–DE2Z Diseases of the digestive system

14.EA00–EM0Z Diseases of the skin

15.FA00–FC0Z Diseases of the musculoskeletal system or connective tissue

16.GA00–GC8Z Diseases of the genitourinary system

17.HA00–HA8Z Conditions related to sexual health

18.JA00–JB6Z Pregnancy, childbirth or the puerperium

19.KA00–KD5Z Certain conditions originating in the perinatal period

20.LA00–LD9Z Developmental anomalies

21.MA00–MH2Y Symptoms, signs or clinical findings, not elsewhere classified

22.NA00–NF2Z Injury, poisoning or certain other consequences of external causes

23.PA00–PL2Z External causes of morbidity or mortality

24.QA00–QF4Z Factors influencing health status or contact with health services

25.RA00–RA26 Codes for special purposes

26.SA00–SJ3Z Supplementary Chapter Traditional Medicine Conditions - Module I

27.VA00–VC50 Supplementary section for functioning assessment (in line with WHO-DAS 2)

28.X...–X... Extension Codes ("terminology component" of ICD-11)

Harrison, James E., et al. "ICD-11: an international classification of diseases for the twenty-first century." *BMC medical informatics and decision making* 21.6 (2021): 1-10.

Search

[Advanced Search]

Browse

Coding Tool

Special Views

Info

- ▼ Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic, central nervous system or related tissues
 - ▶ Malignant mesenchymal neoplasms
 - ▶ Malignant neoplasms of lip, oral cavity or pharynx
 - ▶ Malignant neoplasms of digestive organs
 - ▶ Malignant neoplasms of middle ear, respiratory or intrathoracic organs
 - ▶ Malignant neoplasms of skin
 - ▶ Malignant neoplasms of peripheral nerves or autonomic nervous system
 - ▶ Malignant neoplasms of retroperitoneum, peritoneum or omentum
 - ▶ Malignant neoplasms of breast
 - ▶ Malignant neoplasms of female genital organs

 - ▶ Malignant neoplasms of male genital organs
 - ▼ Malignant neoplasms of urinary tract
 - ▼ 2C90 Malignant neoplasms of kidney, except renal pelvis
 - 2C90.0 Renal cell carcinoma of kidney, except renal pelvis

Foundation URI : <http://id.who.int/icd/entity/825917541>

2C90.0 Renal cell carcinoma of kidney, except renal pelvis

2022 ICD-10-CM Diagnosis Code C64.9

2C90 Malignant neoplasms of kidney, except renal pelvis

Show all ancestors

Description

A carcinoma arising from the renal parenchyma. The incidence of renal cell carcinoma has increased by 35% from 1973 to 1991. There is a strong correlation between cigarette smoking and the development of renal cell carcinoma. The clinical presentation includes : haematuria, flank pain and a palpable lumbar mass. A high percentage of renal cell carcinomas are diagnosed when an ultrasound is performed for other purposes. Diagnostic procedures include: ultrasound, intravenous pyelography and computed tomography (CT).

Postcoordination

Add detail to **Renal cell carcinoma of kidney, except renal pelvis**

Laterality (use additional code, if desired .)

XK9J	Bilateral
XK8G	Left
XK9K	Right
XK70	Unilateral, unspecified

<https://icd.who.int/browse11>

Heterogeneity

- Moving toward standardization
 - International Classification of Diseases (ICD) codes
 - Classification of diseases
 - Current Procedural Terminology (CPT) codes
 - Procedures/interventions
 - Logical Observation Identifiers Names and Codes (LOINC)
 - Labs
 - Sometimes, multiple systems exist
 - Medications: NDC, MedDRA, CPT, Healthcare Common Procedure Coding System

Heterogeneity

- Moving toward standardization
 - Clinical notes
 - A critical component of clinical data
 - Written text remains the most natural and expressive method to document clinical events
 - A significant portion of clinical data is in clinical notes
 - **Inconsistent descriptions of identical data**

Radical prostatectomy pathology report

- **Inconsistent descriptions of identical data**

- 10 different ways of describing pT2, without specifying pT2
 1. "confined to the prostate"
 2. "invades into but not through the prostatic capsule"
 3. "invades but does not transgress the capsule"
 4. "tumor does not transgress the "prostatic capsule""
 5. "extends to, but not through the prostatic capsule."
 6. "not infiltrating periprostatic adipose tissue"
 7. "apparently localized"
 8. "no capsular penetration demonstrated"
 9. "no capsular invasion is present."
 10. "extracapsular extension is not identified."

Heterogeneity

- Moving toward standardization
 - Clinical notes
 - **Synoptic reporting** is being increasingly adopted in some fields (radiology and pathology (remember CoC))

Accession: AAAA0000

Procedure: radical prostatectomy

Histologic type: acinar adenocarcinoma

Grade group: 2

Margins: uninvolved by invasive carcinoma

Number of lymph nodes involved: 0

Number of lymph nodes examined: 3

Pathologic stage classification (AJCC 8th edition): **Primary tumor: pT2**

Heterogeneity

- Moving toward standardization
 - Provider notes are lagging behind



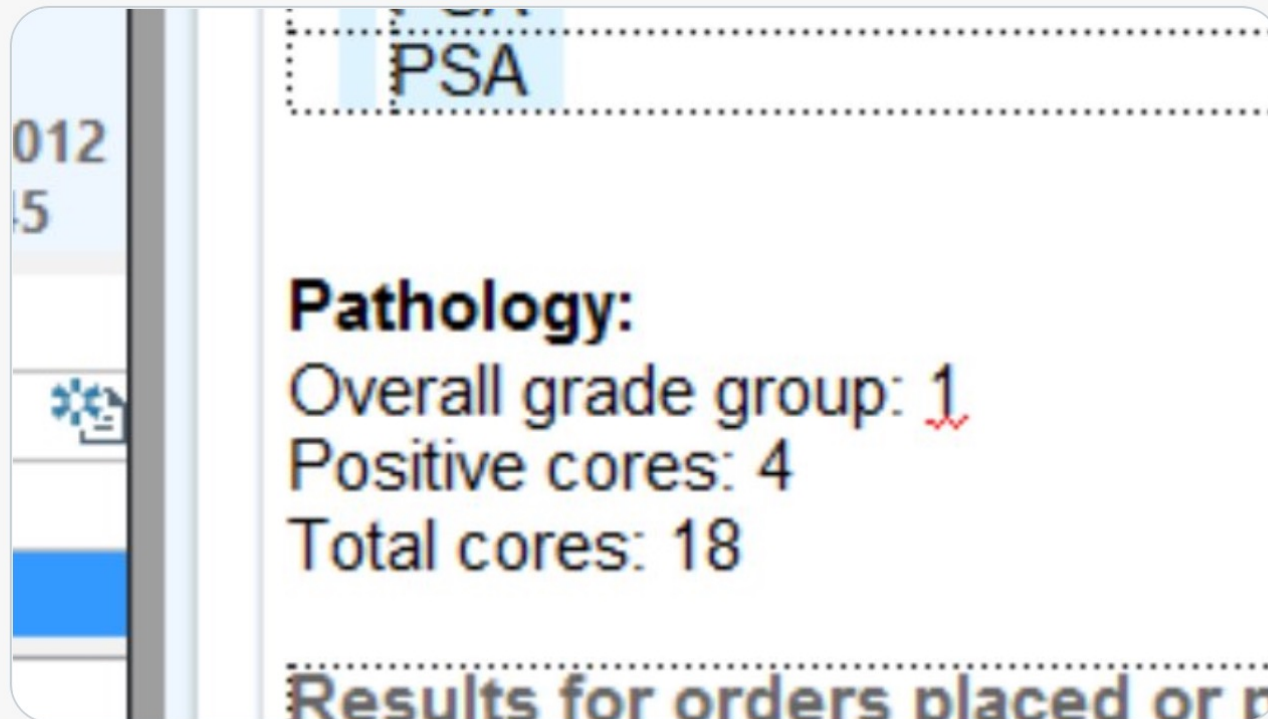
n different ways of describing PSA

1. psa rose to XX.XX in YYYY
2. psa rose to XX.XX in MM/YY
3. psa rose from XX.XX ng/ml in YYYY
4. psa rose from XX.XX in MM/YYYY to XX.XX₂ in MM/YY₂
5. psa increased to XX.XX (MM/YY)
6. psa increased to XX.XX Month YYYY
7. psa increased to XX.XX Month YYYY from XX.XX₂ in YYYY₂
8. psa elevation to XX.XX in Month YYYY
9. psa from MM/DD/YYYY is XX.XX
10. psa from XX.XX in MM/YYYY to XX.XX₂ in MM/YYYY₂
11. psa from XX.XX in MM/YYYY to XX.XX₂ (MM/YYYY₂)
12. psa from Month DD, YYYY was XX.XX
13. ...



Madhur Nayan @DrMadhurNayan · May 6

Until there are standardized methods to report [#clinical](#) data, I will write my notes as if they will eventually be abstracted for [#machinelearning](#) 😊 I am kindly requesting others to do the same 🙏 [#digitalhealth](#) [#bigdata](#) [#DeepLearning](#) [#ml4h](#)



Noisy data

- Recorded data may not reflect ground truth

Circumcision and Risk of HIV among Males from Ontario, Canada

of THE JOURNAL
UROLOGY®
www.auajournals.org/journal/juro

Madhur Nayan,^{1,*} Robert J. Hamilton,¹ David N. Juurlink,^{2,3} Peter C. Austin^{2,†} and Keith A. Jarvi^{4,5,6,‡}

- We used physician claims data to identify receipt of circumcision.
 - Residents of Ontario have universal access to physician services and hospital care.

What are potential limitations of using physician claims to identify receipt of circumcision?

Nayan, Madhur, et al. "Circumcision and risk of HIV among males from Ontario, Canada." *The Journal of urology* 207.2 (2022): 424-430.

Noisy data

- **Recorded data may not reflect ground truth**
- **Validation study of ICD codes**
 - 4,008 randomly selected charts for patients from four teaching hospitals in Alberta, Canada
 - ICD coding from 4 professionally trained health record coders compared to chart review by 2 nurses for 32 conditions

Quan, Hude, et al. "Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database." *Health services research* 43.4 (2008): 1424-1441.

Noisy data

Conditions	Chart review	ICD-10 Data	Difference Chart— ICD-10
Myocardial infarction	12.8	8.4	+4.4
Cardiac	21.8	9.1	+12.7
The ICD-10 data underreported 31 conditions and slightly over-reported one condition (renal failure).			
Obesity	5.5	4.7	+0.8
Depression	11.9	7.3	+4.6
Renal failure	4.0	4.9	-0.9

Quan, Hude, et al. "Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database." *Health services research* 43.4 (2008): 1424-1441.

Noisy data

- Multiple codes may represent a common feature

Most common prescriptions in MIMIC-III

	NDC Code	count
Iso-Osmotic Dextrose	0	86935
Sodium Chloride 0.9% Flush	0	83392
Insulin	0	81356
SW	0	72458
Magnesium Sulfate	409672924	55211
D5W	0	54938
Furosemide	517570425	53073
Potassium Chloride	338070341	47968
D5W	338001702	43038
LR	338011704	35407
Vancomycin	338355248	34741
0.9% Sodium Chloride	338004904	34682
Potassium Chloride	456066270	32533
Heparin	63323026201	31413
NS	338004902	30815

Noisy data

- **Messy data**

- Mass General Brigham Research Patient Data Registry
 - Health history
 - Concept_name contains 'height'
 - Different values on same date
 - Impossible values (0, 789)
 - Clinical context important for less obvious features
 - Text data (5ft 10 inch, 262 pounds, “eats 2x a day, skips breakfast some morning”)

Missing Data

- Missing data is ubiquitous in clinical data
 - Why?
- Mechanism of missingness may be important
- When reporting analysis of clinical data, important to
 - Quantify missing data
 - How missing data was accounted for

Haukoos, Jason S., and Craig D. Newgard. "Advanced statistics: missing data in clinical research—part 1: an introduction and conceptual framework." *Academic Emergency Medicine* 14.7 (2007): 662-668.

Patient/Provider Goals of Clinical Data Science

- Mrs. Patel is a 65 year old who was recently diagnosed with kidney cancer. She returns to your office to discuss treatment and has some questions.
 - After treatment, **what is the risk** of my cancer coming back before the Ultimate World Cruise (December 2023)?
 - **Will the risk** of my cancer coming back **change** if I get a partial nephrectomy instead of a radical nephrectomy?

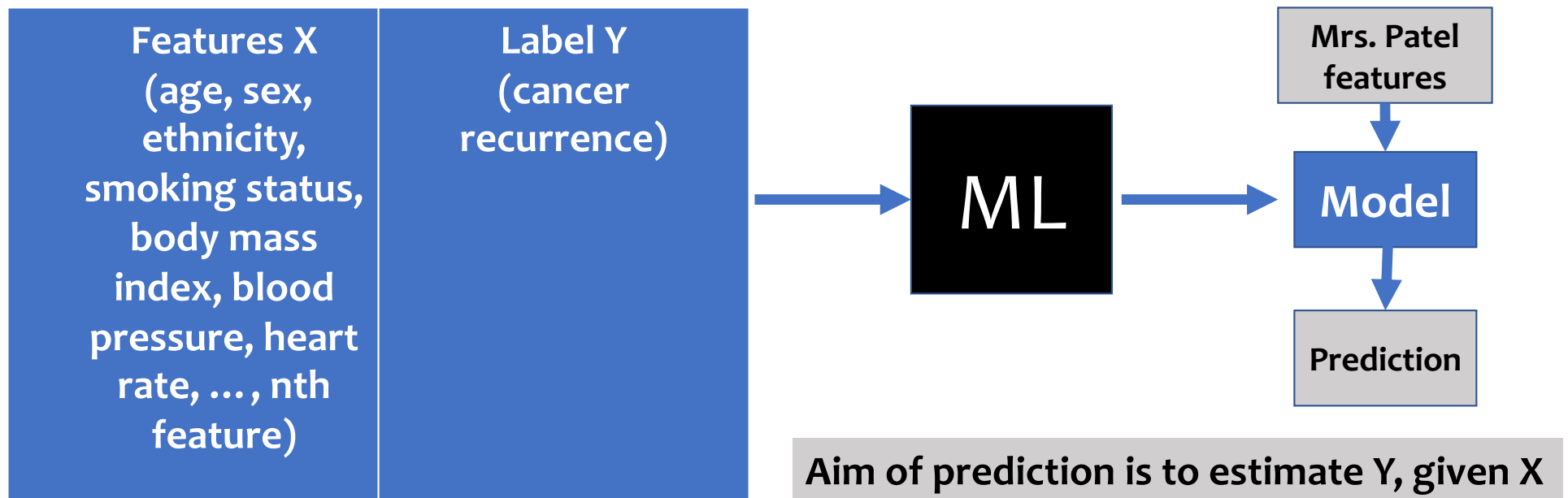
How would you answer these questions using clinical data science?

Will my cancer come back?

- We cannot know the ground truth (will Mrs. Patel's cancer recur before December 2023)
- At best, we can **estimate** her risk
 - Population average of patients treated for kidney cancer
 - In this patient?
 - Sex
 - Ethnicity
 - Etc.

Will my cancer come back?

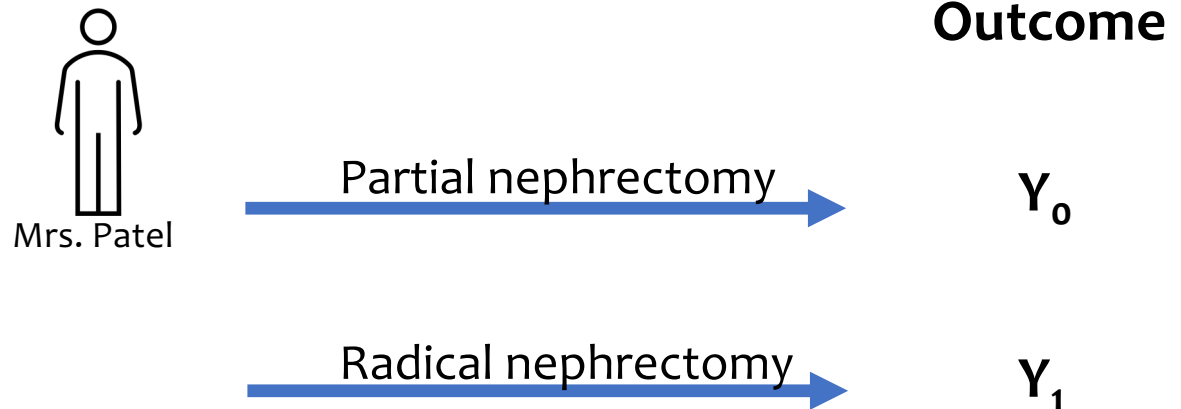
- How would you estimate the risk of Mrs. Patel's cancer recurrence?



Change the risk of my cancer coming back?

- You hypothesize that type of surgery (partial vs. radical) will change her risk of cancer recurrence.

- Ground truth



- Reality: We cannot know the ground truth

Change the risk of my cancer coming back?

- Since we cannot know the ground truth, at best, we can **estimate** her risk under the two conditions (partial vs. radical nephrectomy) with **causal inference**

Aim of causal inference is to estimate the effect of T on Y

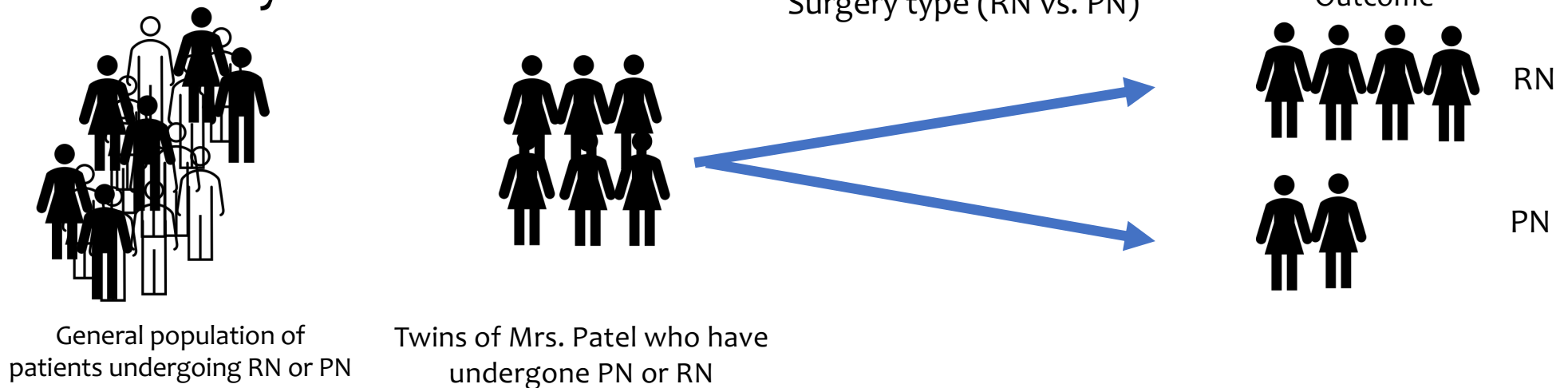
Clinical Data Science for Prediction vs. Causal Inference

- Aim of prediction is to estimate Y , given X
- Aim of causal inference is to determine the effect on Y , given a change in T
- A model may be predictive, but causal inference can help determine why
 - Understanding why is often important in healthcare
 - What are potential harms of adopting a highly-predictive black-box model?

Change the risk of my cancer coming back?

- You hypothesize that type of surgery (partial vs. radical) will change her risk of cancer recurrence. How do you evaluate this hypothesis?

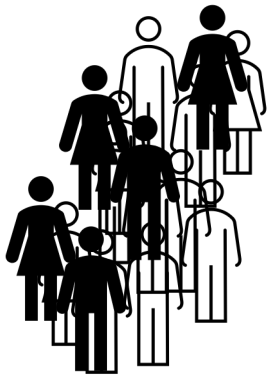
- Ideally



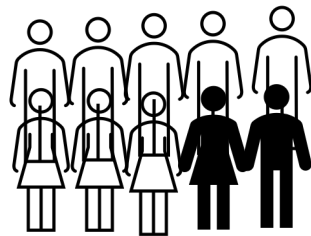
Change the risk of my cancer coming back?

- You hypothesize that type of surgery (partial vs. radical) will change her risk of cancer recurrence. How do you evaluate this hypothesis?

- Reality

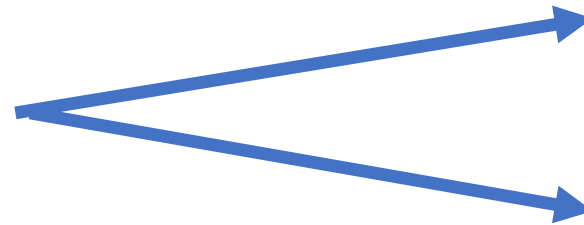


Selected population of patients undergoing RN or PN

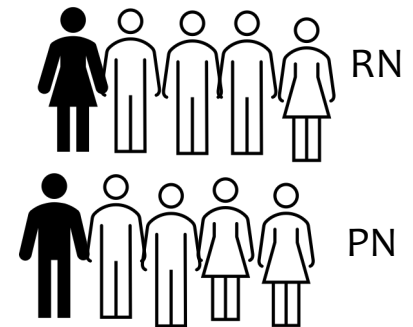


Patients **“similar”** to Mrs. Patel who have undergone PN or RN

Surgery type (RN vs. PN)



Outcome



RN

PN

Clinical Research Study Designs

Descriptive

- Case report
- Case series
- Survey

Analytic

Observational

- Cohort studies
- Cross sectional
- Case-control

Experimental

- Randomized controlled trials

Change the risk of my cancer coming back?

- You perform a retrospective cohort study of patients that have recently been diagnosed with kidney cancer and have undergone either partial or radical nephrectomy.
 - What are some differences between retrospective and prospective data collection?

Change the risk of my cancer coming back?

- Population: 1454 patients with pT1 (tumor \leq 7cm)
 - Partial nephrectomy n=379 (26.1%)
 - Radical nephrectomy n=1075 (37.9%)
- Results
 - Recurrence rate lower in partial nephrectomy group
 - Why?

Patard, Jean-Jacques, et al. "Safety and efficacy of partial nephrectomy for all T1 tumors based on an international multicenter experience." *The Journal of urology* 171.6 Part 1 (2004): 2181-2185.

TABLE 2. *Comparison of tumor characteristics*

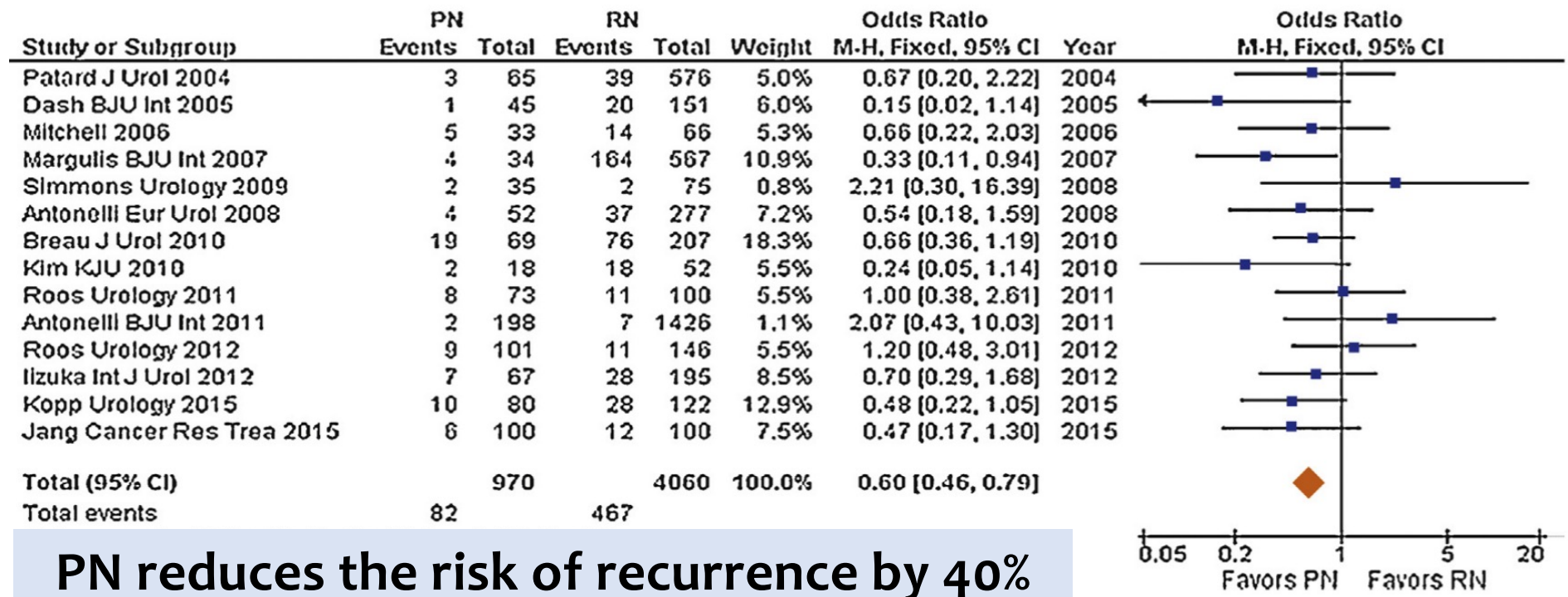
	T1a Tumors	
	Partial	Radical
Mean cm tumor \pm SD	2.5 \pm 0.8	3.2 \pm 0.8
	T1b Tumors	
	Partial	Radical
	5.3 \pm 0.8	5.6 \pm 0.8
		0

Population characteristics should be noted in all clinical data science studies

Patard, Jean-Jacques, et al. "Safety and efficacy of partial nephrectomy for all T1 tumors based on an international multicenter experience." *The Journal of urology* 171.6 Part 1 (2004): 2181-2185.

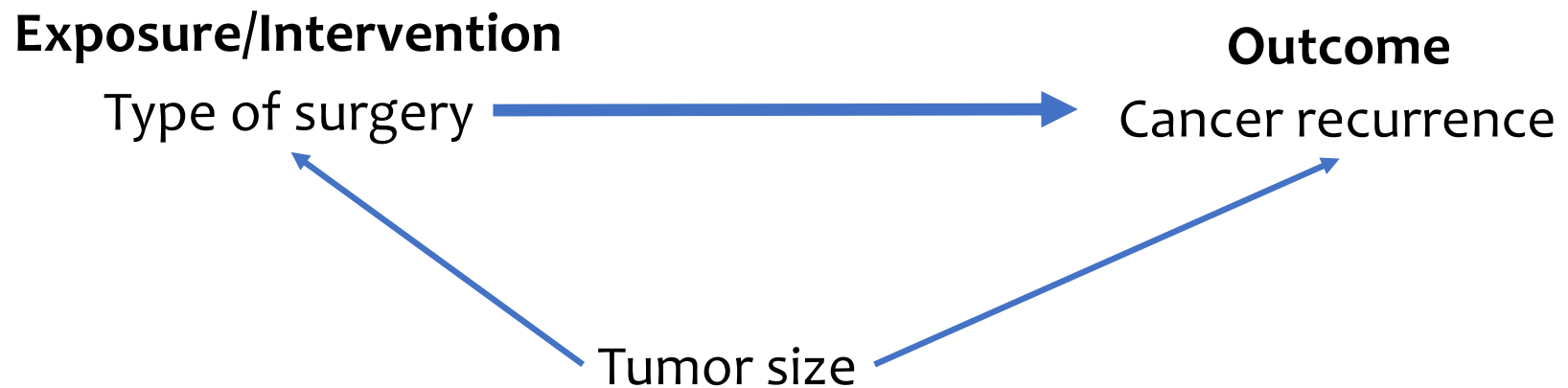
Meta-analysis

- Method of aggregating results from different studies



Disadvantages of Observational Studies

- Groups may differ in more ways than just exposure/treatment status



Disadvantages of Observational Studies

- Methods of addressing **MEASURED** differences in baseline characteristics
 - Covariate adjustment
 - Propensity score weighting
 - Matching
 - More to come on this later!
- **UNMEASURED differences can persist**

Clinical Research Study Designs

Descriptive

- Case report
- Case series
- Survey

Analytic

Observational

- Cohort studies
- Cross sectional
- Case-control

Experimental

- Randomized controlled trials

Clinical Research Study Designs

- Randomized controlled trial vs. Cohort studies
 - **Similarity:** compare outcomes in a population with similar characteristics, that differ by exposure/treatment status
 - **Difference:** Exposure/treatment assignment is random in experimental studies, and observed/non-random in observational/non-experimental studies

“The beauty of randomization is that it assures, if sample size is sufficiently large, that both known and **unknown** determinants are evenly distributed between treatment and control groups”

Guyatt, Gordon H., and Drummond Rennie. "Users' guides to the medical literature." *Jama* 270.17 (1993): 2096-2097.

Randomized Controlled Trials

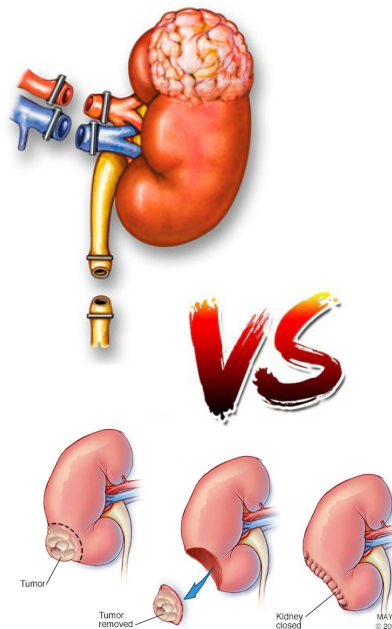
- Advantages
 - Comparable groups at baseline
- Disadvantages
 - Sometimes impractical
 - Randomizing males born in Ontario to circumcision vs. not
 - Does not ensure generalizability/external validity

RCT: Radical vs. Partial Nephrectomy

- EORTC 30904



Population: 541 patients with tumors <5cm suspicious for kidney cancer



Randomized to
RN vs. PN

Results

Local recurrence

RN 1/273 = 0.37%

PN 6/278 = 2.16%

Van Poppel, Hendrik, et al. "A prospective, randomised EORTC intergroup phase 3 study comparing the oncologic outcome of elective nephron-sparing surgery and radical nephrectomy for low-stage renal cell carcinoma." *European urology* 59.4 (2011): 543-552.

<https://www.fairbanksurology.com/robotic-radical-nephrectomy>
<https://www.mayoclinic.org/tests-procedures/nephrectomy/multimedia/img-20332175>

Questions

6.871/HST.956: Machine Learning for Healthcare

