



# Current Practice and Frontiers in Rule-Based Electronic Phenotyping

---

Fabrício Kury, MD

UAB Informatics Institute PowerTalk Series

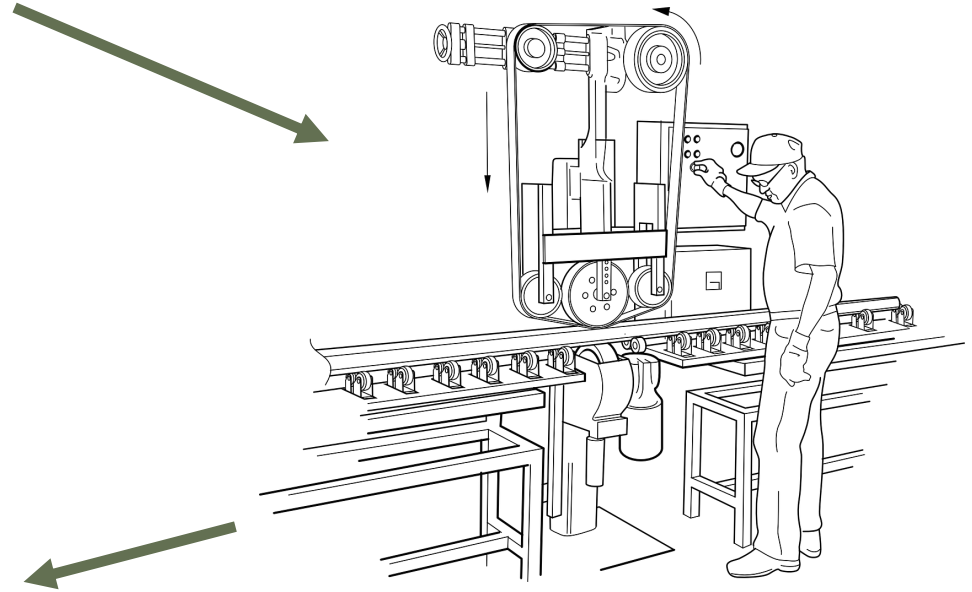
April 1<sup>st</sup>, 2022



medical data



observational study  
patients and covariates  
(features)



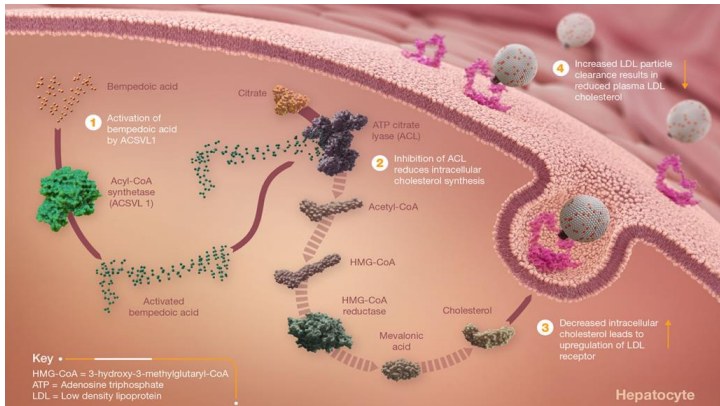
phenotyping

"the process of defining the necessary and sufficient criteria that a patient's record must satisfy to consider an exposure or outcome to have occurred for that patient" (Callahan, 2021, doi: 10.1093/jamia/ocab027)

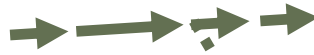


“raw” data records

↑ manual or automated record



true clinical event



Extract-  
Transform-  
Load



data warehouse

phenotyping



observational study

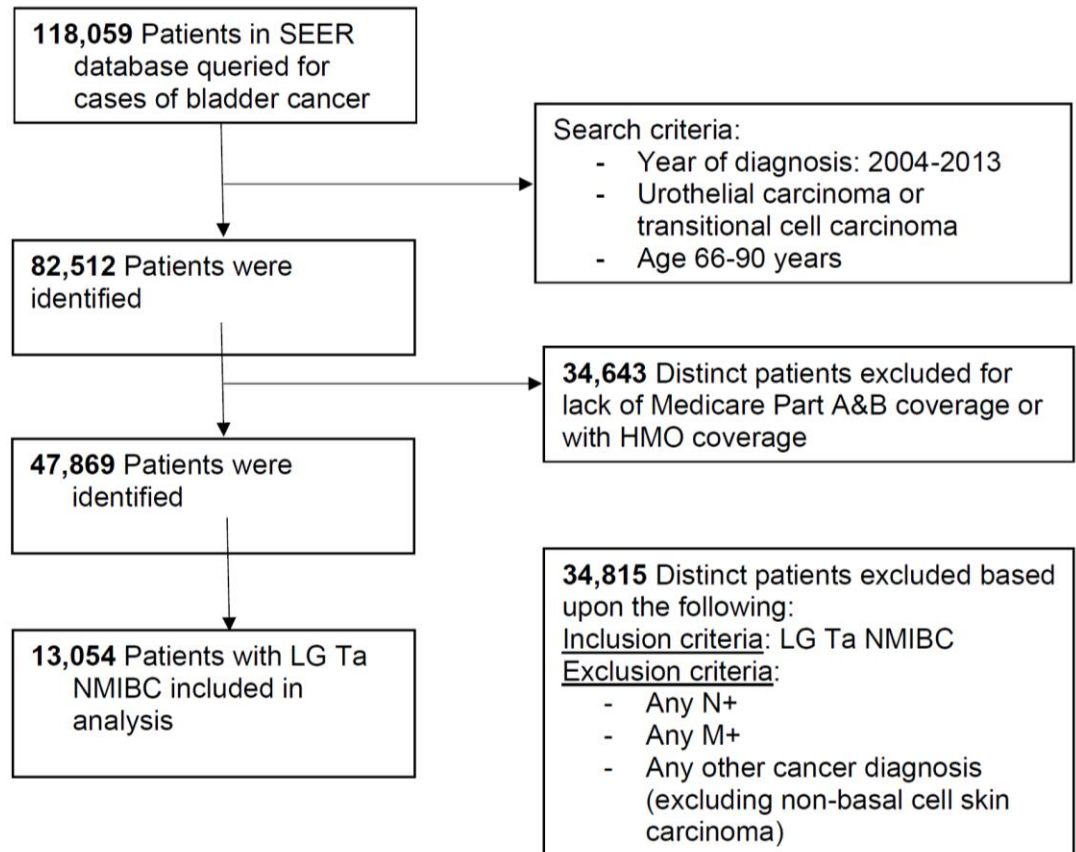
# Management, Surveillance Patterns, and Costs Associated With **Low-Grade Papillary Stage Ta Non-Muscle-Invasive Bladder Cancer Among Older Adults, 2004-2013**

March 18, 2022

**eFigure.** Study Flow Diagram Illustrating Cohort Selection

## Study Cohort

The study included older adults (aged 66-90 years) with a diagnosis of low-grade Ta urothelial bladder cancer between January 1, 2004, and December 31, 2013 (eFigure in the [Supplement](#)).



**eligibility criteria**  
**phenotype**

Abbreviations: HMO, health maintenance organization; NMIBC, non-muscle invasive bladder cancer; SEER, surveillance, epidemiology, and end results.

Original Investigation | Surgery

# Association of Trauma Center Designation With Postdischarge Survival Among Older Adults With Injuries

Molly P. Jarman, PhD, MPH; Ginger Jin, MS; Joel S. Weissman, PhD; Arlene S. Ash, PhD; Jennifer Tjia, MD, MSCE; Ali Salim, MD; Adil Haider, MD, MPH; Zara Cooper, MD, MSc

March 16, 2022

## Methods

### Population and Data Sources

Using Medicare claims from Inpatient and Outpatient Research Identifiable Files,<sup>10,11</sup> we identified 433 169 fee-for-service beneficiaries aged 65 years or older diagnosed with traumatic injury between January 1, 2014, and December 31, 2015, resulting in inpatient admission (**Figure 1**). Traumatic injury was defined based on the 2015 National Trauma Data Standard,<sup>12</sup> using *International Classification of Diseases, Ninth Revision (ICD-9)* *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)* (eTable 1 in the Supplement). We excluded patients who died in the emergency department because these early trauma deaths are likely attributable to nonsurvivable injuries.<sup>13</sup> We also excluded patients with a primary noninjury diagnosis, with unknown county of residence, and patients treated at hospitals with unknown TC status because of missing or invalid facility identification numbers. Beneficiaries were included based on the first observed qualifying injury in the study period (ie, the index event). We used Medicare claims data from January 1, 2013, to December 31, 2014, to estimate preinjury health status and claims through December 31, 2016, to assess 365-day mortality.

Original Investigation | Surgery

# Association of Trauma Center Designation With Postdischarge Survival Among Older Adults With Injuries

Molly P. Jarman, PhD, MPH; Ginger Jin, MS; Joel S. Weissman, PhD; Arlene S. Ash, PhD; Jennifer Tjia, MD, MSCE; Ali Salim, MD; Adil Haider, MD, MPH; Zara Cooper, MD, MSc

March 16, 2022

## Methods

### Population and Data Sources

Using Medicare claims from Inpatient and Outpatient Research Identifiable Files,<sup>10,11</sup> we identified 433 169 fee-for-service beneficiaries aged 65 years or older diagnosed with traumatic injury between

We excluded patients older than 90 years, those with node-positive and/or metastatic disease, those with tumor stage of Tis/T1 or greater, those without continuous Medicare fee-for-service coverage, and those without available Medicare Part A and Part B claims data for 12 months before and after diagnosis.

injuries.<sup>13</sup> We also excluded patients with a primary noninjury diagnosis, with unknown county of residence, and patients treated at hospitals with unknown TC status because of missing or invalid facility identification numbers. Beneficiaries were included based on the first observed qualifying injury in the study period (ie, the index event). We used Medicare claims data from January 1, 2013, to December 31, 2014, to estimate preinjury health status and claims through December 31, 2016, to assess 365-day mortality.

Original Investigation | Surgery

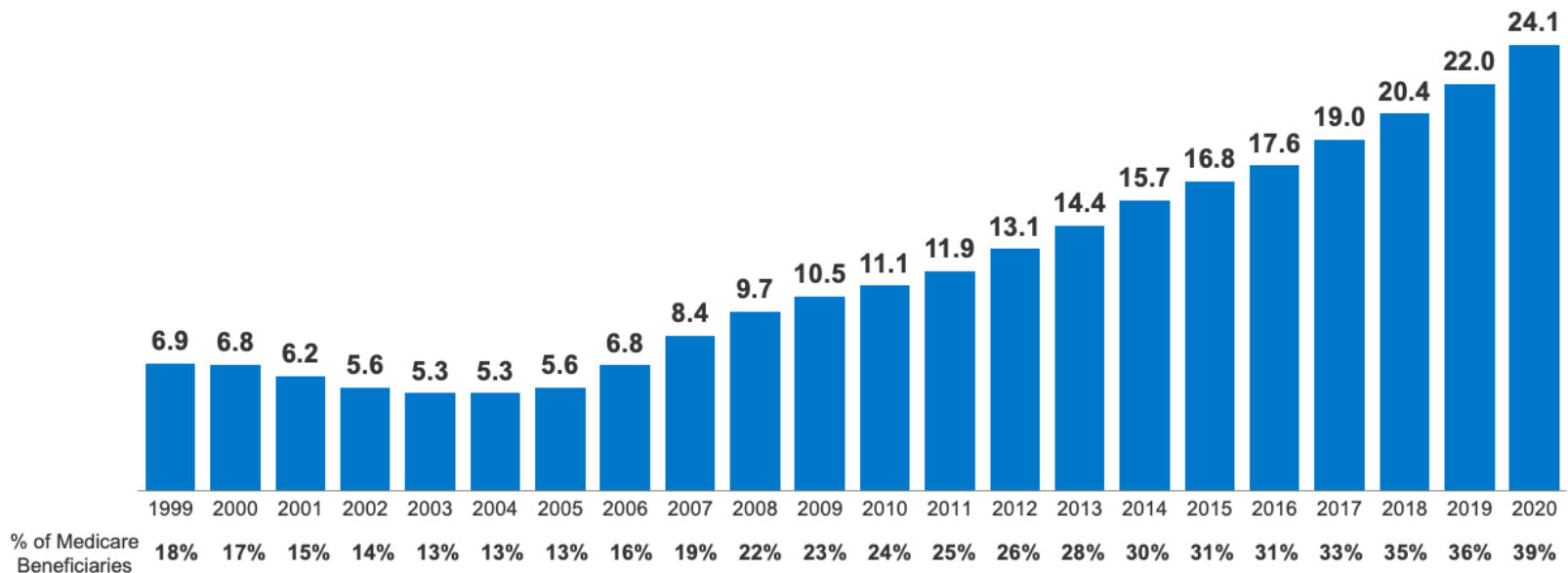
# Association of Trauma Center Designation With Postdischarge Survival Among Older Adults With Injuries

Molly P. Jarman, PhD, MPH; Ginger Jin, MS; Joel S. Weissman, PhD; Arlene S. Ash, PhD; Jennifer Tjia, MD, MSCE; Ali Salim, MD; Adil Haider, MD, MPH; Zara Cooper, MD, MSc

March 16, 2022

Figure 1

## Total Medicare Advantage Enrollment, 1999-2020 (in millions)



NOTE: Includes cost plans as well as Medicare Advantage plans. About 62 million people are enrolled in Medicare in 2020.

SOURCE: KFF analysis of CMS Medicare Advantage Enrollment Files 2008-2020, and MPR, 1999-2007; enrollment numbers from March of the respective year, with the exception of 2006, which is from April. Number of people eligible for Medicare comes from the CMS Medicare Advantage Penetration Files for years 2008-2009; for years 2010-2020, number of people eligible for Medicare comes from the Medicare Enrollment Dashboard.

Original Investigation | Surgery

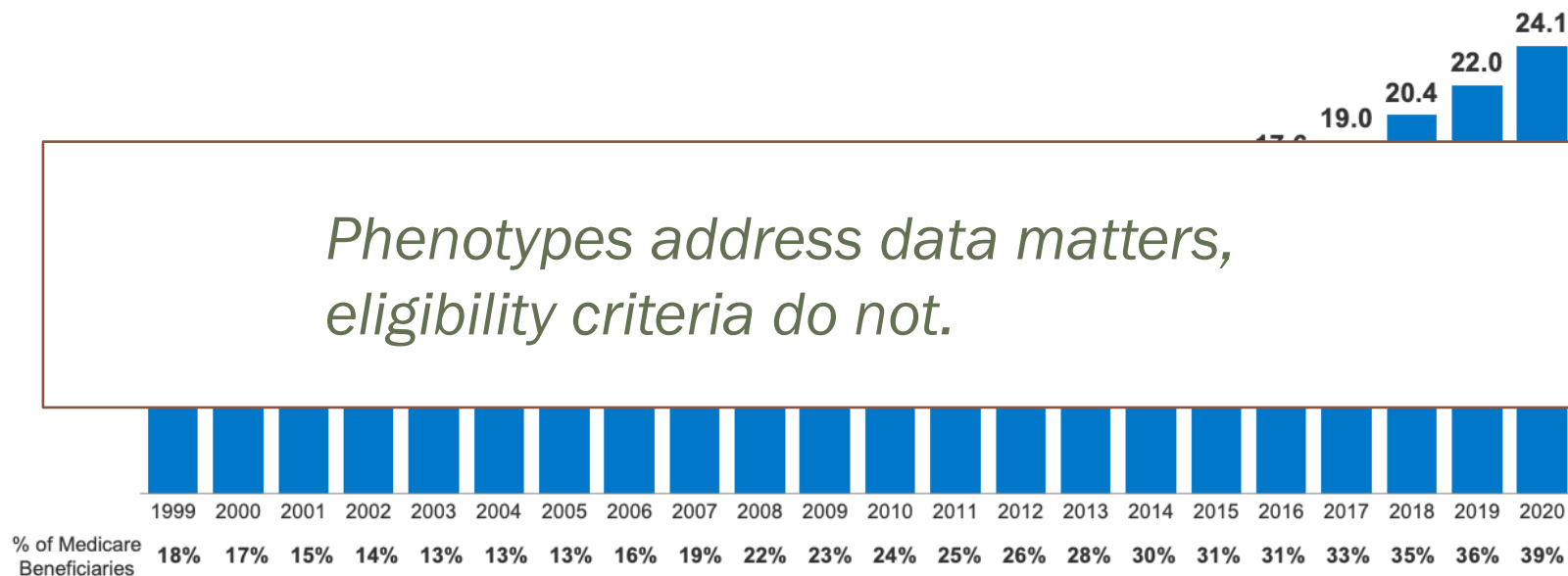
# Association of Trauma Center Designation With Postdischarge Survival Among Older Adults With Injuries

Molly P. Jarman, PhD, MPH; Ginger Jin, MS; Joel S. Weissman, PhD; Arlene S. Ash, PhD; Jennifer Tjia, MD, MSCE; Ali Salim, MD; Adil Haider, MD, MPH; Zara Cooper, MD, MSc

March 16, 2022

Figure 1

## Total Medicare Advantage Enrollment, 1999-2020 (in millions)



NOTE: Includes cost plans as well as Medicare Advantage plans. About 62 million people are enrolled in Medicare in 2020.

SOURCE: KFF analysis of CMS Medicare Advantage Enrollment Files 2008-2020, and MPR, 1999-2007; enrollment numbers from March of the respective year, with the exception of 2006, which is from April. Number of people eligible for Medicare comes from the CMS Medicare Advantage Penetration Files for years 2008-2009; for years 2010-2020, number of people eligible for Medicare comes from the Medicare Enrollment Dashboard.



## eligibility criteria

Starts from medical knowledge, models a cohort in clinical terms.

*“has prostate cancer”*

## hypothesis-first phenotyping = rule-based phenotyping

Starts from medical knowledge and data representation knowledge, models a cohort via how it could/should appear in the data.

*“has any record of ICD-10 code C61”*

## data-first phenotyping

Starts from the data, and possibly some kind of gold standard, and uses algorithms to derive a model.

*“this artificial neural network will tell you which patients are in”*

## eligibility criteria

Starts from medical knowledge, models a cohort in clinical terms.

*“has prostate cancer”*

## hypothesis-first phenotyping = rule-based phenotyping

Starts from medical knowledge and data representation knowledge, models a cohort via how it could/should appear in the data.

*“has any record of ICD-10 code ~~C61~~ prostatectomy with chemical castration”*

## data-first phenotyping

Starts from the data, and possibly some kind of gold standard, and uses algorithms to derive a model.

*“this artificial neural network will tell you which patients are in”*



reality →

data

phenotype  
developer

*Allegory of the Cave / Plato's Cave*



**Seth Rosen**

@sethrosen



Them: Can you just quickly pull this data for me?

Me: Sure, let me just:

```
SELECT * FROM  
some_ideal_clean_and_pristine.table_that_you_think_exists
```

[Traduire le Tweet](#)

1:42 PM · 20 avr. 2020 · Twitter Web App

*As the data model determines the algorithm,  
data representation determines the phenotype.*

# eMerge depression phenotype

<https://phekb.org/phenotype/depression>

## eMERGE-3 Depression Phenotype Pseudo Code

Primary site: Kaiser Permanente Washington & University of Washington

### Primary site contacts:

Aaron Scrol (aaron.scrol@kp.org)

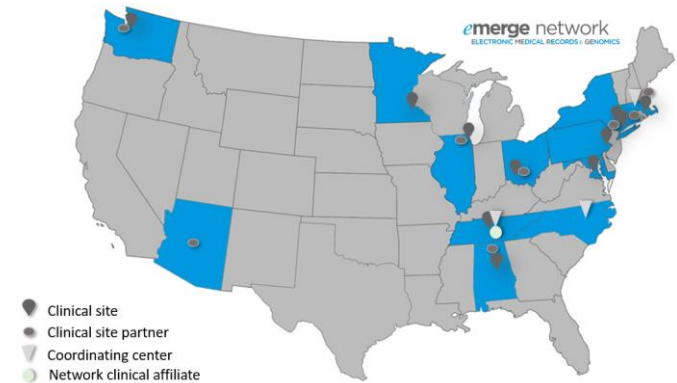
Arvind Ramaprasan (Arvind.ramaprasan@kp.org)

David Carrell (david.s.carrell@kp.org)

Version: April 17, 2019

## Contents

Background & significance.....	3
Measurement approach .....	3
Granular longitudinal data.....	4
Strategy for implementing this phenotype.....	4
Depression types.....	4
What to do about bipolar disorder? .....	4
The “2/30/180” rule.....	5
Exclusion criteria.....	5
Phenotype logic and flow diagram .....	5
Narrative summary of phenotype logic .....	7
Depression diagnosis codes .....	8
Depression diagnosis code granular data .....	8
Table DEP-2. Format of de-identified, granular longitudinal depression diagnosis code data to be provided for each patient over for all available time periods.....	8

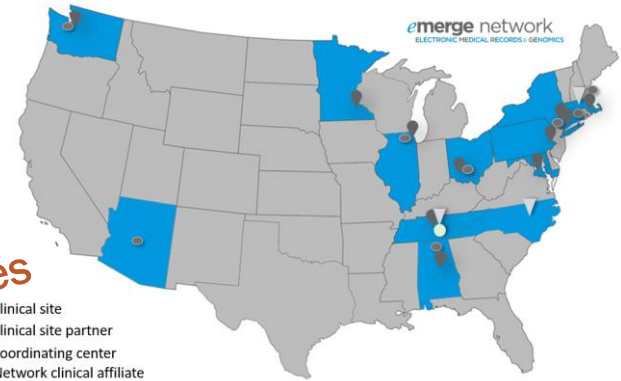


“designed to accommodate [...] variability across [...] sites with respect to:

- 1) overall duration of patient enrollment in the health system,
- 2) frequency of health system encounters,
- 3) availability of information from which diagnoses may be inferred,
- 4) diagnostic coding practices, and
- 5) treatment modalities.”

# eMerge depression phenotype

<https://phekb.org/phenotype/depression>



## Granular longitudinal data

- depression diagnoses, **ICD-9 or 10**
- antidepressant medications, **drug generic names**
- patient-reported depression scale measures, and **PHQ-9**
- psychotherapy **ICD or CPT codes**

Specifications for these four types of data are provided in (...) **data dictionaries**

All data and logic needed to implement this phenotype algorithm can be derived from the granular longitudinal data described above.

## The “2/30/180” rule

requires evidence to be present on at least two (2) distinct calendar days that are at least thirty (30) days apart and not more than one hundred and eighty (180) days apart

is intended to avoid (...) administrative artifacts of a diagnostic process that resulted in the subject **not** being diagnosed with depression.

# eMerge depression phenotype

## Case type 1

$$Ca_1 = C * B$$

## Case type 2

$$Ca_2 = D * !C * B$$

## Case type 3

$$Ca_3 = E * !D * !C * B$$

## Control

$$Ct = !I * !G * H * !F * !B$$

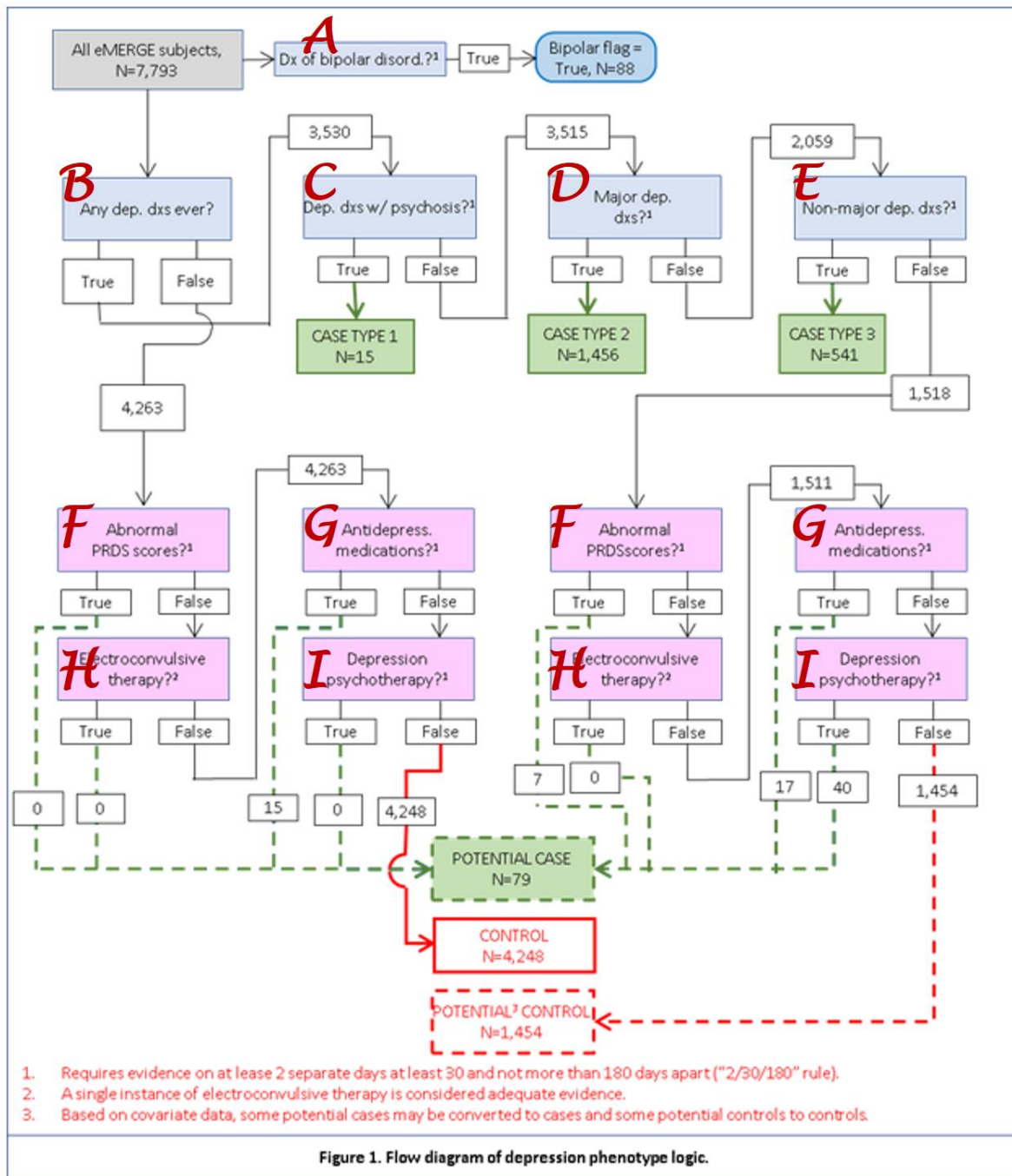


Figure 1. Flow diagram of depression phenotype logic.

...how long will it take  
run this phenotype on  
87 million patients,  
billions of records?

# eMerge depression phenotype

## Case type 1

$$Ca_1 = C * B$$

## Case type 2

$$Ca_2 = D * !C * B$$

## Case type 3

$$Ca_3 = E * !D * !C * B$$

## Control

$$Ct = !I * !G * H * !F * !B$$

Kleene's three-valued logic ( $K_3$ ) truth tables:

NOT(A)		AND(A, B)			OR(A, B)		
A	$\neg A$	A $\wedge$ B		B			
A	$\neg A$	A	B	F	U	T	
F	T	F	F	F	F	F	
U	U	A	U	F	U	U	
T	F	T	F	U	U	T	
			T	F	T	T	

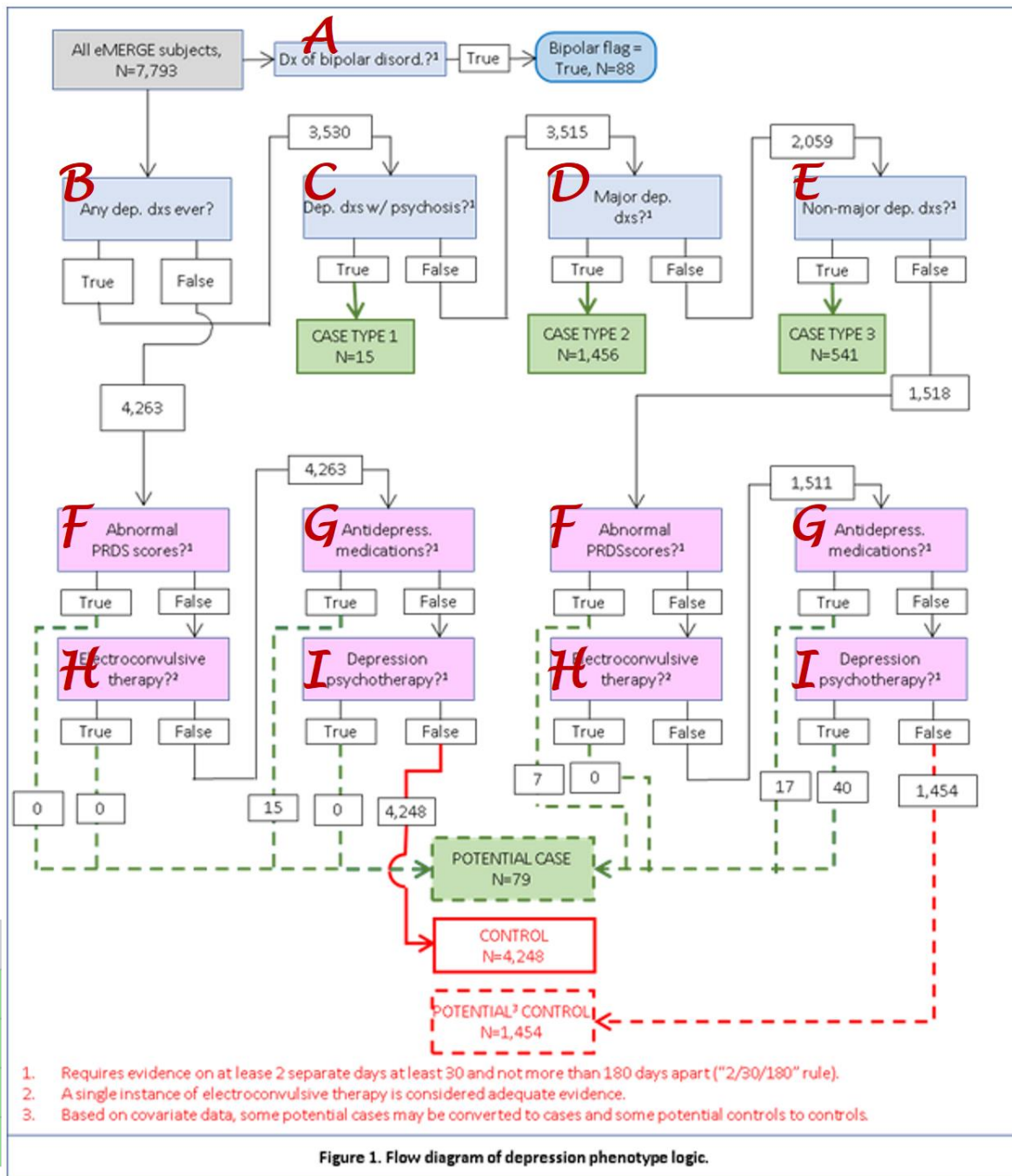
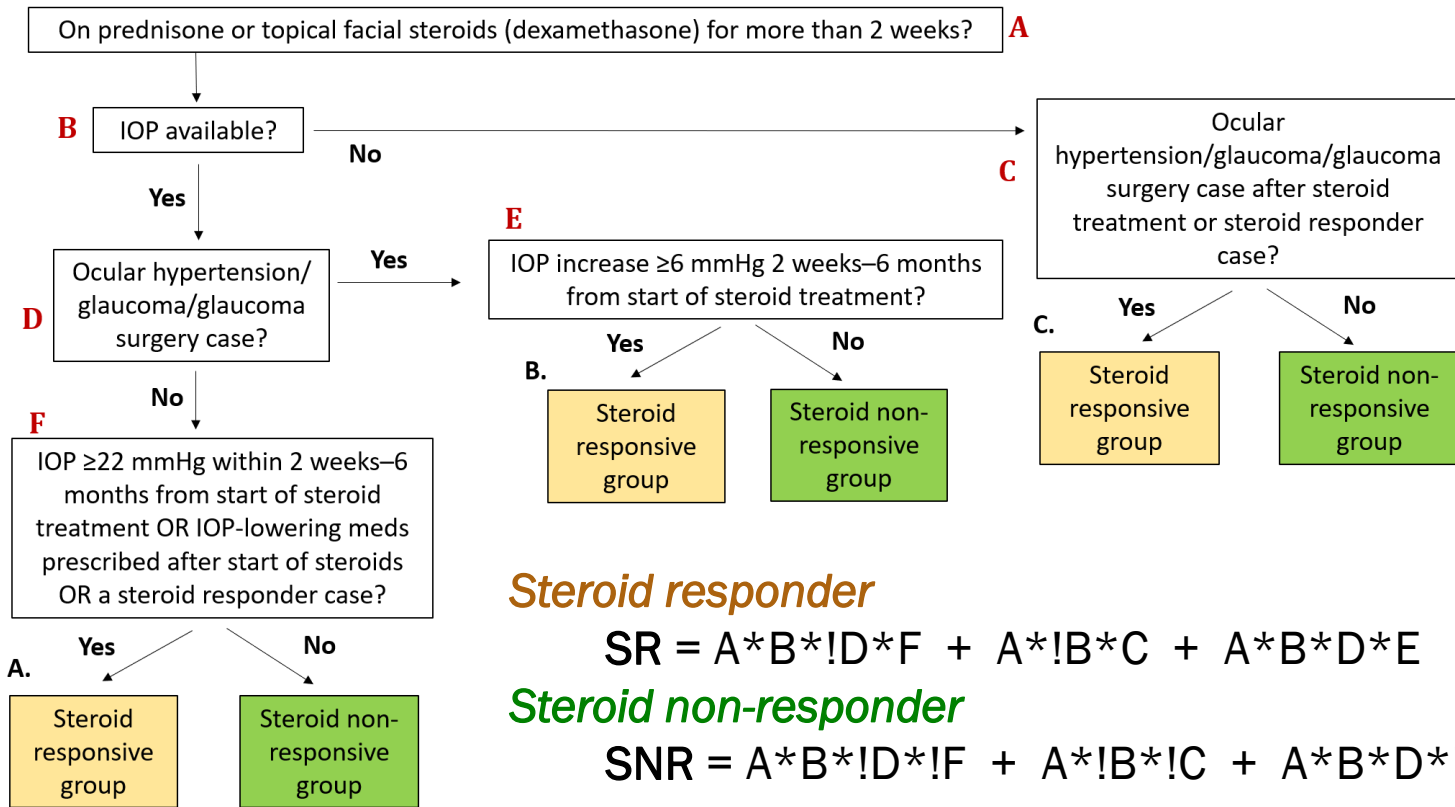


Figure 1. Flow diagram of depression phenotype logic.



# Steroid-induced glaucoma



## Steroid responder

$$SR = A * B * !D * F + A * !B * C + A * B * D * E$$

## Steroid non-responder

$$SNR = A * B * !D * !F + A * !B * !C + A * B * D * !E$$

## Neither

$$N = !(SR) * !(SNR)$$

$$N = !(A * B * !D * F + A * !B * C + A * B * D * E) * !(A * B * !D * !F + A * !B * !C + A * B * D * !E)$$

## BOOLEAN EXPRESSIONS SIMPLIFIER

★ LOGICAL EXPRESSION CALCULATOR/SIMPLIFIER/MINIFIER

$!(a*b*!d*f + a*!b*c + a*b*d*e) * !(a*b*!d*!f$

★ RESULT FORMAT

ANY FORMAT

DISJUNCTIVE NORMAL FORM DNF (SUM OF PRODUCTS/SOP/MINTERMS)

CONJUNCTIVE NORMAL FORM CNF (PRODUCT OF SUMS/POS/MAXTERMS)

ONLY NAND GATES (NOT-AND  $\bar{\wedge}$ )

ONLY NOR GATES (NOT-OR  $\bar{\vee}$ )

★ NOTATION  ALGEBRAIC (\*, +, !)

LOGICAL ( $\wedge$ ,  $\vee$ ,  $\neg$ )

PROGRAMMING (&&, ||, ~)

LITERAL (AND, OR, NOT)

▶ CALCULATE

See also: [Truth Table](#) – [Equation Solver](#) – [Binary Code](#)

<https://www.dcode.fr/>

or

<https://www.boolean-algebra.com/>

or

<https://www.symbolab.com/solver/boolean-algebra-calculator>

or...

### Results

$!(a * b * !d * f + a * !b * c + a * b * d * e) * !(a * b * !d * !f + a * !b * !c + a * b * d * !e)$

↓

$! a$

Boolean Expressions Calculator - [dCode](#)

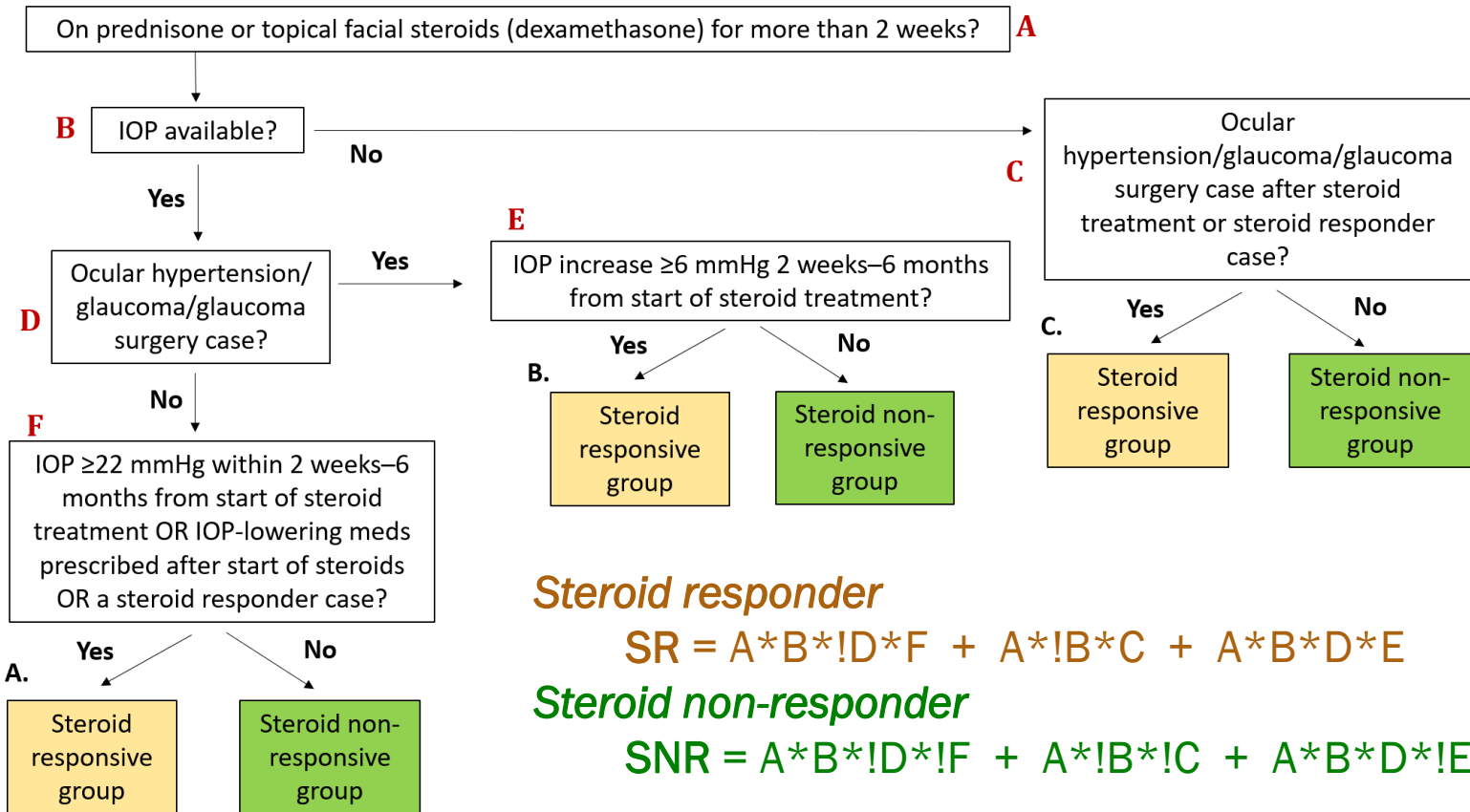
**Neither**

$$N = !(SR) * !(SNR)$$

$$N = !(A*B*!D*F + A*!B*C + A*B*D*E) * !(A*B*!D*!F + A*!B*!C + A*B*D*!E)$$

$$N = !A$$

# Steroid-induced glaucoma



## Steroid responder

$$SR = A*B*!D*F + A*!B*C + A*B*D*E$$

## Steroid non-responder

$$SNR = A*B*!D*!F + A*!B*!C + A*B*D*!E$$

## Neither

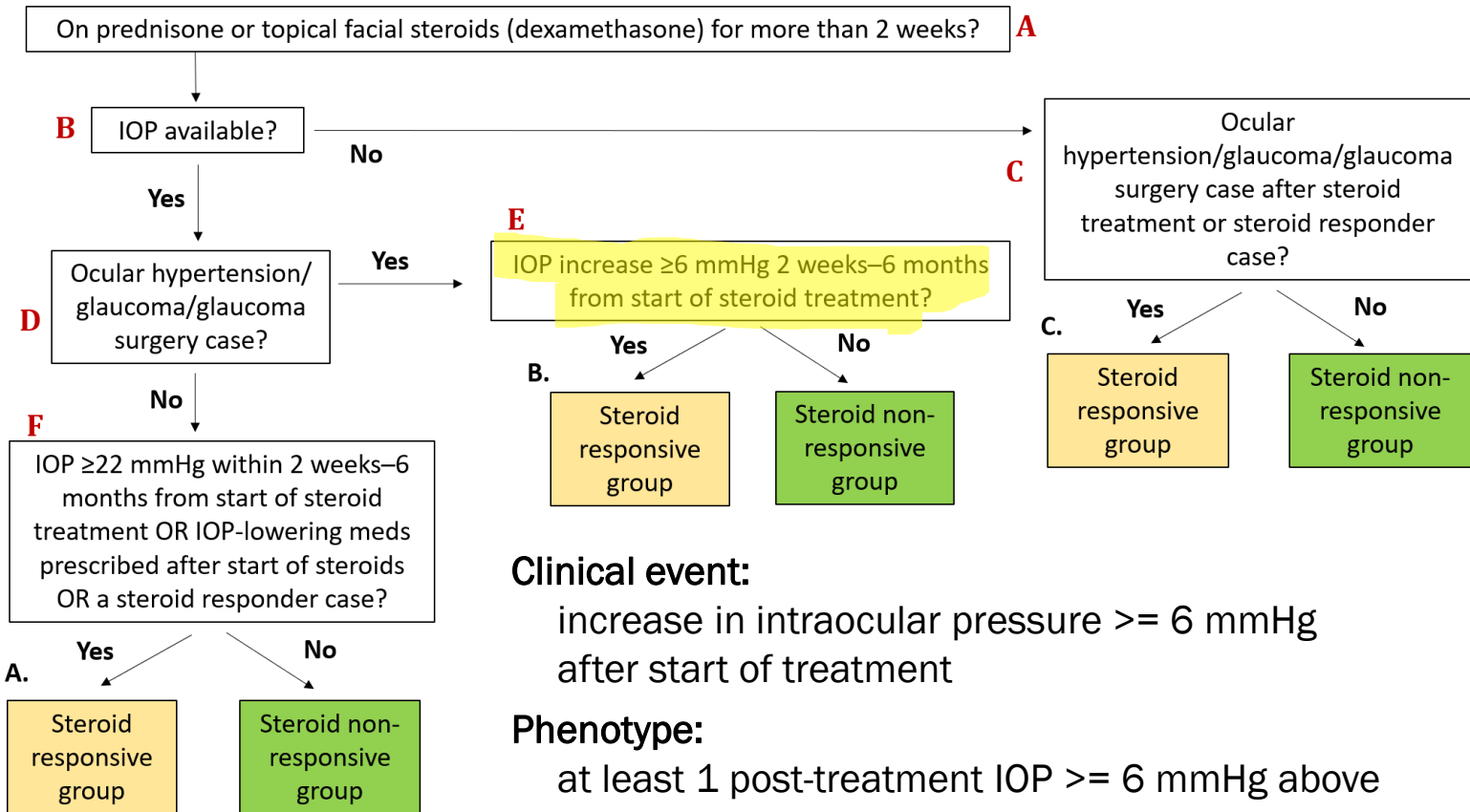
$$N = !(SR) * !(SNR)$$

$$N = !(A*B*!D*F + A*!B*C + A*B*D*E) * !(A*B*!D*!F + A*!B*!C + A*B*D*!E)$$

$$N = !A$$

Covid Measurement		Covid Diagnosis	Temporal	variant	Cohort Id
+ve, no -ve	AND	Diagnosis	-3d to 3d/0d to 3d	2	C47
+ve	AND	Diagnosis	-3d to 3d/0d to 3d	2	C46
+ve, no -ve		-----	-3d to 3d/0d to 3d	2	
+ve		-----		1	C58
+ve, no -ve	OR	Diagnosis, (no -ve/missing)	-3d to 3d/0d to 3d	2	
+ve, no -ve (2)	OR	Diagnosis, (no -ve) (2)	-3d to 3d/0d to 3d	4	
+ve (1)	OR	Diagnosis, (no -ve/tested-missing) (2)	-3d to 3d/0d to 3d	2	
+ve (1)	OR	Diagnosis, (no -ve) (2)	-3d to 3d/0d to 3d	2	C56/C84/C85
+ve (1)	OR	Diagnosis (1)		1	C55
-----		Diagnosis (1), (no -ve/tested-missing) (2)		1	
-----		Diagnosis (1), (no -ve) (2)		1	
-----		Diagnosis (1)		1	C44

# Steroid-induced glaucoma

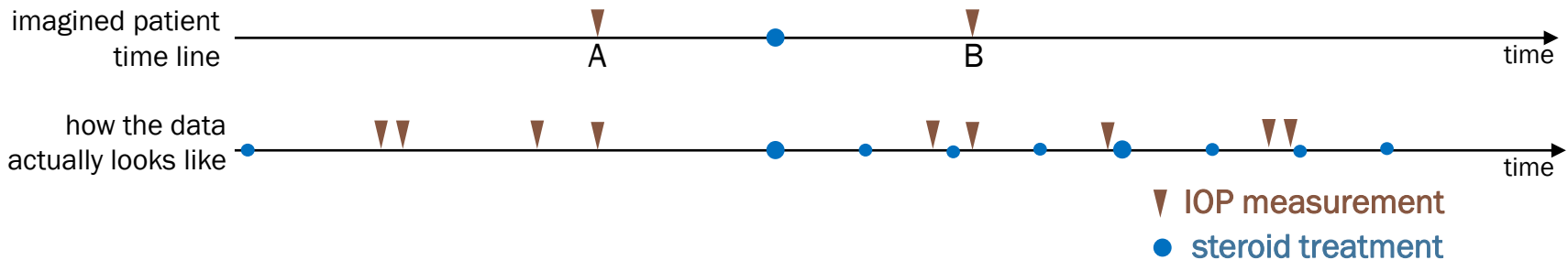


## Clinical event:

increase in intraocular pressure  $\geq 6$  mmHg after start of treatment

## Phenotype:

at least 1 post-treatment IOP  $\geq 6$  mmHg above the *pre-treatment average*





data warehouse

phenotype:

*patients with disseminated intracascular coagulation*

phenotyping



observational study

...platelet count?

...patient age?

...ISTH score?

	Overt DIC scoring system in JMHW		Overt DIC scoring system in ISTH	
Platelet count ( $\times 10^9/l$ )	$\leq 50 \uparrow$	3	$\leq 50 \uparrow$	2
	50 - 80	2	50 - 100	1
	80 - 120	1	$\geq 100$	0
	$\geq 120$	0		
Prothrombin time (PT)	PT-INR $\geq 1.67$	2	PT-prolongation $\geq 6$ sec	2
	1.25 - 1.67	1	3 - 6 sec	1
	$\leq 1.25$	0	$\leq 3$ sec	0
Fibrinogen (mg/dl)	$\leq 100 \uparrow$	2	$\leq 100 \uparrow$	1
	100 - 150	1	$\geq 100$	0
	$\geq 150$	0		
Fibrin-related marker	FDP (mg/l) $\geq 40$	3	D-dimer ( $\mu\text{g/ml}$ ) $\geq 4$	3
	20 - 40	2	1 - 4	2
	10 - 20	1	$\leq 1$	0
	$\leq 10$	0		
Symptoms and underlying diseases	Bleeding symptoms* or organ dysfunction†	1	underlying disease known to be associated with DIC	need to diagnose as DIC
Definition of DIC	Seven points or more		Five points or more	



data warehouse

phenotyping



observational study

phenotype:  
*patients with  
 ISTH score > 4*

	Overt DIC scoring system in JMW		Overt DIC scoring system in ISTH	
Platelet count ( $\times 10^9/l$ )	$\leq 50 \dagger$	3	$\leq 50 \dagger$	2
	50 – 80	2	50 – 100	1
	80 – 120	1	$\geq 100$	0
	$\geq 120$	0		
Prothrombin time (PT)	PT-INR		PT-prolongation	
	$\geq 1.67$	2	$\geq 6$ sec	2
	1.25– 1.67	1	3 – 6 sec	1
	$\leq 1.25$	0	$\leq 3$ sec	0
Fibrinogen (mg/dl)	$\leq 100 \dagger$	2	$\leq 100 \dagger$	1
	100 – 150	1	$\geq 100$	0
	$\geq 150$	0		
Fibrin-related marker	FDP (mg/l)		D-dimer ( $\mu\text{g/ml}$ )	
	$\geq 40$	3	$\geq 4$	3
	20 - 40	2	1 – 4	2
	10 – 20	1	$\leq 1$	0
	$\leq 10$	0		
Symptoms and underlying diseases	Bleeding symptoms* or organ dysfunction†	1	underlying disease known to be associated with DIC	need to diagnose as DIC
Definition of DIC	Seven points or more		Five points or more	

...platelet count?

...patient age?

...ISTH score?



Current Age \*

Age must be between 20-79

Sex \*

Male Female

Race \*

White African American Other

Systolic Blood Pressure (mm Hg) \*

Value must be between 90-200

Diastolic Blood Pressure (mm Hg) \*

Value must be between 60-130

Total Cholesterol (mg/dL) \*

Value must be between 130 - 320

HDL Cholesterol (mg/dL) \*

Value must be between 20 - 100

LDL Cholesterol (mg/dL)

Value must be between 30-300

History of Diabetes? \*

Yes No

Smoker?

Current Former Never

On Hypertension Treatment? \*

Yes No

On a Statin?

Yes No

On Aspirin Therapy?

Yes No

...inner join X on...



Current Age  \*

Age must be between 20-79

Sex \*

Race \*

Systolic Blood Pressure (mm Hg) \*

Value must be between 90-200

Diastolic Blood Pressure (mm Hg) \*

Value must be between 60-130

Total Cholesterol (mg/dL) \*

Value must be between 130 - 320

HDL Cholesterol (mg/dL) \*

Value must be between 20 - 100

LDL Cholesterol (mg/dL)  \*

Value must be between 30-300

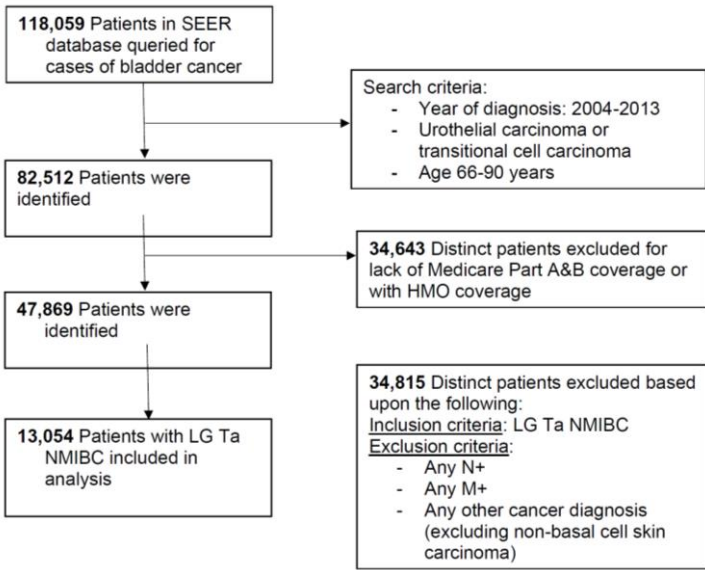
History of Diabetes? \*

 Smoker?  \*"/> "/> 

On Hypertension Treatment? \*

On a Statin?  \*On Aspirin Therapy?  \*

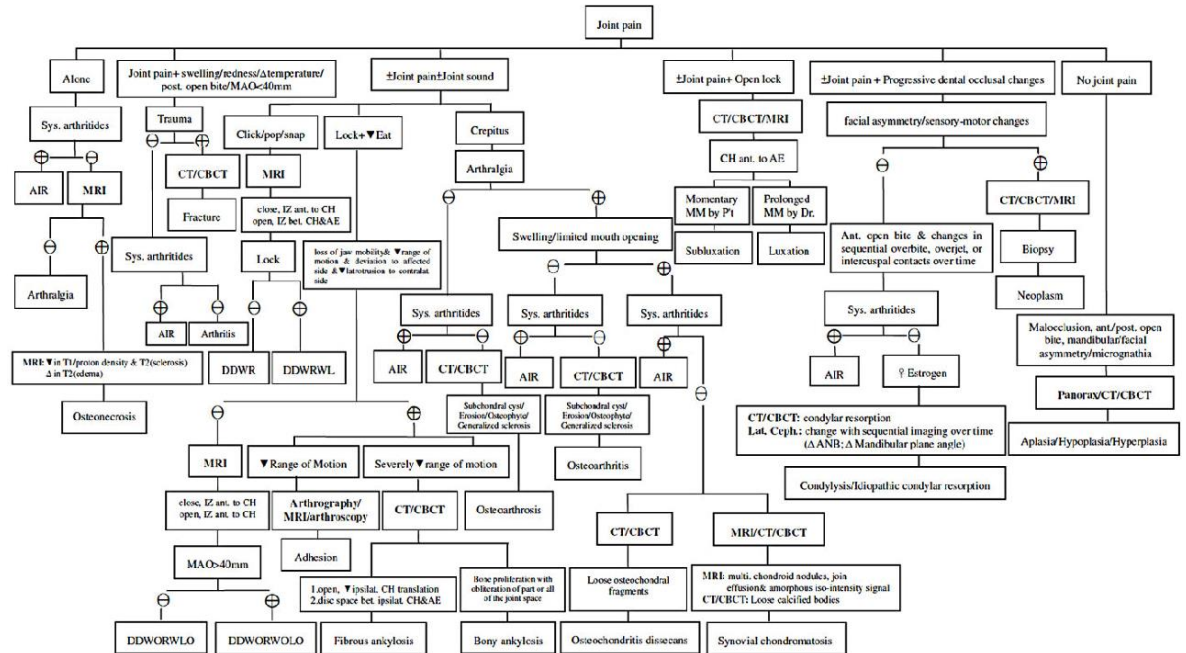
Computational complexity can grow exponentially when data points are interdependent.



Complexity level seen in today's rule-based phenotypes

**NoSQL phenotyping:**  
*ACE: the Advanced Cohort Engine for searching longitudinal patient records.*  
 Callahan et al, 2021. doi: 10.1093/jamia/ocab027

Complexity level seen in today's medical guidelines



# GUI-based phenotyping tools

New Query  
0 patients

+ New Query Databases

All Concepts Search...

Demographics 58176

Current Age 58176

Ethnicity 58176

Gender 58176

Race 48176

Vital Status 58176

Diseased 2669

Living 55507

Encounters 49640

My Saved Cohorts

Procedures 49521

Learn More

Limit to

Patients Who Anytime At Least 1x

And Anytime At Least 1x

And Anytime At Least 1x

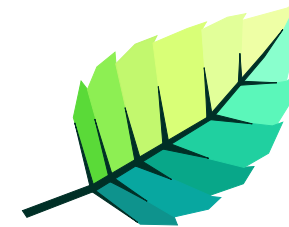
Had Procedure for OPERATIONS ON THE ENDOCRINE SYSTEM (ICD9.06.01-07.99)

Are aged BETWEEN 65 and 80 years

In the Same Encounter

In the Same Encounter

Run Query



University of Washington' Leaf

ATLAS

Home

Data Sources

Search

Concept Sets

Cohort Definitions

Characterizations

Cohort Pathways

Incidence Rates

Profiles

Estimation

Prediction

Jobs

Configuration

Feedback

Apache 2.0  
open source software

provided by  
**OHDSI**  
join the journey.

Cohort #1770710

New users of ACE inhibitors as first-line monotherapy for hypertension

Definition Concept Sets Generation Reporting Export Messages 3

enter a cohort definition description here

Cohort Entry Events

Events having any of the following criteria:

+ Add Initial Event

a drug exposure of ACE inhibitors

+ Add attribute...

Delete Criteria

for the first time in the person's history

with continuous observation of at least 365 days before and 0 days after event index date

Limit initial events to: earliest event per person.

Restrict initial events

Inclusion Criteria

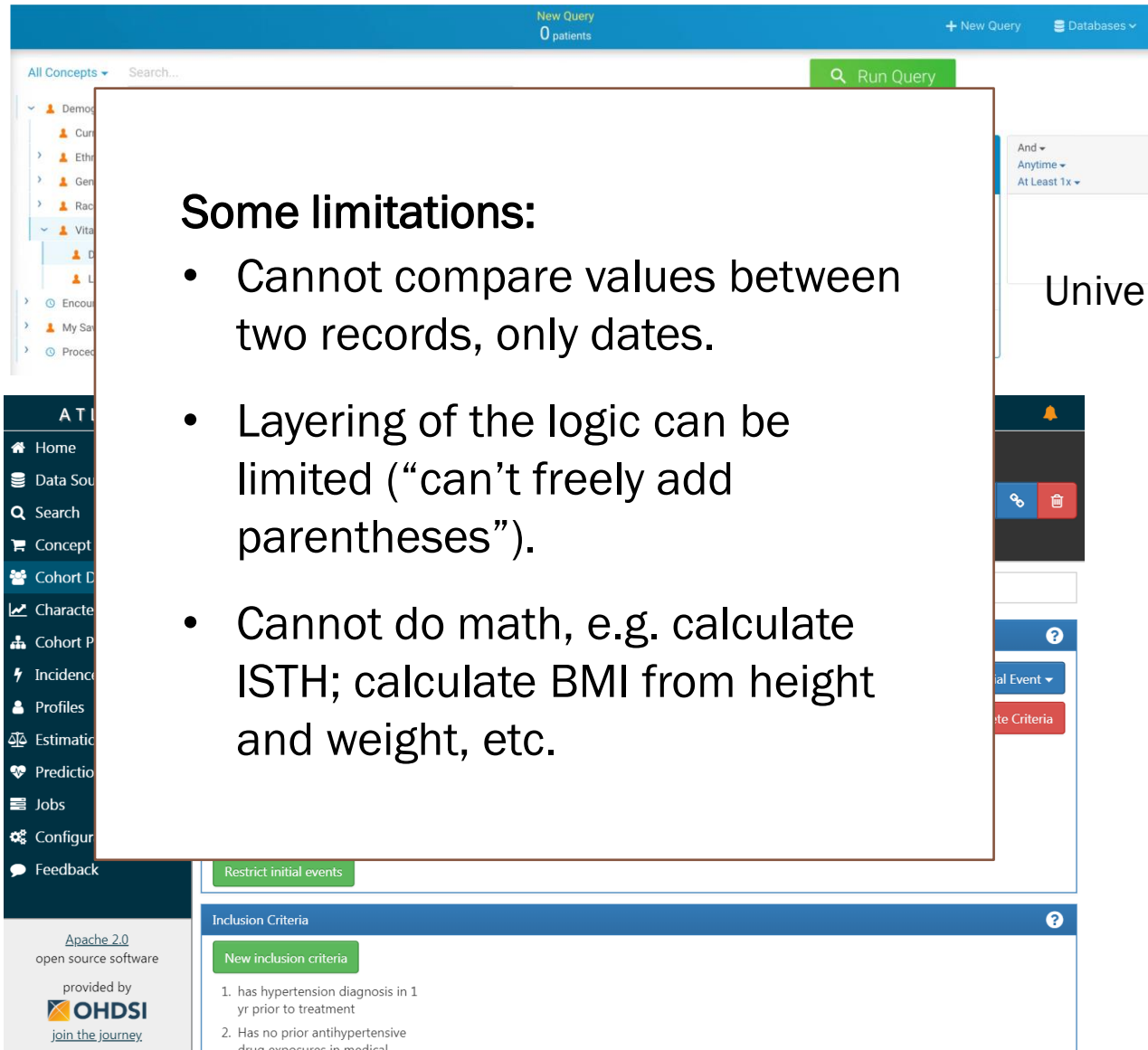
New inclusion criteria

1. has hypertension diagnosis in 1 yr prior to treatment
2. Has no prior antihypertensive drug exposures in medical



OHDSI' Atlas

# GUI-based phenotyping tools



The screenshot shows the OHDSI Atlas interface. At the top, it says "New Query" and "0 patients". A search bar is visible. On the left, there is a navigation menu with options like "Home", "Data Sources", "Search", "Concepts", "Cohort Definition", "Characteristics", "Cohort Profiles", "Incidence", "Profiles", "Estimation", "Prediction", "Jobs", "Configuration", and "Feedback". The central area contains a text box with the following text:

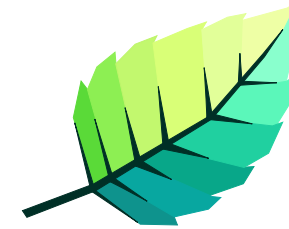
**Some limitations:**

- Cannot compare values between two records, only dates.
- Layering of the logic can be limited (“can’t freely add parentheses”).
- Cannot do math, e.g. calculate ISTH; calculate BMI from height and weight, etc.

At the bottom of the screenshot, there is a section for "Inclusion Criteria" with a "New inclusion criteria" button and a list of criteria:

1. has hypertension diagnosis in 1 yr prior to treatment
2. Has no prior antihypertensive drug exposures in medical

At the bottom left, it says "Apache 2.0 open source software provided by OHDSI join the journey."



University of Washington' Leaf



OHDSI' Atlas

# Conclusions

---

- A phenotype is an algorithm that abstracts patients and covariates from data records left over by medical events.
- The job of a phenotype is to look at patient data, and infer truths about the patient and the care delivered.
- Phenotypes address data issues, unlike eligibility criteria.

# Conclusions

---

- Phenotyping requires you to know, or require, a particular data representation.
- Boolean algebra can help investigate a phenotype at conceptual level.

# Conclusions

---

- Computational complexity can impart insidious bias on phenotypes and warrants further research.
- The greatest computational complexity in rule-based phenotyping is usually when many data points are interdependent and part of the “entry event.”
- Novel database technologies (“NoSQL”) can reduce computational complexity by orders of magnitude.

# Conclusions

---

- Open-source tools can model the logic of most phenotypes, but still lack major features and cannot compete with the expressive power of a programming language.



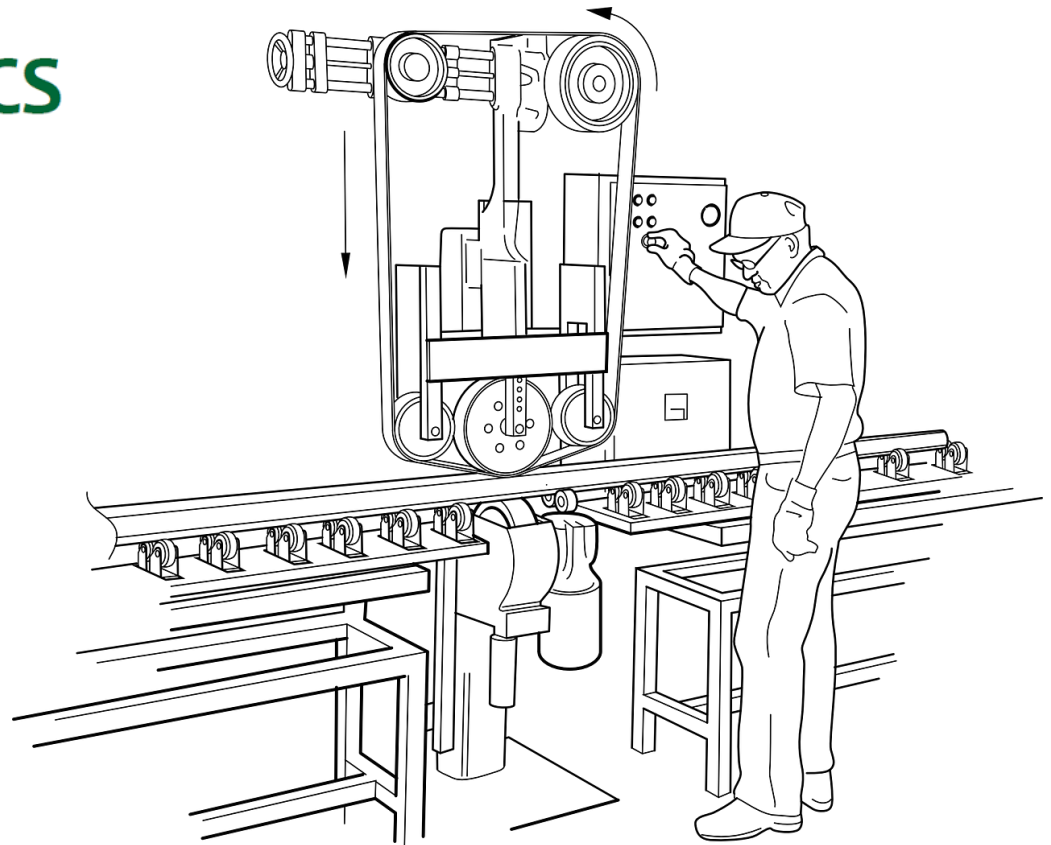


*Thank you!*

*Fabrício Kury, MD*

[fab@kury.dev](mailto:fab@kury.dev)

[github.com/fabkury](https://github.com/fabkury)



Current Practice and Frontiers in  
Rule-Based Electronic Phenotyping

April 1st, 2022

**POWERTALK**  
**SEMINAR SERIES**  
**CLINICAL INFORMATICS**