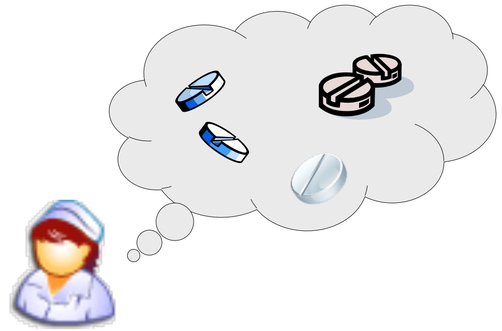
Data-Driven Healthcare: Visual Analytics for Exploration and Prediction of Clinical Data

Adam Perer

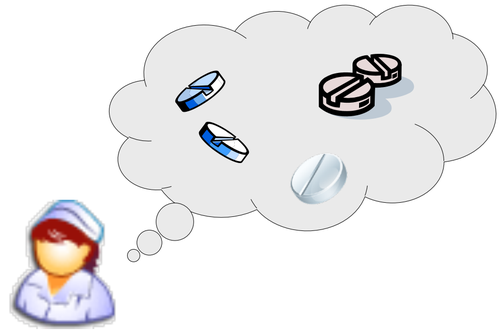
IBM Research

Patient

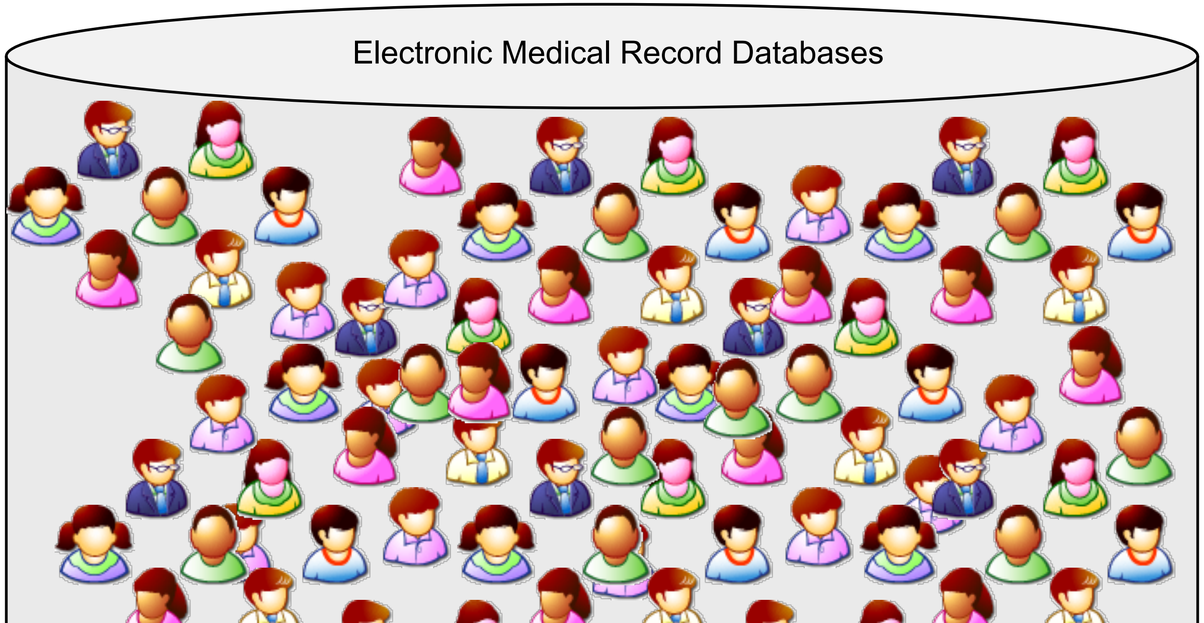


Clinician

Patient



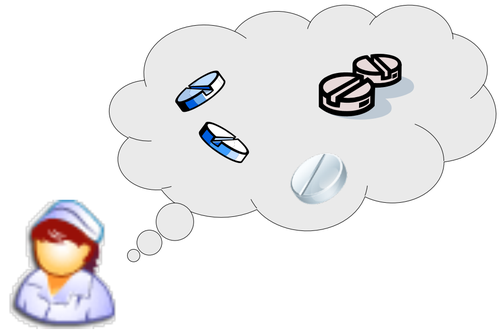
Clinician



Thousands or Millions of Patients

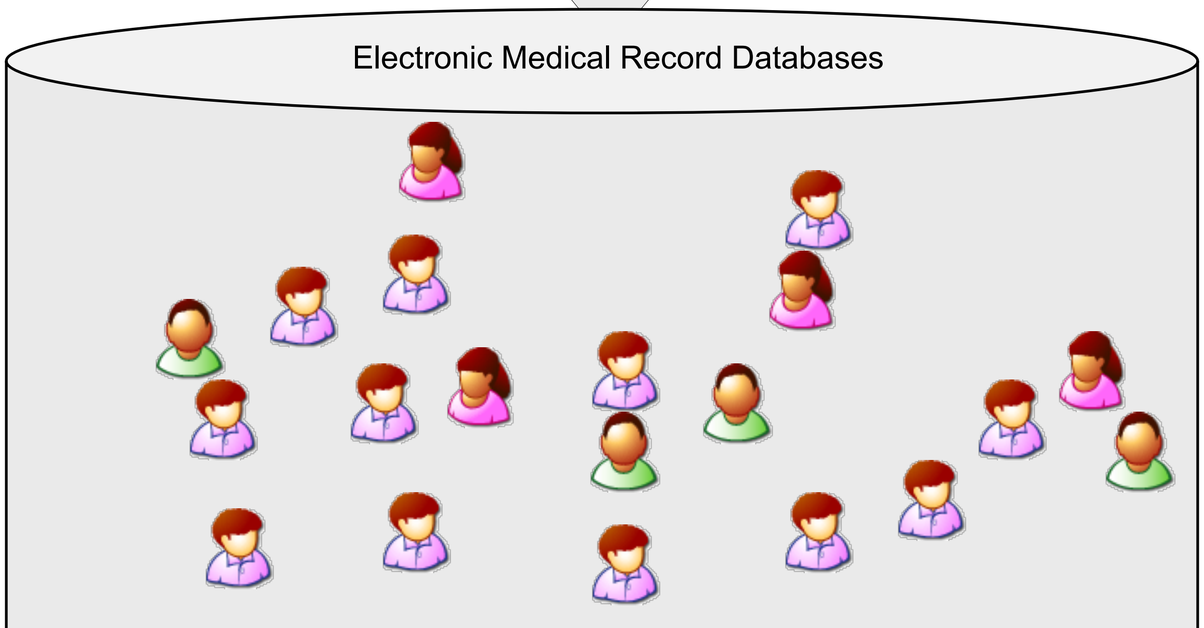
- 10+ Years of Data Per Patient
- Tens of Thousands of Features
  - Demographics
  - Diagnoses
  - Labs
  - Procedures
  - Claims
- Unstructured Physician Notes

Patient



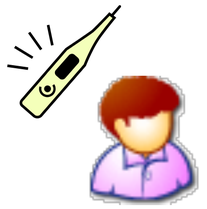
Clinician

Search and Analysis

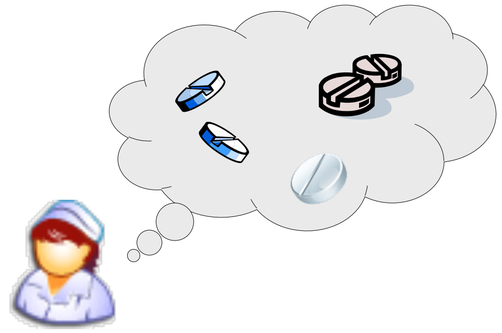


Thousands or Millions of Patients

- 10+ Years of Data Per Patient
- Tens of Thousands of Features
  - Demographics
  - Diagnoses
  - Labs
  - Procedures
  - Claims
- Unstructured Physician Notes

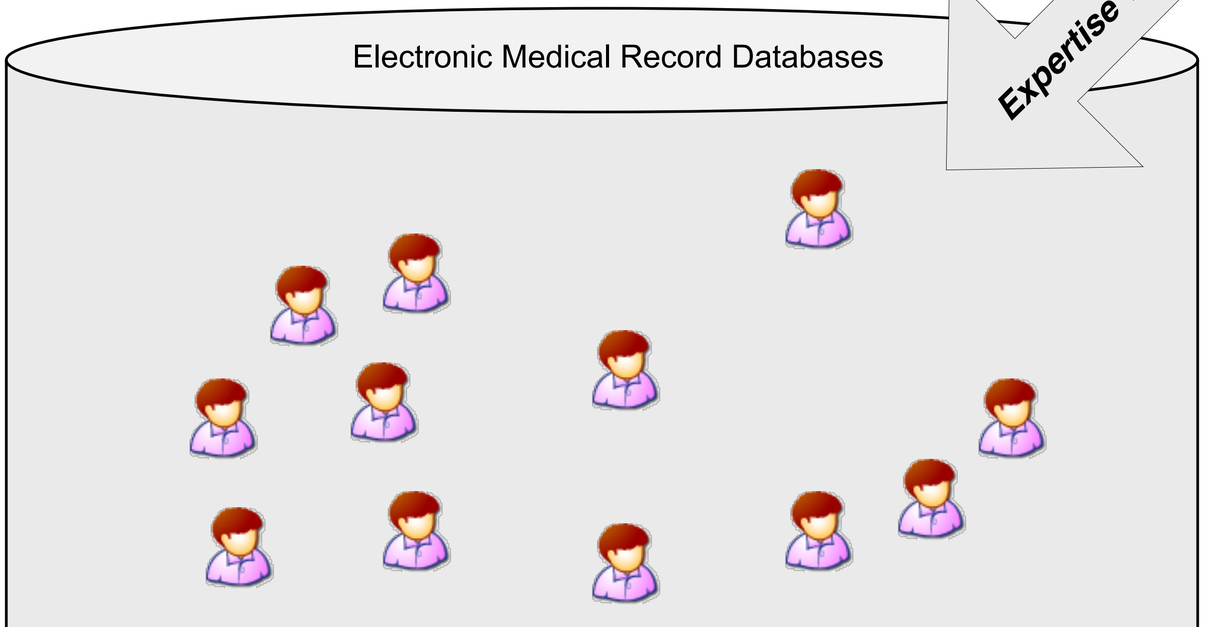


Patient



Clinician

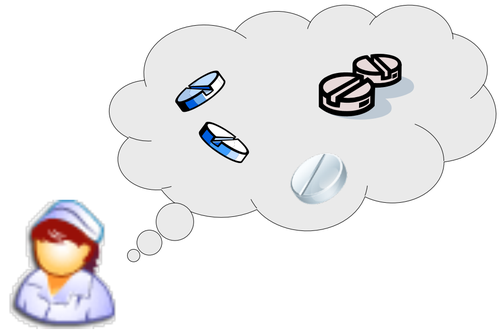
Expertise via Interaction



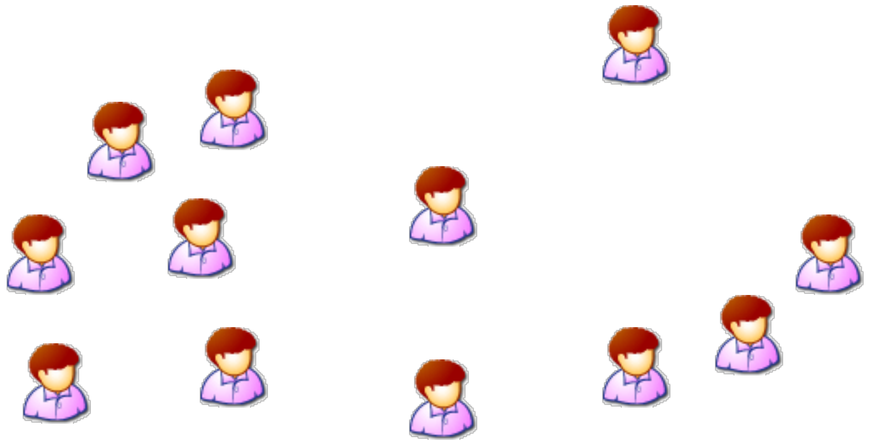
Thousands or Millions of Patients

- 10+ Years of Data Per Patient
- Tens of Thousands of Features
  - Demographics
  - Diagnoses
  - Labs
  - Procedures
  - Claims
- Unstructured Physician Notes

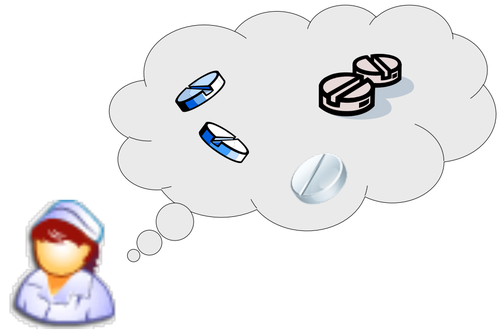
Patient



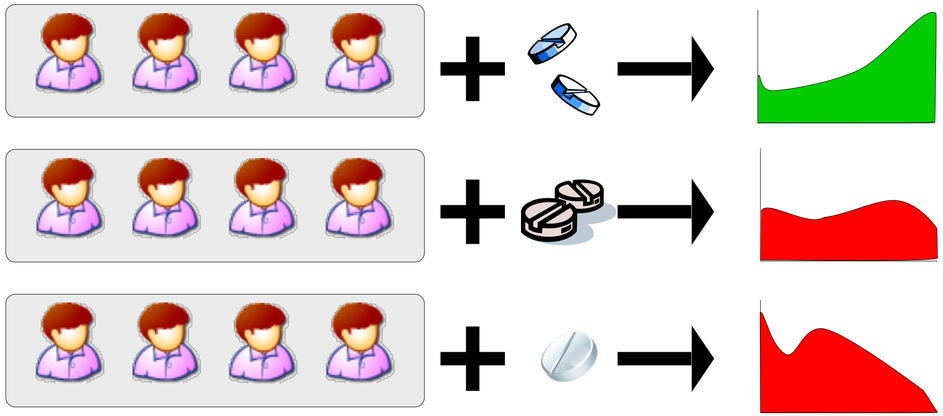
Clinician

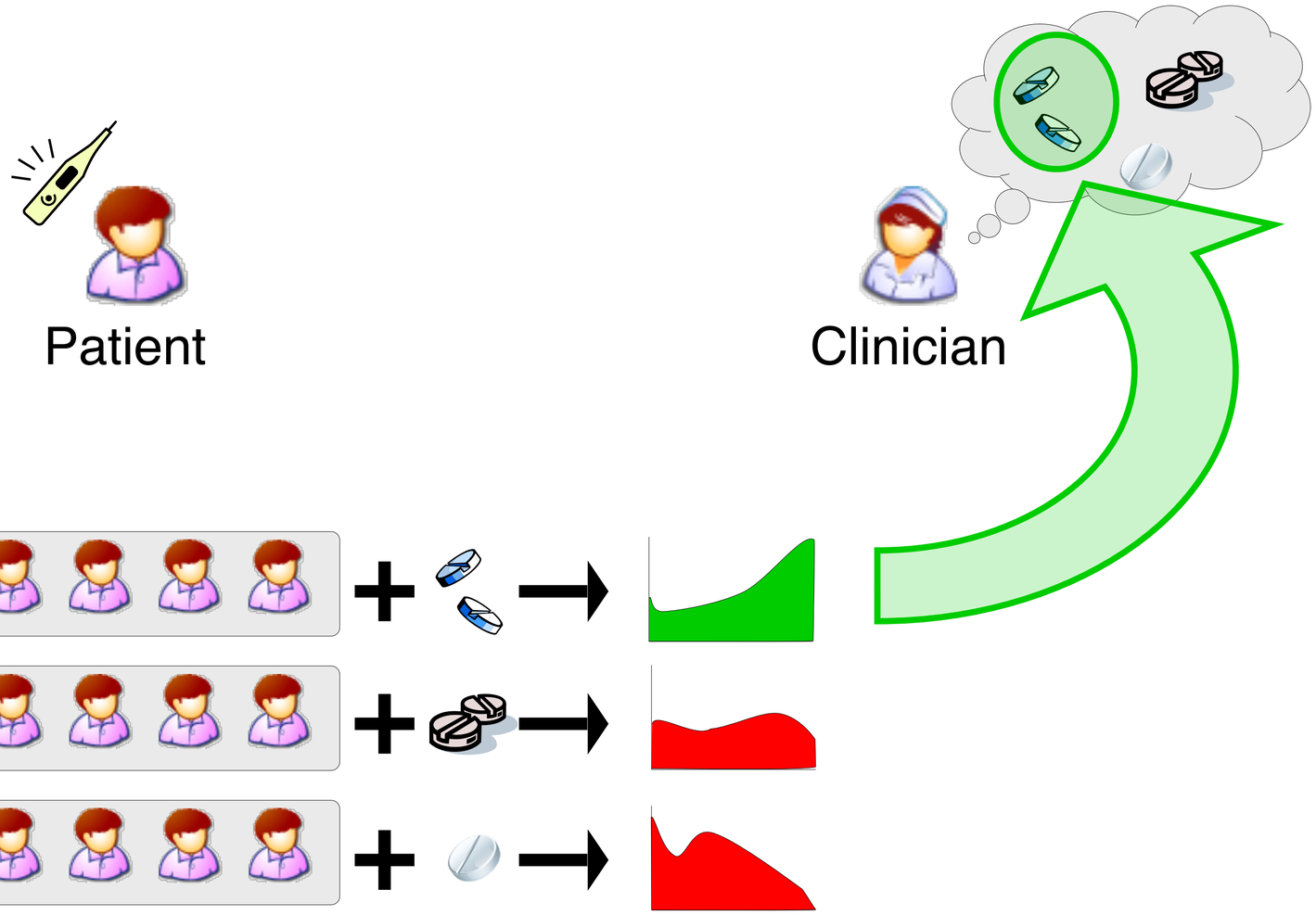


Patient



Clinician




Patient
Clinician

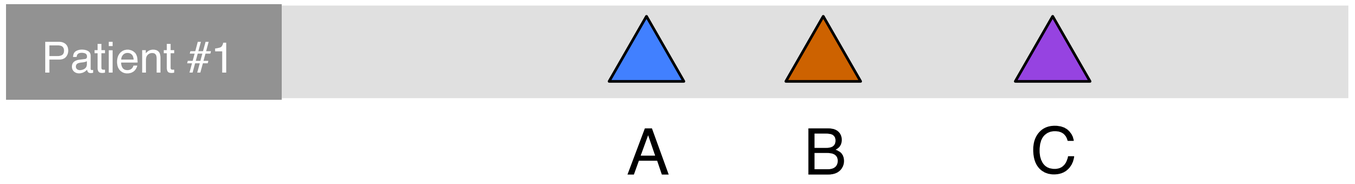
outline

- Visual tools for exploring clinical data support unearthing insights from clinical records
  - **CareFlow**
- Beyond exploration, clinical researchers often want predictions, too.
  - **Coquito**, **Prospector**

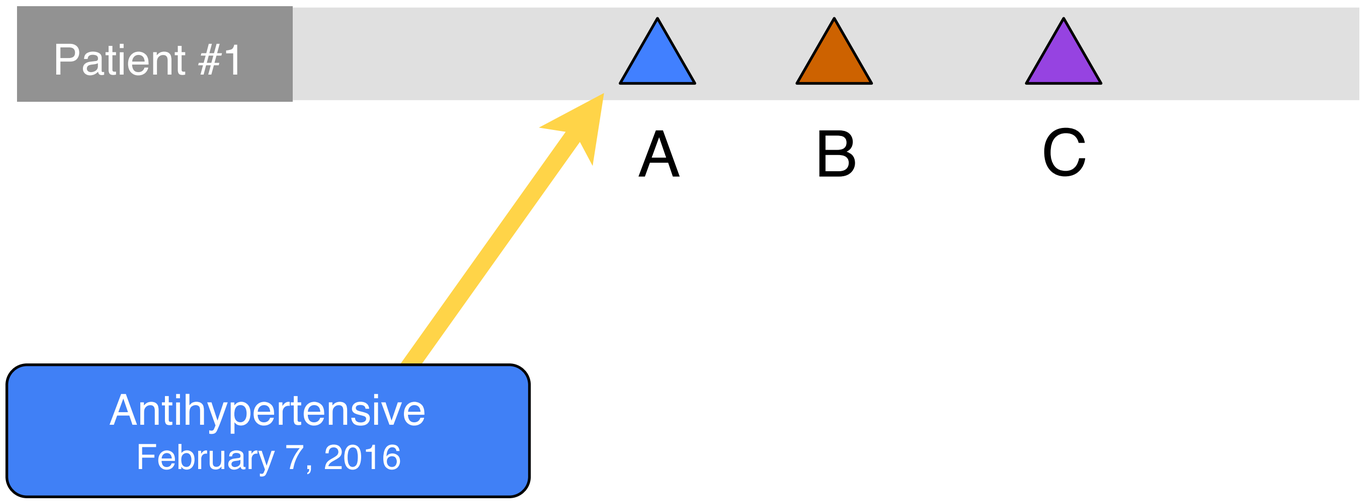
CareFlow

IBM **WatsonHealth**

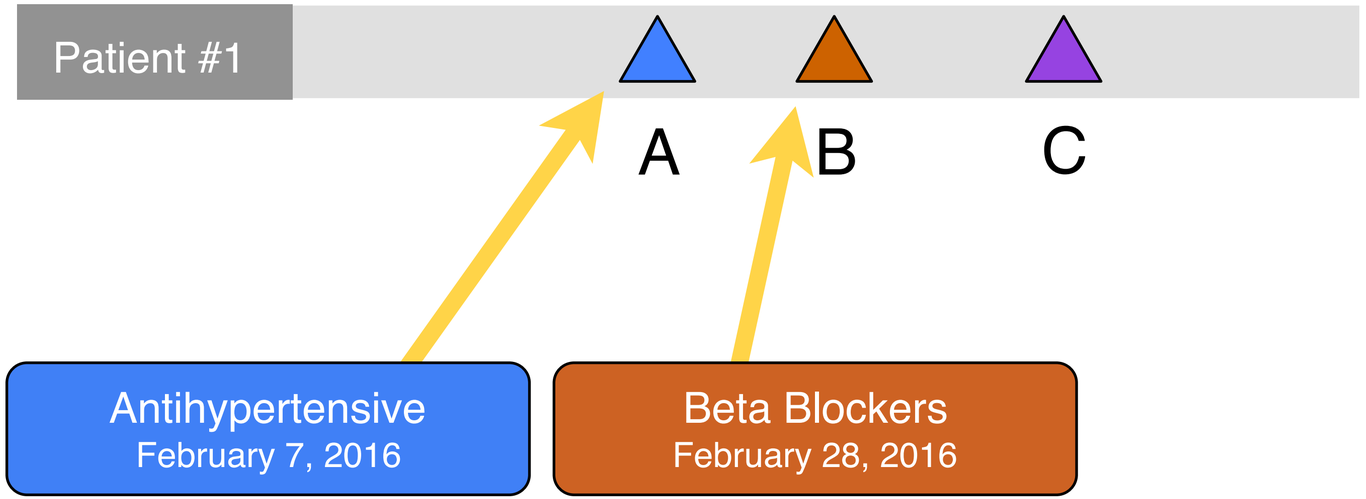
electronic medical records



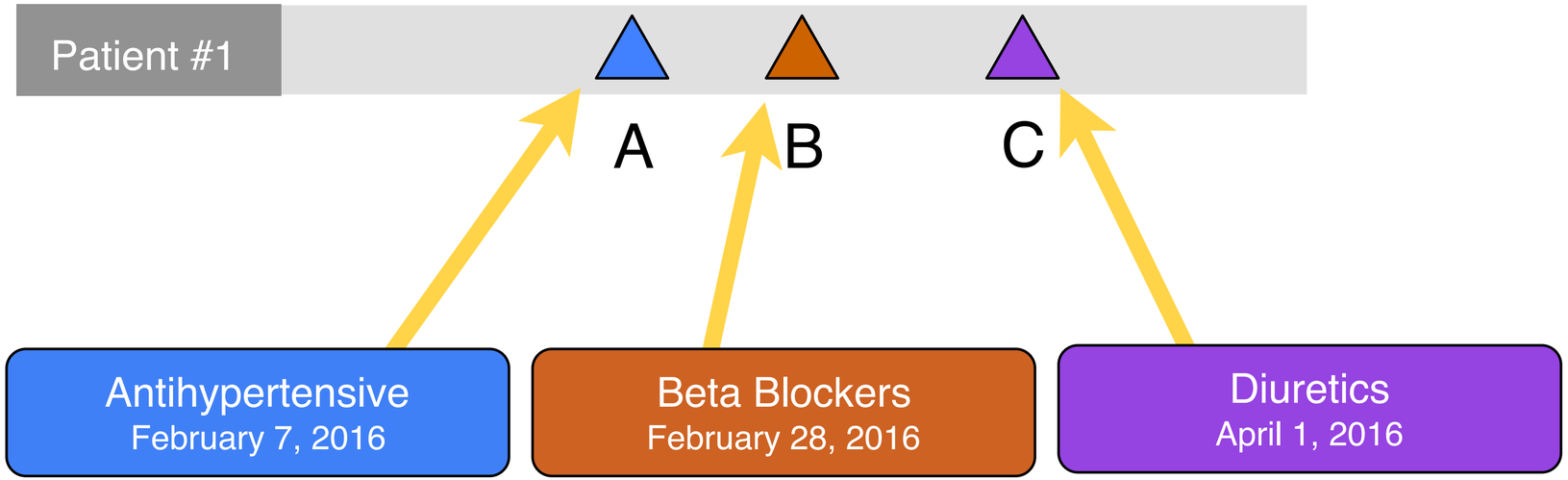
electronic medical records



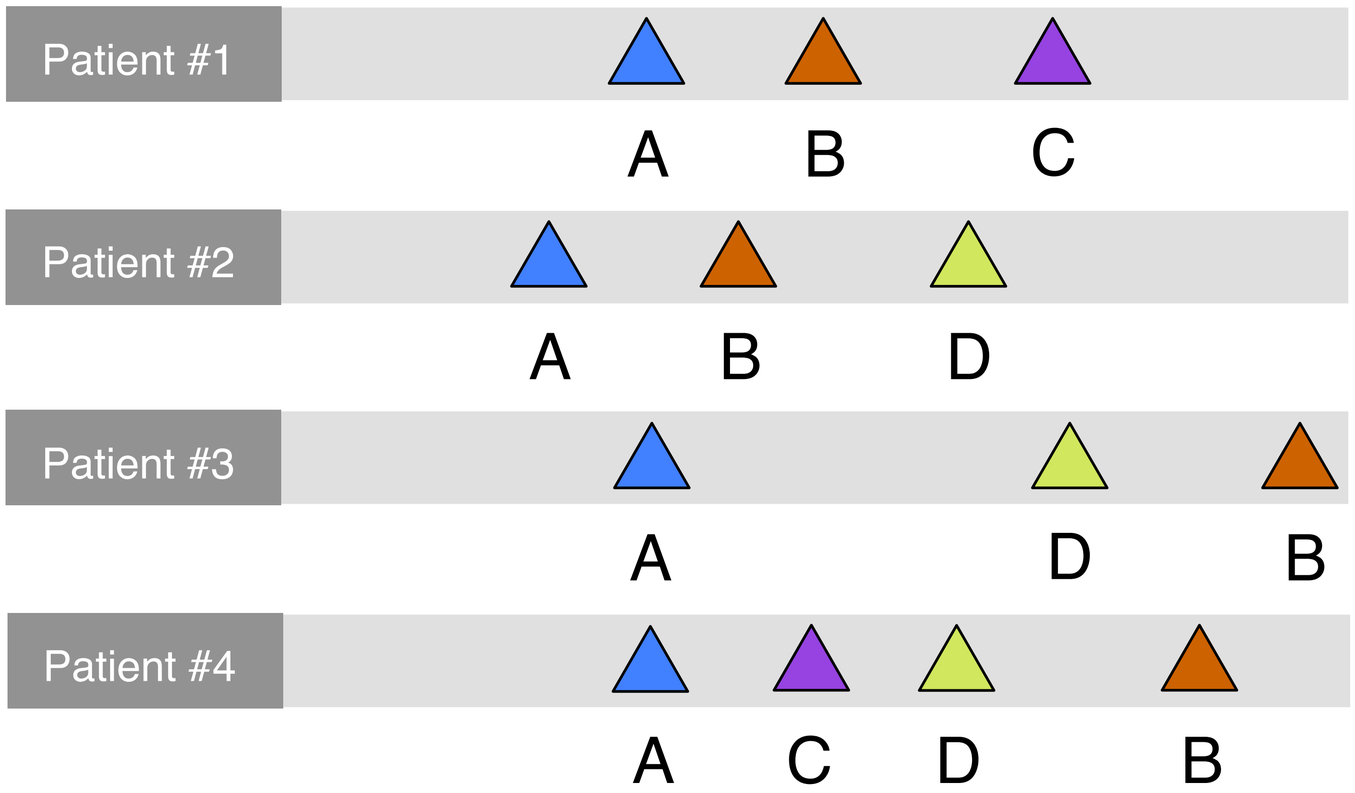
electronic medical records



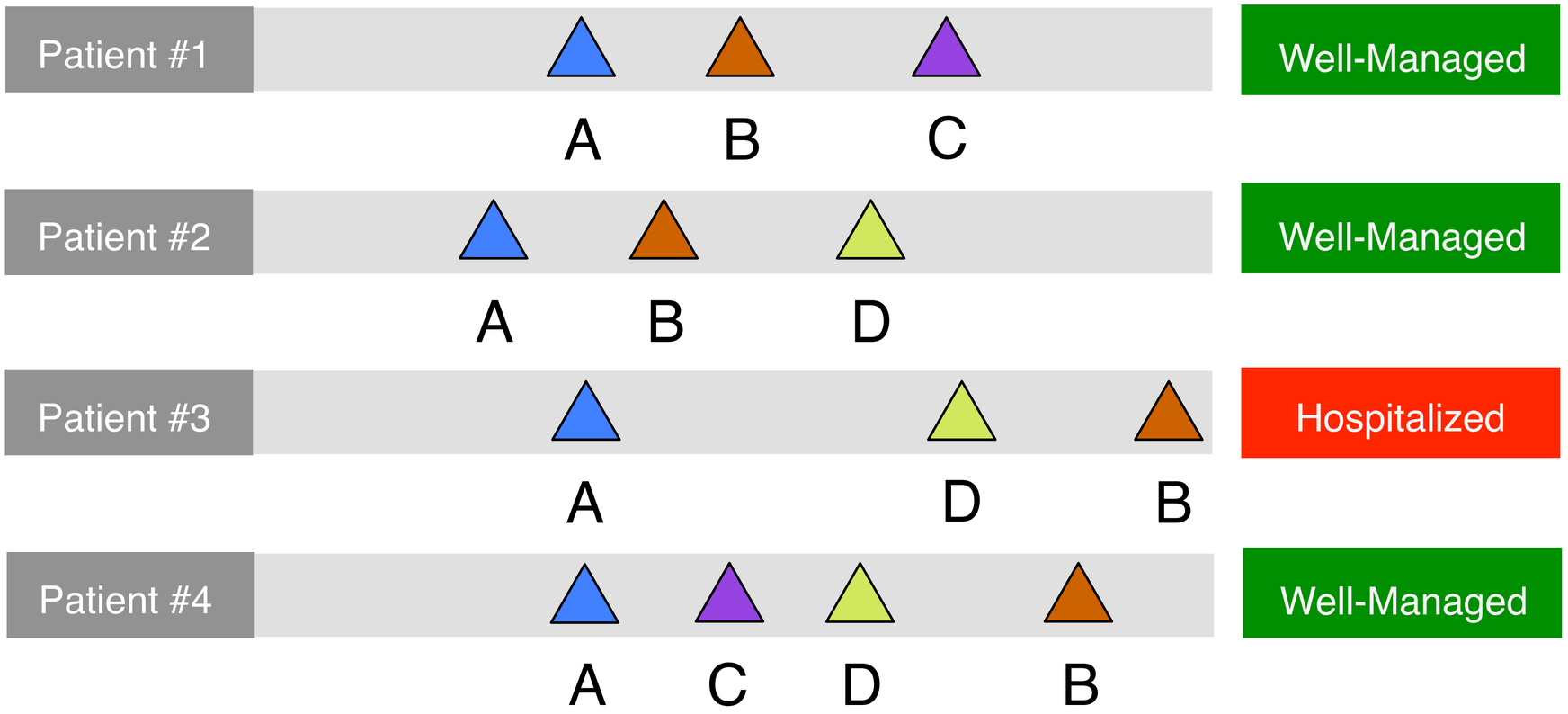
electronic medical records



electronic medical records

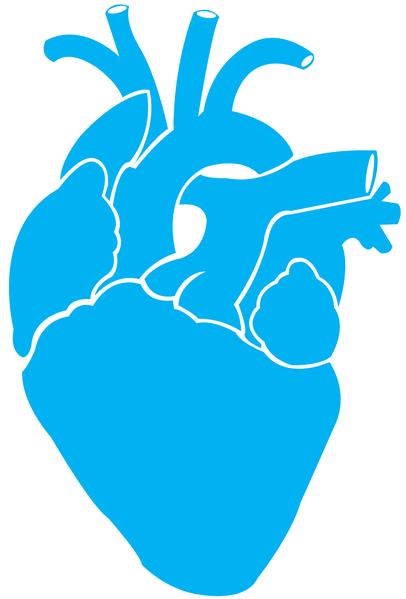


electronic medical records



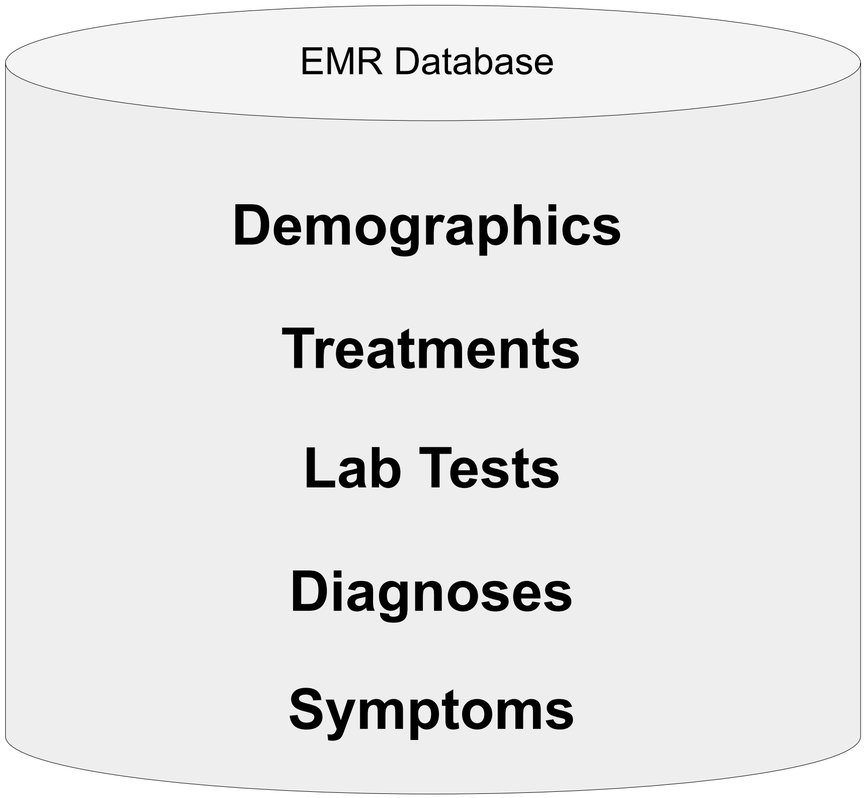
heart failure

- Potentially fatal disease that affects **2%** of adults in developed countries
  - Difficult to manage
  - No systematic clinical guidelines for treating Heart Failure
  - Presence of co-morbidities affects treatment recommendations.



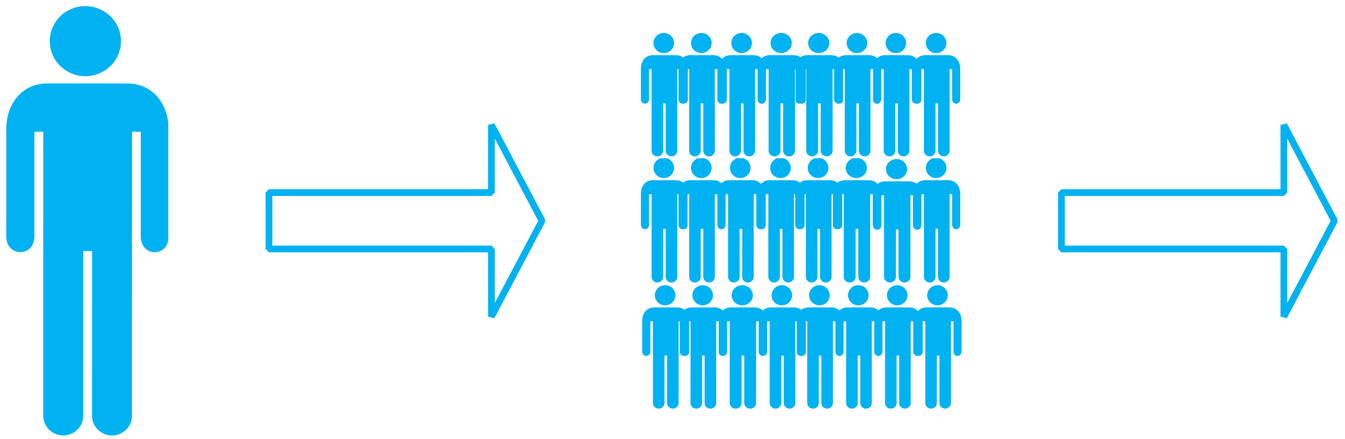
population

- Hundreds of Thousands of Patients diagnosed with Congestive Heart Failure



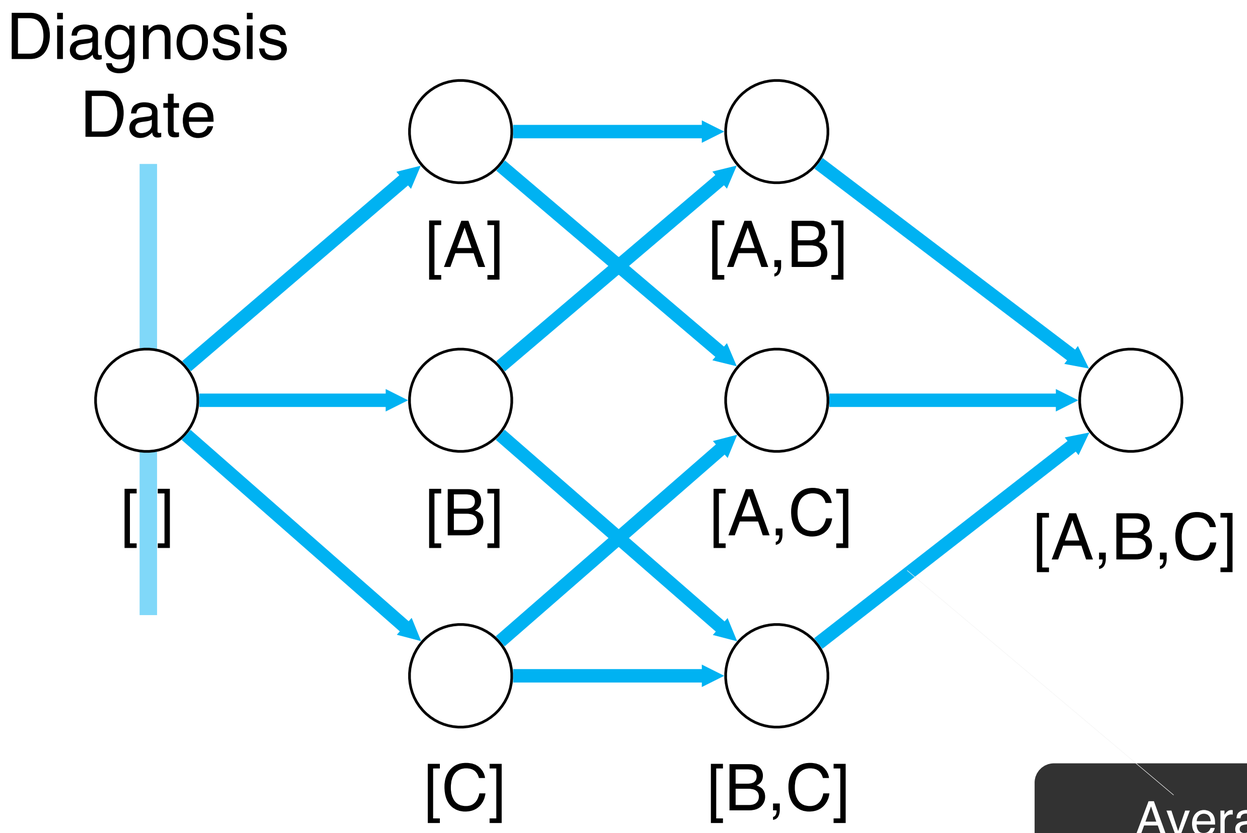
aggregation

- Start with target patient
- Find similar patients
  - Using our similarity analytics on relevant data
    - Features include medications, symptoms, and diagnoses, and lab tests
- Align all patients by disease diagnosis

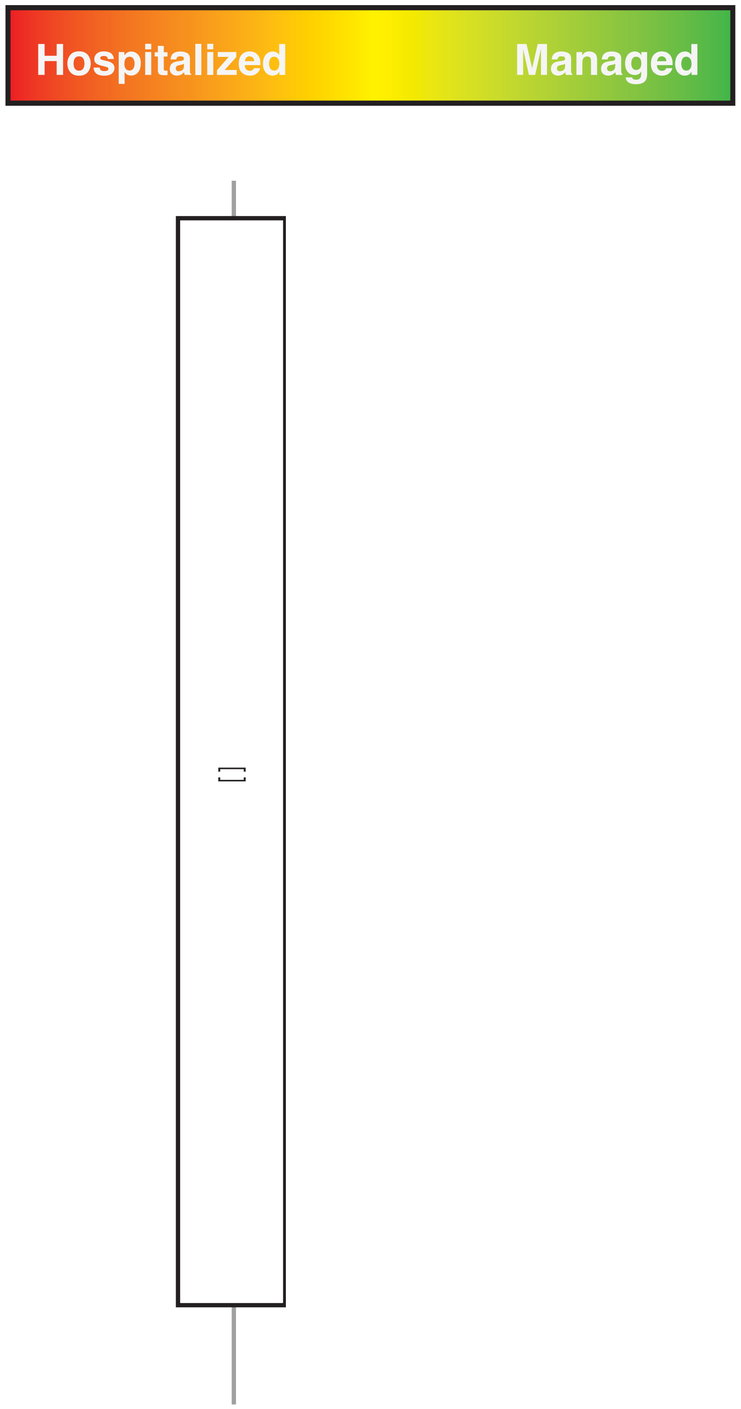


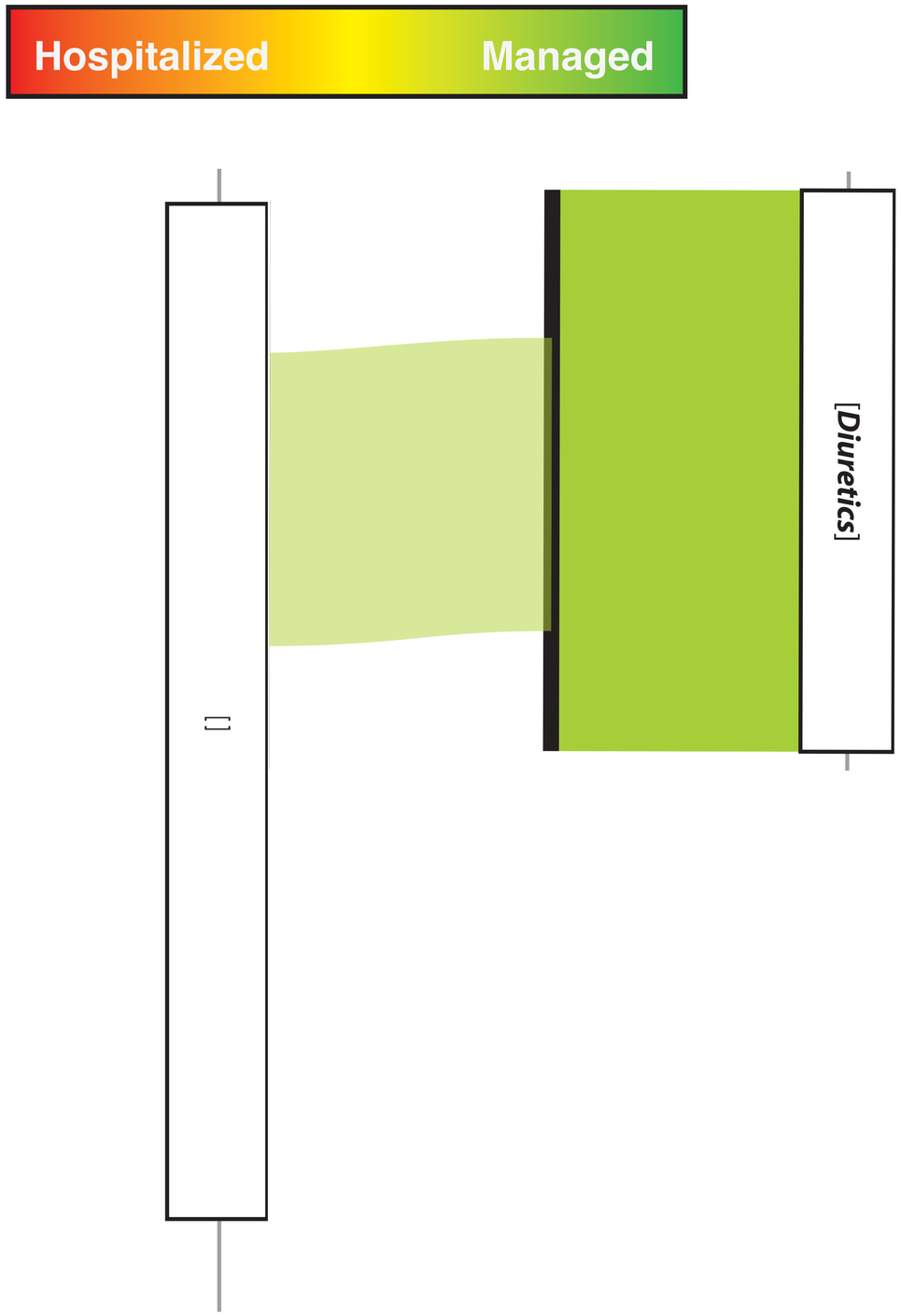
What are the treatment pathways after diagnosis?

aggregation



Average outcome = 0.4
Average time = 10 days
Number of patients = 10


Hospitalized
Managed
[ ]

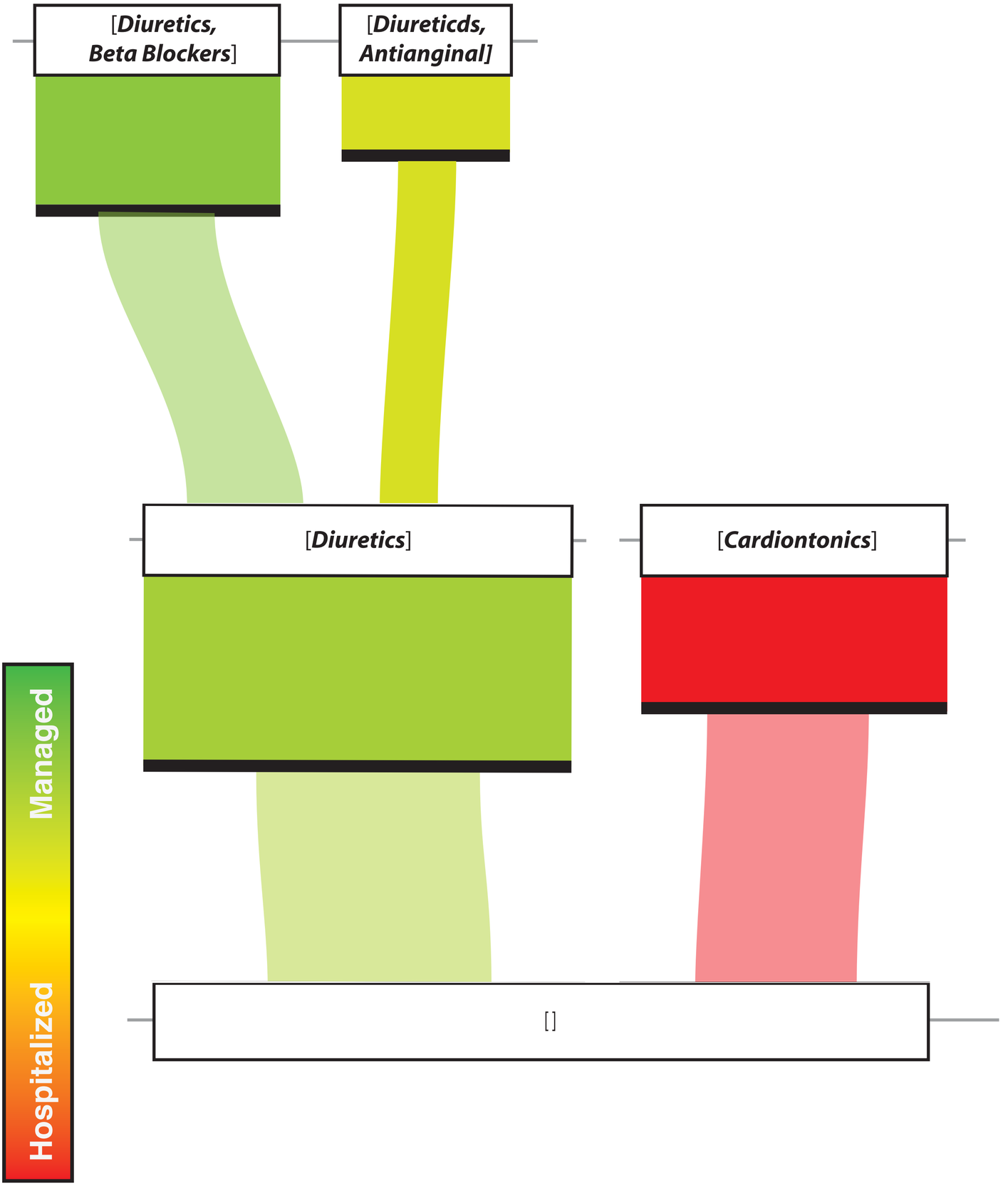

Hospitalized
Managed
[]
[Diuretics]

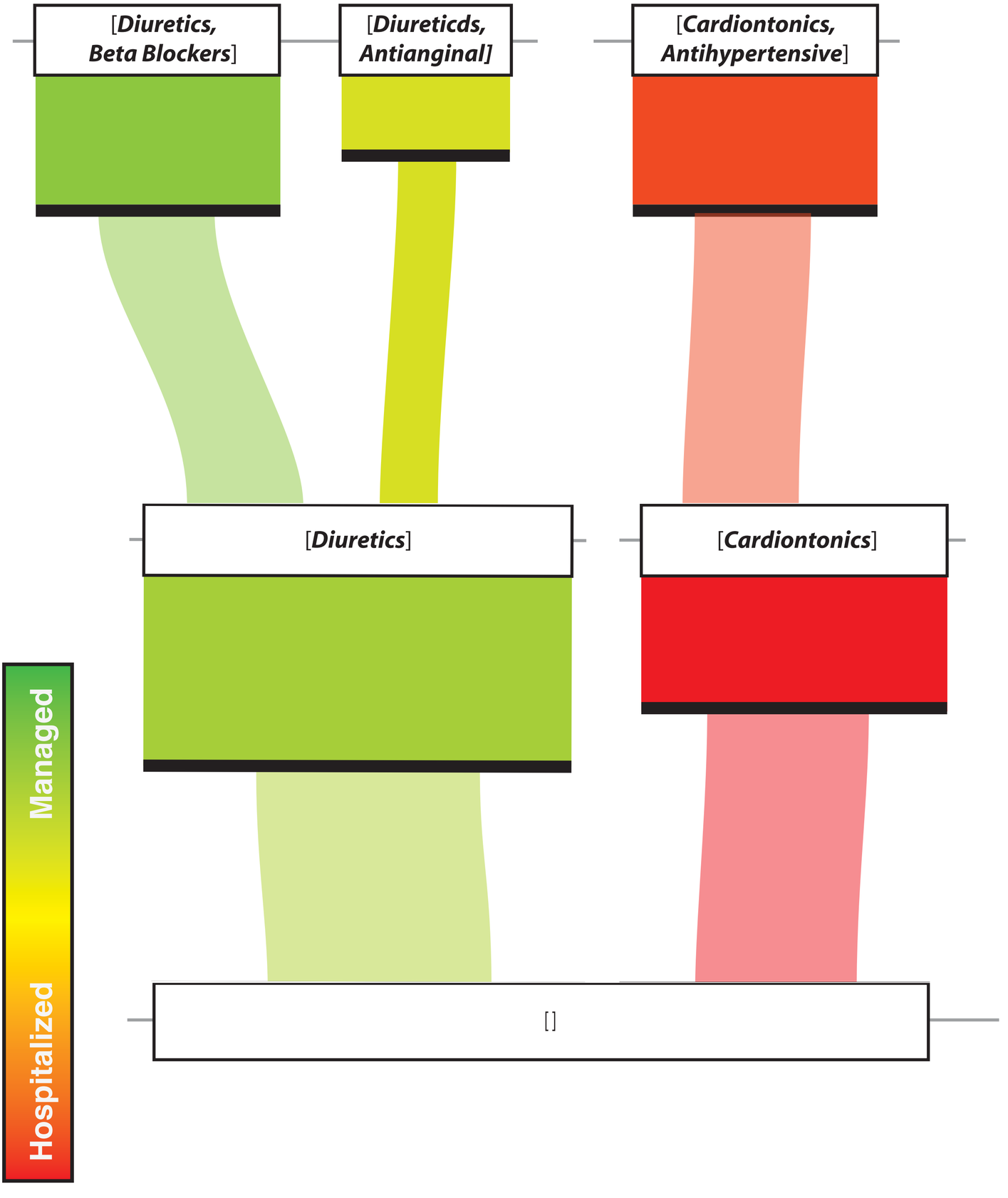
Hospitalized Managed

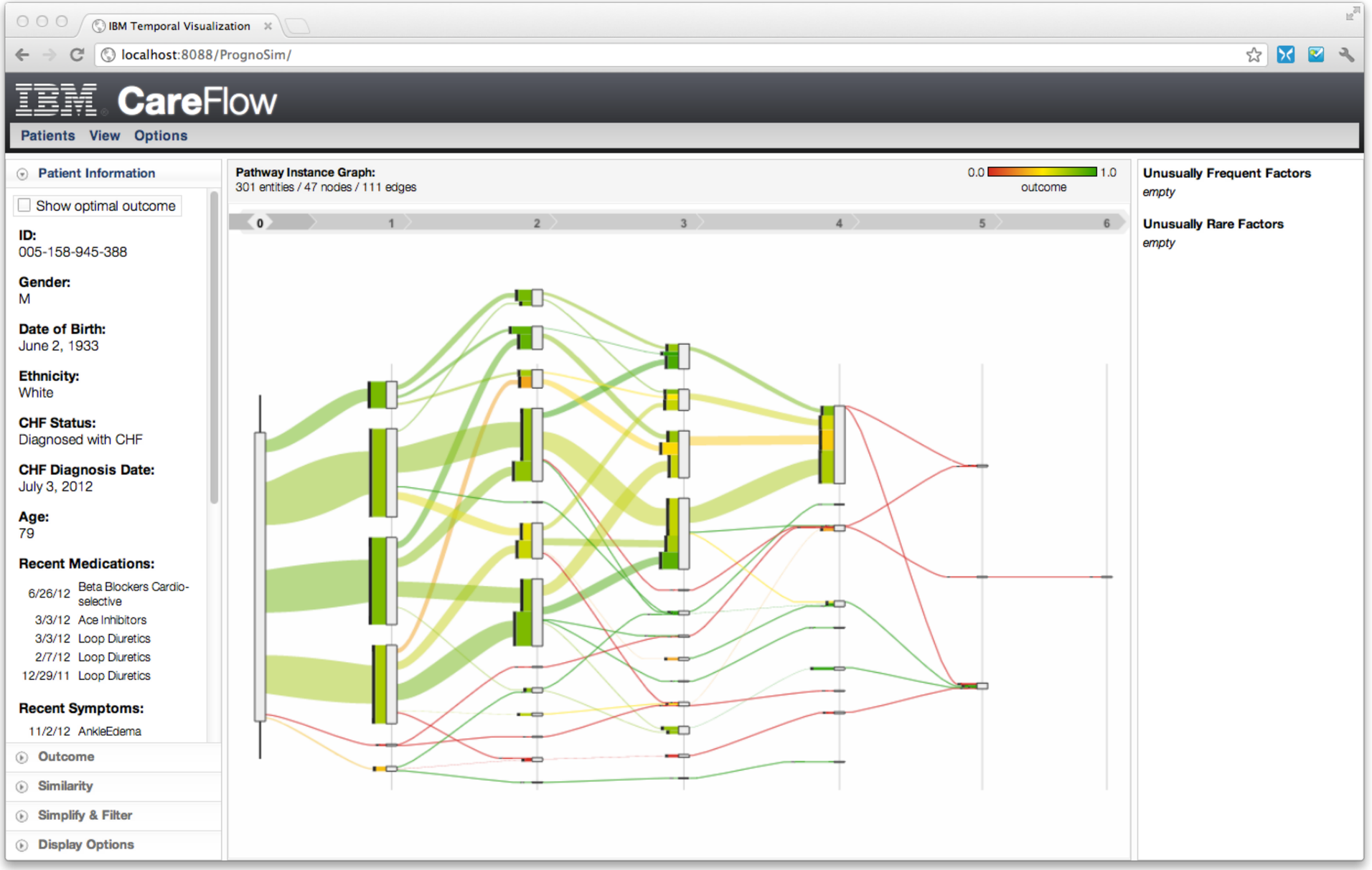
[]

[*Diuretics*]

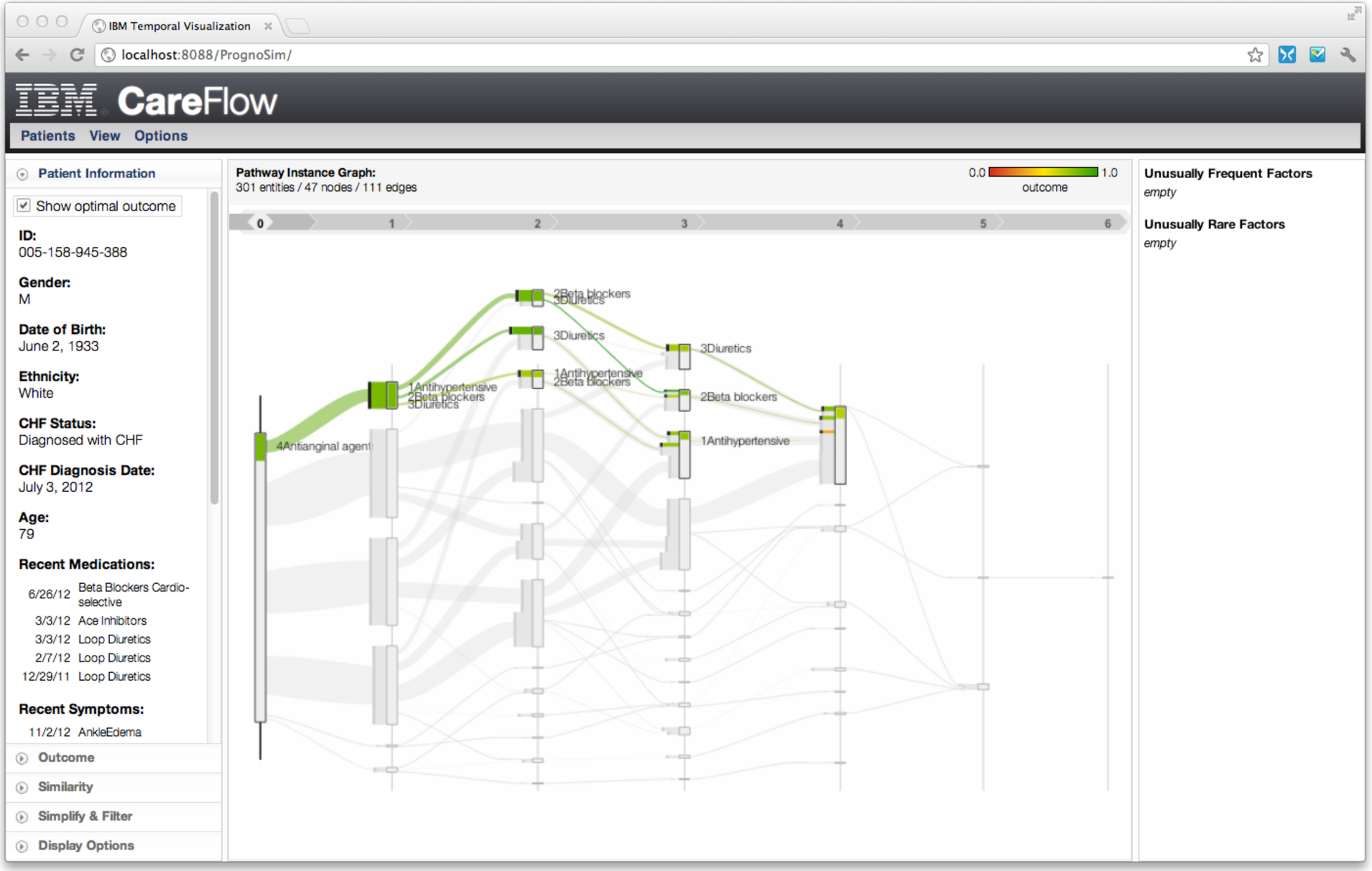
[*Cardiontonics*]


[Diuretics, Beta Blockers]
[Diureticds, Antianginal]
[Diuretics]
[Cardiontonics]
[]
Managed
Hospitalized


[Diuretics, Beta Blockers]
[Diureticds, Antianginal]
[Cardiontonics, Antihypertensive]
[Diuretics]
[Cardiontonics]
[]
Managed
Hospitalized

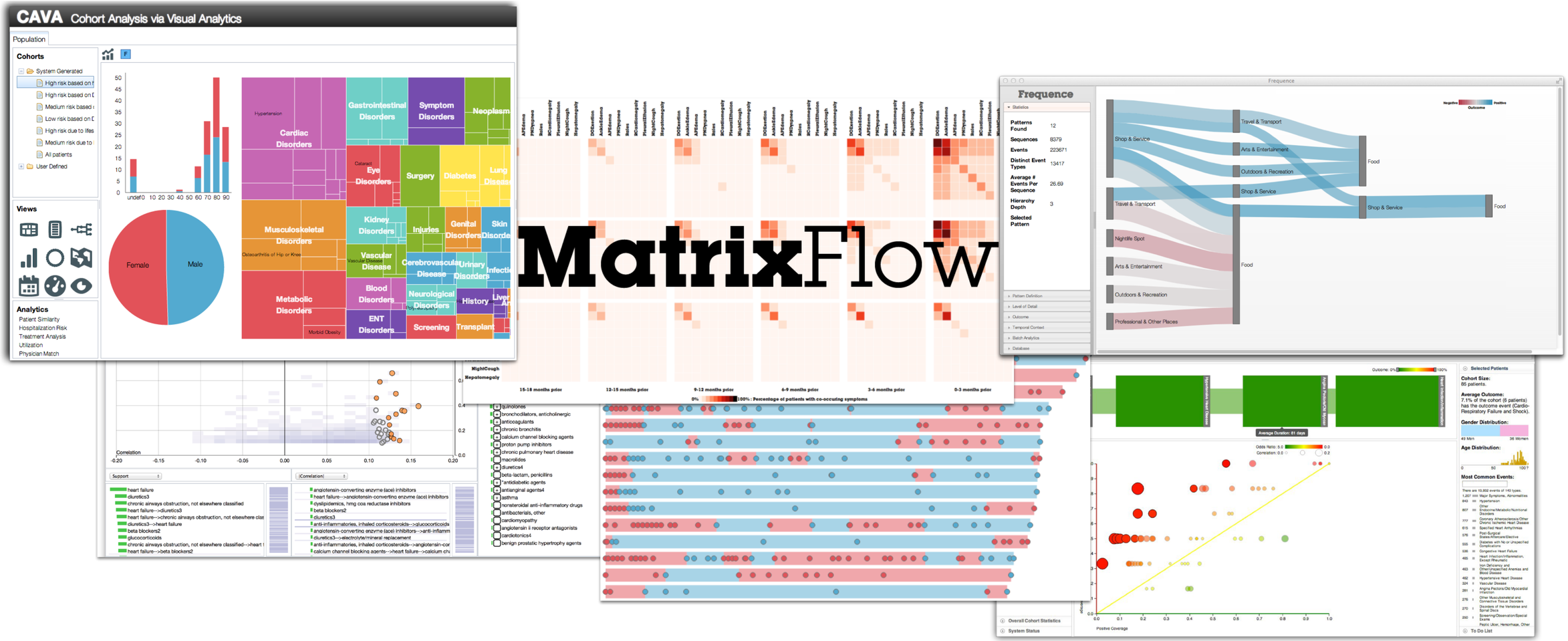


Care Pathways of 300 similar patients



Optimal Care Pathway among 300 similar patients

other tools for clinical exploration



Videos and Papers at **http://perer.org**

clinical researchers:

clinical researchers:

clinical researchers:

clinical researchers:

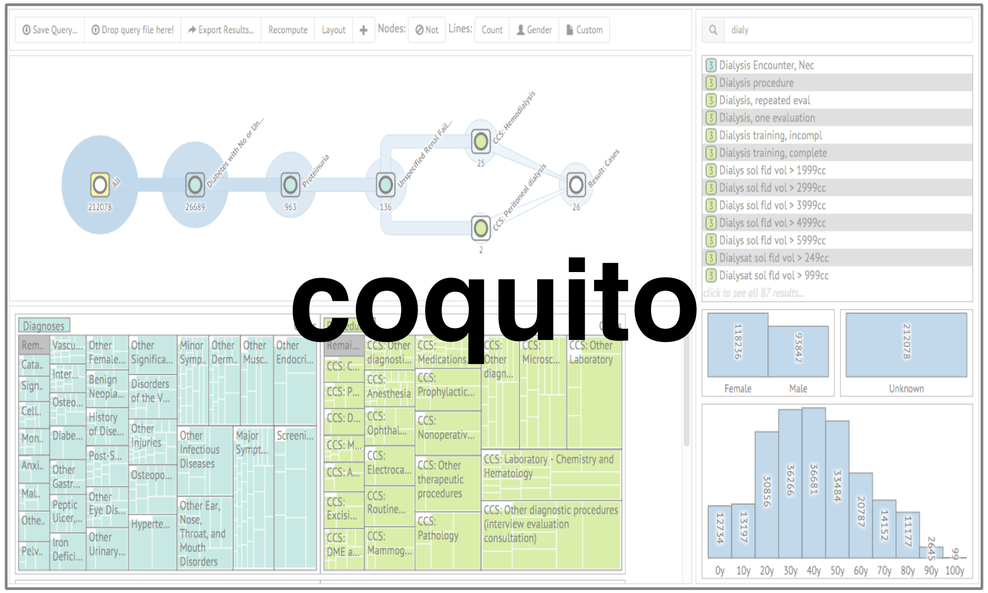
clinical researchers:

the **role**
of **visualization**
in **prediction**

what can visualization do?

START

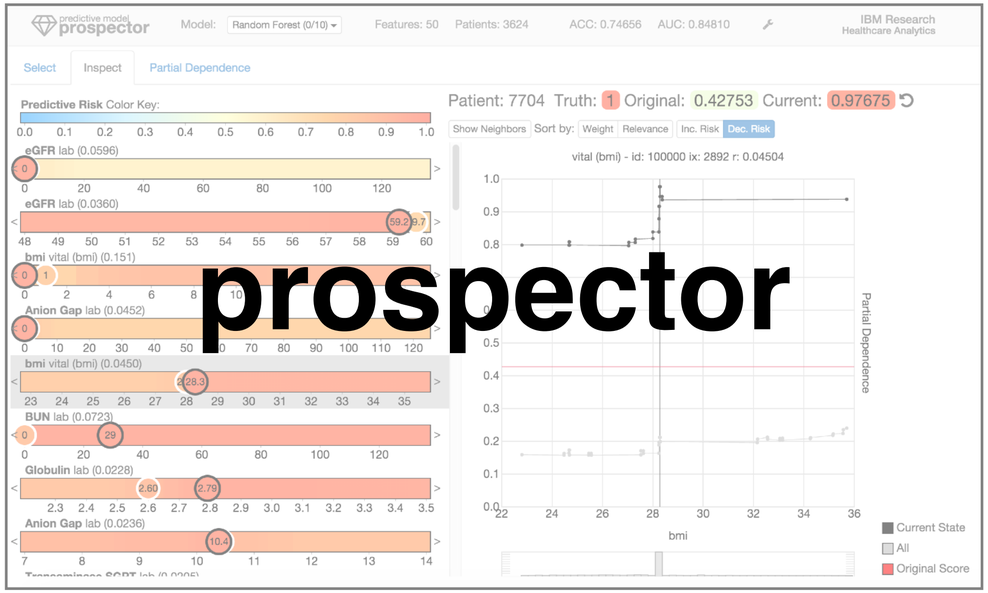
Cohort Definition



coquito

END

Model Interpretability

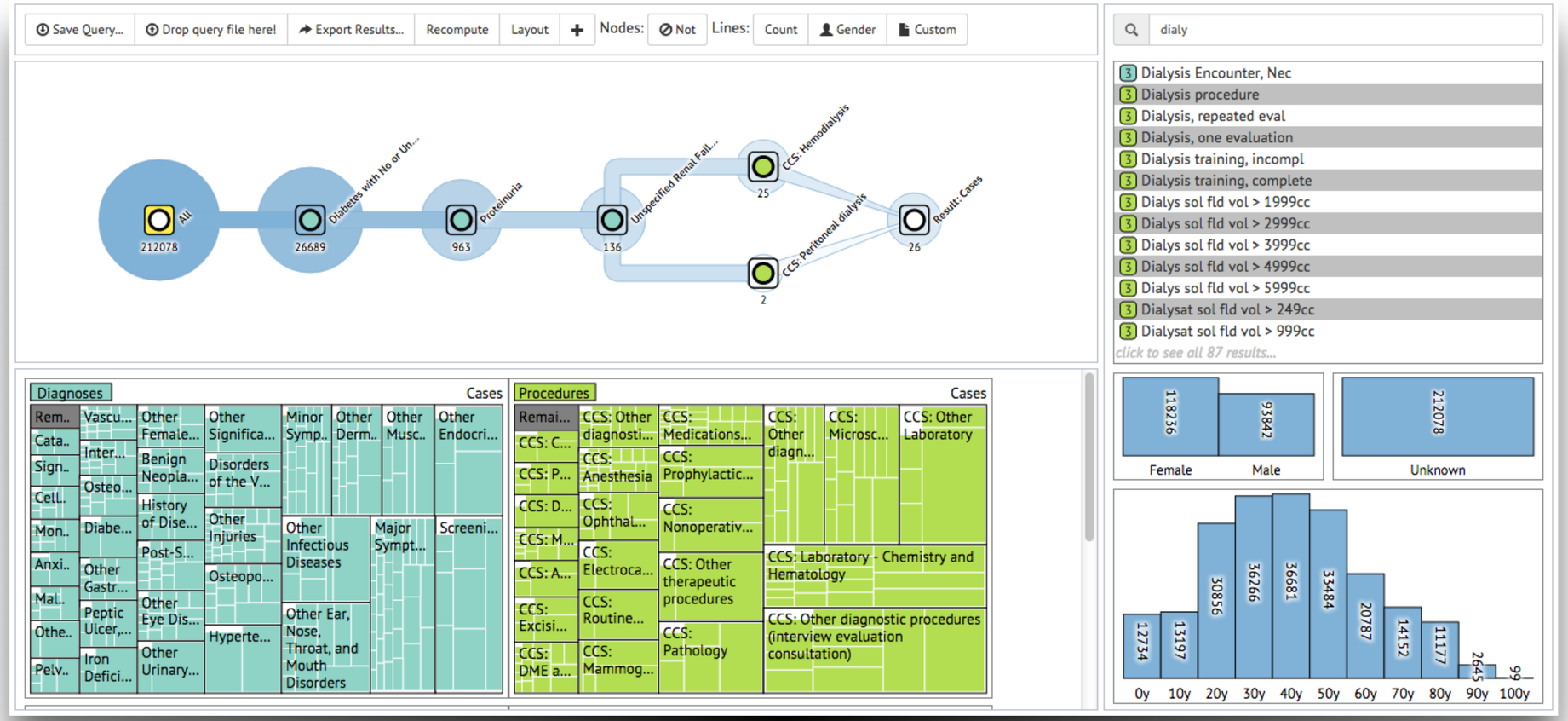


prospector

1 **Cohort Construction**

coquito

cohort **qu**eries with **i**terative **o**verviews



Josua Krause, **Adam Perer**, and Harry Stavropolous. Supporting Iterative Cohort Construction with Visual Temporal Queries. IEEE Visual Analytics Science and Technology (VAST 2015).

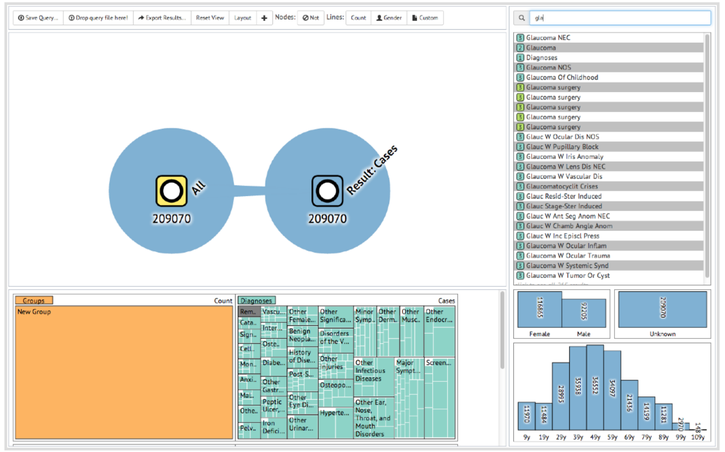
defining cohorts

- Typically, defining cohorts is a slow process:
  - First, medical researchers define requirements.
  - Then, Technologists write SQL queries and deliver them to medical researchers.
  - But, often too many patients or too few patients, and the process must restart.

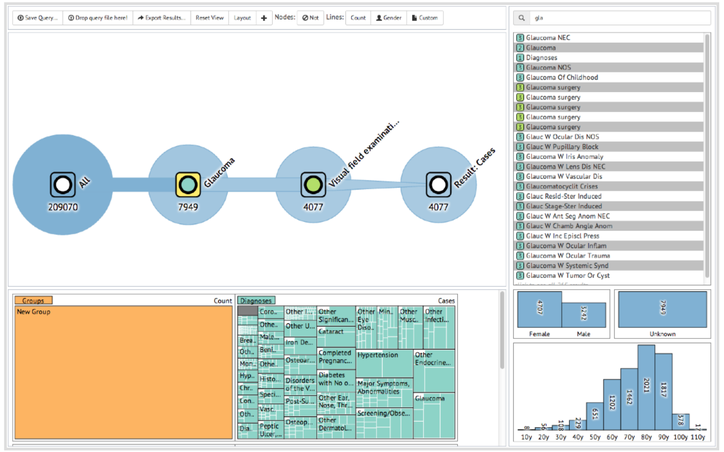
```
215862, 1123300800000), (232190, 1112932800000), (237154, 1113105600000), (248873, 1117252800000), (249428, 1092196800000), (250877, 1112331600000), (251852, 1116
269990, 1109048400000), (283194, 1051070400000), (297322, 1123646400000), (304412, 1119067200000)) as tmp (pid, from_time) where table.level_1_id is not null  and
level_2_id is not null  and level_3_id is not null  and level_4_id is not null  and table.gender_description in ('Female','Male') and table.ethnic_group in ('Unkn
table.actual_day > 1115956800000 and tmp.pid = table.patient_id and table.actual_day > tmp.from_time) as res group by res.level_1_id, res.level_2_id, res.level_3_
level_4_id

select count(*), res.level_1_id, res.level_2_id, res.level_3_id, res.level_4_id from (select distinct table.patient_id, table.level_1_id, table.level_2_id, table.
table.level_4_id from visual_model_creation_workbench.hierarchies_dates_demographics_sho as table, (values (181, 1030507200000), (381, 1116388800000), (1111, 1056
2057, 1033704000000), (2197, 1049256000000), (2689, 1039064400000), (3617, 1071550800000), (3962, 1120795200000), (5752, 1070686800000), (6448, 1046840400000), (6
1089432000000), (7785, 1102914000000), (7881, 1080709200000), (7883, 1056427200000), (8245, 1080882000000), (8561, 1075352400000), (8877, 1091592000000), (9081, 1
, (9092, 1073106000000), (9413, 1084680000000), (11480, 1118203200000), (12340, 1112932800000), (13146, 1063339200000), (15220, 1116820800000), (16851, 1124337600
1054180800000), (19207, 1074402000000), (19330, 1031284800000), (19880, 1050724800000), (19937, 1124942400000), (20838, 1046754000000), (21927, 1119931200000), (2
1081915200000), (22270, 1045976400000), (23595, 1121140800000), (23651, 1047531600000), (24554, 1048050000000), (25251, 1056600000000), (26443, 1106888400000), (2
1052712000000), (29043, 1083643200000), (30803, 1116734400000), (31693, 1095307200000), (32170, 1092888000000), (32777, 1048741200000), (34373, 1116216000000), (3
```

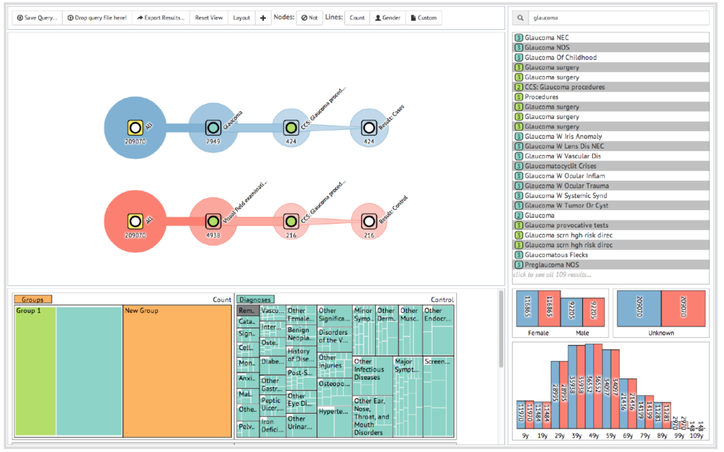
defining cohorts with coquito



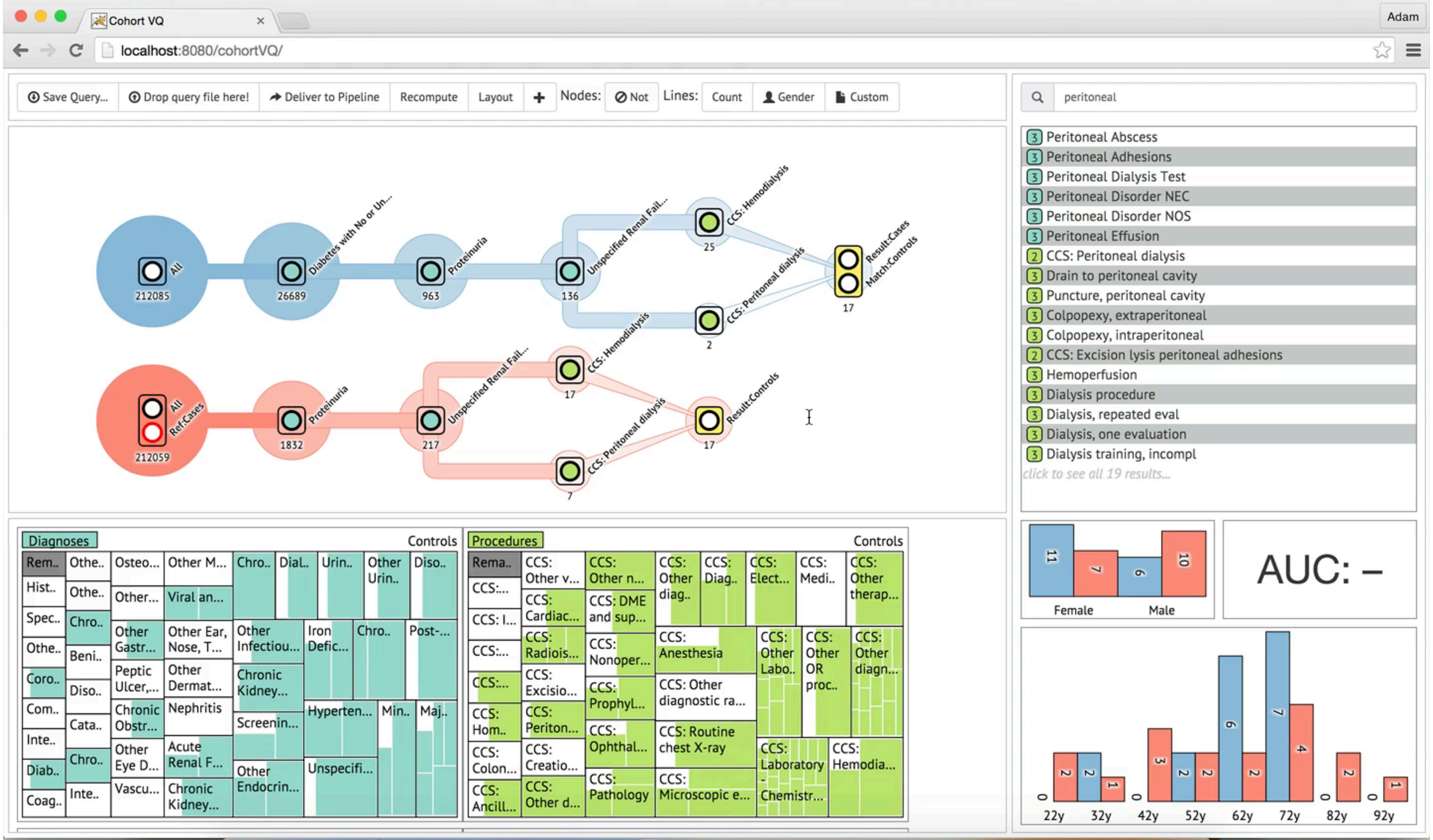
drag and drop constraints with immediate feedback and hints for query refinement

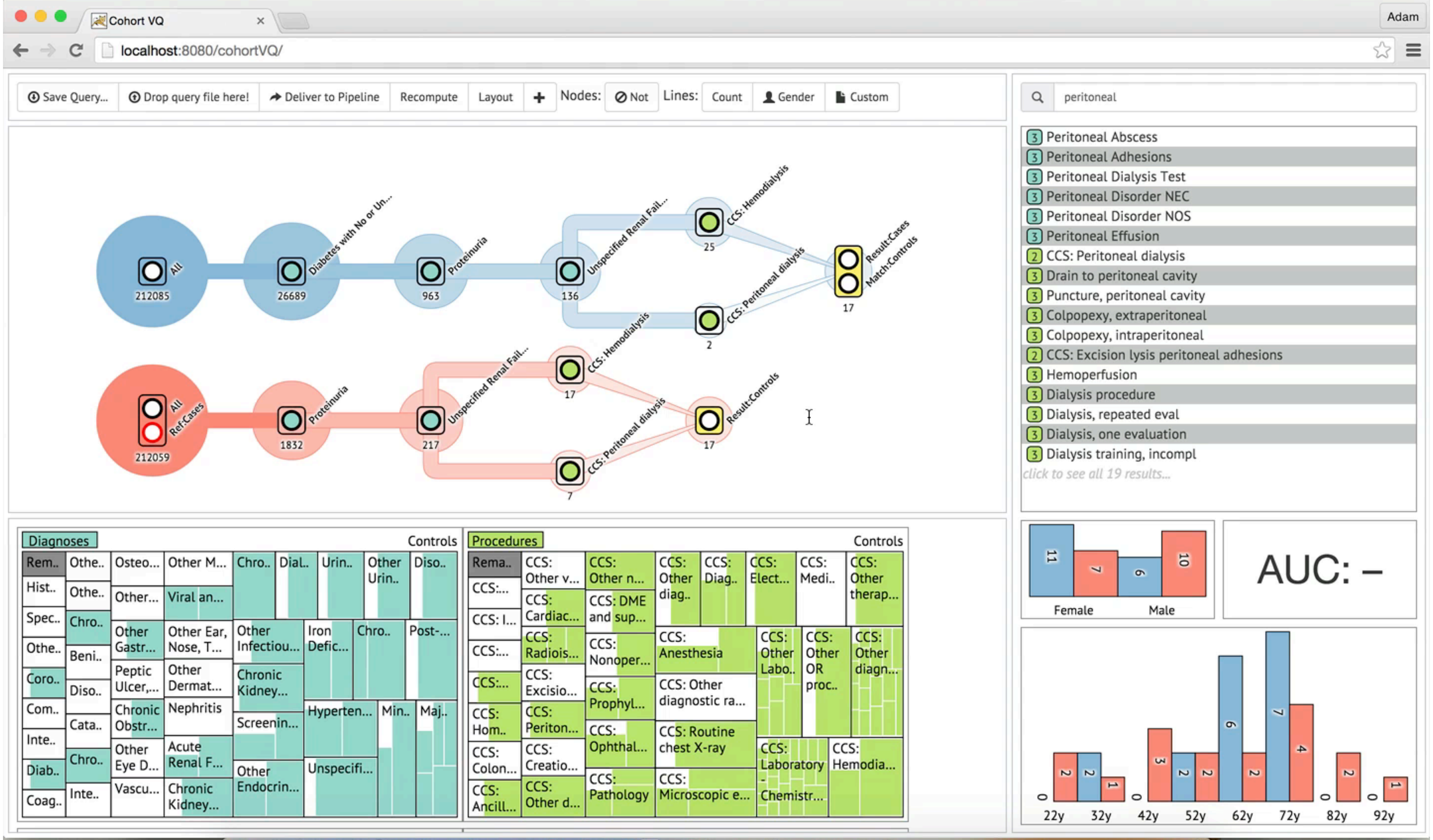


supports using complex temporal logic



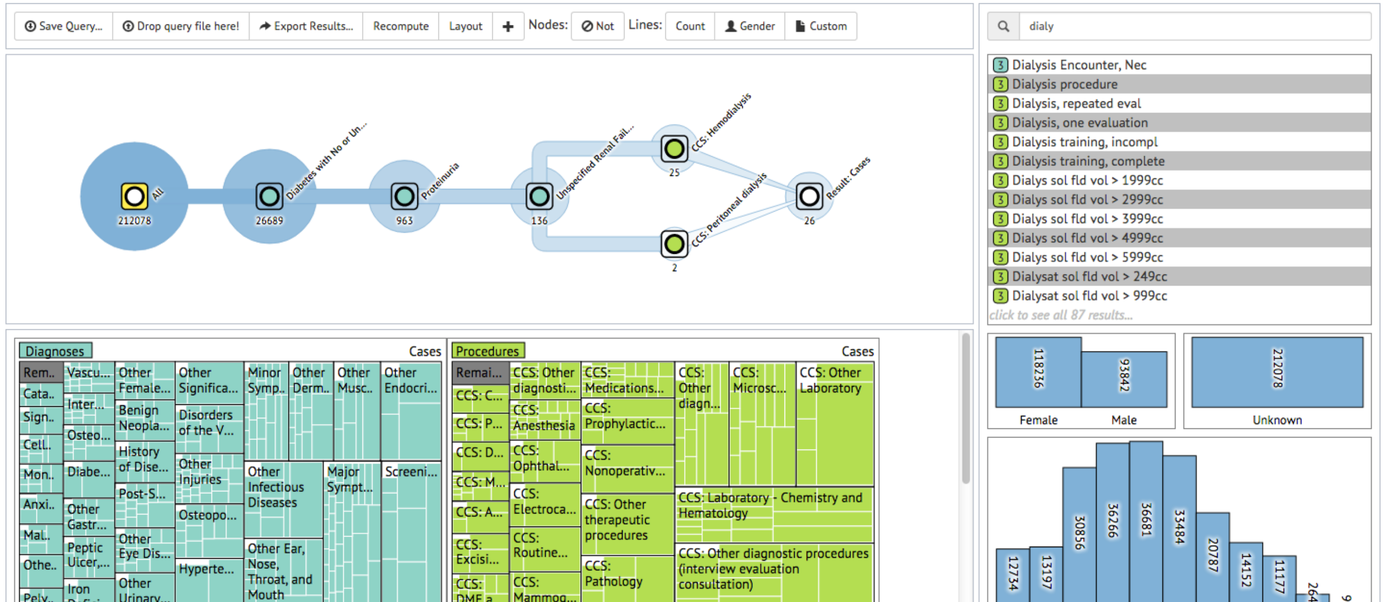
support for multiple queries side-by-side (for cases and controls)


Cohort VQ
Adam
localhost:8080/cohortVQ/
Save Query...
Drop query file here!
Deliver to Pipeline
Recompute
Layout
Nodes:
Not
Lines:
Count
Gender
Custom
peritoneal
All
212085
Diabetes with No or Un...
26689
Proteinuria
963
Unspecified Renal Fail...
136
CCS: Hemodialysis
25
CCS: Peritoneal dialysis
2
Result:Cases Match:Controls
17
All Ref:Cases
212059
Proteinuria
1832
Unspecified Renal Fail...
217
CCS: Hemodialysis
17
CCS: Peritoneal dialysis
7
Result:Controls
17
Peritoneal Abscess
Peritoneal Adhesions
Peritoneal Dialysis Test
Peritoneal Disorder NEC
Peritoneal Disorder NOS
Peritoneal Effusion
CCS: Peritoneal dialysis
Drain to peritoneal cavity
Puncture, peritoneal cavity
Colpopexy, extraperitoneal
Colpopexy, intraperitoneal
CCS: Excision lysis peritoneal adhesions
Hemoperfusion
Dialysis procedure
Dialysis, repeated eval
Dialysis, one evaluation
Dialysis training, incompl
click to see all 19 results...
Diagnoses
Controls
Procedures
Controls
11
7
6
10
Female
Male
AUC: –
22y
32y
42y
52y
62y
72y
82y
92y


Cohort VQ
Adam
localhost:8080/cohortVQ/
Save Query...
Drop query file here!
Deliver to Pipeline
Recompute
Layout
Nodes:
Not
Lines:
Count
Gender
Custom
peritoneal
All
212085
Diabetes with No or Un...
26689
Proteinuria
963
Unspecified Renal Fail...
136
CCS: Hemodialysis
25
CCS: Peritoneal dialysis
2
Result:Cases Match:Controls
17
All Ref:Cases
212059
Proteinuria
1832
Unspecified Renal Fail...
217
CCS: Hemodialysis
17
CCS: Peritoneal dialysis
7
Result:Controls
17
Peritoneal Abscess
Peritoneal Adhesions
Peritoneal Dialysis Test
Peritoneal Disorder NEC
Peritoneal Disorder NOS
Peritoneal Effusion
CCS: Peritoneal dialysis
Drain to peritoneal cavity
Puncture, peritoneal cavity
Colpopexy, extraperitoneal
Colpopexy, intraperitoneal
CCS: Excision lysis peritoneal adhesions
Hemoperfusion
Dialysis procedure
Dialysis, repeated eval
Dialysis, one evaluation
Dialysis training, incompl
click to see all 19 results...
Diagnoses
Controls
Procedures
Controls
11
7
6
10
Female
Male
AUC: –
22y
32y
42y
52y
62y
72y
82y
92y

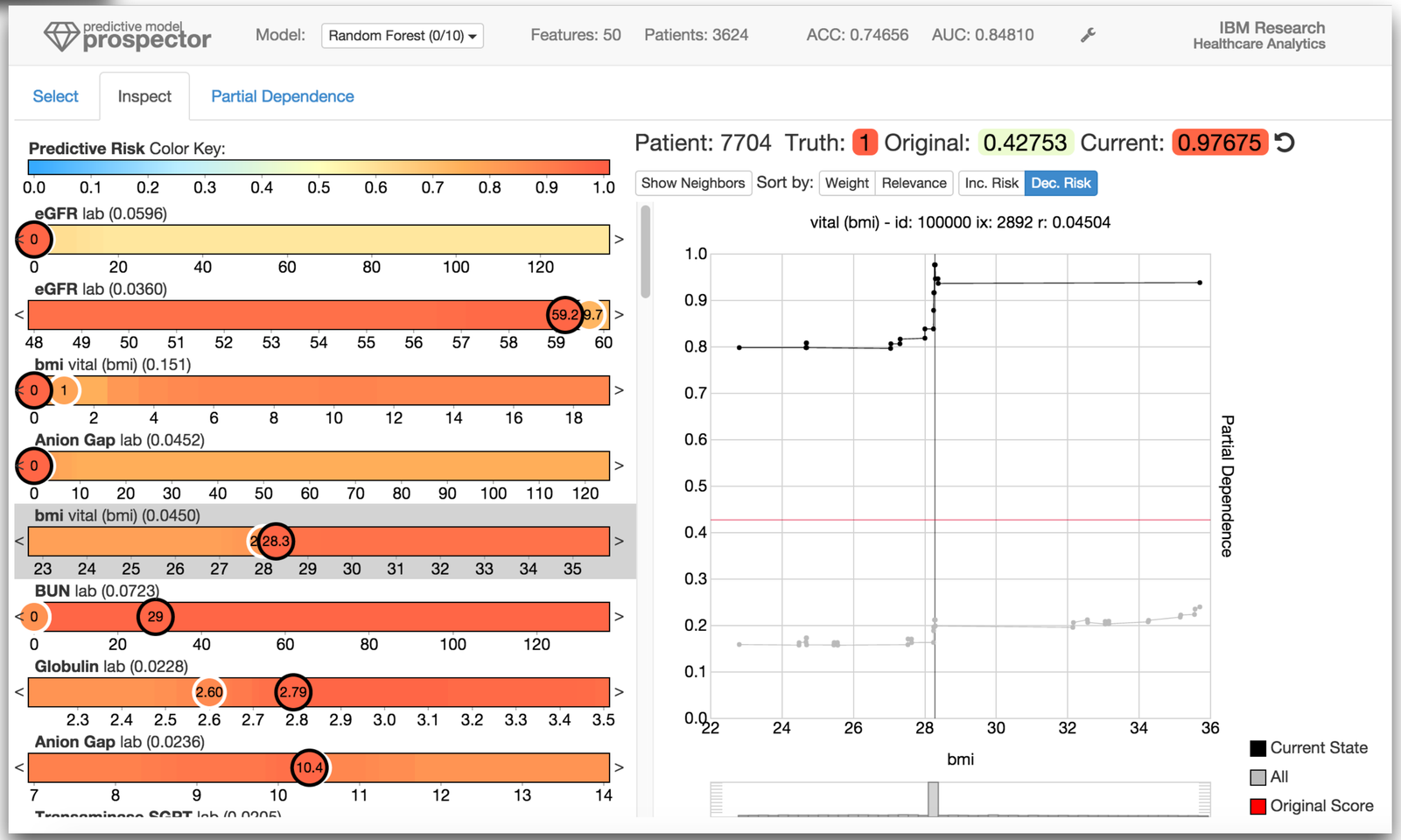
coquito lessons

- Easy and interactive query formulation lets domain experts explore the data
- Visible intermediate results provide critical feedback
- Hints for query refinements are helpful in improving queries



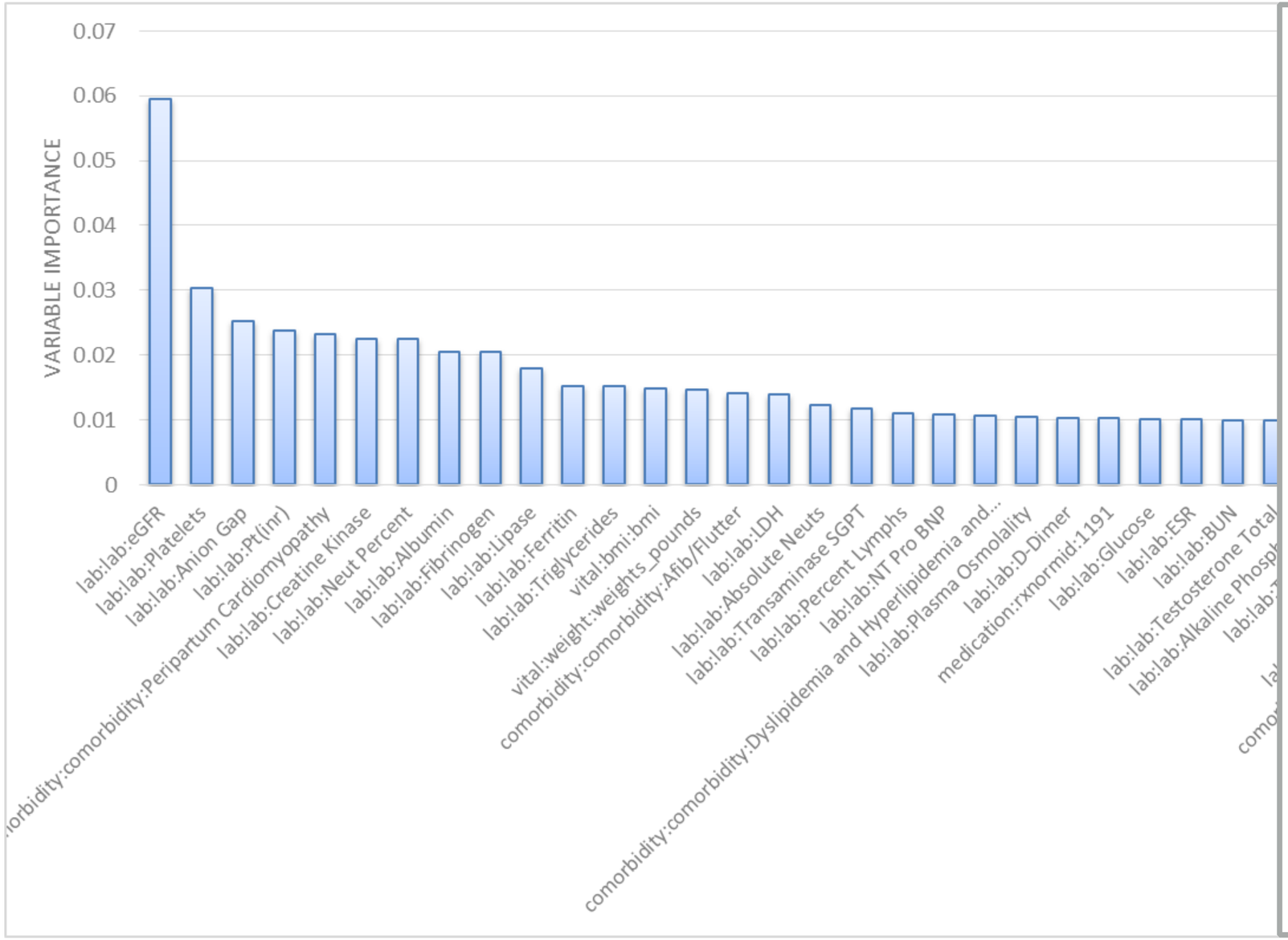
Model Interpretability

predictive model prospector



Josua Krause, Adam Perer, and Kenney Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. ACM Conference on Human Factors in Computing Systems (CHI 2016). San Jose, California. (2016).

typical predictive model report



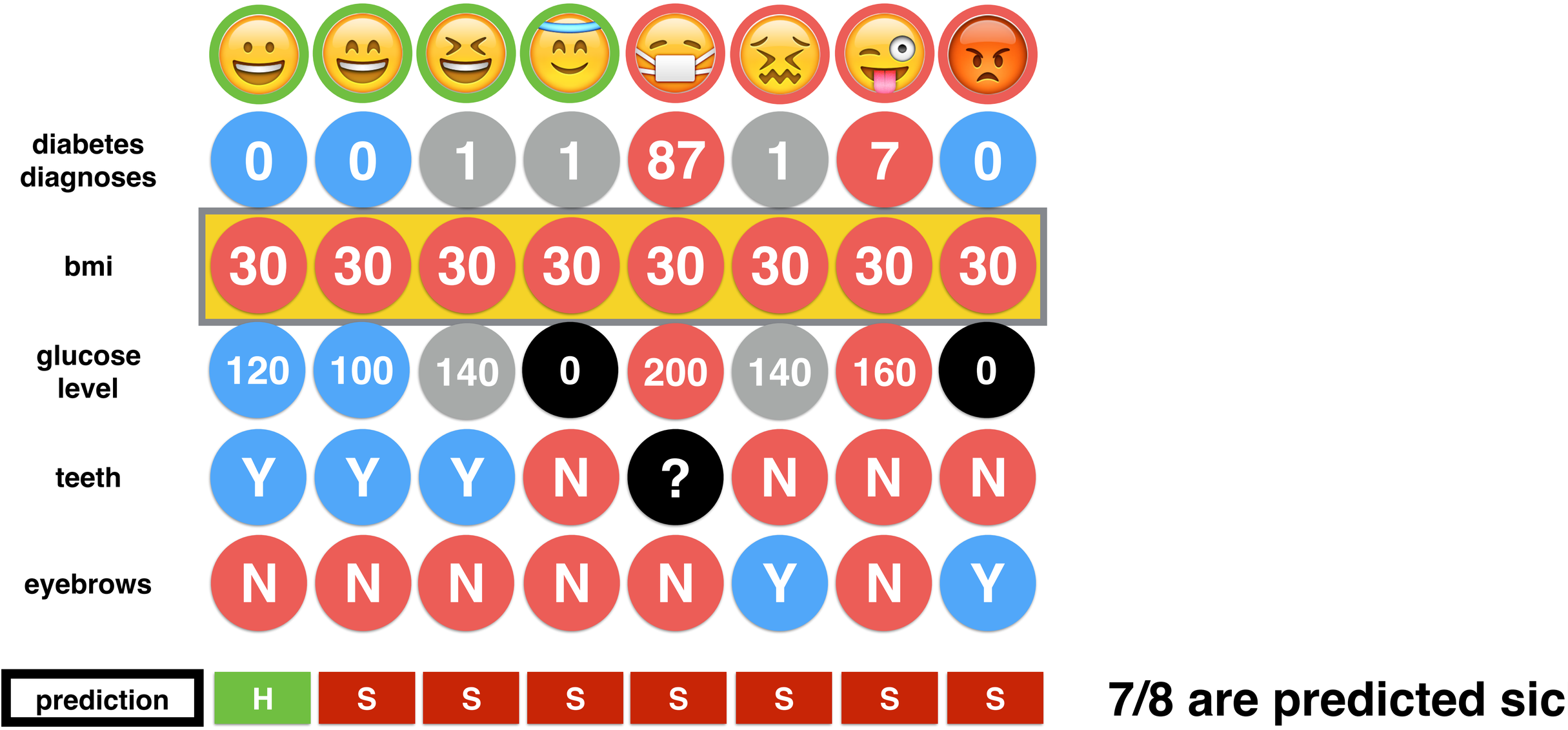
Typically simply a list of top features and their weights

Why?
Difficult to summarize complex models

Issues
One cannot interpret how the values of each feature impact the prediction

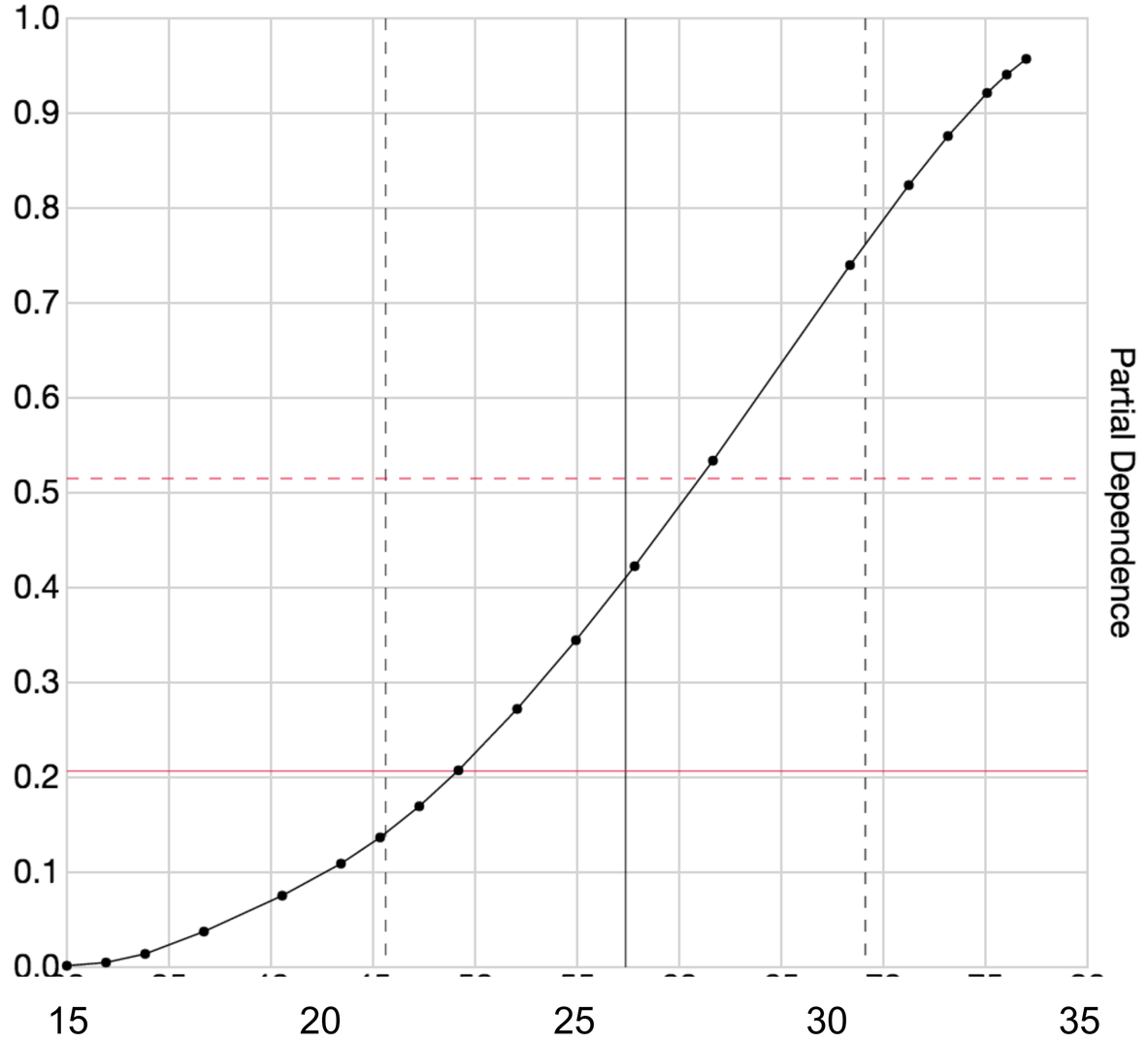
One cannot interact with the model to test hypotheses

partial dependence

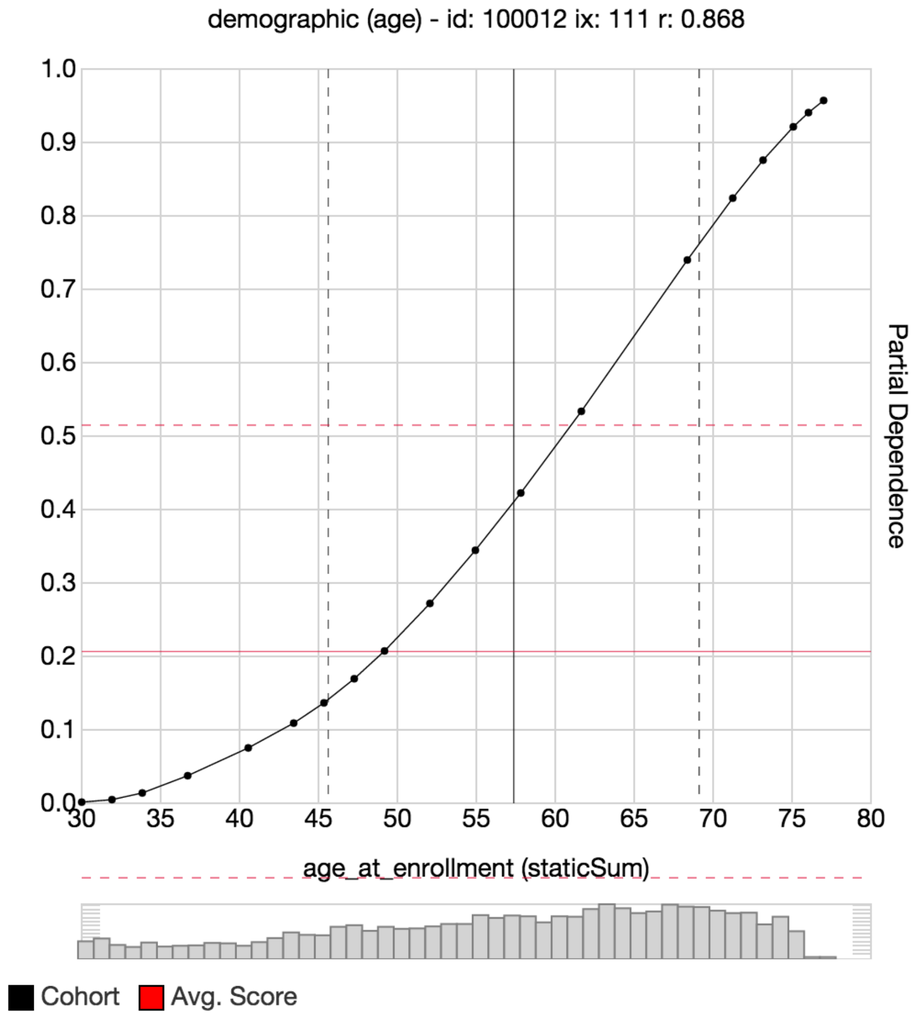


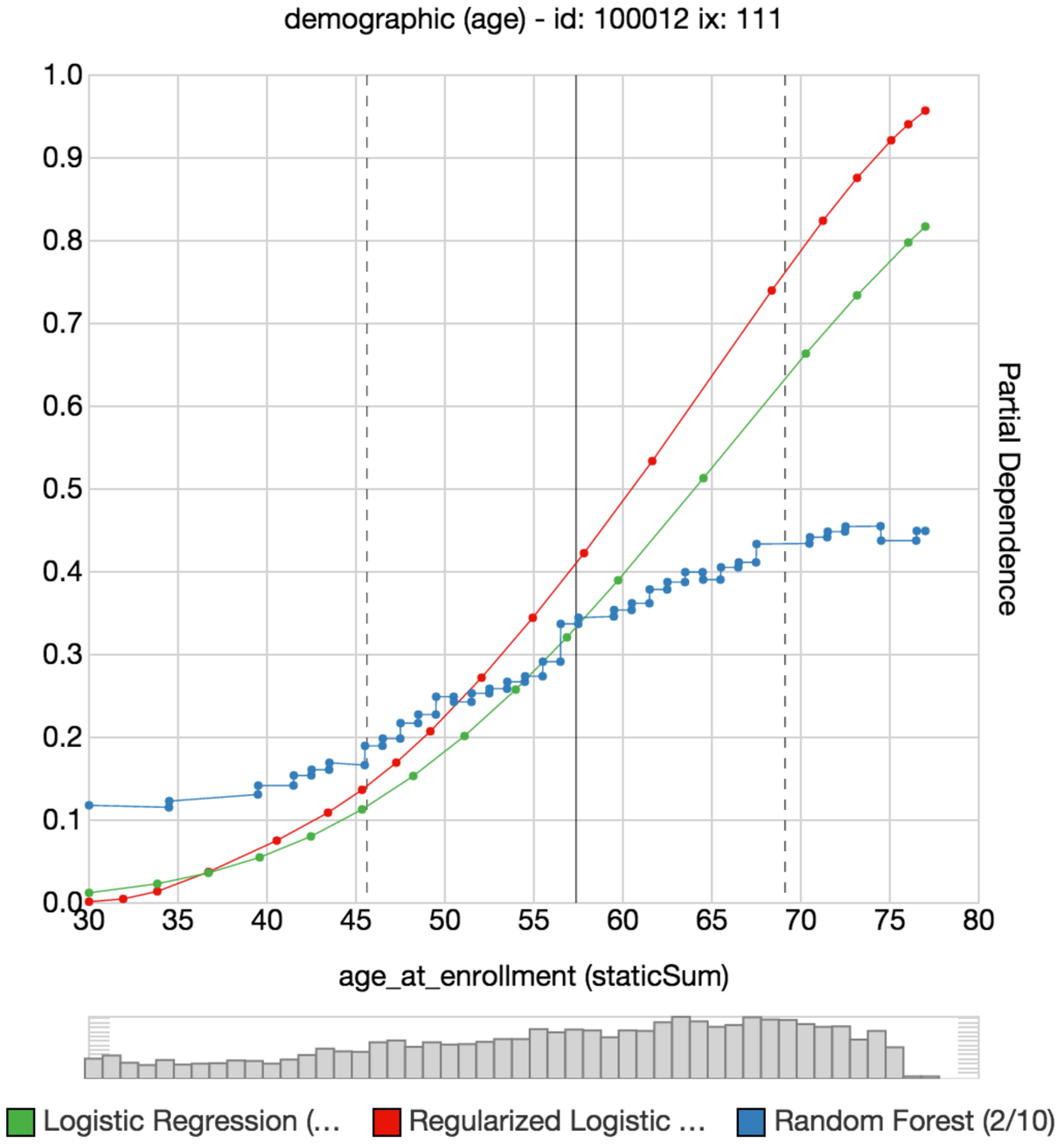
7/8 are predicted sick

partial dependence

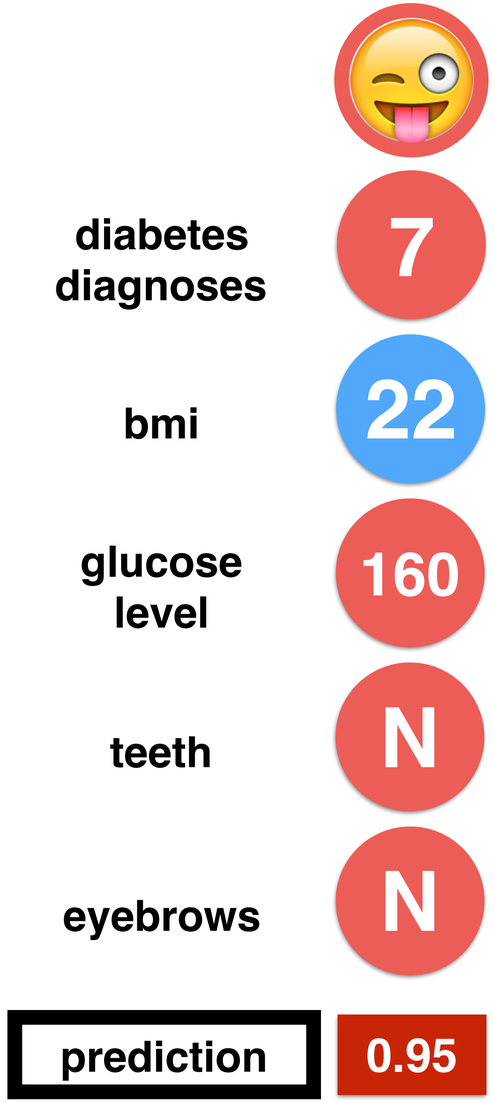


partial dependence

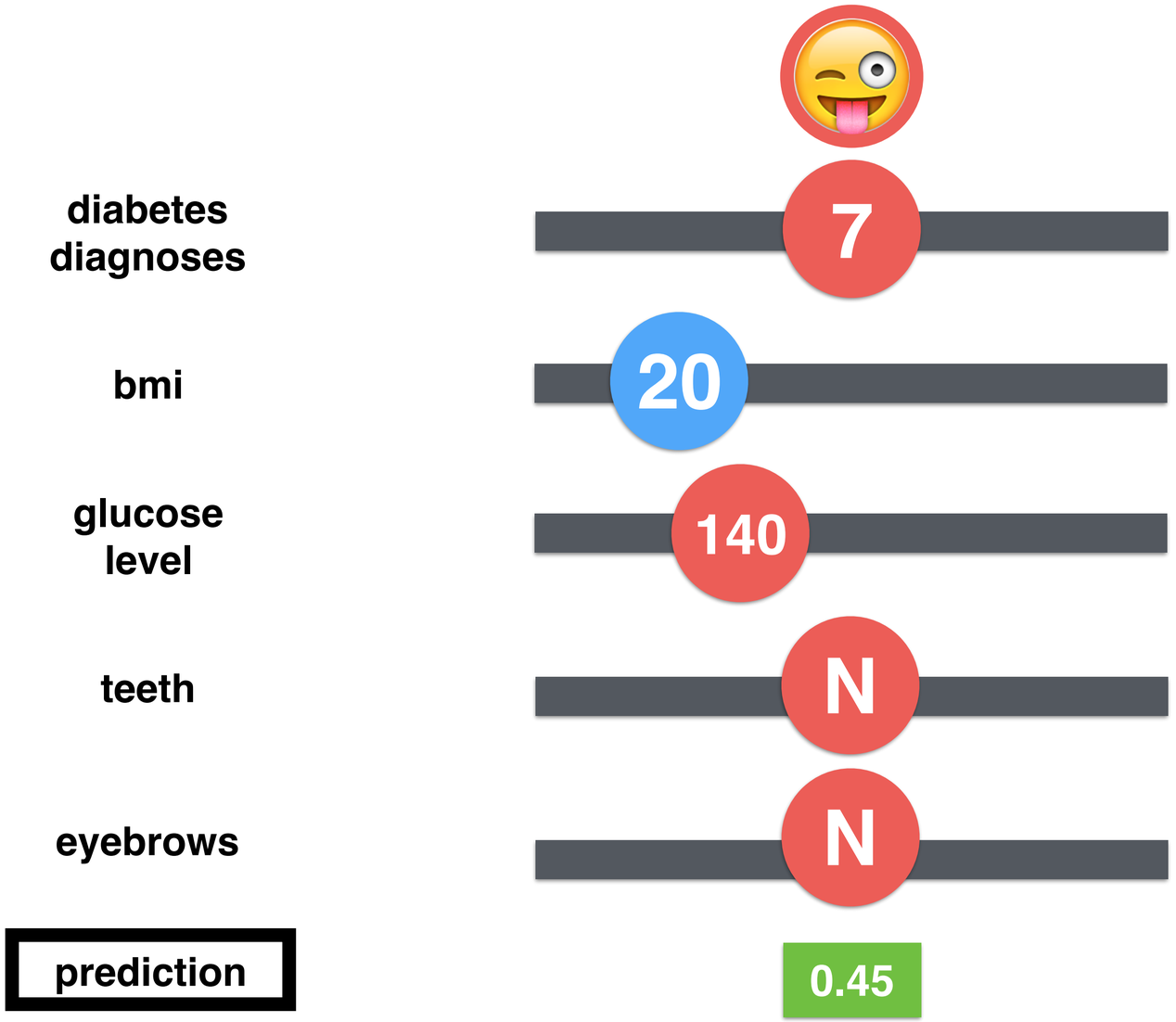


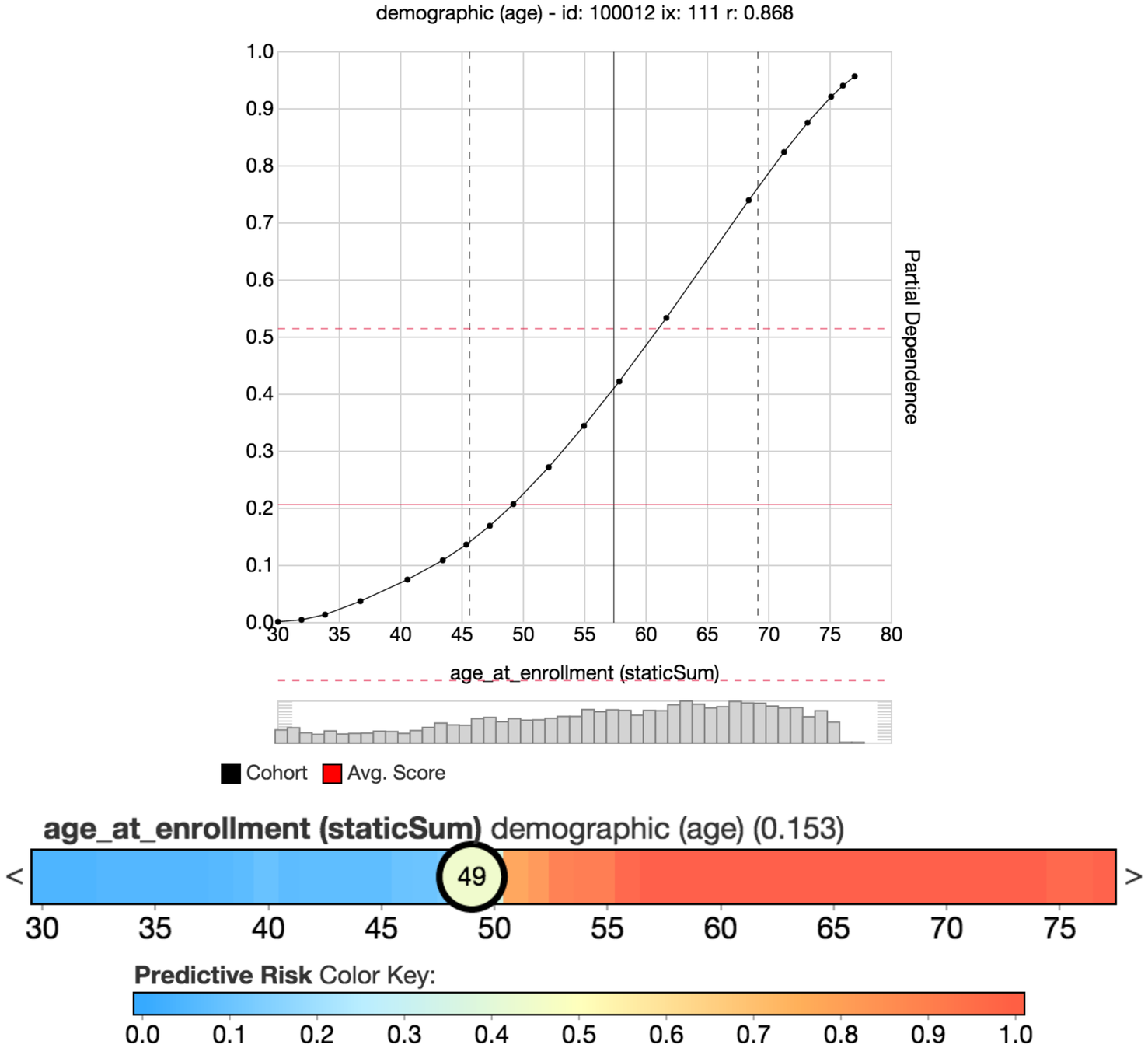

demographic (age) - id: 100012 ix: 111
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
30
35
40
45
50
55
60
65
70
75
80
Partial Dependence
age_at_enrollment (staticSum)
Logistic Regression (...
Regularized Logistic ...
Random Forest (2/10)

localized inspection



localized inspection




demographic (age) - id: 100012 ix: 111 r: 0.868
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
30
35
40
45
50
55
60
65
70
75
80
Partial Dependence
age_at_enrollment (staticSum)
Cohort
Avg. Score
age_at_enrollment (staticSum) demographic (age) (0.153)
<
49
>
30
35
40
45
50
55
60
65
70
75
Predictive Risk Color Key:
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

- predicting onset of diabetes for 4000 patients
- 4 month long term case study with 5 data scientists
- stories of visualization-driven insights in the paper

predictive model prospector

Model: Random Forest (0/10)

Features: 50 Patients: 59 / 3624 ACC: 0.74656 AUC: 0.84810

Select | Inspect | Partial Dependence

Predictive Risk Color Key:

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Globulin lab (0.0228)

< 3.01 3.20 >

2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5

bmi (count) vital (bmi) (0.151)

< 0 1 >

0 2 4 6 8 10 12 14 16 18

Calcium lab (0.0222)

< 9.77 9.95 >

8.6 8.8 9.0 9.2 9.4 9.6 9.8 10.0

bmi vital (bmi) (0.0450)

< 32.3 33.9 >

23 24 25 26 27 28 29 30 31 32 33 34 35

Glucose lab (0.0457)

< 96.5 107 >

90 100 110 120 130 140 150

eGFR lab (0.0360)

< 59.0 60 >

48 49 50 51 52 53 54 55 56 57 58 59 60

Alkaline Phosphatase lab (0.0163)

< 74.1 82 >

50 60 70 80 90 100 110

Anion Gap lab (0.0236)

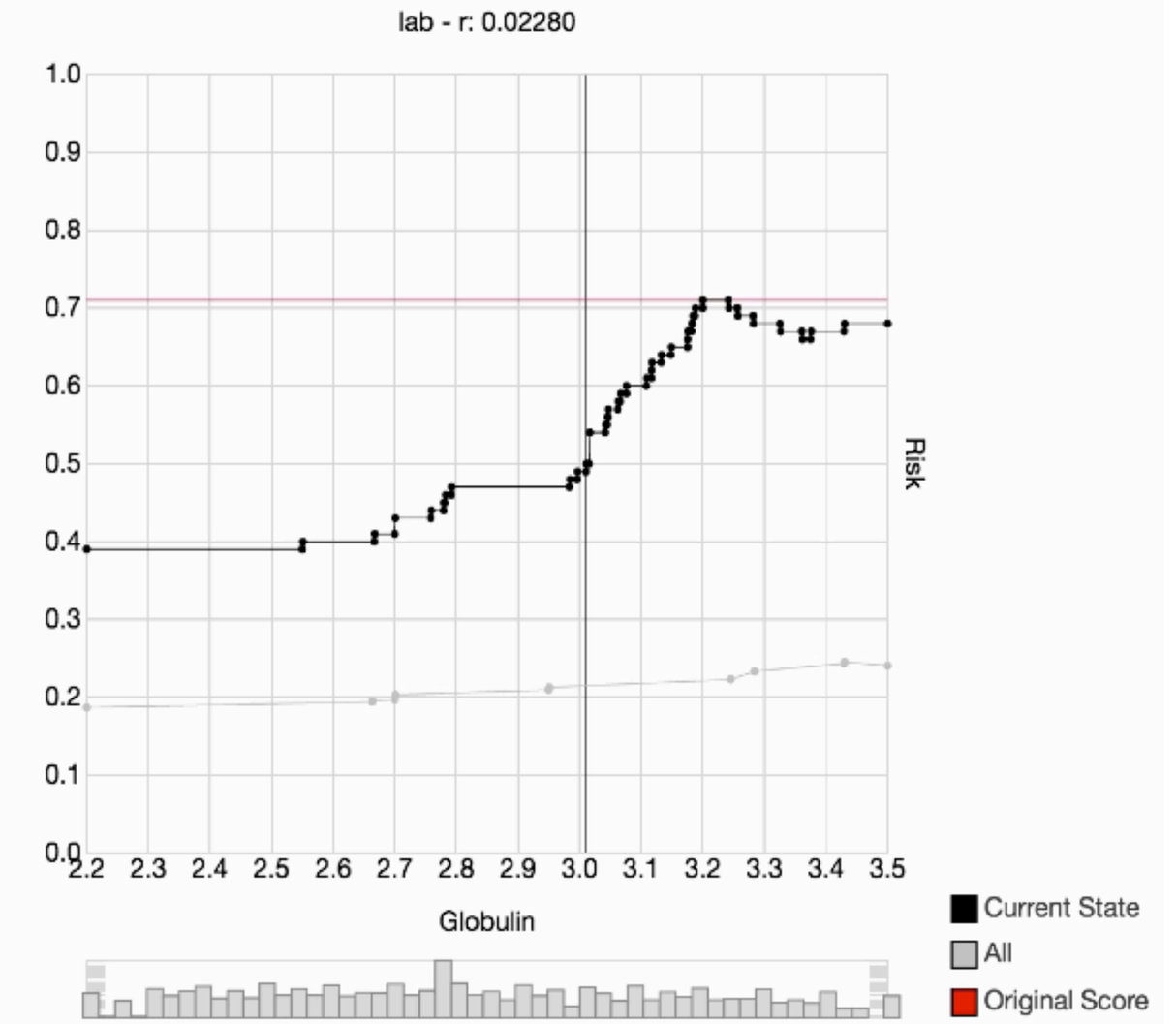
< 10.3 12 >

7 8 9 10 11 12 13 14

Platelets lab (0.0218)

Patient: 5754 Truth: 1 Original: 0.71000 Current: 0.49000

Show Neighbors | Sort by: Weight | Relevance | Inc. Risk | Dec. Risk



predictive model prospector

Model: Random Forest (0/10)

Features: 50 Patients: 59 / 3624 ACC: 0.74656 AUC: 0.84810

Select | Inspect | Partial Dependence

Predictive Risk Color Key:

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Globulin lab (0.0228)

< 3.01 3.20 >

2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5

bmi (count) vital (bmi) (0.151)

< 0 1 >

0 2 4 6 8 10 12 14 16 18

Calcium lab (0.0222)

< 9.77 9.95 >

8.6 8.8 9.0 9.2 9.4 9.6 9.8 10.0

bmi vital (bmi) (0.0450)

< 32.3 33.9 >

23 24 25 26 27 28 29 30 31 32 33 34 35

Glucose lab (0.0457)

< 96.5 107 >

90 100 110 120 130 140 150

eGFR lab (0.0360)

< 59.0 60 >

48 49 50 51 52 53 54 55 56 57 58 59 60

Alkaline Phosphatase lab (0.0163)

< 74.1 82 >

50 60 70 80 90 100 110

Anion Gap lab (0.0236)

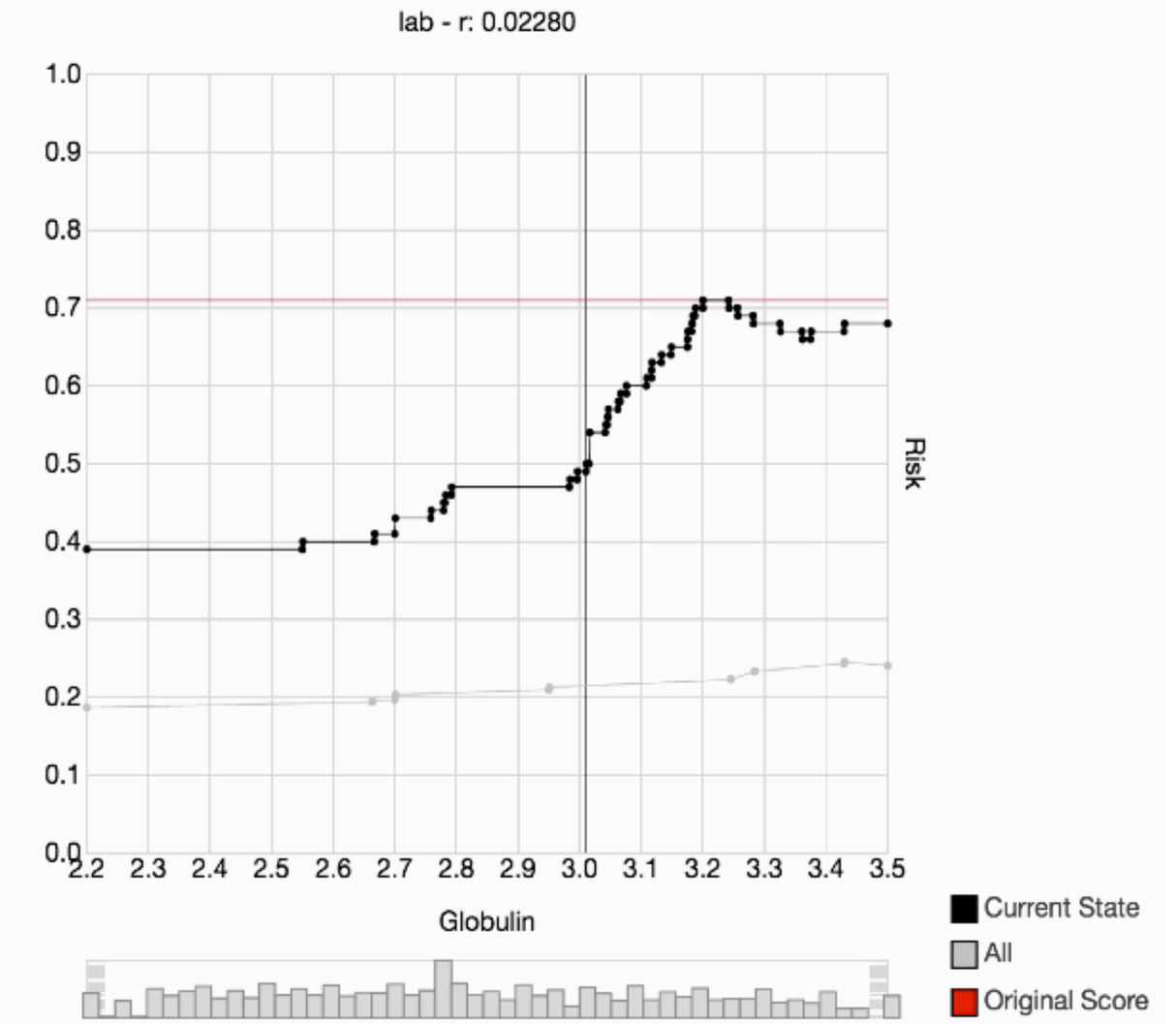
< 10.3 12 >

7 8 9 10 11 12 13 14

Platelets lab (0.0218)

Patient: 5754 Truth: 1 Original: 0.71000 Current: 0.49000

Show Neighbors | Sort by: Weight | Relevance | Inc. Risk | Dec. Risk



take-aways

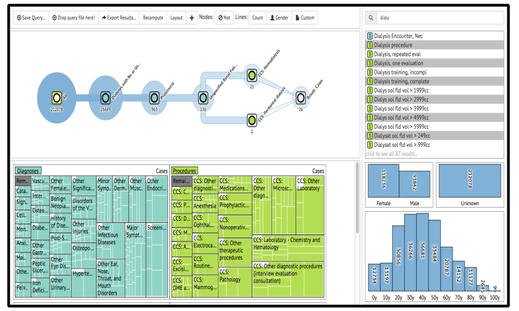
Clinical Data is complex and messy.

Exploratory visual analytics tools fill a much needed gap.

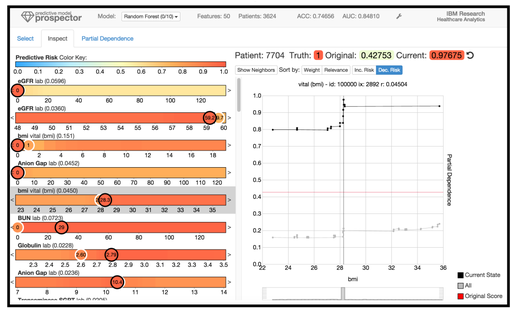
However, exploratory tools alone do not address their predictive desires.

There is a strong role for visualization in predictive tasks.

CareFlow (CHI 2013)



COQUITO (VAST 2015)



Prospector (CHI 2016)

Adam Perer

[papers and videos at http://perer.org]