DIHI Data Engineering

Pipelines for EHR Analytics
Outline

1. DIHI Efforts & RFAs
2. EHR Data Barriers
3. Sepsis Watch
4. Data Pipelines for Learning Health
DIHI Core

EFFORTS & RFAS
DIHI
Duke Institute for Health Innovation

Implementation and Health Delivery Science
- Catalyze multidisciplinary teamwork
- New care models
- Structured interface to Duke Health System
- Living laboratory to incubate, refine, validate and scale new ideas

Health Technology Innovation
- Leverage a growing health data infrastructure
- Connect to the digital health ecosystem
  - Collaboration and co-development of technology

Health Leadership and Workforce Development
- Goal to train current and future leaders across health care in
  - Leadership
  - Management
  - Innovation
  - Quantitative health sciences
- Contribute to developing the workforce of the future

Applied Health Policy
- Analyze effects of policies and health innovations
- An outlet to support the other DIHI innovation domains
- Develop innovation-focused health care reforms and programs
Our evolving approach

• **Innovation by design:**

  - understanding **user workflow**, desired **outcomes** and problems (needs) and then collaboratively develop concepts and prototypes, and **iterate through** to finalize **solution**.
Sourcing Innovation - DIHI Annual RFA

• Proposals in partnership with an implementation lead
• Full alignment with Duke Health Strategy
• 70+ applications per year
• Top applications reviewed and funded by Duke Health Leadership
Training Innovators - Education

• Med Student Data Scientists

• Data Scientist Healthcare Immersion

• Project Leadership and Responsibility

• Code – Model – Collaborate
Data Engineering

- Clinical and Claims Data Pipelines
- Clinical Data Monitor
- Data Aggregation & Atomization Systems
- De-identification Framework
- ML Model Deployment Framework
- User Interfaces
- Web and Mobile Applications for Health
DIHI Data Science

EHR DATA BARRIERS
Key Barriers

• Access
  – Limited reporting tools in EHR to extract data
  – Clarity database access requires training, investment and institutional knowledge
  – Cost of data extraction services
  – Tools and application outputs not reproducible, not transparent, doesn’t support deployment of apps

• Timeliness
  – Demand >>> supply of data extraction and curation services
  – Each extraction effort is re-built from ground up
  – Review and prioritization of requests is opaque, slow, plagued with politics
  – Require multiple iterations to obtain usable data

• Quality
  – No standard quality assurance process; variation in data quality attributable to analyst
  – Exploratory data analysis and evaluation of data left to user
  – Extraction and curation processes are static, despite dynamic data model
  – Data source expertise is fragmented and siloed across organizations
Scenario: Faculty member can’t rapidly access data

- **Wish:** I need a data set of all patients of a certain cohort over the last 2 years in our EHR. **I would like this by the end of this week.**
- **Reality:**
  - Only a few people in the institution can collate data sets at the institution
  - Faculty member speaks to EHR team at institution about project (1 week)
  - Faculty member socializes the need with their Chair (3 weeks)
  - Chair puts in a request with the EHR team for this data (1 day)
  - EHR team prioritizes request (4 weeks..if lucky or heavy bribe was paid)
  - Faculty member is allowed to speak to someone on the EHR team and they discuss the data needs (1 week)
  - EHR team pulls the data (12 weeks)
  - Faculty member realizes that EHR team pulled the wrong labs (1 week)
  - EHR team finds the right labs and re pulls data (6 weeks)

Process took 28 weeks and 1 day!
Scenario: Hospitals can’t deploy new data-driven apps in clinical care

• Hundreds of high-quality models are developed and validated in the medical literature every year
• Mapping local data sources to model predictors and validating extract, transform, and load pipeline can cost >$200,000 per model
• After taking 6-12 months to successfully implement and operationalize a model, hospital leadership prioritizes new use case
• New models with improved performance are published and hospital operations lags behind machine learning capabilities by years to decades
Scenario: Lab names change and it causes increase in cost and morbidity

- In the detection of Sepsis a blood culture is needed
- Over a three month period of time, label results for blood culture were renamed as Report as opposed to Bld_cult. Nobody knew.
- Physicians were ordering blood culture tests and not receiving results. At least they thought they weren’t.
- This resulted in 30% more blood culture tests being done, resulting in unnecessary costs and needle sticks to the patient,
- When timelines for Sepsis are counted in minutes instead of hours, this becomes a truly deadly error in the data.
Disappearing Blood Cultures

Blood Culture Component Name Trends Over Time

Blood Culture Order Description Trends Over Time
Lab Name Variability
Lab Test naming disparities, years after phase out
Live and Archival Systems

Chronicles
- Live data capture system
- *Cache* based hierarchical data store
- Complex in structure & Code
- Backed up to Clarity 12a-5a daily
- De-Normalized

Clarity
- ~24-hour stale Chronicles data warehouse
- Relational
- Complex (12k+ tables)
- Oracle (@Duke)
- Mostly Normalized
Security

Security

- PACE – Isolated Analytic Environment
- Elevated VPN
- SSL
- LDAP
- DHTS Vulnerability Scans
- ‘Production-like’ environment
Benefit to Duke

- Close the loop and give feedback to clinics in aggregate
- Enable Learning Health Units
- Simplify aggregate reporting
- Care Prioritization Modeling
- Simplify contributions to Common Data Models
- QA / QC / QI
- Dashboards
- Applications
DIHI Data Engineering

SEPSIS WATCH
The Problem: Sepsis

Percent of patients who received appropriate care for severe sepsis and septic shock\(^1\)

- 750,000 cases in US annually, high mortality (30-50%)
- $18,000 per hospital admission, $23B across all payers
- No clear time of onset, no clear biomarker
- 3-hour treatment bundle failure increases inpatient mortality 14%

\(^1\)Per CMS 2015 sepsis data
Transforming Sepsis Care Through Deep Learning

Past: Alarm Fatigue

NEWS Score fired BPA 447 times/day;
Average of 42 unique patients/day.
~100 times/patient

- Admitted patients
- Elevated NEWS score
- Best Practice Advisory
- RRT-Sepsis called
- Order set initiated

Low PPV: Only 6.8% of patients who had a NEWS-based BPA had a discharge diagnosis of sepsis

~63% of BPAs cancelled

Bringing Deep Learning to Duke Health Sepsis Care

DIHI Innovation Project:
Implementation of a Novel Duke-Specific Model to Detect and Treat Sepsis
Pis: Cara O’Brien, Katherine Heller, Armando Bedoya, Meredith Clement

<table>
<thead>
<tr>
<th>National Early Warning Score (NEWS)</th>
<th>Reclassification Rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1 0 1? 2? 3?</td>
<td>40</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>40 30 20 10 0 1 2 3</td>
<td>60</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>20 10 0 1 2 3</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>0 1 2 3</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>

Recurrent Neural Network

- Deep learning! Automatically learn rich features!
- Naturally handles variable-length sequences!
- Requires a complete dataset with no missing values.
- Requires regularly spaced inputs.

Multitask Gaussian Process

- Model multivariate time series.
- Handles irregularly spaced observation times.
- Imputes missing values on a regular grid, along with an estimate of uncertainty.

Futoma et al. International Conference on Machine Learning, 2017
## Defining Adult Sepsis at Duke

<table>
<thead>
<tr>
<th>2 or more SIRS criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Temperature &gt;38°C or &lt;36°C (6 hours)</td>
<td></td>
</tr>
<tr>
<td>• HR &gt;90 (6 hours)</td>
<td></td>
</tr>
<tr>
<td>• RR &gt;20 (6 hours)</td>
<td></td>
</tr>
<tr>
<td>• WBC count &gt;12, &lt;4, or % bandemia &gt;10% (24 hours)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suspicion for infection</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Blood culture order (24 hours)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1 element of end organ failure</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Creatinine &gt;2.0 (24 hours)</td>
<td></td>
</tr>
<tr>
<td>• INR &gt;1.5 (24 hours)</td>
<td></td>
</tr>
<tr>
<td>• Total bilirubin &gt;2.0 (24 hours)</td>
<td></td>
</tr>
<tr>
<td>• SBP &lt;90 or decrease in SBP by &gt;40 (6 hours)</td>
<td></td>
</tr>
<tr>
<td>• Platelets &lt;100 (24 hours)</td>
<td></td>
</tr>
<tr>
<td>• Lactate ≥2 (24 hours)</td>
<td></td>
</tr>
</tbody>
</table>
## Adult Sepsis Definition Analysis

<table>
<thead>
<tr>
<th></th>
<th>SIRS ≥2</th>
<th>qSOFA ≥2</th>
<th>SIRS ≥2 + any culture ordered</th>
<th>qSOFA ≥2 + any culture ordered</th>
<th>SIRS ≥2 + bacteremia</th>
<th>SIRS ≥2 + any culture ordered + element of organ damage</th>
<th>SIRS ≥2 + blood culture ordered + element of organ damage</th>
<th>ICD diagnosis code associated with sepsis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of encounters</td>
<td>32928</td>
<td>17423</td>
<td>14327</td>
<td>7110</td>
<td>1419</td>
<td>13358</td>
<td>9184</td>
<td>2884</td>
<td>43046</td>
</tr>
<tr>
<td>Median length of stay in days (lower-upper quartiles)</td>
<td>4.6 (2.8-8.1)</td>
<td>5.9 (3.2-10.7)</td>
<td>6.4 (3.7-12.1)</td>
<td>8.3 (4.5-16.3)</td>
<td>11.0 (5.9-23.7)</td>
<td>6.9 (3.9-12.8)</td>
<td>7.3 (4.1-14.6)</td>
<td>7.5 (4.1-15.4)</td>
<td>4.0 (2.4-7.0)</td>
</tr>
<tr>
<td>Inpatient mortality rate (%)</td>
<td>3.7%</td>
<td>6.7%</td>
<td>6.9%</td>
<td>12.6%</td>
<td>15.0%</td>
<td>7.4%</td>
<td>9.7%</td>
<td>16.3%</td>
<td>2.9%</td>
</tr>
<tr>
<td>ICU requirement rate (%)</td>
<td>21.3%</td>
<td>32.0%</td>
<td>28.7%</td>
<td>45.0%</td>
<td>38.9%</td>
<td>30.0%</td>
<td>34.5%</td>
<td>46.4%</td>
<td>18.9%</td>
</tr>
<tr>
<td>Antibiotic administration rate (%)</td>
<td>62.4%</td>
<td>69.0%</td>
<td>82.8%</td>
<td>85.5%</td>
<td>97.8%</td>
<td>83.2%</td>
<td>90.0%</td>
<td>98.5%</td>
<td>63.2%</td>
</tr>
<tr>
<td>IV fluid administration rate (%)</td>
<td>38.0%</td>
<td>37.8%</td>
<td>47.4%</td>
<td>49.6%</td>
<td>67.1%</td>
<td>48.5%</td>
<td>56.7%</td>
<td>86.7%</td>
<td>42.4%</td>
</tr>
<tr>
<td>Vasopressor administration rate (%)</td>
<td>10.2%</td>
<td>17.1%</td>
<td>15.0%</td>
<td>27.3%</td>
<td>28.8%</td>
<td>16.0%</td>
<td>19.4%</td>
<td>32.8%</td>
<td>9.6%</td>
</tr>
</tbody>
</table>
Model Performance

C-statistics by Method

- Our Model
- SIRS
- Random Forest
- NEWS
- Cox Regression
- QSOFA
- Logistic Regression

C-statistic (AU-ROC) vs Hours prior to sepsis
Model Performance

Model Operating Alarms

Avg. Cases Detected Early / 24 Hrs

Avg. Alarms / Hr

- Our Model
- Random Forest
- Cox Regression
- Logistic Regression
- SIRS
- NEWS
- QSOFA
Model Performance

Model Operating Alarms

Captures ~7 more cases of sepsis early

Median time of 5 hours prior to clinical presentation!
Data Science Project Process

Raw Files → Build Features → Processed Files

Transformations → Analytic Files

Predictive Modeling

\[
[\beta_1 \ \beta_2 \ ... \ \beta_n]
\begin{bmatrix}
X_1 \\
X_2 \\
... \\
X_n 
\end{bmatrix}
\]
Data Science Project Process – Sepsis

**Raw Files**
- Duke EHR

**Transformations**

**Build Features**

**Processed Files**
- Analyte results
- Vital signs
- Medication admins
- Demographics

**Analytic Files**
- Analyte results
- Vital signs
- Medication admins
- Demographics

**Predictive Modeling**
- Predict risk of sepsis within next 24 hours
Roles & Responsibilities

- **RRT Nurse:** Primary user of Sepsis Watch, initiates communication with ED staff, documents that communication via significant event note, rounds on patient once admitted.

- **ED Patient Flow Coordinator:** Secondary user of Sepsis Watch (non-interactive user), helps coordinate patient transfer destination and handoff with ED Nurse and IP Nurse.

- **ED Nurse:** Aware of Sepsis Watch, administers bundle treatments, hands off patient to IP Nurse.

- **ED Physician:** Aware of Sepsis Watch, primary communication point with RRT Nurse, orders Sepsis Bundle requirements in Maestro Care.

- **Inpatient Nurse:** Aware of Sepsis Watch, receives patient from ED Nurse handoff, administers bundle treatments.
Sepsis Watch Web Application

Each “card” represents a single patient at Duke Hospital.

Sepsis Criteria Met
- Time when patient met sepsis criteria (black cards only)

Labs and Vitals
- Temperature, Pulse, Blood Pressure, Respirations, White Blood Cell Count, Lactate level, Mean Arterial Pressure

Bundle Items in Past 3 Hrs
- Indicates whether any of the bundle requirements have been acted on in the last 3 hours

Disclaimer: This is not PHI. These are test patients and fake data, and so may show incorrect values (e.g., MAP calculation)
Sepsis Watch Application Performance Monitoring

Click a tab below to see information regarding daily and weekly statistics!

| Encounter | Height & Weight | Sepsis Events | Risk Scores | Analytes | Medications | Vitals | Blood Cultures |

Vitals Daily Count

- blood_pressure
- pulse
- pulse_oximetry
- r_duhs_ip_supplemental_oxygen
- r_duhs_level_of_consciousness
- r_oxygen_device
- respirations
- temperature
- urine_output

Date:
- Sep 2 2018
- Sep 9
- Sep 16
- Sep 23
Version Controlled Infrastructure

- Docker
- Vagrant
- Airflow
- RabbitMQ
- Ansible
- Gitlab

Sepsis Watch is deployed with 1 line of code:

- 2 Load Balanced Webapps
- 1 Airflow Code orchestration UI
- 1 Message Queue UI
- 1 Message Queue
- 1 Task Scheduler
- Up to 6 Workers, half extracting data and half running models.

All running on combination of 6 DHTS VMs

Test environment is the same, with an extra Webapp for development. All deployed with the same code.

Test env watches All ED Contacts across all Hospitals
Docker & Docker-Compose

Docker  Docker-Compose
Docker & Docker-Compose

Docker

Docker-Compose
Docker-compose

- Containerized Software
- Disposability
- Replication across environments
- Operating System Agnostic
- Build it once, ship it anywhere

Control of Environment!
Airflow

• Software Scheduling of Tasks that may, or may not, depend on one another.

Task: Send an e-mail every morning at 10am

1. Set alarm for 10:00am EST
2. Send pre-composed e-mail
Airflow

• Software Scheduling of Tasks that may, or may not, depend on one another.
  
  Task: Send an e-mail every morning at 10am

  1. Set alarm for 10:00am EST
  2. Send pre-composed e-mail

10 am
Airflow

• Software Scheduling of Tasks that may, or may not, depend on one another.

Task: Send an e-mail every morning at 10am

1. Set alarm for 10:00am EST
2. Send pre-composed e-mail

*5000?
# Airflow - Control Dashboard

<table>
<thead>
<tr>
<th>DAG</th>
<th>Schedule</th>
<th>Owner</th>
<th>Recent Tasks</th>
<th>Last Run</th>
<th>DAG Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>airflow_db_cleanup</td>
<td>Daily</td>
<td>operations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clean-demographics-task</td>
<td>None</td>
<td>airflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fetch-active-orders-task</td>
<td>None</td>
<td>airflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fetch-census-task</td>
<td>Weekly</td>
<td>airflow</td>
<td></td>
<td></td>
<td>2019-02-25 15:25</td>
</tr>
<tr>
<td>fetch-encounters-task</td>
<td>None</td>
<td>airflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fetch-flowsheet-data-task</td>
<td>None</td>
<td>airflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fetch-lab-results-task</td>
<td>None</td>
<td>airflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fetch-med-admins-task</td>
<td>None</td>
<td>airflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sepsis_model</td>
<td>Weekly</td>
<td>airflow</td>
<td></td>
<td></td>
<td>2019-02-25 14:00</td>
</tr>
<tr>
<td>time_of_sepals</td>
<td>Weekly</td>
<td>airflow</td>
<td></td>
<td></td>
<td>2019-02-25 15:25</td>
</tr>
</tbody>
</table>
Airflow - Pull and Clean Data
Airflow – Monitor Tasks

Run: scheduled_2019-02-25T15:15:00
Layout: Left->Right

Task_ID: fetch-encounters-task
Run: 2019-02-25T15:15:00
Operator: PythonOperator
Started: 2019-02-25T15:20:50.861753
Duration: 17.98s
State: success

- fetch-encounters-task
- fetch-active-orders-task
- fetch-flowsheet-data-task
- fetch-lab-results-task
- fetch-med-admins-task
- clean-demographics-task
Airflow

- Elegant Code Orchestration with Directed Acyclic Graphs (DAGs)
- Task Idempotency
- Logging
- Task Execution Metadata

Control of Code / Models!
Real-Time Model Serving Architecture
Real-Time Model Serving Architecture

- **Airflow Webapp**
- **Task Scheduler**
- **TaskQueue**
- **Data Extraction**
- **Data Cleaning**
- **SepsisWatch Webapp**
- **Sepsis Risk Scores**
- **Sepsis Model Run**
- **Sepsis Definition Run**
- **Sepsis Status Call**
Real-Time Model Serving Architecture
Real-Time Model Serving Architecture
Real-Time Model Serving Architecture

Diagram showing the architecture involving Airflow Webapp, Task Scheduler, TaskQueue, Data Extraction, Data Cleaning, ScribeDB, Sepsis Risk Scores, Sepsis Model Run, Sepsis Definition Run, and Sepsis Status Call.
Real-Time Model Serving Architecture

Airflow Webapp

Task Scheduler

TaskQueue

Data Extraction

Data Cleaning

Chronicles (Streamed EHR)

ScribeDB

Sepsis Risk Scores

Sepsis Model Run

Sepsis Definition Run

Sepsis Status Call
Real-Time Model Serving Architecture
Real-Time Model Serving Architecture
Real-Time Model Serving Architecture
Real-Time Model Serving Architecture
DIHI Data Engineering

DATA PIPELINES FOR LEARNING HEALTH
Serving Sepsis

**Tools:**
- Docker
- Vagrant
- Airflow
- RabbitMQ
- Ansible
- Gitlab
Serving Generic Models - Lightweight

**Tools:**
- Docker Compose
- Vagrant
- Airflow
- RabbitMQ
- Ansible
- Gitlab
Design Goals & Requirements

Goals:
• Build on and Improve SepsisWatch Backend
• Version Controlled Infrastructure
• Dev/Prod Parity
• Ease of development
• Ease of deployment
• Ease of maintenance
• Ease of extensibility

Reqs:
• 24-hour stale ‘clean’ data
• Historical ‘clean’ data 7/2014 to Present
• All major EHR data represented
• Deployment in PACE
• Designed for iteration and improvement
• Built *agnostic* to source data model
Early Evidence
Concept Validation with Ebony Boulware

2015-2016 DCC Chronic Kidney Disease Pilot

- **Fall 2015**: DIHI and DTRI Grant Submissions
- **Winter 2015**: DIHI and DTRI Grant Submissions
- **Spring 2016**: $115,000 Awarded
- **Summer 2016**: Initial Feature Construction and Model Deployment
- **Spring 2017**: Features, Model, Workflow Validated and Operationalized
2015-2016 DCC Chronic Kidney Disease Pilot

Concept Validation with Ebony Boulware

Initial Feature Construction and Model Deployment

Features, Model, Workflow Validated and Operationalized

Fall 2015
Winter 2015
Spring 2016
Summer 2016
Spring 2017

DIHI and DTRI Grant Submissions

$115,000 Awarded

Total Time: 1.5 years
## 2015-2016 DCC Chronic Kidney Disease Pilot

<table>
<thead>
<tr>
<th>Category</th>
<th>Tasks</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Extraction</strong></td>
<td>SQL queries, Exploratory data analysis, Data warehouse validation, Documentation</td>
<td>40,000</td>
</tr>
<tr>
<td><strong>Data Transformation</strong></td>
<td>Deploy models, transform lab values</td>
<td>1,000</td>
</tr>
<tr>
<td><strong>Application Development</strong></td>
<td>Requirements gathering, design, back end development, front end development, hardware, product launch</td>
<td>130,000</td>
</tr>
<tr>
<td><strong>Clinical Validation</strong></td>
<td>Data element validation, workflow validation</td>
<td>50,000</td>
</tr>
<tr>
<td><strong>Total Cost</strong></td>
<td></td>
<td>220,000</td>
</tr>
</tbody>
</table>

Predicting Colorectal Cancer from Complete Blood Counts

Request from NYU Data Science Lead with Model Citation and Clarity SQL

May 22, 2018

Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study

Yaron Kinar, Nir Kalkstein, Pinchas Akiva, Bernard Levin, Elizabeth E Haff, Inbal Goldshtein, Gabriel Chodick, and Varda Shalev
Predicting Colorectal Cancer from Complete Blood Counts

Hemoglobin, Hematocrit, WBCs, Platelets, CRC Outcome Curated in PACE for DUHS Patients

Request from NYU Data Science Lead with Model Citation and Clarity SQL

May 22, 2018

NYU Langone Health

CENTER FOR HEALTHCARE INNOVATION AND DELIVERY SCIENCE (CHIDS)

~$700k

May 28, 2018

Model Performance: AUC 0.72

Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study

Yaron Kinar,1 Nir Kalkstein,1 Pinchas Akiva,1,8 Bernard Levin,2 Elizabeth E Half,3 Inbal Goldshtein,4 Gabriel Chodick,4 and Varda Shalev4,5

RECEIVED 20 May 2015
REVISED 1 November 2015
ACCEPTED 7 November 2015
PUBLISHED ONLINE FIRST 15 February 2016

DIHI

Proprietary and Confidential
Predicting Colorectal Cancer from Complete Blood Counts

Hemoglobin, Hematocrit, WBCs, Platelets, CRC Outcome Curated in PACE for DUHS Patients

Request from NYU Data Science Lead with Model Citation and Clarity SQL

Patient Age, Gender, Curated in PACE for DUHS Patients

May 22, 2018

May 28, 2018

Model Performance: AUC 0.72

May 29, 2018

Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study

Yaron Kinar,1 Nir Kalkstein,1 Pinchas Akiva,1,4 Bernard Levin,2 Elizabeth E Half,3 Inbal Goldshtein,4 Gabriel Chodick,4 and Varda Shalev1,4

To: Michael Gao

5/29/18, 10:19 PM

.81 auc

Just like their papers
Predicting Colorectal Cancer from Complete Blood Counts

Hemoglobin, Hematocrit, WBCs, Platelets, CRC Outcome Curated in PACE for DUHS Patients

Request from NYU Data Science Lead with Model Citation and Clarity SQL

Patient Age, Gender, Curated in PACE for DUHS Patients

May 22, 2018

May 28, 2018

Model Performance: AUC 0.72

May 29, 2018

Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study

Yaron Kiner,1 Nir Kalkstein,1 Pinchas Akiva,1,∗ Bernard Levin,2 Elizabeth E Half,3 Inbal Goldshtein,4 Gabriel Chodick,4 and Varda Shalev4,5

To: Michael Gao

5/29/18, 10:19 PM

.81 auc

Just like their papers

Total Time: 48 hours
## Predicting Colorectal Cancer from Complete Blood Counts

<table>
<thead>
<tr>
<th>Category</th>
<th>Tasks</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Extraction</strong></td>
<td>SQL queries, Exploratory data analysis, Data warehouse validation, Documentation</td>
<td>80</td>
</tr>
<tr>
<td><strong>Data Transformation</strong></td>
<td>Develop models, transform lab values</td>
<td>200</td>
</tr>
<tr>
<td><strong>Application Development</strong></td>
<td>Requirements gathering, design, back end development, front end development, hardware, product launch</td>
<td>200</td>
</tr>
<tr>
<td><strong>Clinical Validation</strong></td>
<td>Data element validation, workflow validation</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total Cost</strong></td>
<td></td>
<td>480</td>
</tr>
</tbody>
</table>
Outpatient Watch

- Monitor disease trends, utilization patterns, and epidemiologic variables of interest
- Identify subgroups of patients who have unmet needs
- Assist program managers to identify appropriate patients for interventions
- Monitor effectiveness of treatment for cohort
DIHI Data Engineering

CURRENT & FUTURE DIRECTIONS
Expansion & Improvement

• Improve EHR abstraction layer
• Add Monitoring and Content Aggregation as services
• Further Simplify model portability and testing
• Migrate codebase to Go where performance is a concern

• Begin simplification of code migration out of PACE for DHTS deployment at scale.
• Begin work with Learning Health Units
Model + Workflow -> Scale -> Repeat

- Improve Patient Health and Interaction with the Health system through new models and improved workflows.

- Improve discovery rate by Lowering/Removing the barrier to entry for Near-Real-Time EHR Analytics & Model development in a secure PACE environment.
Team Data Science

**Data + Quantitative + Software**

- Joe Futoma
- Marshall Nichols
- Michael Gao
- Mark Sendak
- Mike Revoir
- Katherine Heller
- Bryce Wolery
- Sanjay Hariharan
- Anthony Lin
- Michael Kahl
- Suresh Balu

**Clinical Design**

- Mark Sendak
- Cara O’Brien
- Armando Bedoya
- Meredith Clement
- Nathan Brajer
- Anthony Lin
- Suresh Balu

**Clinical Implementation**

- Mary Ann Fuchs
- Cara O’Brien
- Alan Kirk
- Armando Bedoya
- Mary Ann Fuchs
- Mark Sendak
- Suresh Balu
- Will Ratliff

**Technology**

- Eric Poon
- Jeff Ferranti
- Susan Engelbosch
- Tres Brown
- Pedro Borghes
- Suresh Balu
- Mike Revoir
- Marshall Nichols

**Leadership**

- Tom Owens
- Bill Faulkerson
- Jeff Ferranti
- Eric Poon
- Alan Kirk
- Mary Ann Fuchs

---

DIHI

---

Duke Institute for Health Innovation

Duke University Health System

Duke Statistical Science

DukeHealth

Technology Services
# Data Pipeline key stats

<table>
<thead>
<tr>
<th>oid</th>
<th>table_schema</th>
<th>table_name</th>
<th>row_estimate</th>
<th>total_bytes</th>
<th>index_bytes</th>
<th>toast_bytes</th>
<th>table_bytes</th>
<th>total</th>
<th>index</th>
<th>toast</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>16419</td>
<td>public</td>
<td>act</td>
<td>11740800</td>
<td>5936373760</td>
<td>3243999520</td>
<td>8192</td>
<td>2692866048</td>
<td>5661 MB</td>
<td>3093 MB</td>
<td>8192 bytes</td>
<td>2568 MB</td>
</tr>
<tr>
<td>16436</td>
<td>public</td>
<td>allergy</td>
<td>2082590</td>
<td>686824800</td>
<td>385351680</td>
<td>8192</td>
<td>300924928</td>
<td>654 MB</td>
<td>368 MB</td>
<td>8192 bytes</td>
<td>287 MB</td>
</tr>
<tr>
<td>16450</td>
<td>public</td>
<td>analyte</td>
<td>199500000</td>
<td>17213788972</td>
<td>7616695008</td>
<td>8192</td>
<td>95970926592</td>
<td>160 GB</td>
<td>71 GB</td>
<td>8192 bytes</td>
<td>89 GB</td>
</tr>
<tr>
<td>16467</td>
<td>public</td>
<td>demographic</td>
<td>3102200</td>
<td>1767350272</td>
<td>1051467776</td>
<td>8192</td>
<td>715874304</td>
<td>1685 MB</td>
<td>1003 MB</td>
<td>8192 bytes</td>
<td>683 MB</td>
</tr>
<tr>
<td>16481</td>
<td>public</td>
<td>icd9_diagnosis</td>
<td>501242000</td>
<td>17652645888</td>
<td>12245286912</td>
<td>0</td>
<td>5407358976</td>
<td>16 GB</td>
<td>11 GB</td>
<td>0</td>
<td>5157 MB</td>
</tr>
<tr>
<td>16495</td>
<td>public</td>
<td>icd10_diagnosis</td>
<td>51237100</td>
<td>17570816000</td>
<td>12184846336</td>
<td>0</td>
<td>5385969664</td>
<td>16 GB</td>
<td>11 GB</td>
<td>0</td>
<td>5136 MB</td>
</tr>
<tr>
<td>16507</td>
<td>public</td>
<td>encounter</td>
<td>20726500</td>
<td>11312575600</td>
<td>3318767616</td>
<td>8192</td>
<td>7993802752</td>
<td>11 GB</td>
<td>3165 MB</td>
<td>8192 bytes</td>
<td>7623 MB</td>
</tr>
<tr>
<td>16518</td>
<td>public</td>
<td>ip_encounter</td>
<td>1002070</td>
<td>400703488</td>
<td>147824640</td>
<td>8192</td>
<td>252870656</td>
<td>382 MB</td>
<td>141 MB</td>
<td>8192 bytes</td>
<td>241 MB</td>
</tr>
<tr>
<td>16530</td>
<td>public</td>
<td>med_admin</td>
<td>79611200</td>
<td>38668269440</td>
<td>22808346624</td>
<td>8192</td>
<td>15877914624</td>
<td>36 GB</td>
<td>21 GB</td>
<td>8192 bytes</td>
<td>15 GB</td>
</tr>
<tr>
<td>16547</td>
<td>public</td>
<td>med_list</td>
<td>426513000</td>
<td>19570411104</td>
<td>121368035328</td>
<td>8192</td>
<td>74336067584</td>
<td>182 GB</td>
<td>113 GB</td>
<td>8192 bytes</td>
<td>69 GB</td>
</tr>
<tr>
<td>16563</td>
<td>public</td>
<td>note</td>
<td>178318000</td>
<td>356872798208</td>
<td>6316466336</td>
<td>80754884608</td>
<td>21295322764</td>
<td>332 GB</td>
<td>59 GB</td>
<td>75 GB</td>
<td>198 GB</td>
</tr>
<tr>
<td>16580</td>
<td>public</td>
<td>order</td>
<td>139902000</td>
<td>69870641152</td>
<td>44706488320</td>
<td>8192</td>
<td>25164144640</td>
<td>65 GB</td>
<td>42 GB</td>
<td>8192 bytes</td>
<td>23 GB</td>
</tr>
<tr>
<td>16595</td>
<td>public</td>
<td>patient</td>
<td>1530870</td>
<td>176594944</td>
<td>63946752</td>
<td>0</td>
<td>112648192</td>
<td>168 MB</td>
<td>61 MB</td>
<td>0</td>
<td>107 MB</td>
</tr>
<tr>
<td>16602</td>
<td>public</td>
<td>patient_mrn</td>
<td>1454680</td>
<td>186875904</td>
<td>106717184</td>
<td>0</td>
<td>80158720</td>
<td>178 MB</td>
<td>102 MB</td>
<td>0</td>
<td>76 MB</td>
</tr>
<tr>
<td>16612</td>
<td>public</td>
<td>pat_loc_hx</td>
<td>4600480</td>
<td>2617434112</td>
<td>1455259648</td>
<td>8192</td>
<td>1162166272</td>
<td>2496 MB</td>
<td>1388 MB</td>
<td>8192 bytes</td>
<td>1108 MB</td>
</tr>
<tr>
<td>16629</td>
<td>public</td>
<td>problem_list</td>
<td>8959800</td>
<td>5305212928</td>
<td>3395190784</td>
<td>8192</td>
<td>1910013952</td>
<td>5059 MB</td>
<td>3238 MB</td>
<td>8192 bytes</td>
<td>1822 MB</td>
</tr>
<tr>
<td>16647</td>
<td>public</td>
<td>procedure</td>
<td>990623</td>
<td>504733696</td>
<td>253231104</td>
<td>8192</td>
<td>251494400</td>
<td>481 MB</td>
<td>242 MB</td>
<td>8192 bytes</td>
<td>240 MB</td>
</tr>
<tr>
<td>16661</td>
<td>public</td>
<td>provider</td>
<td>93578</td>
<td>30900224</td>
<td>8060928</td>
<td>0</td>
<td>22383296</td>
<td>29 MB</td>
<td>7872 kB</td>
<td>0</td>
<td>22 MB</td>
</tr>
<tr>
<td>16668</td>
<td>public</td>
<td>social_hx</td>
<td>12370100</td>
<td>4115201240</td>
<td>2674622464</td>
<td>8192</td>
<td>1440579584</td>
<td>3925 MB</td>
<td>2551 MB</td>
<td>8192 bytes</td>
<td>1374 MB</td>
</tr>
<tr>
<td>16684</td>
<td>public</td>
<td>flowsheet</td>
<td>755320000</td>
<td>321500430336</td>
<td>184668102656</td>
<td>57344</td>
<td>136832270336</td>
<td>299 GB</td>
<td>172 GB</td>
<td>56 kB</td>
<td>127 GB</td>
</tr>
</tbody>
</table>

Showing 1 to 20 of 20 entries
Pipeline execution statistics

[Graph showing pipeline execution statistics over time]
Thank you!

-DIHI Team