Research IT Information & Exchange Series

Large Data Set Mining to Answer Research Questions

March 24, 2017

Research IT Information & Exchange Series

- **Goal:** To educate pediatric researchers on the Research IT and Informatics resources and expertise available to facilitate their research and to identify areas where we can enhance IT methods to better support research.
- **Format:** One hour sessions led by subject matter experts to present information on the current services and expertise available.
- **Intended audience:** Researchers with an interest in capitalizing on Research IT tools to make their research better. Also, researchers who are interested in using Big Data and Healthcare Analytic approaches in their research.

Research IT Information & Exchange Series:

Learn about using different clinical datasets in your research. Included in the discussion today:

- Children's Clinical Data Warehouse
- Hadoop
- Medicaid data at GT

Research IT Information & Exchange Series

Our presenters today

- **Tal Senior, RN, BSN,** Manager, IT Analysis, Children's Healthcare of Atlanta
  - Tal.Senior@choa.org
- **Tod Davis,** Business Intelligence Architect/Developer, Children's Healthcare of Atlanta
  - Tod.davis@choa.org
- **Nicoleta Serban, PhD,** Coca Cola Associate Professor, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
  - nicoleta.serban@isye.gatech.edu
- **Richard Starr,** Research Scientist, Institute for People and Technology (IPaT) at Georgia Tech
  - rstarr7@gatech.edu

data@choa.org

Tal Senior, RN BSN

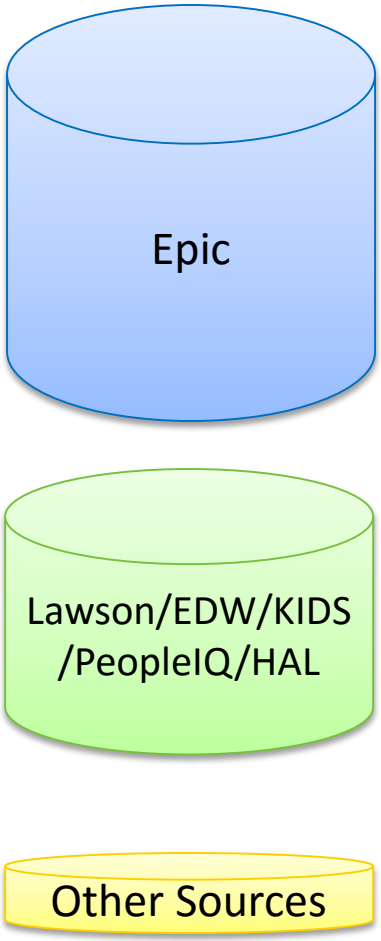
Manager, Reporting and Analytics

Business Intelligence

tal.senior@choa.org

Children's℠
Healthcare of Atlanta
Dedicated to All Better

Children's Data Sources



Clinical Operations
Financial Operations
Research
Outcomes
Meaningful Use
Compliance
Legal
Security

Administration/Employee Compliance
Clinical Operations
Financial Operations
Research

Department Specific

Preparation for Data Request

- IRB Letter OR IRB Exemption letter (if process improvement or quality initiative).

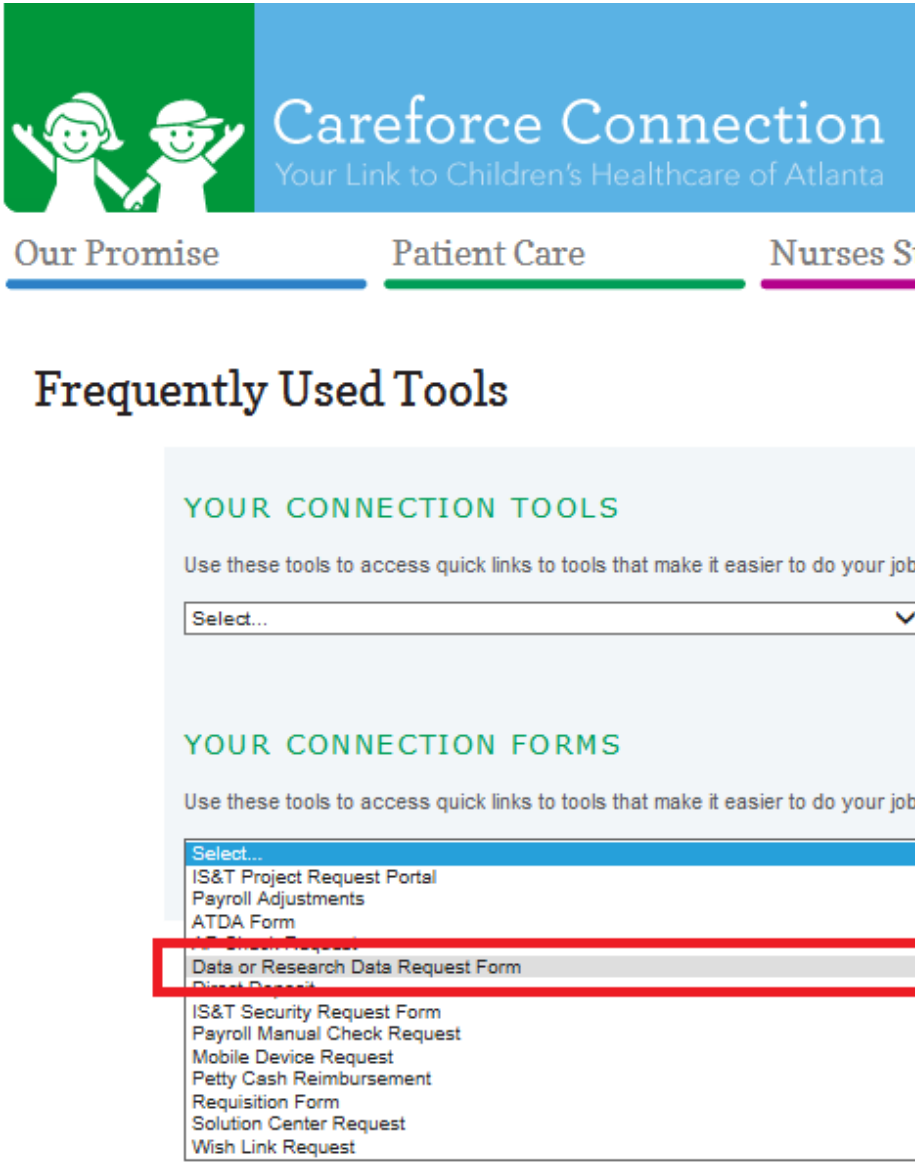
- Identify your Children's data sponsor: Children's employee responsible for this data request. Required to sign and return Children's Data Sponsor document. Sponsor can also enter the data request on your behalf (if you don't have access to the Children's network to enter the request).

- If study feasibility request, counts only will be provided but Children's sponsor must be identified.

- Start early.

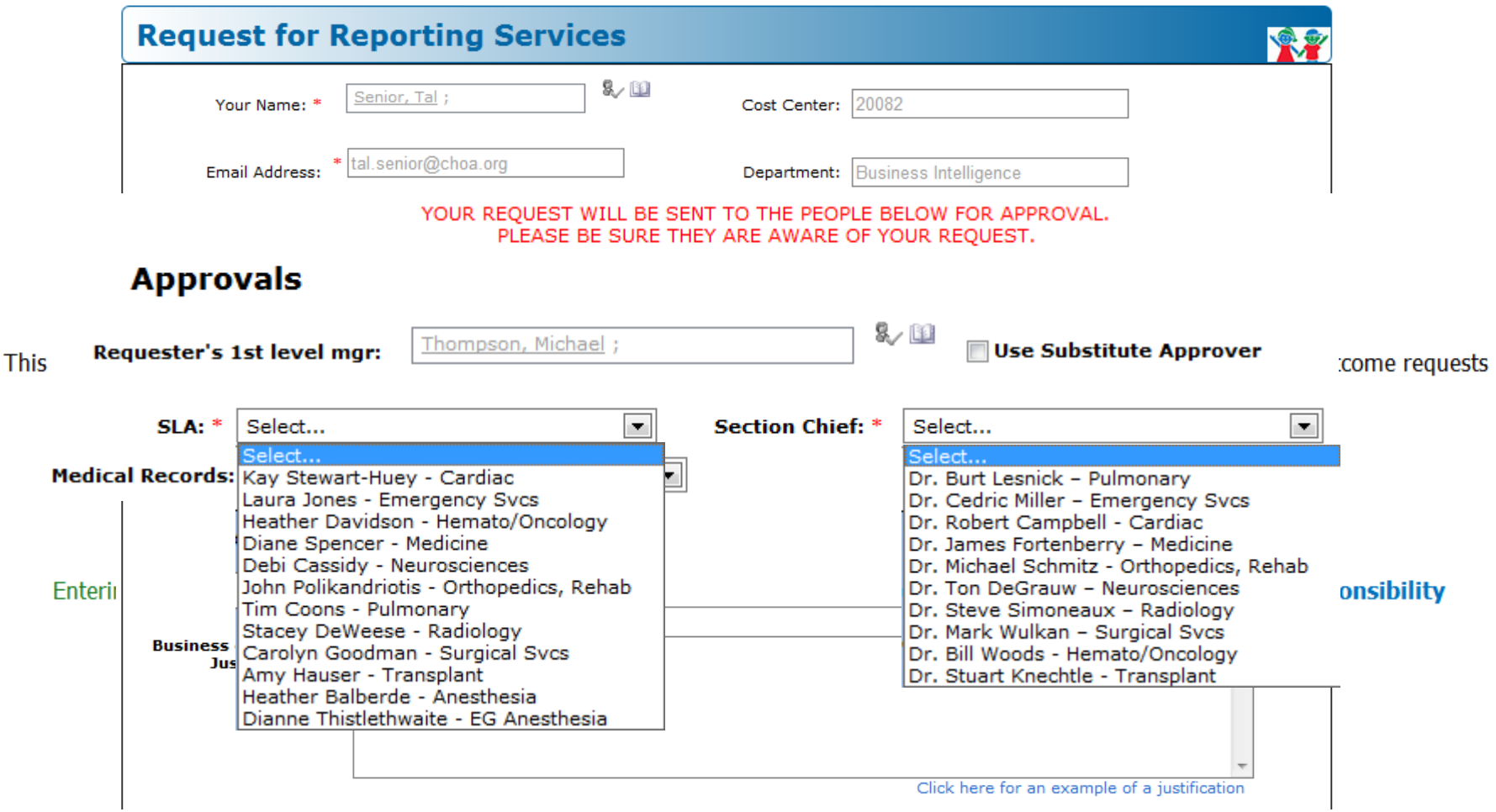
Requesting Data

1. Reporting Services Request:



- email: data@choa.org

Requesting Data



Examples

We would like to know how many occurrences of kangaroo care (aka skin-to-skin care) have taken place in our unit from December 2016 - March 2017. *–NICU Assistant Nurse Manager*

We need to build a report that captures all botox HB charges so that we can see if we also billed a PB administration code. *–Practice Manager*

Need to pull a list of addresses and guarantor names for patients associated with a request for privacy officer. *–Office of General Counsel*

We need a report that provides data on the Rapid Response Team calls. The report should include the patient's name and MRN#, location including campus and unit as well as PEWS score at the time of the Rapid Response team call, admit date and time, attending MD, diagnosis and problem list. Also, include the event notes at the time of the Rapid Response Team call if possible. *–Quality, Code Blue Committee*

Examples con't…

We need a report that displays iNO utilization throughout the organization. This will allow us to review all cost centers that utilize this gas. This data will be used for reconciliation and medical protocol by patient. We need to reconcile on a weekly basis to make sure the clinical team is looking at each patient as they are on and off the gas. –*Financial Operations*

AEA completion is required for all employees. Leaders require sufficient reports to monitor and enforce mandatory compliance.–*Learning Services*

This study aims to review metrics of efficiency in the treatment of children with thyroid nodules and compare the value delivered when preoperative tissue sampling (biopsy) is or is not employed

IRB# 14-197 Issued, 12.18.2014
Expires 12/17/2015, 300 subjects between 01/2007 and 6/2014
Approved by Dr. Wulkan

-General Pediatric Surgeon

When will I get my data?

- Clear requirements
- Provide a mock-up of data layout
- Provide examples

data@choa.org

Children's℠
Healthcare of Atlanta
Dedicated to All Better

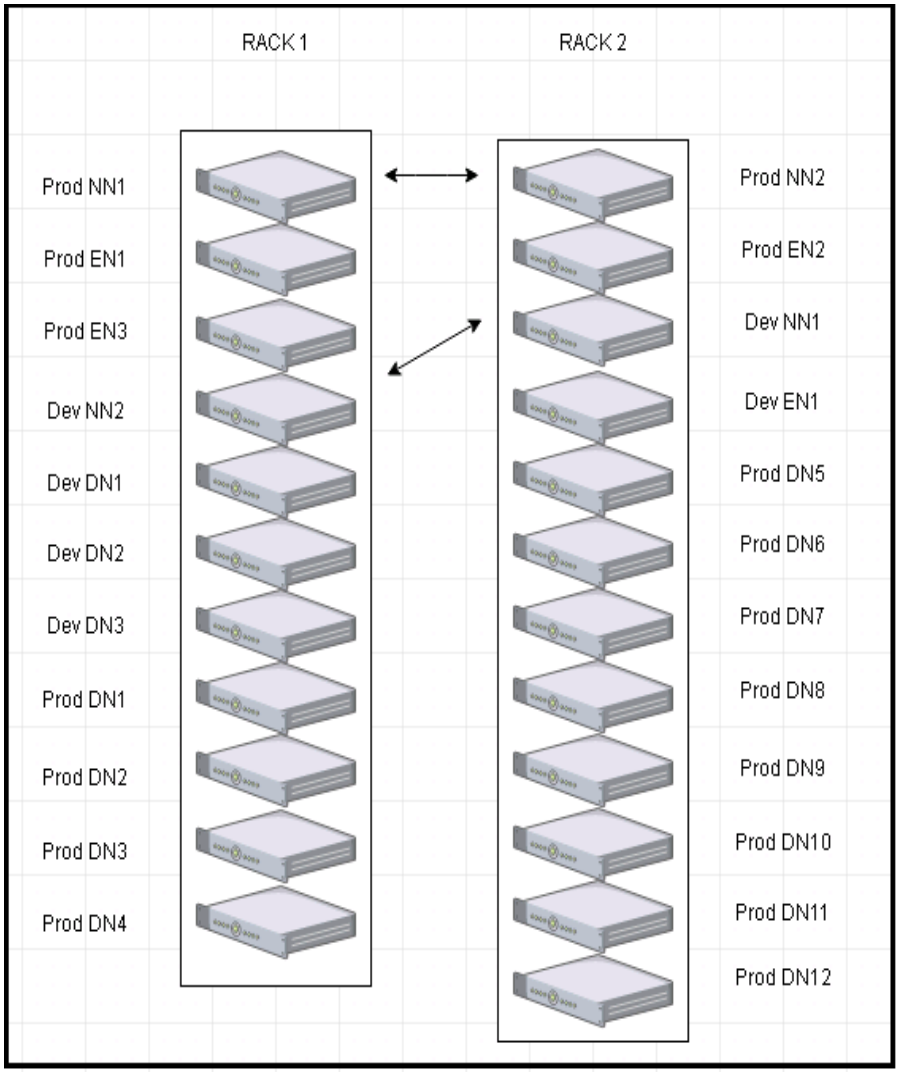
Children's Healthcare Of Atlanta

Tod Davis

CHOA IS&T, Business Intelligence

Hadoop

Hadoop Architecture



- Cisco UCS Big Data Cluster
- Cloudera Data Hub 5.10
- 553TB storage, 6TB Memory, 920 Processors
- Encryption on disk and over the wire. Kerberos and integrated with Active Directory

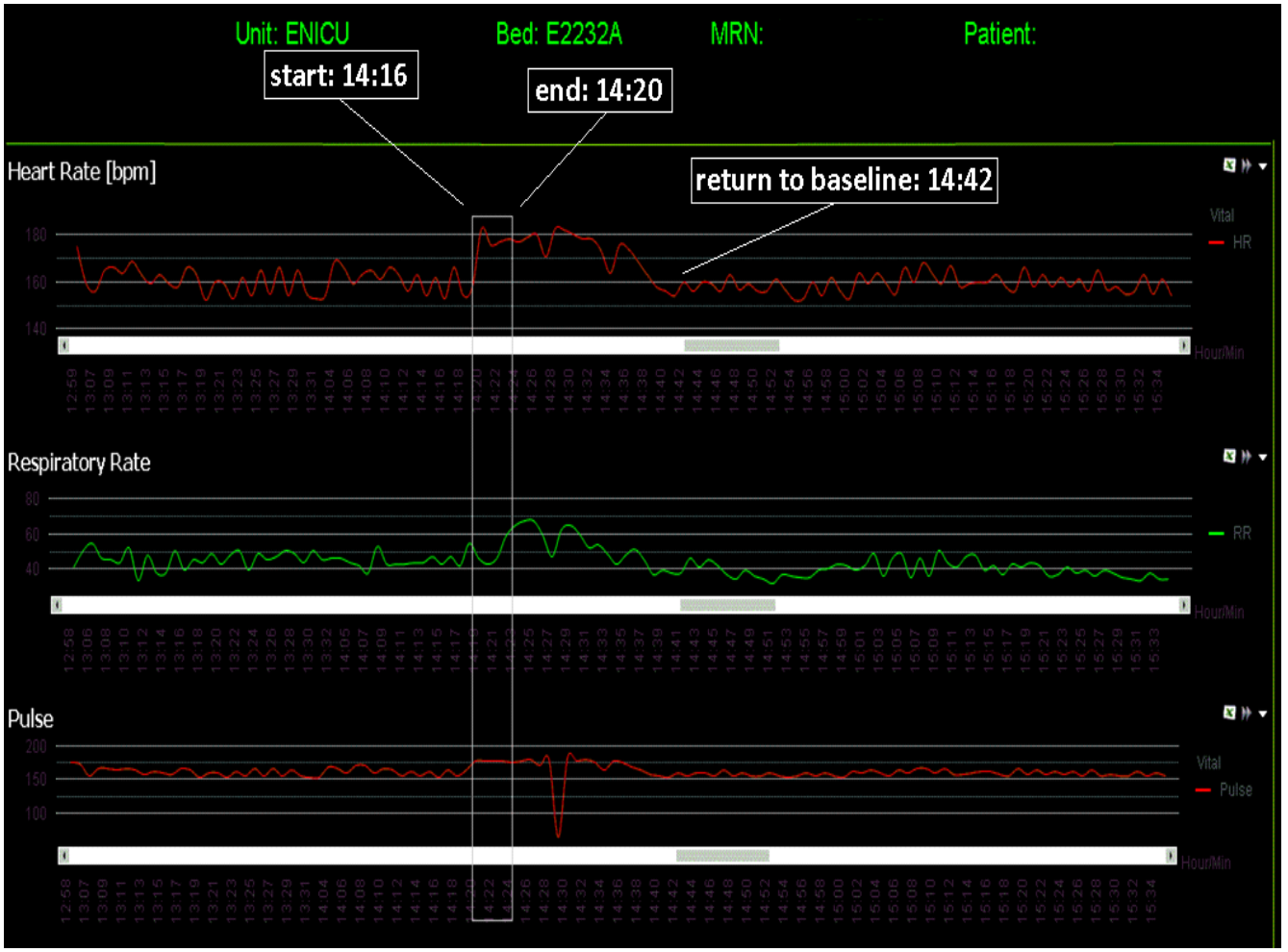
- Hadoop is a distributed storage and compute platform designed to be run on a cluster of "commodity" servers. Data storage and processing are distributed across the cluster. This ensures rapid data processing and reliability in the event of hardware failure.

Hadoop Data

- HBOC: 1985 – 1993
- SMS: 1993 – 2005
- KIDS: 2005 – present
- Epic Clarity – 2005 present
- ICU and OR Vitals 9/2013 - present
- CICU high frequency vitals and waveforms: 6/2016 - present
- EPA Georgia Air Quality: 1985 to present
- Genomic Sequence Data beginning 10/2017
  - Bacterial
  - Pharmacogenomic
  - Tumor

NICU Eye Exam Stress

- Vital Signs
- Epic Data

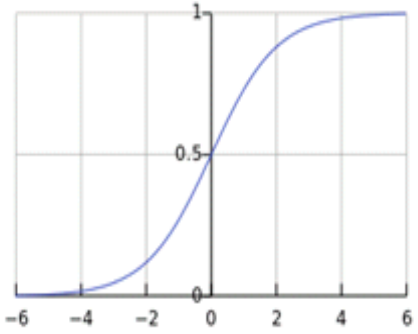


Clinician Notes and Natural Language Processing

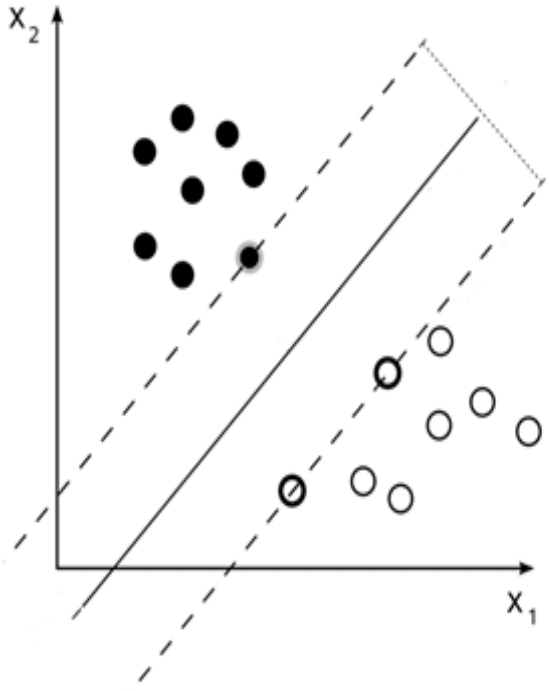
- Naïve Bayes

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}$$

- Logistic Regression



- Support Vector Machine



NLTK

ANACONDA

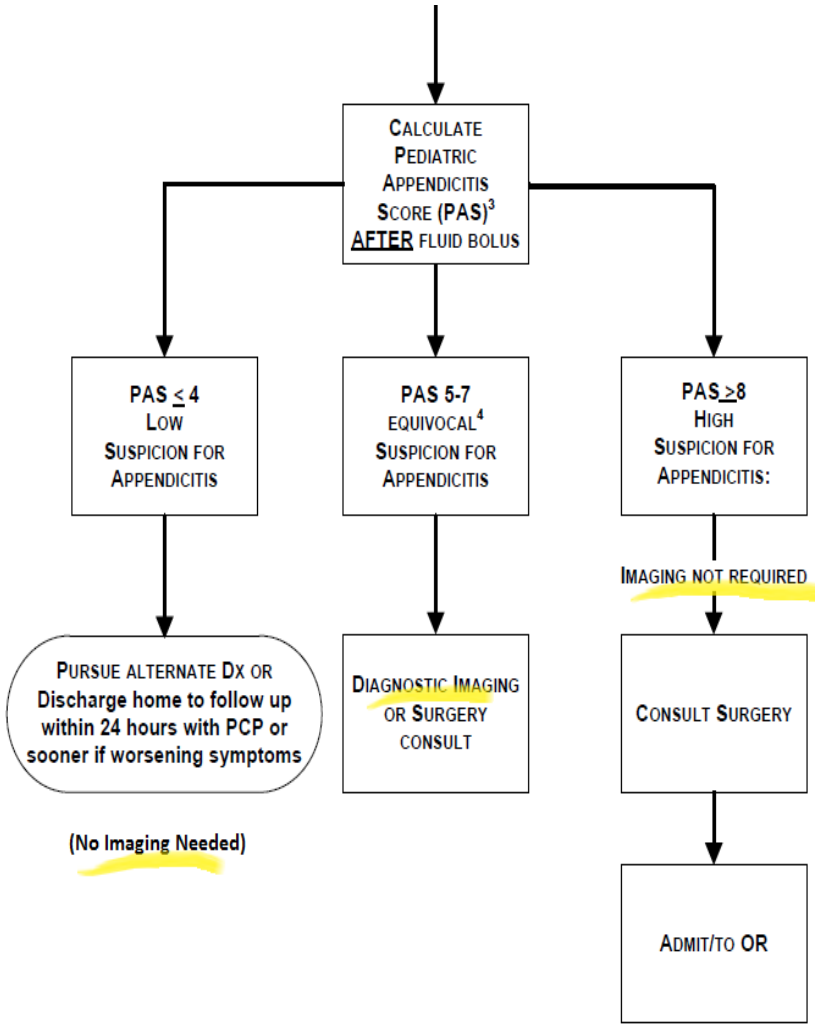
Clinician Notes and Natural Language Processing

- appendicitis score detection
- detect textual features with ML and NLP
- operationalize an NLP pipeline

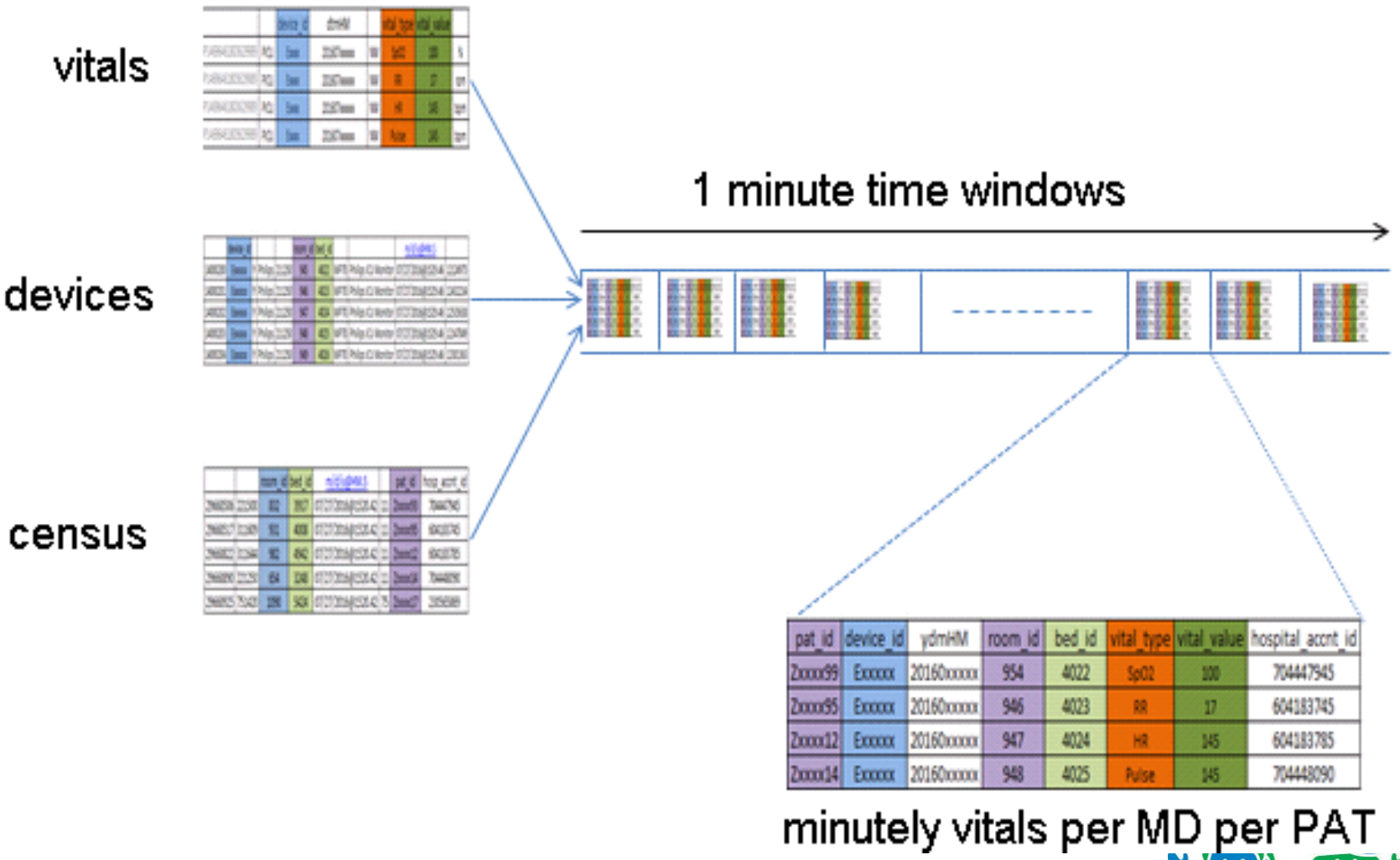
[3]Pediatric Appendicitis Score (PAS)*

TO BE PERFORMED BY MD ONLY

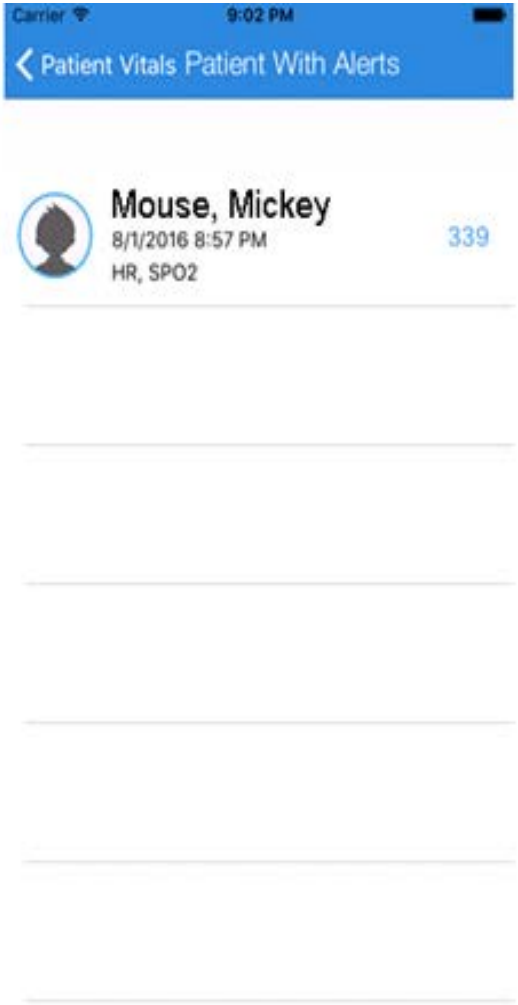
CLINICAL FINDING	POINTS
• MIGRATION OF PAIN FROM UMBILICUS TO RLQ	1
• COUGH/HOPPING/PERCUSSION TENDERNESS IN RLQ	2
• ANOREXIA	1
• ELEVATION OF TEMPERATURE (TEMP ≥38°C)	1
• NAUSEA/VOMITING	1
• LEUKOCYTOSIS (WBC>10,000MM^3)	1
• RLQ TENDERNESS	2
• DIFFERENTIAL WBC W/LEFT SHIFT (POLYMORPHONUCLEAR NEUTROPHILIA >7500/MM^3)	1
• **TOTAL:**	___



Spark Streaming and Near Time Mobile Alerting



Spark Streaming and Near Time Mobile Alerting



Thanks !

Questions ?

Health Analytics Group at Georgia Tech: Data in Action

Nicoleta Serban, PhD
Coca Cola Associate Professor
Julie Swann, PhD
Harold R. and Mary Anne Nash Professor

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Georgia Tech Health Analytics
Stewart School of Industrial & Systems Engineering

Georgia Tech ISyE
Stewart School of Industrial & Systems Engineering

Health Analytics at Georgia Tech

Georgia Tech

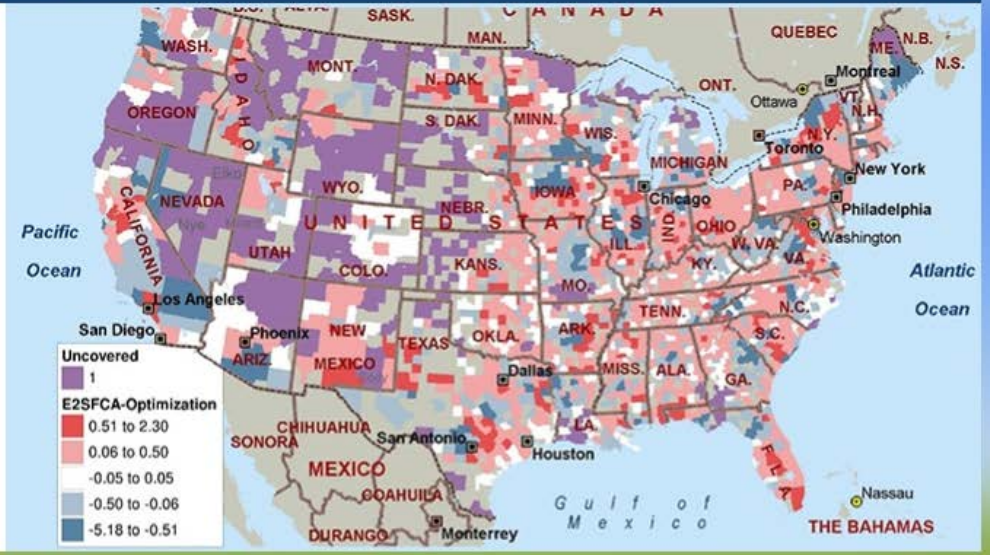
Georgia Tech Health Analytics
Stewart School of Industrial & Systems Engineering

Home About Focus Areas Data Tools Publications People CONTACT SEARCH

An optimization framework for measuring spatial access over healthcare networks

Measurement of healthcare spatial access over a network involves accounting for demand, supply, and network structure.

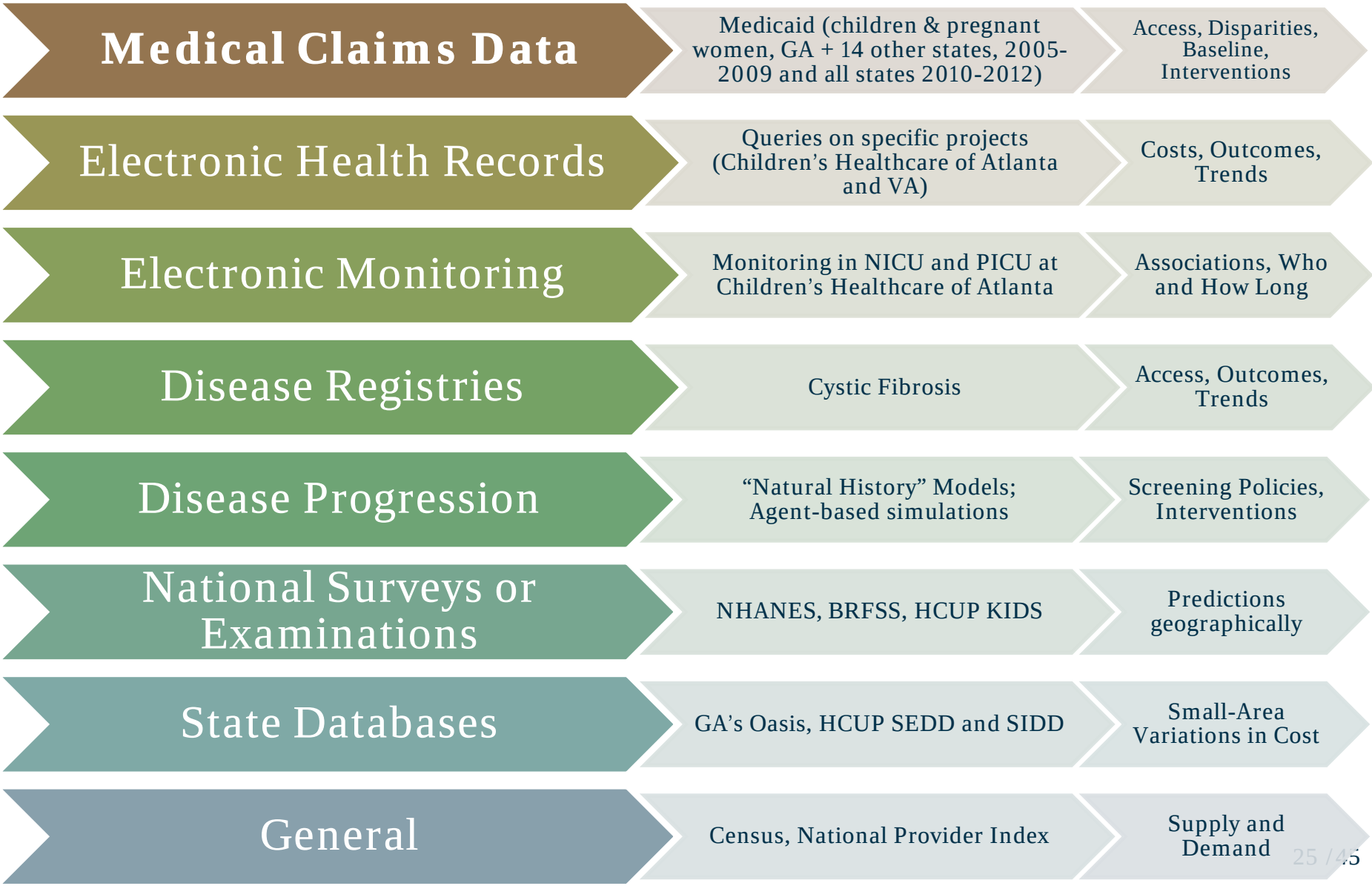
Read More



HEALTH ANALYTICS @ GEORGIA TECH

Health Analytics at Georgia Tech bridges fundamental mathematical and computational modeling with health services research and health economics as a means of translating health and healthcare data into knowledge and decision making.

Health Analytics: Data Landscape



CMS Medicaid Data

- MAX Claims Data Files
  - **Personal Summary**: patients, demographics, birthdate, etc.
  - **Inpatient**: claims, diagnoses, procedures, LOS, payment
  - **Other Therapy**: claims for physician, lab, clinic, outpatient
  - **Long Term Care**: facility type, date of service, etc.
  - **Prescription Drug**: paid drug claims
  - (National Provider ID & Characteristics File for 2009 forward)
- Years 2005 – 2009 for 14 states
  - SE: Georgia, Alabama, Arkansas, Louisiana, Mississippi, N. Carolina, S. Carolina, Tennessee, Texas & Other: California, Minnesota, New York, Pennsylvania
- Years 2010-2012 for all states available (46+)
- Data for 2013 available for purchase soon
- Terabytes of information (100's Billions of claims records), coded in CMS claims terminology, extensive data dictionary (420+ pages)
  - Highly complex: heterogeneous set of patients, multiple hierarchy levels (e.g. states), observational study, compounded dependencies
  - Patient-level Identifiable-Files requiring high levels of safeguards

CMS Medicaid: Approved Topics

1) MEASURING AND EXPLAINING INEQUITIES:

To assess the impact of healthcare system characteristics *vs.* inequities in healthcare, including *geographical*, *use*, *quality*, *expenditure and outcomes* among Medicaid children enrollees, especially in states with historic inequities like in the southeast.

a) To identify geographic areas with widespread and increasing Medicaid healthcare use (status quo and over time) and determine the underlying associative factors (e.g. access to healthcare facilities, race and ethnicity);

b) To investigate geographic variations in healthcare quality indicators (adherence to medications, emergency room visits, and other utilization measures) for high-impact diseases in children such as respiratory deficiencies, obesity, diabetes and other disabilities;

c) To identify geographic subdivisions which have achieved good health outcomes and low disparities despite adverse social determinants, or which have achieved poor health outcomes and high disparities worse than the social inequalities.

CMS Medicaid: Approved Topics

2) OPTIMIZING INTERVENTIONS AND DELIVERY SYSTEMS

To analyze flows and policies across the system, e.g., the match between supply and demand, and financially, both geographically and across time, along with the corresponding costs or outcomes, to analyze improved methods of delivery including medical homes.

a) To examine areas in the children Medicaid expenditure with the greatest costs or utilization, and assess potential interventions for reducing the healthcare costs, especially where interventions may be targeted by patient characteristics such as risk or where chronic issues like pediatric obesity can be addressed;

b) To evaluate the potential costs and benefits to creating a medical home or using telemedicine in the Medicaid system, where the creation may be focused on a subgroup of the Medicaid population or within specific geographical areas with great need;

c) To forecast the available "supply" of general or specialist providers or network services across geographical regions (e.g., counties or census tracts) as a function of socio-economic and other elements, link this factor with the costs or outcomes in the system as measured by the claims data, and examine potential interventions.;

d) To evaluate the impact of various public policies, such as changes in cost-sharing, on the demand for Medicaid coverage.

CMS Medicaid Data: Access to Data

- Data stored in secure location at Georgia Tech, with access to the detailed data by a limited set of GT employees approved by CMS and IRB
- Massive data files, with technology infrastructure for efficient access
- Sharing of aggregated data is allowed with collaborators, if consistent with research purposes and research protocol
  - Cells should have at least 11 entries
  - Data undergoes review process at GT before release from data workstation
- Significant liability involved if breach occurs
- Data management plan revised for easier maintenance

Comments about MAX Data

- Limitations
  - Research must fit within scope proposed to CMS
  - Analysis of raw data must be conducted at GT
  - Process for analyzing data is onerous, time-consuming, and "expensive"
  - Data only covers Medicaid claims
  - Does not include school absenteeism data
  - Data never includes the most recent (~2) years of data
- Positives
  - We can benchmark GA against 13 other states
  - Patients and/or providers can be followed longitudinally
  - Includes provider visits and prescription claims
  - No one has to give permission for us to ask specific questions or publish related answers

Contact

Nicoleta Serban, nserban@isye.gatech.edu

Julie Swann, jswann@isye.gatech.edu

Georgia Tech Health Analytics
Stewart School of Industrial & Systems Engineering

http://healthanalytics.gatech.edu