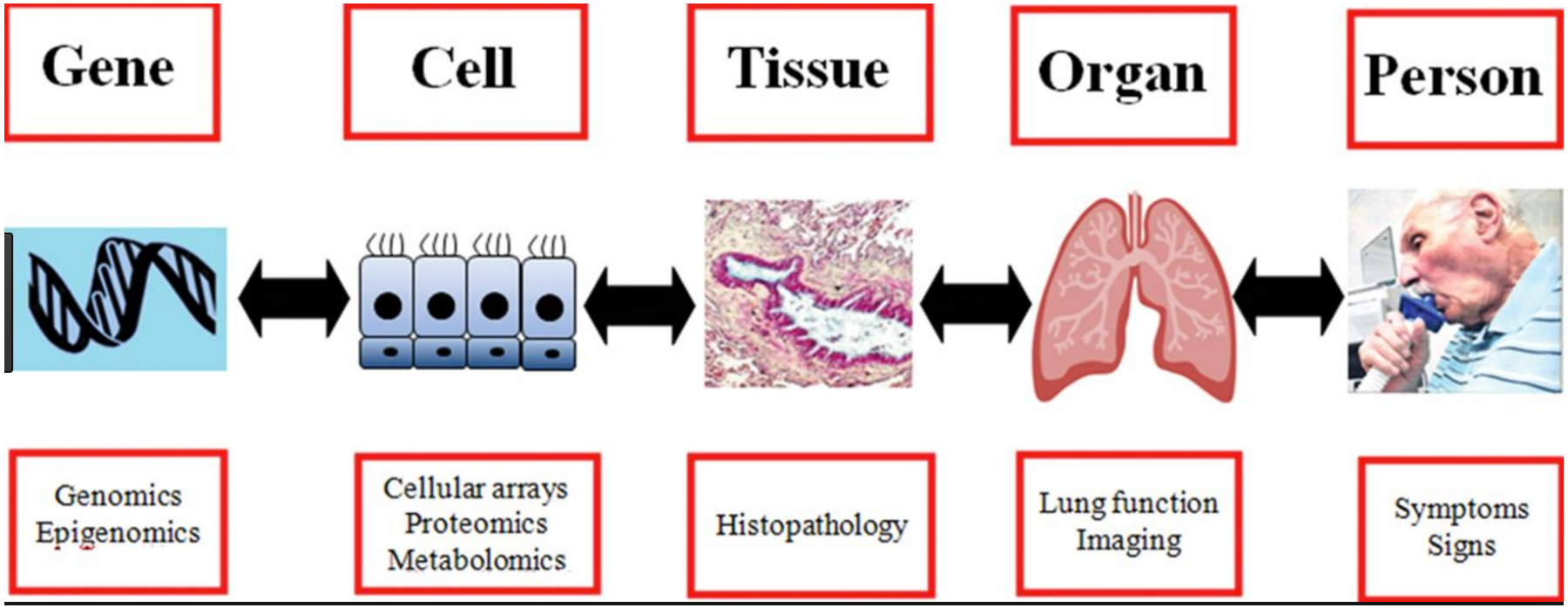
EHR-Based Phenotyping: Bulk Learning and Evaluation (with Infectious Diseases)

Po-Hsiang (Barnett) Chiu

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Phenotypes and phenotyping



Physically observable traits of genotypes (and their interactions with environments)

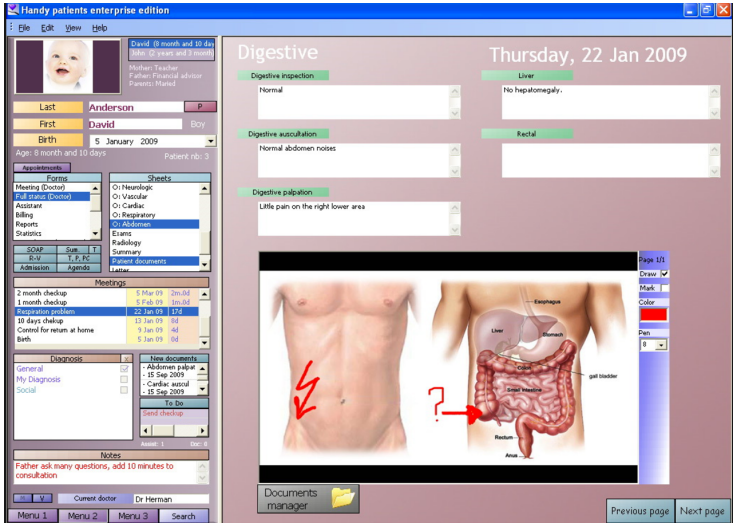
Biochemical or physiological properties, behavior, and products of behavior

Attributions of diseases (e.g. susceptibility)

Diseases (and disease subtypes)

Data-Driven Phenotyping

- Data-driven phenotyping
  - Two main methodologies
    - Rule-based approach (e.g. eMerge, https://emerge.mc.vanderbilt.edu)
    - **Predictive Analytics**
  - Data sources:
    - EHRs/EMRs: Medicinal treatments, diagnoses, lab measurements, etc.
    - Genomic data: SNP arrays, copy number variation (CNVs), etc.
  - Phenotypes
    - Diseases, subtypes, or variables attributed to disease predictions



Diagnostic Concept Units

- Various diseases sharing the same set of diagnostic concept units
- Infectious diseases
  - Lab tests
    - Microorganism, blood, urine, body tissues, stool
  - Medications
    - Antibiotic, antivirus, anthelmintic
- Build statistical models for each diagnostic component and combine them appropriately
  - Ensemble learning

Bulk Learning in a Nutshell …

Bulk Learning is a **batch-phenotyping framework** that uses multiple diseases collectively (i.e. bulk learning set) as a substrate for model learning and evaluation wherein (a given) medical ontology is used to perform feature selection and model stacking is used to construct abstract feature representation of low sample complexity in order to reduce training requirements.

Key Concepts:

1. Build phenotyping models on top of multiple diseases
2. Automatic feature selection using an existing ontology
3. Models are combined via model stacking (a form of ensemble learning)
4. Abstract features

   Dimensionality reduction

5. Less labeled data required for model evaluations

Phenotyping via Bulk Learning

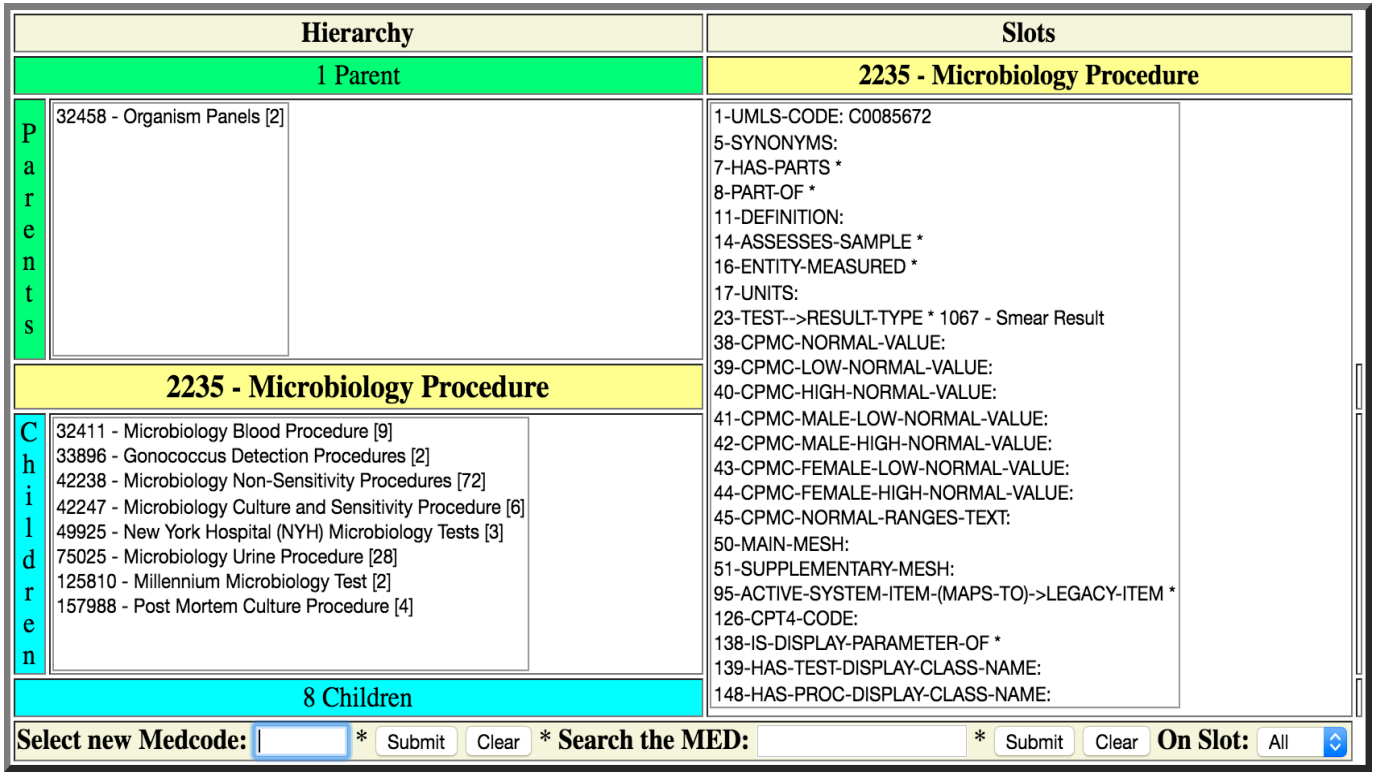
- Under model stacking, we then arrive at the notion of "concept-driven phenotyping"
  - A subset or combinations of lab tests are more attributable to some diseases while the others are better explained by medications
- In this study, infectious diseases associated with 100 ICD-9 codes as the domain of study for bulk learning
  - For simplicity, consider different diagnostic codes as different diseases …
  - Why 100 codes?
  - Code selection strategy?

Bulk Learning Basics I

- Addresses two central issues in predictive analytical approach to computational phenotyping
  - Feature engineering
    - Medical ontology for feature decomposition
    - Medical Entities Dict (http://med.dmi.columbia.edu)
  - Data annotation
    - Ensemble learning (e.g. stacked generalization [Wolpert 1992])
    - Feature abstraction for dimensionality reduction

Medical Ontology for Grouping Features

- Snapshot of Medical Entities Dictionary (http://med.dmi.columbia.edu)



Hierarchy

1 Parent

Parents

32458 - Organism Panels [2]

2235 - Microbiology Procedure

Children

32411 - Microbiology Blood Procedure [9]
33896 - Gonococcus Detection Procedures [2]
42238 - Microbiology Non-Sensitivity Procedures [72]
42247 - Microbiology Culture and Sensitivity Procedure [6]
49925 - New York Hospital (NYH) Microbiology Tests [3]
75025 - Microbiology Urine Procedure [28]
125810 - Millennium Microbiology Test [2]
157988 - Post Mortem Culture Procedure [4]

8 Children

Slots

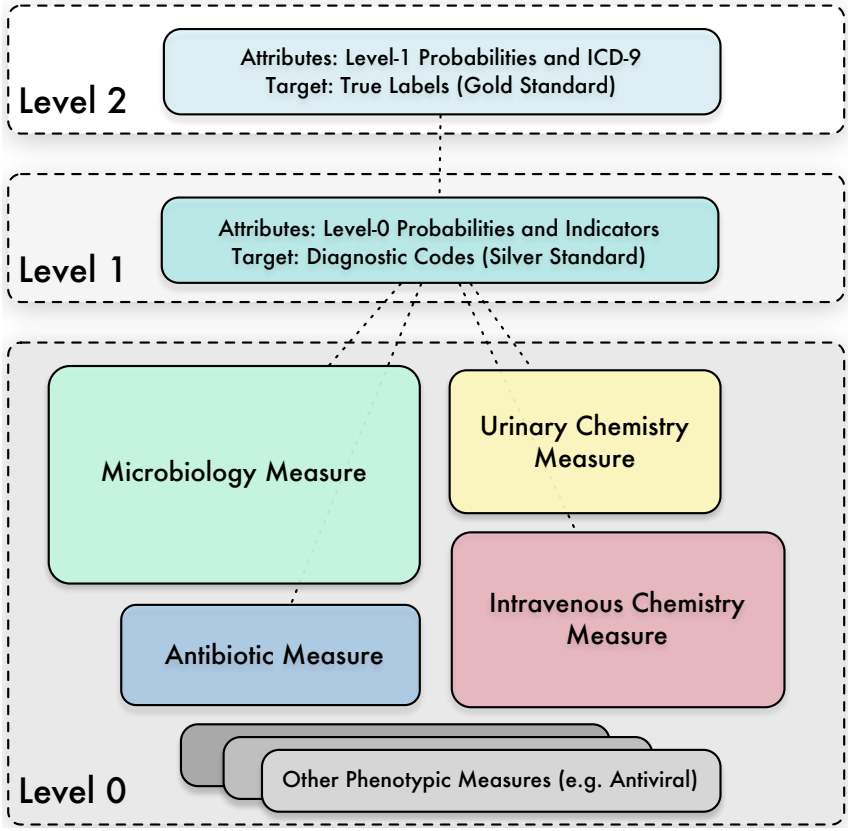
2235 - Microbiology Procedure

1-UMLS-CODE: C0085672
5-SYNONYMS:
7-HAS-PARTS *
8-PART-OF *
11-DEFINITION:
14-ASSESSES-SAMPLE *
16-ENTITY-MEASURED *
17-UNITS:
23-TEST-->RESULT-TYPE * 1067 - Smear Result
38-CPMC-NORMAL-VALUE:
39-CPMC-LOW-NORMAL-VALUE:
40-CPMC-HIGH-NORMAL-VALUE:
41-CPMC-MALE-LOW-NORMAL-VALUE:
42-CPMC-MALE-HIGH-NORMAL-VALUE:
43-CPMC-FEMALE-LOW-NORMAL-VALUE:
44-CPMC-FEMALE-HIGH-NORMAL-VALUE:
45-CPMC-NORMAL-RANGES-TEXT:
50-MAIN-MESH:
51-SUPPLEMENTARY-MESH:
95-ACTIVE-SYSTEM-ITEM-(MAPS-TO)->LEGACY-ITEM *
126-CPT4-CODE:
138-IS-DISPLAY-PARAMETER-OF *
139-HAS-TEST-DISPLAY-CLASS-NAME:
148-HAS-PROC-DISPLAY-CLASS-NAME:

Select new Medcode: * Submit Clear * **Search the MED:** * Submit Clear **On Slot:** All

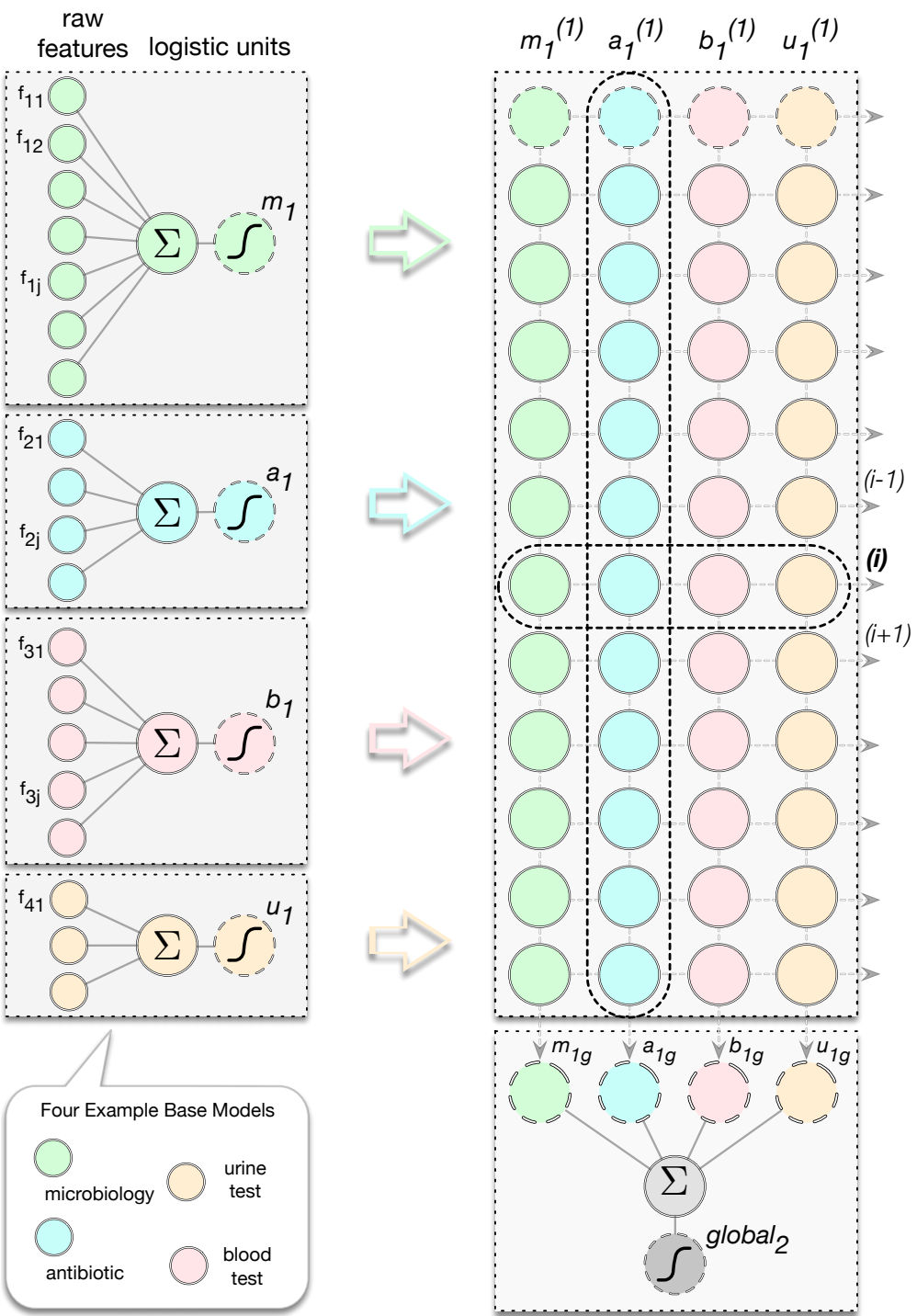
Model Stacking

- Why inspecting multiple (infectious) diseases?
  - Using **multiple diseases** as substrate and identify their common elements
  - Example stacking architecture (under stacked generalization method)



Surrogate Labels vs True Labels

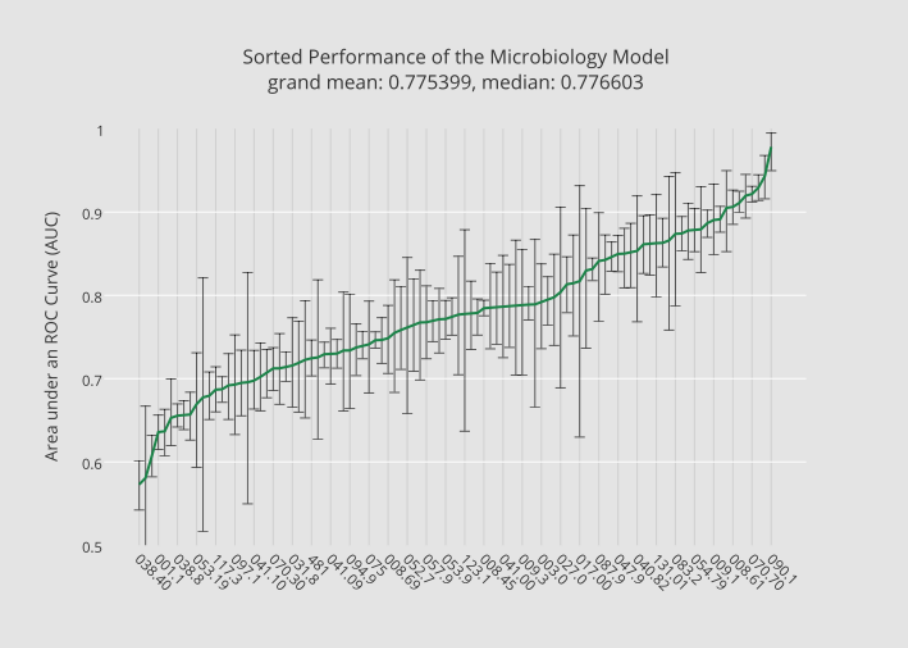
- Model stacking is used to achieve:
  - Improve upon base model performances
  - Transform EHR data to a denser form
- Uses diagnostic codes (e.g. ICD-9) as surrogate labels to establish "approximate predictive models."
- Why surrogate labels (e.g. ICD-9)?
  - Features extracted from EHR can be large
  - Used to derive compact representation of the training data
  - "Free" supervised signals that are sufficiently close but can be obtained without extra work
- Objective: Build statistical models in abstract feature space
  - Create a **sparse** annotation set (i.e. gold standard) that serves a proxy dataset for downstream model evaluations
  - 83 annotated cases

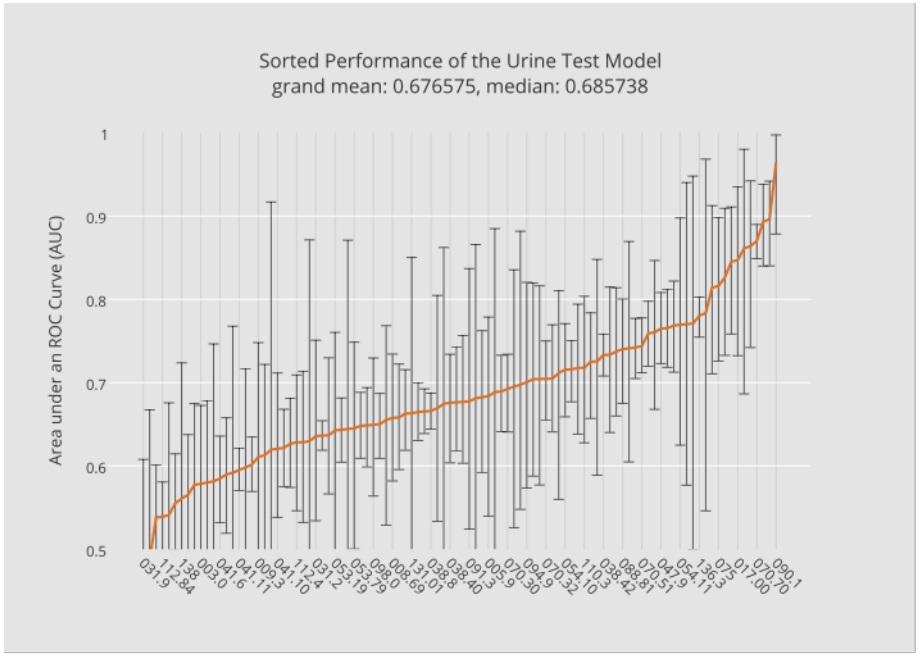

raw features
logistic units
f_{11}
f_{12}
f_{1j}
m_1
f_{21}
f_{2j}
a_1
f_{31}
f_{3j}
b_1
f_{41}
u_1
$m_1^{(1)}$
$a_1^{(1)}$
$b_1^{(1)}$
$u_1^{(1)}$
(i-1)
(i)
(i+1)
m_{1g}
a_{1g}
b_{1g}
u_{1g}
$global_2$
Four Example Base Models
microbiology
urine test
antibiotic
blood test

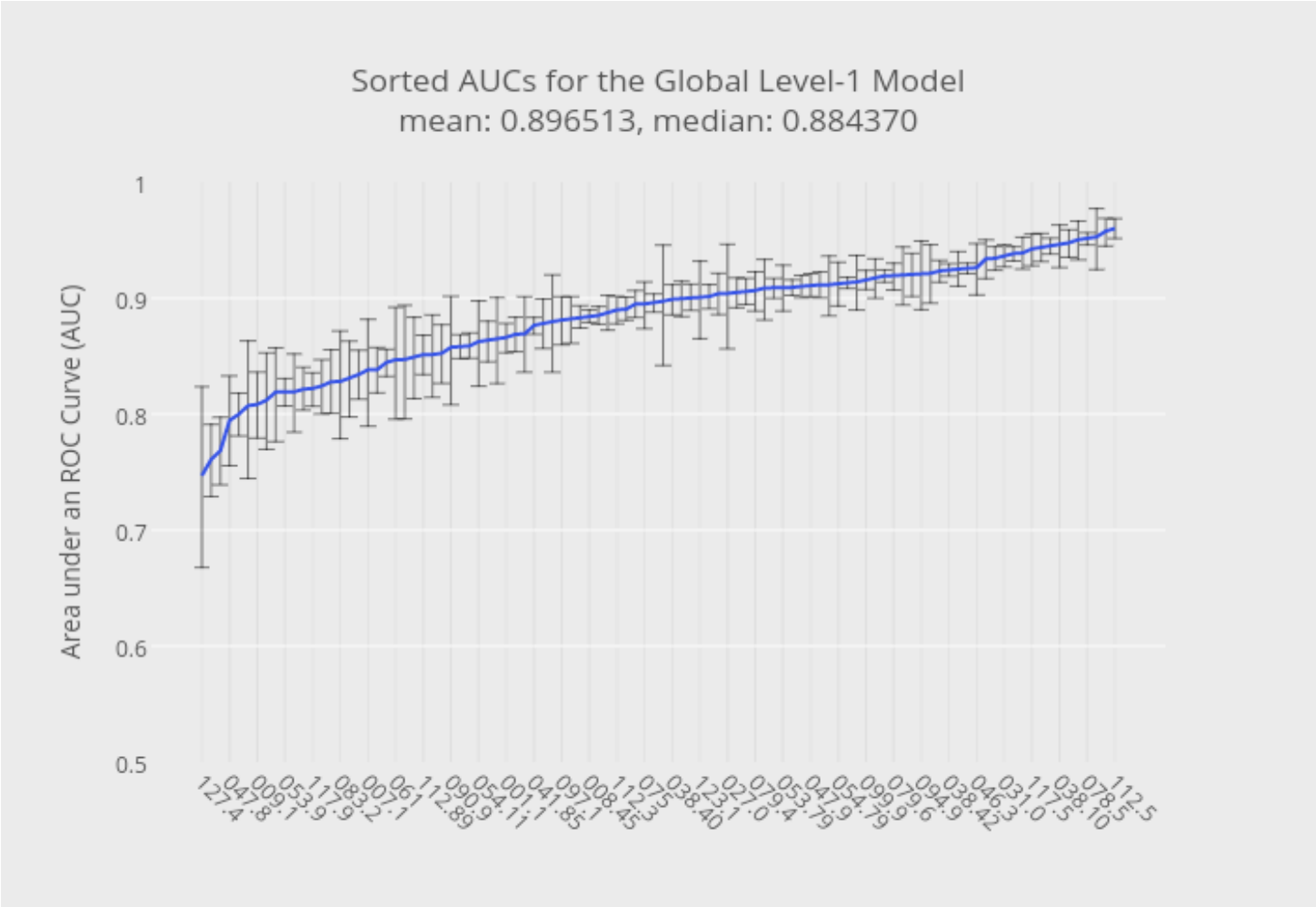
Performance Evaluations

- How well does the model predict ICD-9s (using a separate test data)?
- How well does the model predict annotated data (assoc. with “true labels”)?
  - (Binarized) ICD-9 becomes a candidate feature among abstract features (e.g. probability scores, indicators)
    - Annotated sample consists of randomly selected cases in which errors of ICD-9 coding are corrected
    - Data annotations and coding procedures are two independent processes

Base Level Performances







127.4 Enterobiasis

009.1 Gastroenteritis ...

117.9 Mycoses

047.8 (Other) viral meningitis

053.9 Herpez zoster

Settings	Sensitivity	Specificity	Mean AUC (Repeated 10-fold with 30 cycles)
Level 1 (L1)	1029/1170 (0.88)	212/1320 (0.16)	0.59 (0.51 ~ 0.66)
Level 2 (L2)	812/1170 (0.69)	456/1320 (0.35)	0.52 (0.45 ~ 0.60)
L1 + ICD9	1158/1170 (0.99)	771/1320 (0.58)	0.85 (0.80 ~ 0.89)
L2 + ICD9	910/1170 (0.78)	836/1320 (0.63)	0.74 (0.67 ~ 0.82)
Big Logistic	768/1170 (0.66)	866/1320 (0.66)	0.65 (0.59 ~ 0.72)
Big SVM	784/1170 (0.67)	862/1320 (0.65)	0.53 (0.51 ~ 0.56)

Table 7b. Comparison by annotation types among different meta-classifiers trained by mixing virtual annotations.

Settings	Type TP (39)	Type FP (15)	Type TN (29)	Type FN (0)
Level 1 (L1)	1029/1170 (0.88)	102/450 (0.23)	110/870 (0.13)	n/a
Level 2 (L2)	812/1170 (0.69)	158/450 (0.35)	298/870 (0.34)	n/a
L1 + ICD9	1158/1170 (0.99)	10/450 (0.02)	761/870 (0.87)	n/a
L2 + ICD9	910/1170 (0.78)	104/450 (0.23)	732/870 (0.84)	n/a
Big Logistic	768/1170 (0.66)	276/450 (0.61)	590/870 (0.68)	n/a
Big SVM	784/1170 (0.67)	291/450 (0.65)	571/870 (0.66)	n/a

Other Components

- Semi-supervised learning and virtual annotation set

- The 3rd tier in model stacking hierarchy
  - Trade-off between learned abstract features and the ICD-9 codes as surrogate labels.
  - Performance evaluation on predicting annotated labels

- Ontology-based feature engineering
- Proper design of treatment and control (training) data

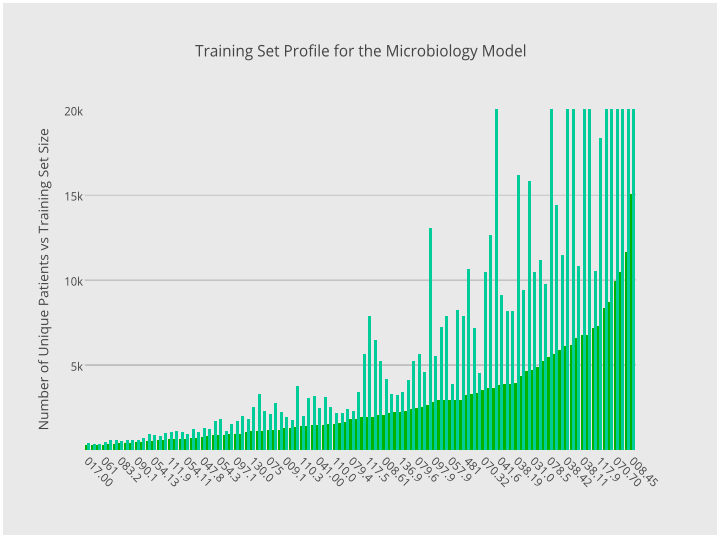
Modeling Perspective

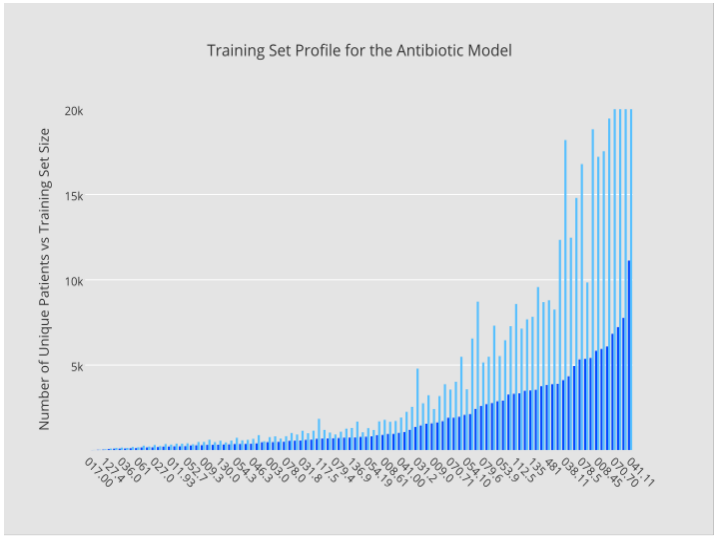
- EHR data consist of observations and latent variables
  - Observations can be directly answered via simple queries
    - Did the patient have tests on E. Coli?
    - Did the patient take Ceftriaxon?
- Latent variables represent quantities that cannot be directly observed in EHR or computed via simple queries
  - Does the patient have an infection?
  - Diagnostic questions: specifically which infections do the patient have?
- Learn classifiers to predict latent variables (with only access to observations)

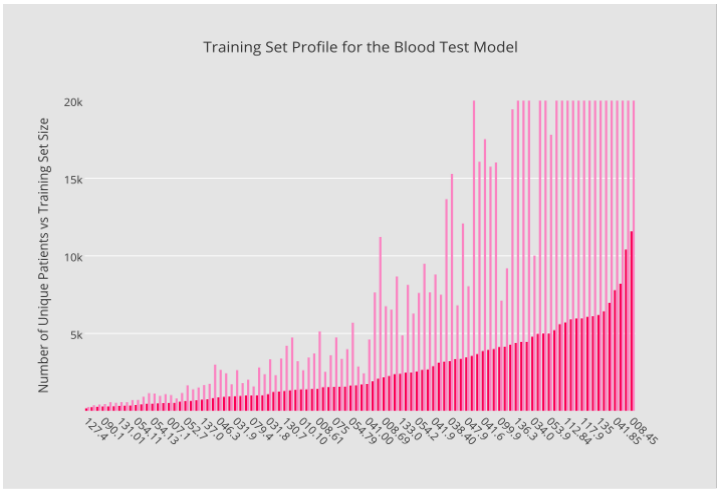
Medical Perspective

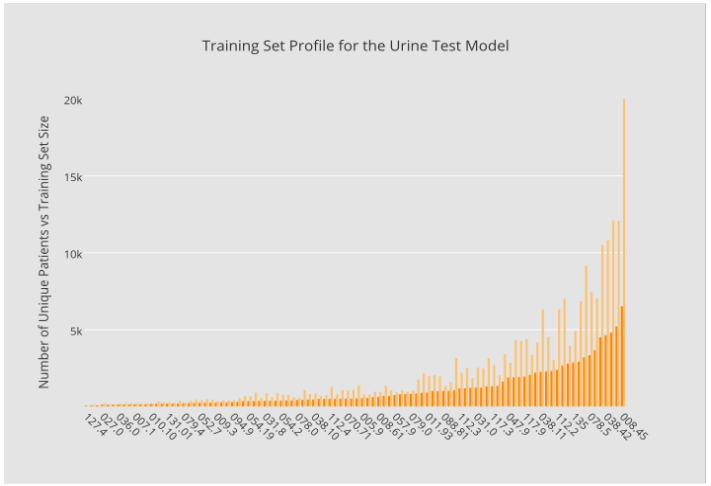
- Seemingly different infectious diseases may share similar sets of lab tests and medications
  - Staph. aureus
    - Skin infections, pneumonia, blood poisoning
  - Ceftriaxone
    - Meningitis
    - Infections at different sites of the body (e.g. bloodstream, lungs, urinary tracts)
- Multiple classifiers for the same disease
  - 4 classifiers per ICD-9 code, each of which is binary classifier
    - 400 classifiers at base level

Data Distribution Perspective









"Can we build a joint model applicable to all diseases?"

Abstract Feature Representation: Design Choices

- Related work in constructing high-level features
  - PCA, unsupervised feature learning, manifold learning, etc.
- Design choices
  - Data characteristics
  - Interpretability

- Deep Neural Network
  - Linear combination
  - Non-linear transformation (e.g. sigmoid, rectifier, etc.)
- Feature set: continuous, dense, and "homogeneous"
  - Image pixels
  - Times series of lab measurements
  - word2vec

- EHR data however are very different
  - sparse and incomplete
  - consist of many different types (binary, categorical, continuous, etc.)
  - Features associated with multiple concepts

Moving Forward …

- Summary
  - Bulk learning is a framework with at least the following system choices
    - The bulk learning set (of target conditions) => base models
    - Classification algorithms (guideline: probabilistic classifiers + well-calibrated)
    - Stacking architecture (multiple tiers => levels of abstractions)
    - Strategy for combining individual (local) disease models to a global model
  - Advantage: Can use a small annotated sample for model construction and evaluation within the abstract feature space (e.g. level-1 data)
    - 83 clinical cases were labeled in this study
  - Challenge: The model involving the interaction between abstract features and ICD-9 do not generalize well into the region of the data where the ICD-9 coding was incorrect
    - Multiple types of surrogate labels

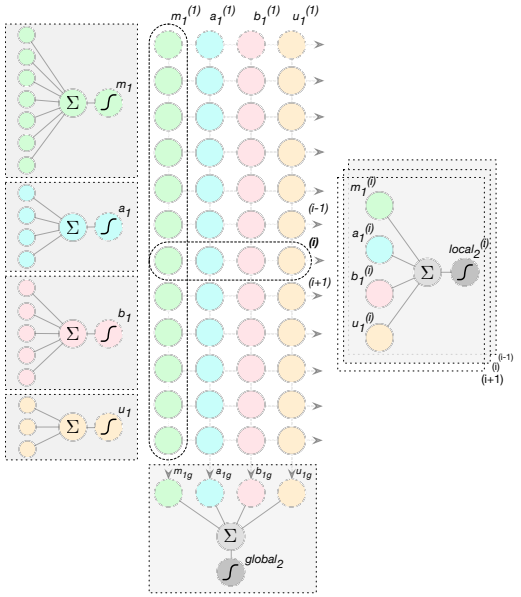
- Ongoing and future work

  Complex decision boundary?

  Other surrogate labels

  Semi-supervised learning

  Active learning



Reference

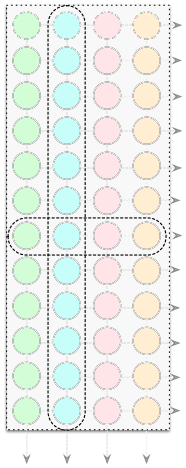
[1] D.H. Wolpert, Stacked generalization, Neural Networks. 5 (1992) 241–259.

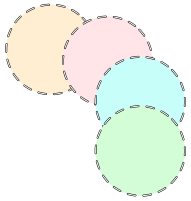
[2] K.M. Ting, I.H. Witten, Issues in stacked generalization, J. Artif. Intell. Res. 10 (1999) 271–289.

[3] J. Jin Chen, C. Cheng Wang, R. Runsheng Wang, Using Stacked Generalization to Combine SVMs in Magnitude and Shape Feature Spaces for Classification of Hyperspectral Data, IEEE Trans. Geosci. Remote Sens. 47 (2009) 2193-2205.

[4] David Baorto, James Cimino, et al.
Available: http://med.dmi.columbia.edu. Access date: Oct 20, 2016.

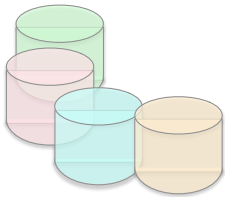
[5] T.A. Lasko, J.C. Denny, M.A. Levy, Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data, PLoS One. 8 (2013) e66341.

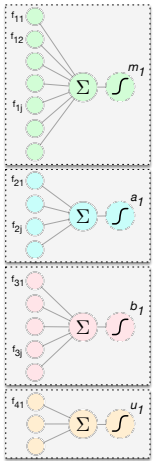


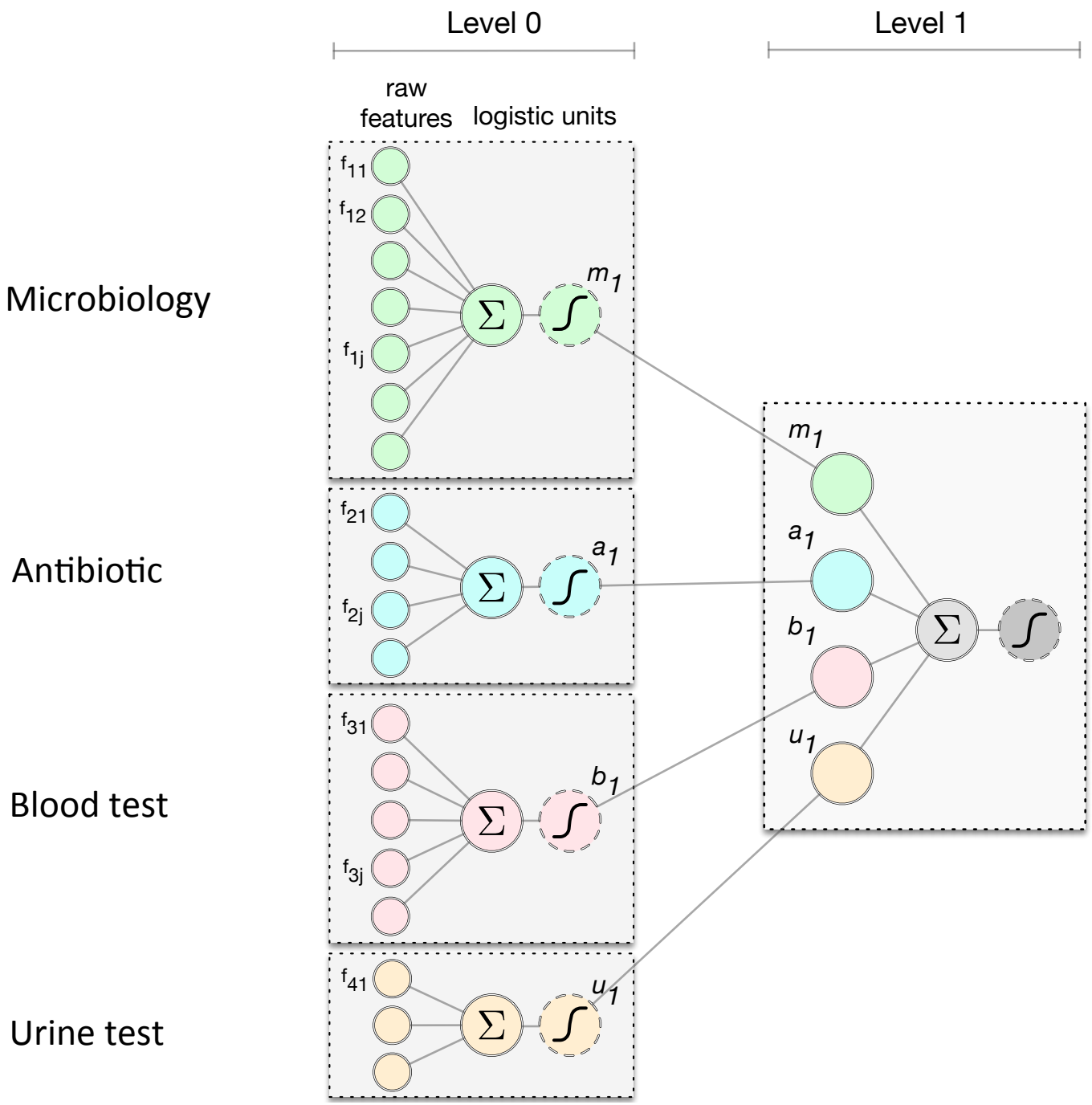


T H A N K

Y O U






Level 0
Level 1
raw features
logistic units
Microbiology
Antibiotic
Blood test
Urine test
f_{11}
f_{12}
f_{1j}
m_1
f_{21}
f_{2j}
a_1
f_{31}
f_{3j}
b_1
f_{41}
u_1
m_1
a_1
b_1
u_1
Σ
$\int$

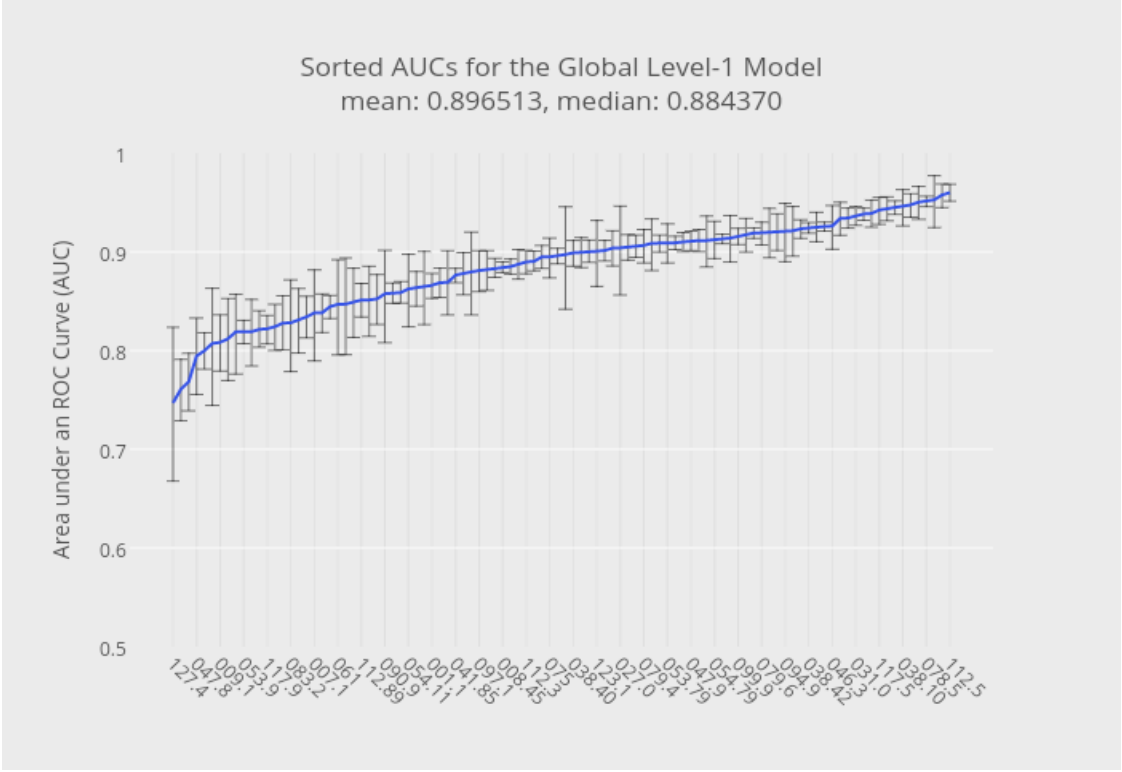
Example Features

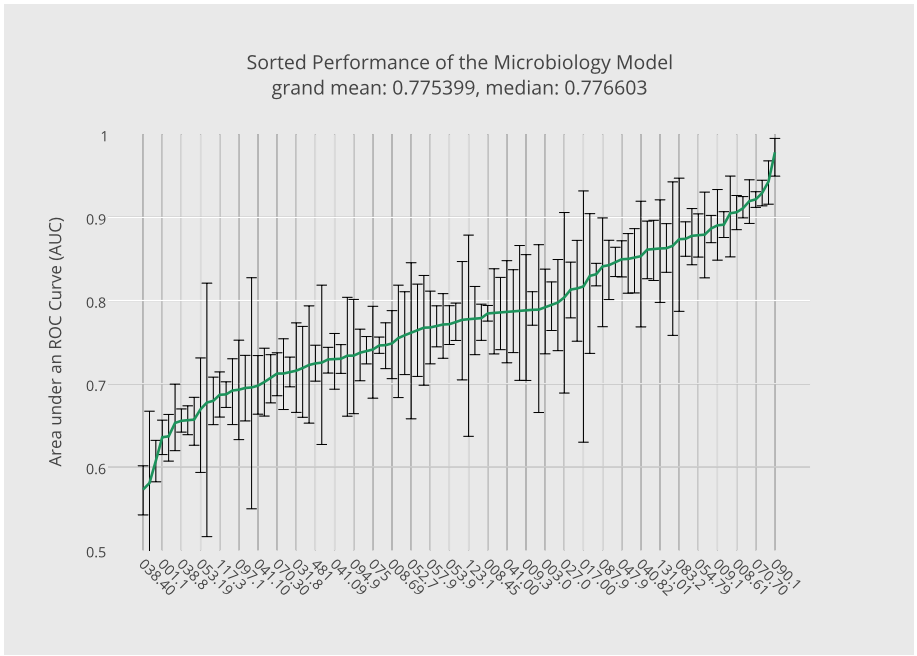
Microorganism Lab Test (Microbiology)	
MedCode	Description
935	Organism Result: Escherichia Coli
799	Organism Result: Candida Albicans
774	Organism Result: Staphylococcus Aureus
910	Organism Result: Klebsiella Pneumoniae
31826	Organism Result: Enterococcus Faecalis
59993	Negative for Clostridium Difficile Toxin A and Toxin B
39576	Rule Out Influenza Virus
316	No Ova or Parasites Found
994	Positive for Gram Negative Rods
36453	Susceptibility Type: Microscan Mic

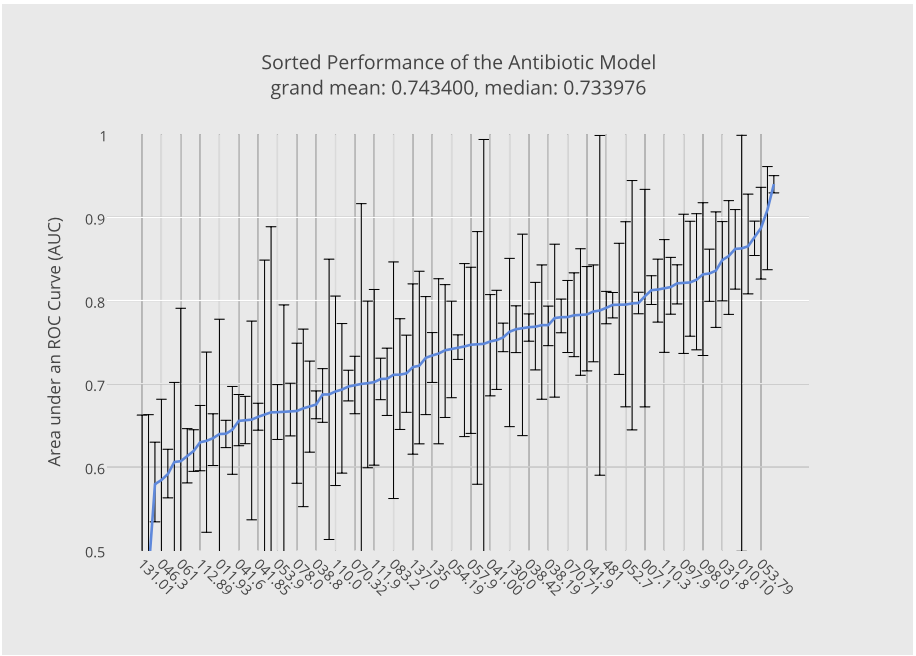
Antibiotic Prescription (Antibiotic)	
MedCode	Description
72900	Piperacillin/Tazobactam
72702	Vancomycin
100198	Ceftriaxone
66042	Levofloxacin
61003	Tobramycin
60671	Azithromycin
62375	Meropenem
61461	Amoxicillin
60918	Dapsone
62879	Cephalexin

Intravenous Chemistry Test (Blood)	
MedCode	Description
69494	Lab Test: Vitamin B12
35995	Lab Test: Lactate, Arterial
39564	Lab Test: Cyclosporine, Whole Blood
65906	Lab Test: Hemoglobin A1c
36300	Lab Test: Vancomycin
59415	Lab Test: Tacrolimus
46418	Blood Bank: ABO Antigen Determination
46421	Blood Bank: Antierythrocyte Antibody Screen
59942	Lab Test: Glucose Wholeblood
59047	Lab Test: Creatine Kinase

Urinary Chemistry Test (Urine)	
MedCode	Description
36265	Lab Test: Ketone
36267	Lab Test: Potassium, Random Urine
36260	Lab Test: Urine Glucose
36269	Lab Test: Urine Leukocyte Esterase
36286	Lab Test: Urine Protein
1390	Urine Blood Test
1395	Urine pH Measurement
1388	Urine Urobilinogen Test
1394	Urine Albumin Test
1392	Urine Acetone Test

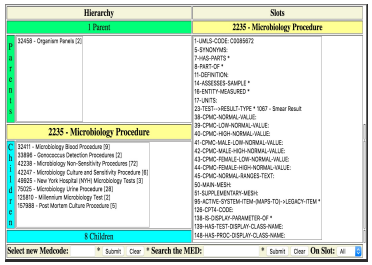

Sorted AUCs for the Global Level-1 Model
mean: 0.896513, median: 0.884370
Area under an ROC Curve (AUC)
1
0.9
0.8
0.7
0.6
0.5

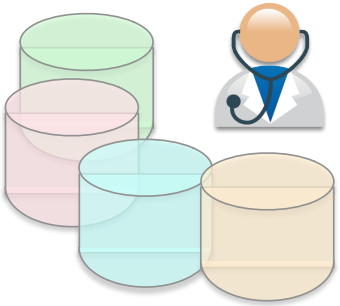

Sorted Performance of the Microbiology Model
grand mean: 0.775399, median: 0.776603
Area under an ROC Curve (AUC)
1
0.9
0.8
0.7
0.6
0.5


Sorted Performance of the Antibiotic Model
grand mean: 0.743400, median: 0.733976
Area under an ROC Curve (AUC)
1
0.9
0.8
0.7
0.6
0.5

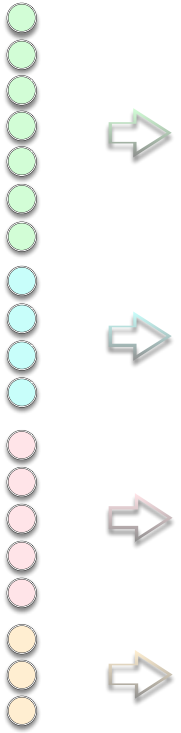
1. Define Feature Groups Using Medical Ontology

1b. Use **Medical Entities Dictionary** to delineate feature scopes



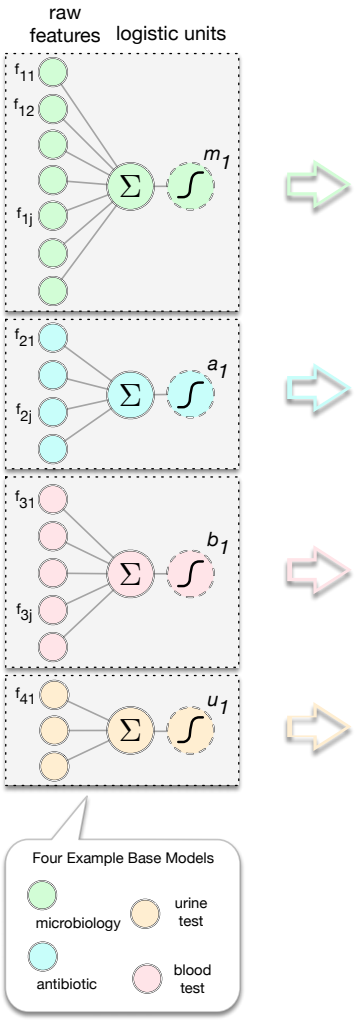


1a. Gather **EHR data** according to medical concepts

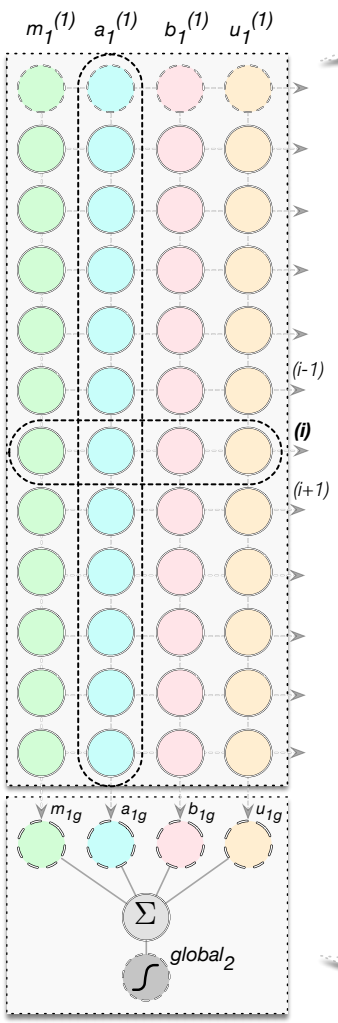


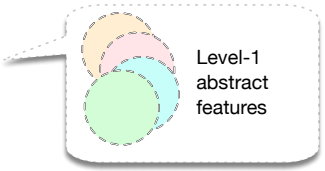
1c. Apply **feature selection** within each concept group

2. Compute Base Models

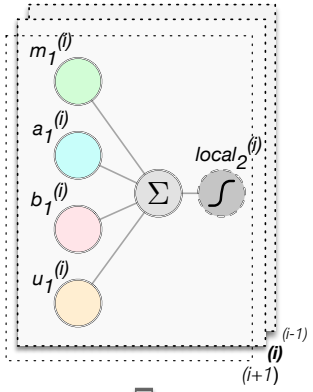


3. Compute Meta Models (via Ensemble Learning)

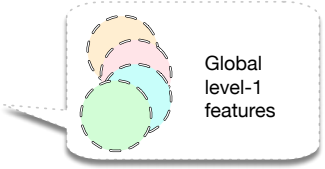




3a. Per-disease ensembles: compute local level-1 models



3b. Cross-disease ensemble: compute a global level-1 model



Individual Level-1 Local Units

Level-1 Global Unit