Machine Learning Case Study:

Prediction of Neurodevelopmental Disorders from the Electronic Health Record

June 27, 2020

MMCi Applied Data Science

Matthew Engelhard

Perinatal and Neonatal Factors (# studies)	Results Across Studies	Summary Effect Estimate (95% CI)
Presentation		
Abnormal presentation (15)	10–, 5↑	1.44 (1.07–1.94)
Breech (4)		1.81 (1.21–2.71)
Other perinatal factors		
Cord complications (14)	13–, 1↑	1.50 (1.00–2.24)
Fetal distress (4)	3–, 1↑	1.52 (1.09–2.12)
Birth injury or trauma (6)	6–	4.90 (1.41–16.94)
Twins or multiple birth (10)	7–, 3↑	1.77 (1.23–2.55)
Maternal hemorrhage (4)	3–, 1↑	2.39 (1.35–4.21)
Birth weight and size		
Total birth weight (decreased) (15)	12–, 2↑, 1↓	
Low birth weight (<2500 g) (15)	8–, 7↑	1.63 (1.19–2.33)
Small for gestational age (10)	7–, 3↑	1.35 (1.14–1.61)
Clinical impression		
Congenital malformation (11)	4–, 7↑	1.80 (1.42–2.82)
Apgar score		
Low 5-minute Apgar score (8)	6–, 2↑	1.67 (1.24–2.26)
Neonatal Status		

PEDIATRICS®
OFFICIAL JOURNAL OF THE AMERICAN ACADEMY OF PEDIATRICS

Review Article

Perinatal and Neonatal Risk Factors for Autism: A Comprehensive Meta-analysis

Hannah Gardener, Donna Spiegelman and Stephen L. Buka

Pediatrics August 2011, 128 (2) 344-355; DOI: https://doi.org/10.1542/peds.2010-1036

Research Letter

FREE

Nov 2012

Prenatal and Perinatal Risk Factors for Attention-Deficit/Hyperactivity Disorder

Jochen Schmitt, MD, MPH; Marcel Romanos, MD

Author Affiliations | Article Information

Arch Pediatr Adolesc Med. 2012;166(11):1074-1075. doi:10.1001/archpediatrics.2012.1078

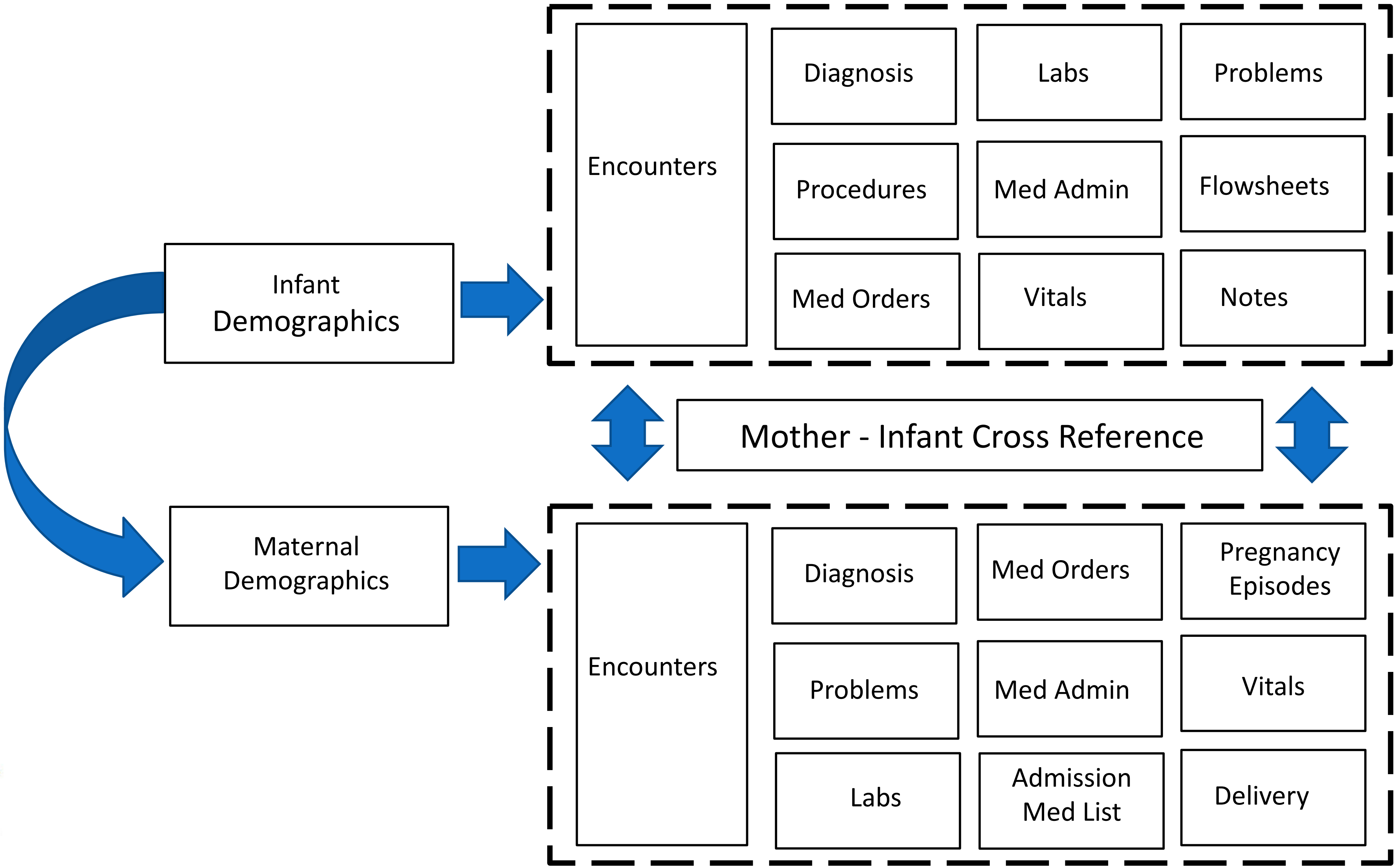
Table. Sample Characteristics and Risk Factors of ADHD in Children and Adolescents

Characteristic/Exposure (Reference for Regression Analyses)	Sample Characteristics, No. (%)[a]		Logistic Regression Analysis, OR (95% CI)	
	Children With ADHD (n = 660)	Children Without ADHD (n = 12 828)	Bivariable (Unadjusted) Analysis	Multivariable (Adjusted) Analysis[b]
Sex (reference: female)	133 (20.2)	6604 (51.5)	4.20 (3.47-5.10)	4.42 (3.56-5.49)
Age, y, mean (SD)	9.8 (4.3)	11.3 (3.4)	1.08 (1.06-1.11)	1.09 (1.07-1.11)
Socioeconomic position[c]				
Upper class (reference)	114 (17.4)	3486 (27.3)	1 [Reference]	1 [Reference]
Middle class	325 (49.5)	6087 (47.7)	1.63 (1.31-2.03)	1.57 (1.23-2.00)
Lower class	218 (33.2)	3202 (25.1)	2.08 (1.65-2.62)	2.04 (1.56-2.68)
Maternal gestational diabetes mellitus (reference: absent)	24 (4.1)	256 (2.2)	1.93 (1.26-2.95)	1.91 (1.21-3.01)
Maternal smoking during pregnancy (reference: never)	158 (24.6)	2081 (16.4)	1.66 (1.38-2.00)	1.48 (1.19-1.84)
Maternal alcohol consumption during pregnancy (reference: never)	96 (14.8)	1775 (14.0)	1.07 (0.86-1.34)	1.02 (0.79-1.33)
Perinatal health problems (reference: absent)[d]	235 (36.2)	2955 (23.2)	1.88 (1.60-2.22)	1.69 (1.40-2.03)
Breastfeeding (ever vs never fully breastfeeding)	345 (56.7)	7943 (67.5)	0.63 (0.54-0.74)	0.83 (0.69-0.996)
Atopic eczema (ever vs never)	132 (20.2)	1820 (14.4)	1.51 (1.24-1.84)	1.62 (1.30-2.02)

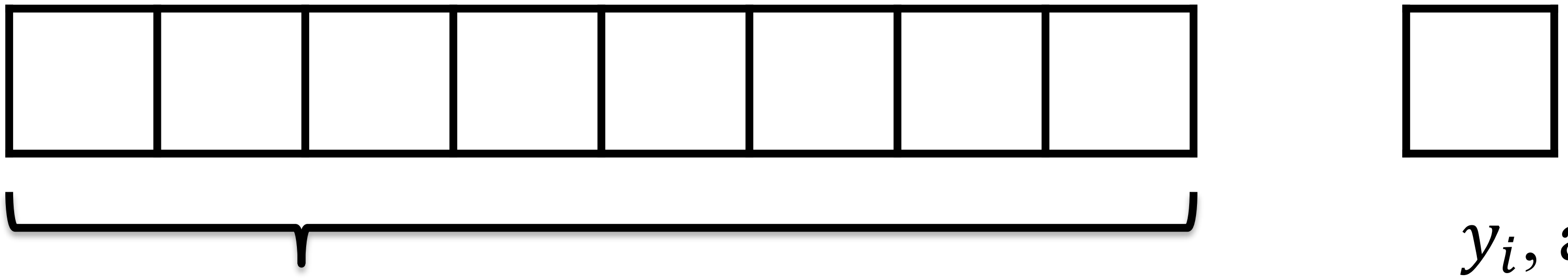
Abbreviations: ADHD, attention-deficit/hyperactivity disorder; OR, odds ratio.
[a] Numbers represent number (proportion) of children per exposed for discrete variables and means (SD) for continuous variables.
[b] Adjusted for all exposures listed in the Table; analysis based on 11 222 observations without any missing data.
[c] Classified based on parental education, professional qualification, professional status, and family net income according to Winkler and Stolzenberg.[4]
[d] Breathing problems, maladaptation, infections, icterus, low birth weight/premature delivery, and/or inpatient treatment.

DukeHealth
ADHD Program

ML of EHR for ASD and ADHD Risk Assessment



A Simplified Approach…

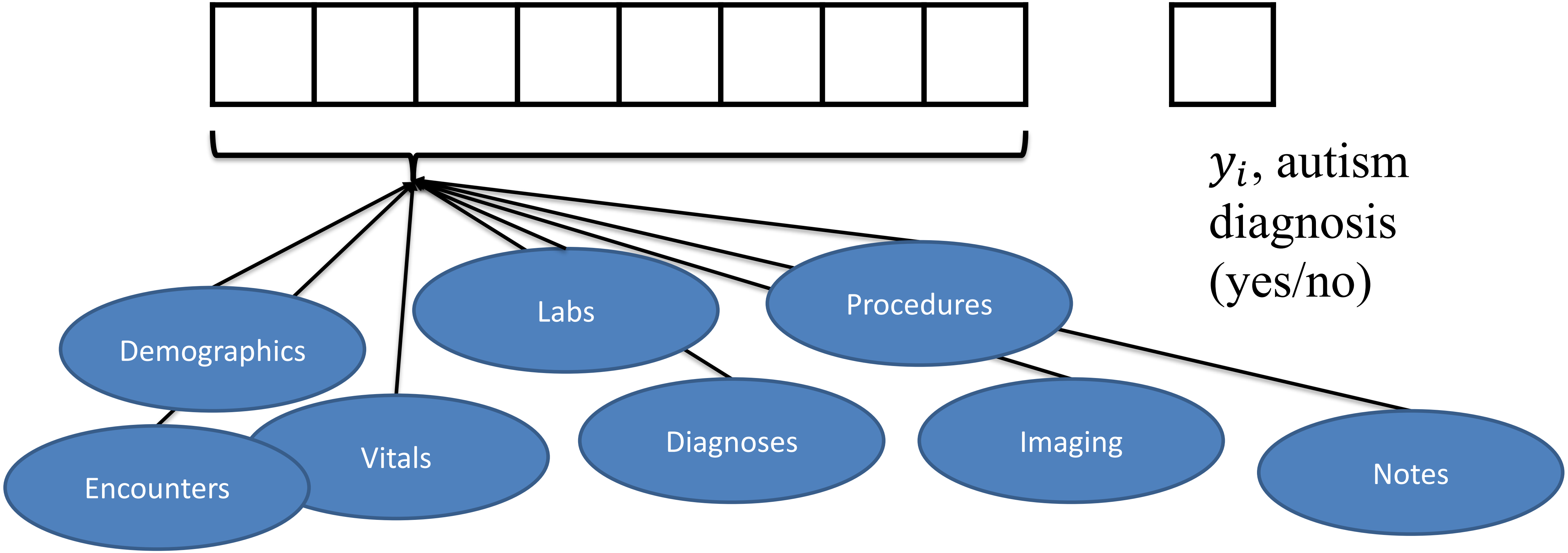


x_i, data/features for patient i

y_i, autism diagnosis (yes/no)

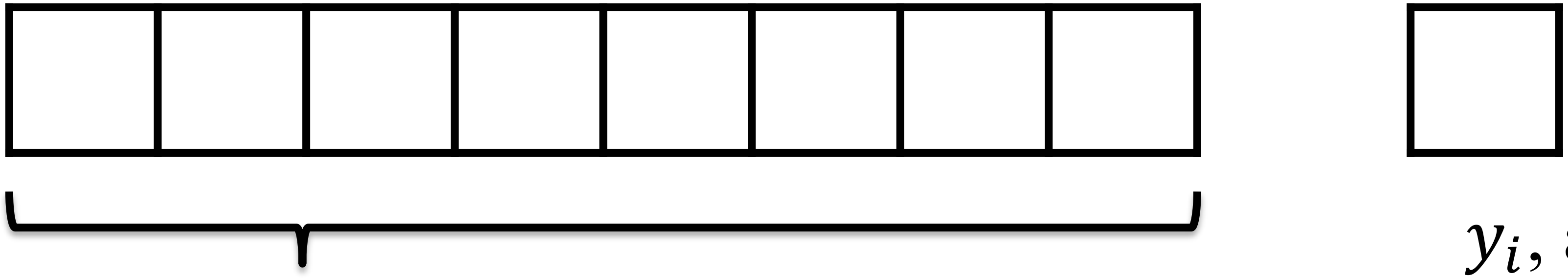
End goal: predict y_i from x_i

A Simplified Approach…



y_i, autism diagnosis (yes/no)

What simplifications are we making?

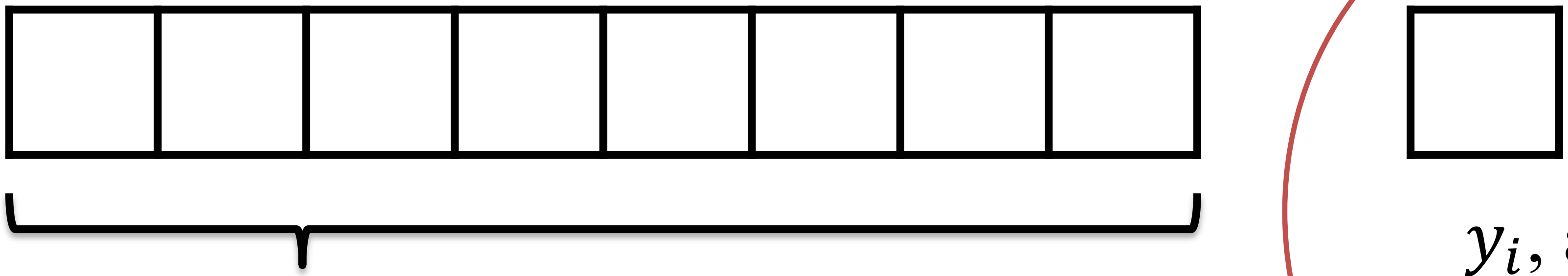


x_i, data/features for patient i

y_i, autism diagnosis (yes/no)

End goal: predict y_i from x_i

What simplifications are we making?

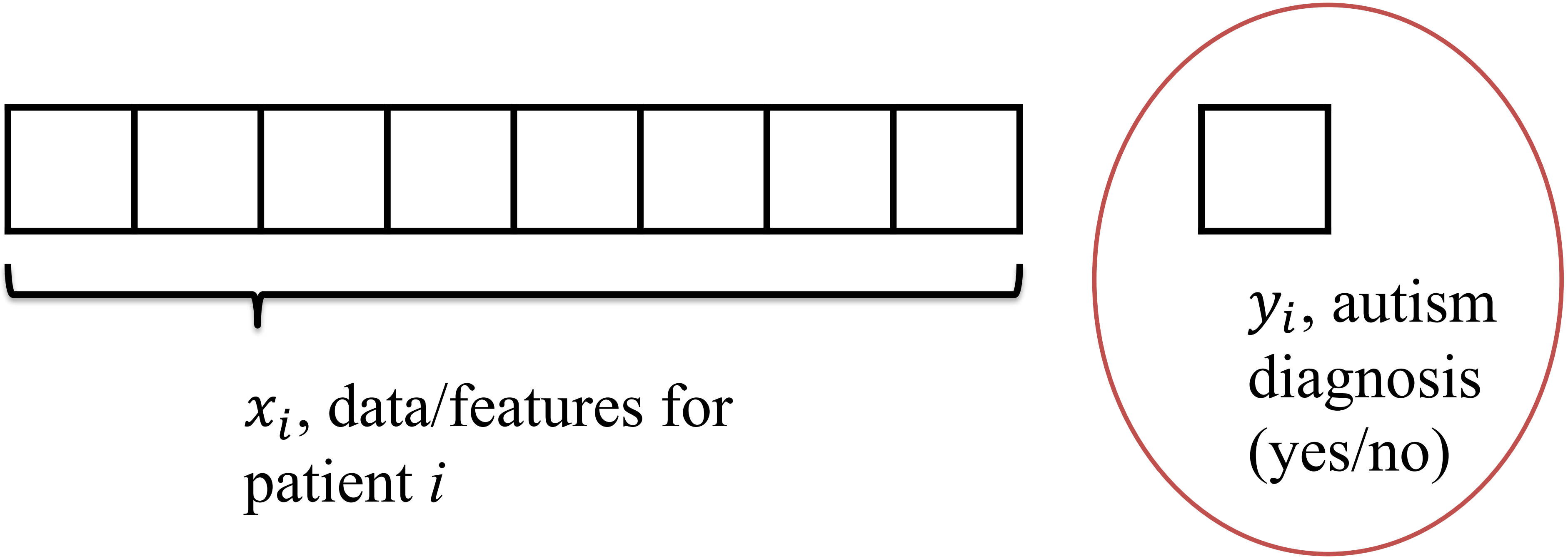


x_i, data/features for patient i

y_i, autism diagnosis (yes/no)

Diagnoses happen at a particular time…

What simplifications are we making?



Diagnoses happen at a particular time…
and identifying them is not at all straightforward

ADHD Computable Phenotype


Validation of the Use of Electronic Health Records for Classification of ADHD Status

Siobhan M Gruschow [1], Benjamin E Yerys [1] [2], Thomas J Power [1] [2], Dennis R Durbin [1] [2], Allison E Curry [1]

Affiliations + expand

PMID: 28112025 PMCID: PMC5843549 DOI: 10.1177/1087054716672337

Free PMC article

Abstract

Objective: To validate an electronic health record (EHR)-based algorithm to classify ADHD status of pediatric patients.

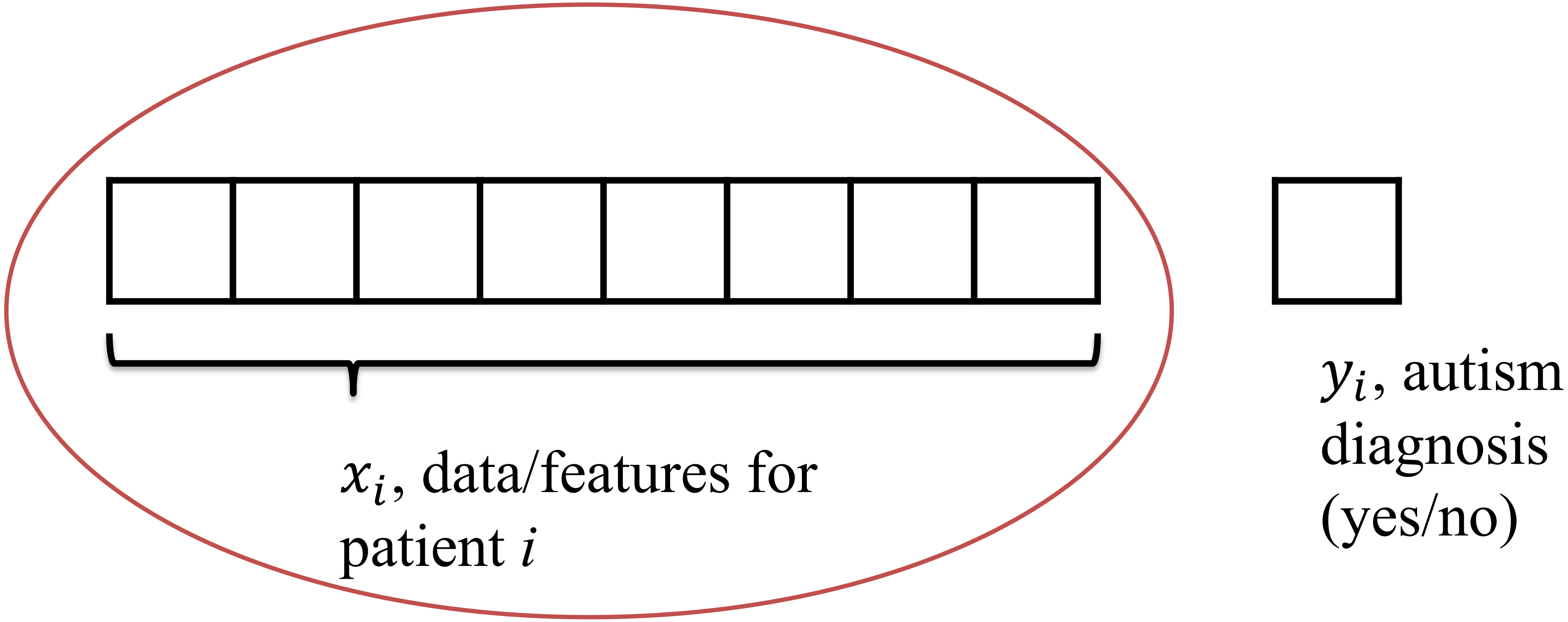
Method: As part of an applied study, we identified all primary care patients of The Children's Hospital of Philadelphia [CHOP] health care network who were born 1987-1995 and residents of New Jersey. Patients were classified with ADHD if their EHR indicated an International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis code of "314.x" at a clinical visit or on a list of known conditions. We manually reviewed EHRs for ADHD patients (n = 2,030) and a random weighted sample of non-ADHD patients (n = 807 of 13,579) to confirm the presence or absence of ADHD.

Results: Depending on assumptions for inconclusive cases, sensitivity ranged from 0.96 to 0.97 (95% confidence interval [CI] = [0.95, 0.97]), specificity from 0.98 to 0.99 [0.97, 0.99], and positive predictive value from 0.83 to 0.98 [0.81, 0.99].

Conclusion: EHR-based diagnostic codes can accurately classify ADHD status among pediatric patients and can be used by large-scale epidemiologic and clinical studies with high sensitivity and specificity.

Keywords: accuracy; adolescents; attention deficit disorder; medical records; sensitivity.

What simplifications are we making?



Data are also collected over time…

Many values for a given feature. Which ones, and how many, do we use?

What's in the EHR?

1. Structured Data
2. Numeric or categorical values, present once
3. Numeric values, present many times
4. Categorical values, present many times

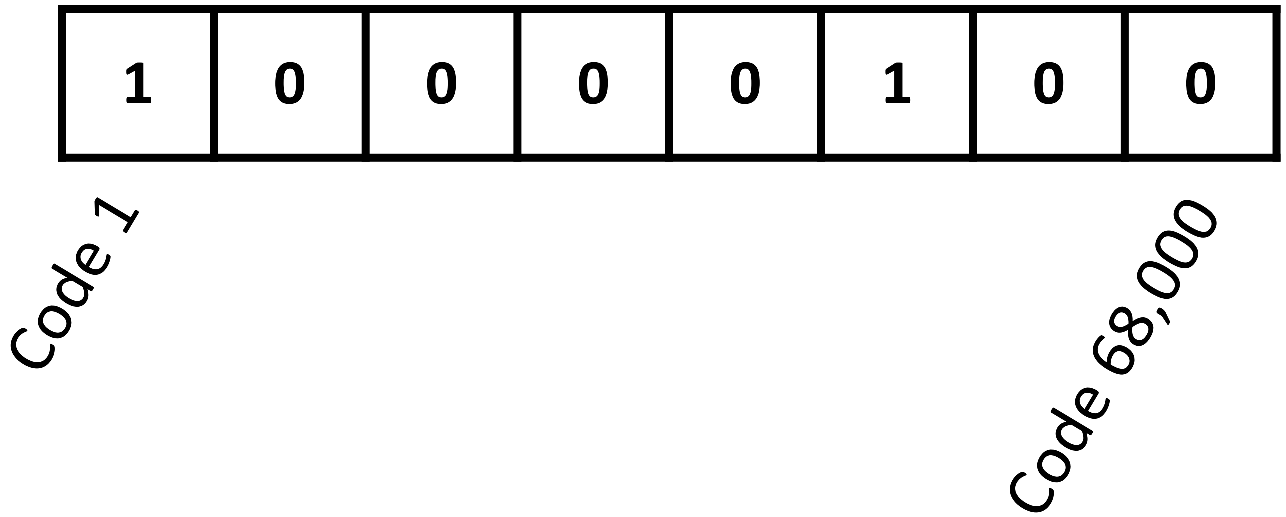
What's in the EHR?

1. **Structured Data**
   - Demographics: don't change over time -> EASY

What’s in the EHR?

1. Structured Data

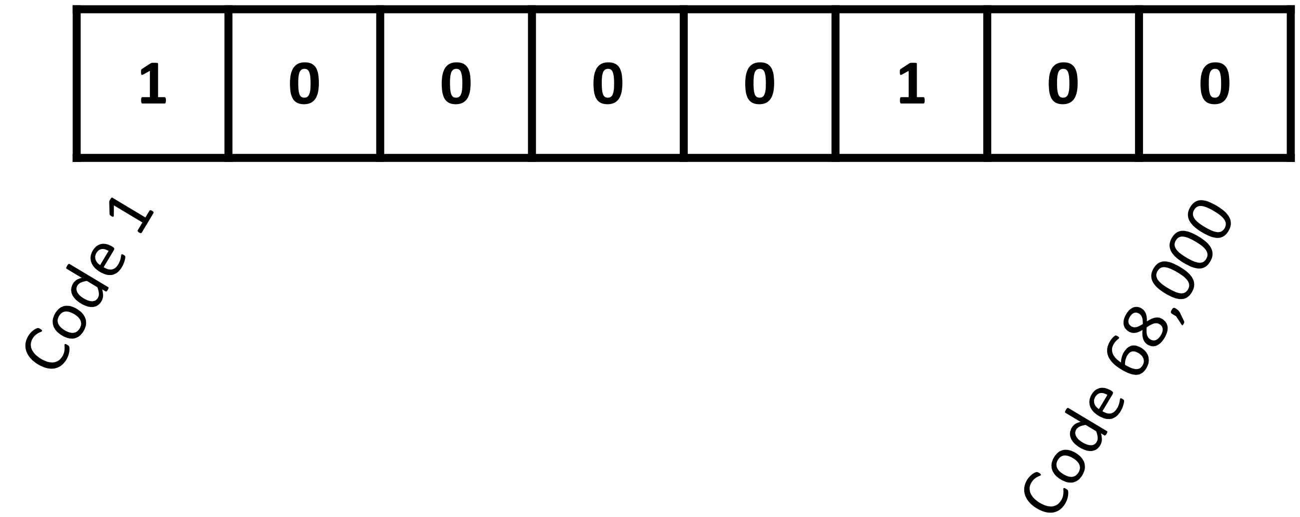
- Demographics: don’t change over time -> EASY
- Dx Codes: observed over time; thousands of codes



What’s in the EHR?

1. Structured Data

- Demographics: don’t change over time -> EASY
- Dx Codes: observed over time; thousands of codes



What's in the EHR?

1. Structured Data

– Demographics: don't change over time -> EASY

– Dx Codes: observed over time; thousands of codes

H·CUP

HEALTHCARE COST AND UTILIZATION PROJECT

Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses

The CCSR is one of the HCUP tools that can be applied to HCUP and other similar databases. These tools are created by AHRQ through a Federal-State-Industry partnership.

What’s in the EHR?

1. **Structured Data**
   - Demographics: don’t change over time -> EASY
   - Dx Codes: observed over time; thousands of codes
   - Encounter details

What’s in the EHR?

1. Str

– D ASY

– D f codes

– E

```
In [18]: df['Clinic Service or Specialty'].value_counts()
Out[18]:
Pediatrics                                   405265
PED. BEHAVIORAL DEVELOPMENT & GEN            315404
Urgent Care                                   44272
Physical and Occupational Therapy             27803
Speech Pathology                              27114
CHART RESPONSIBLE MD                          23164
Ophthalmology                                 15562
Primary Care                                  12959
COMMUNITY AND FAMILY MEDICINE                 11053
Radiology                                     10307
Pediatric Psychiatry                          10027
General Surgery                                9005
Lab                                            8417
Missing or invalid                             7793
```

What's in the EHR?

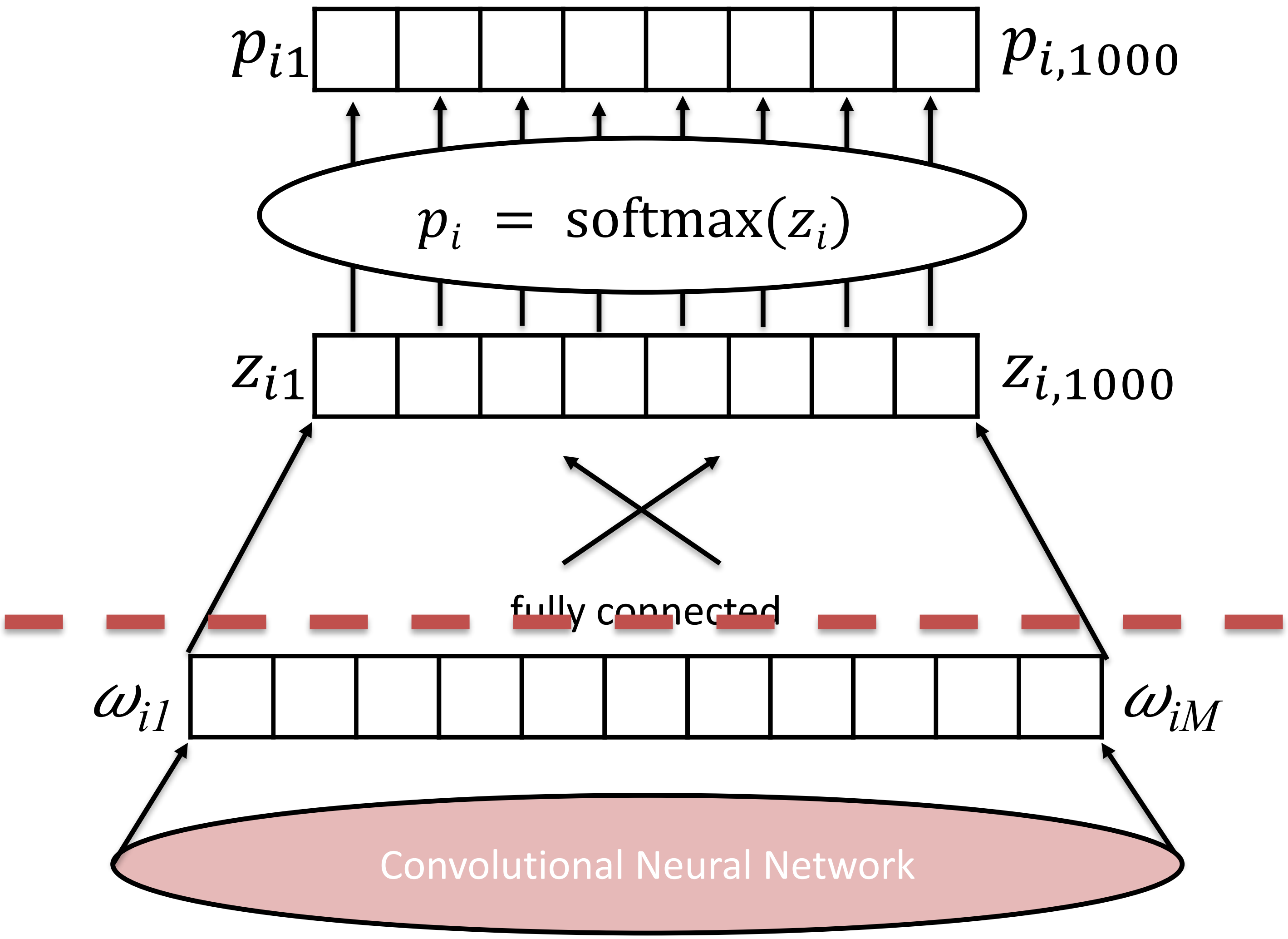
1. **Structured Data**
   - Demographics: don't change over time -> EASY
   - Dx Codes: observed over time; thousands of codes
   - Encounter details
   - Procedure Codes
   - Vitals
   - Labs

What's in the EHR?

1. Structured Data

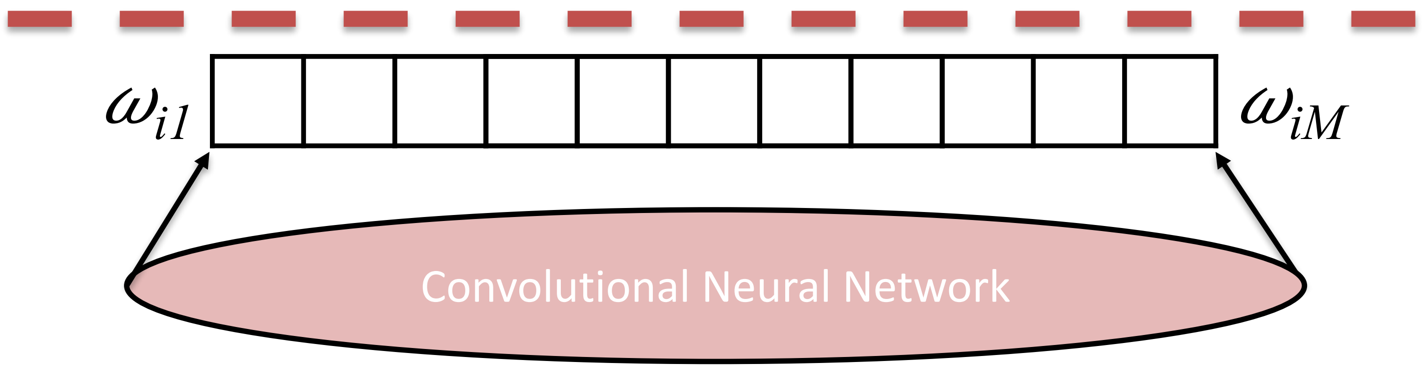
2. **Images***

Use our CNN-based image representations…



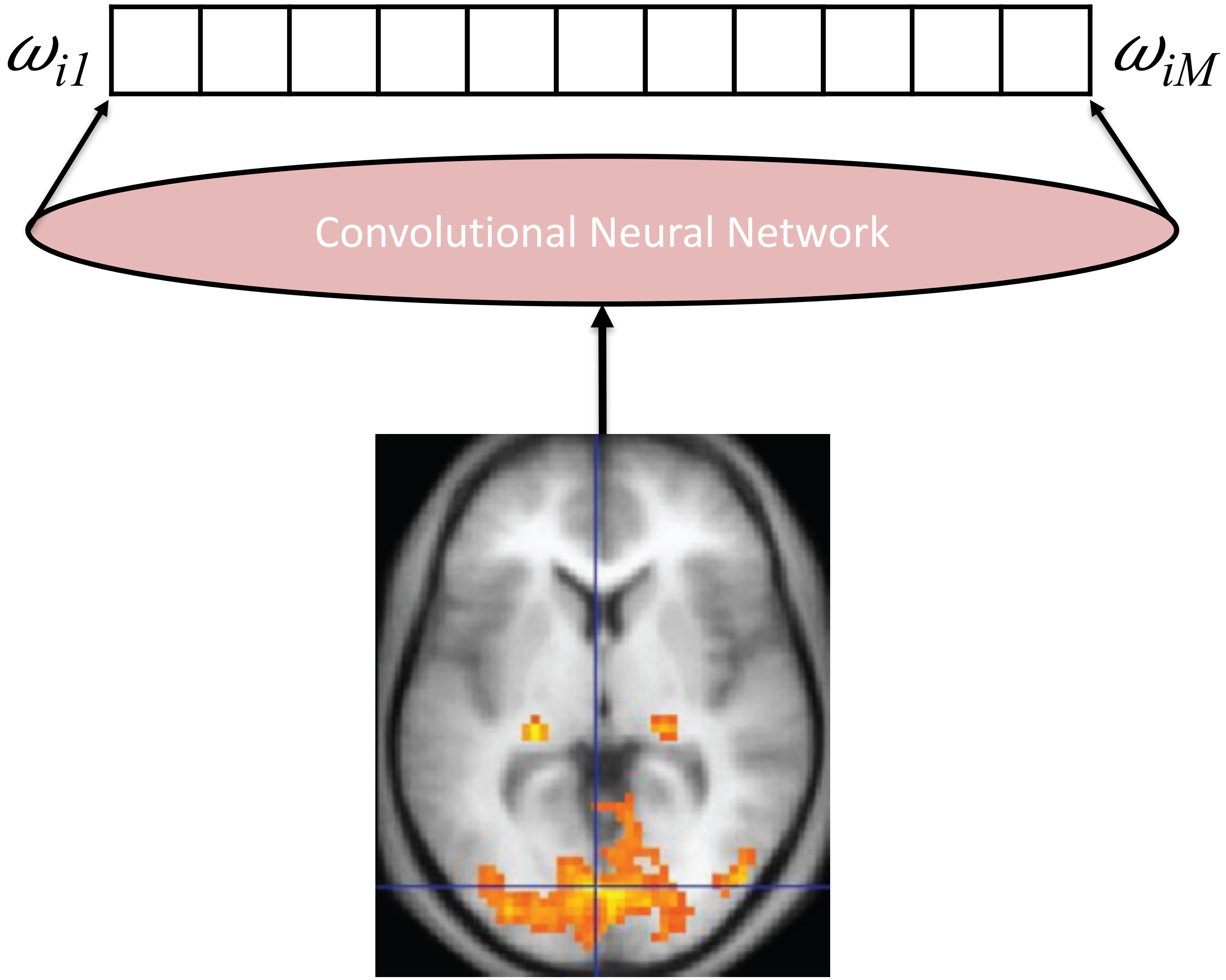
ω_i is a vector of high-level features extracted by the CNN

Use our CNN-based image representations…



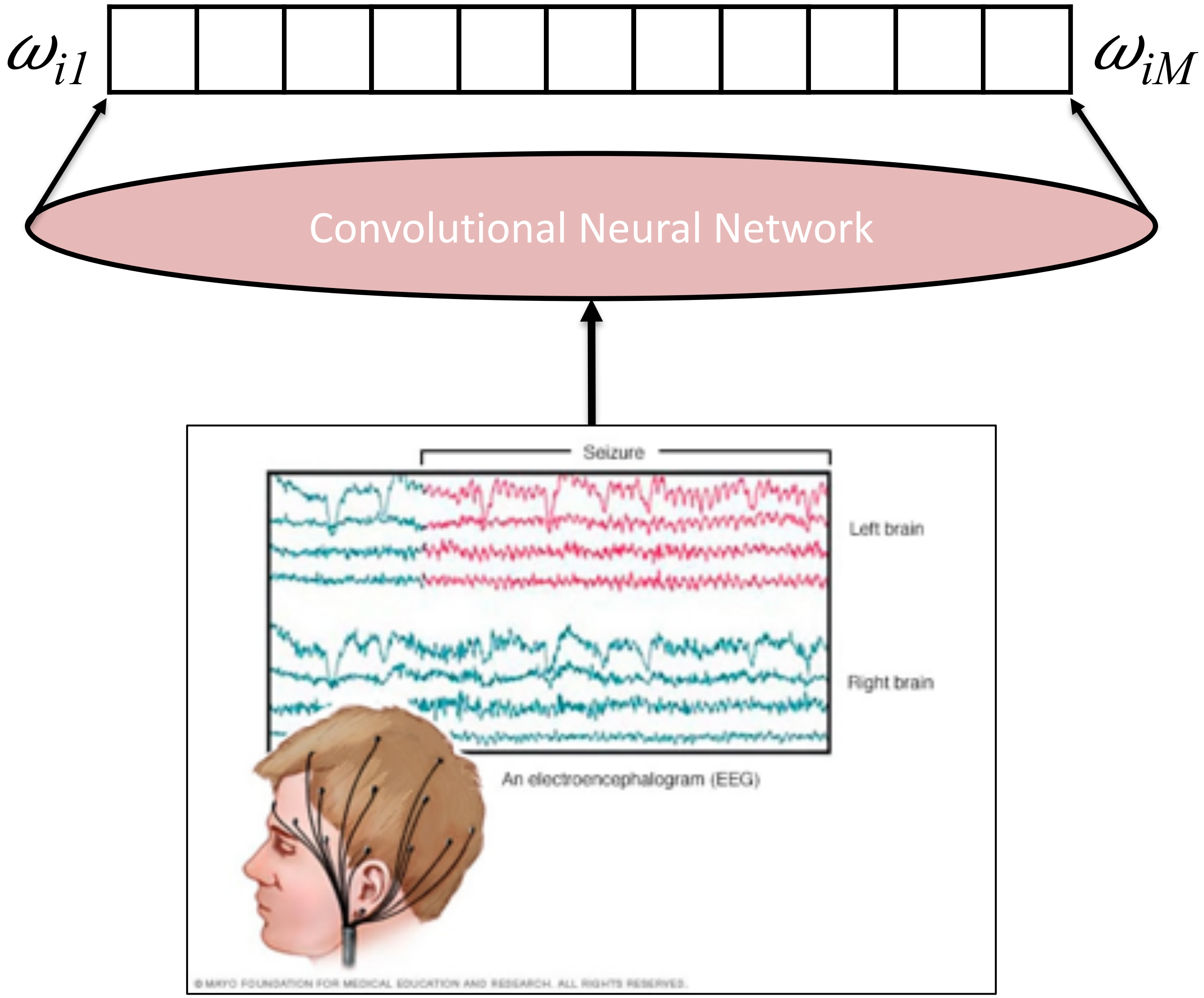
ω_i is a vector of high-level features extracted by the CNN

Use our CNN-based image representations…



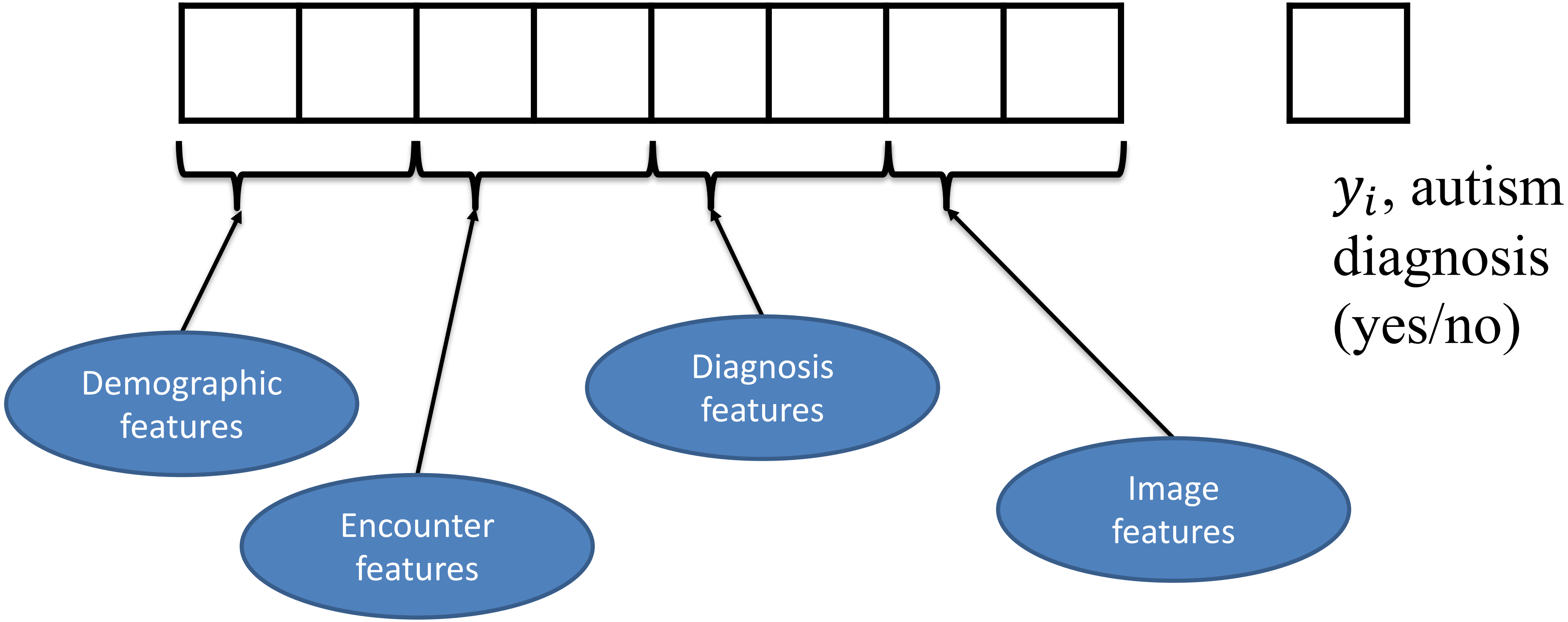
ω_i is a vector of high-level features extracted by the CNN

Or CNN-based EEG representations…



ω_i is a vector of high-level features extracted by the CNN

A Simplified Approach…



What’s in the EHR?

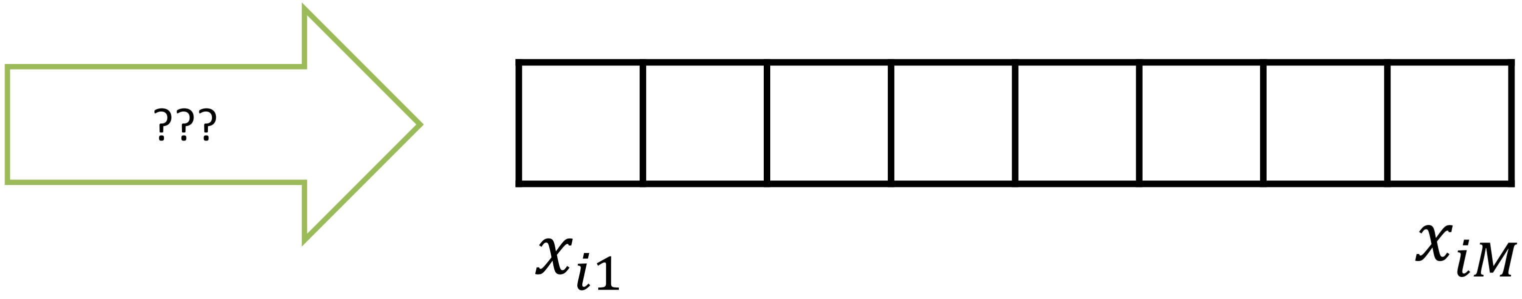
1. Structured Data

2. Images*

3. **Clinical Notes**

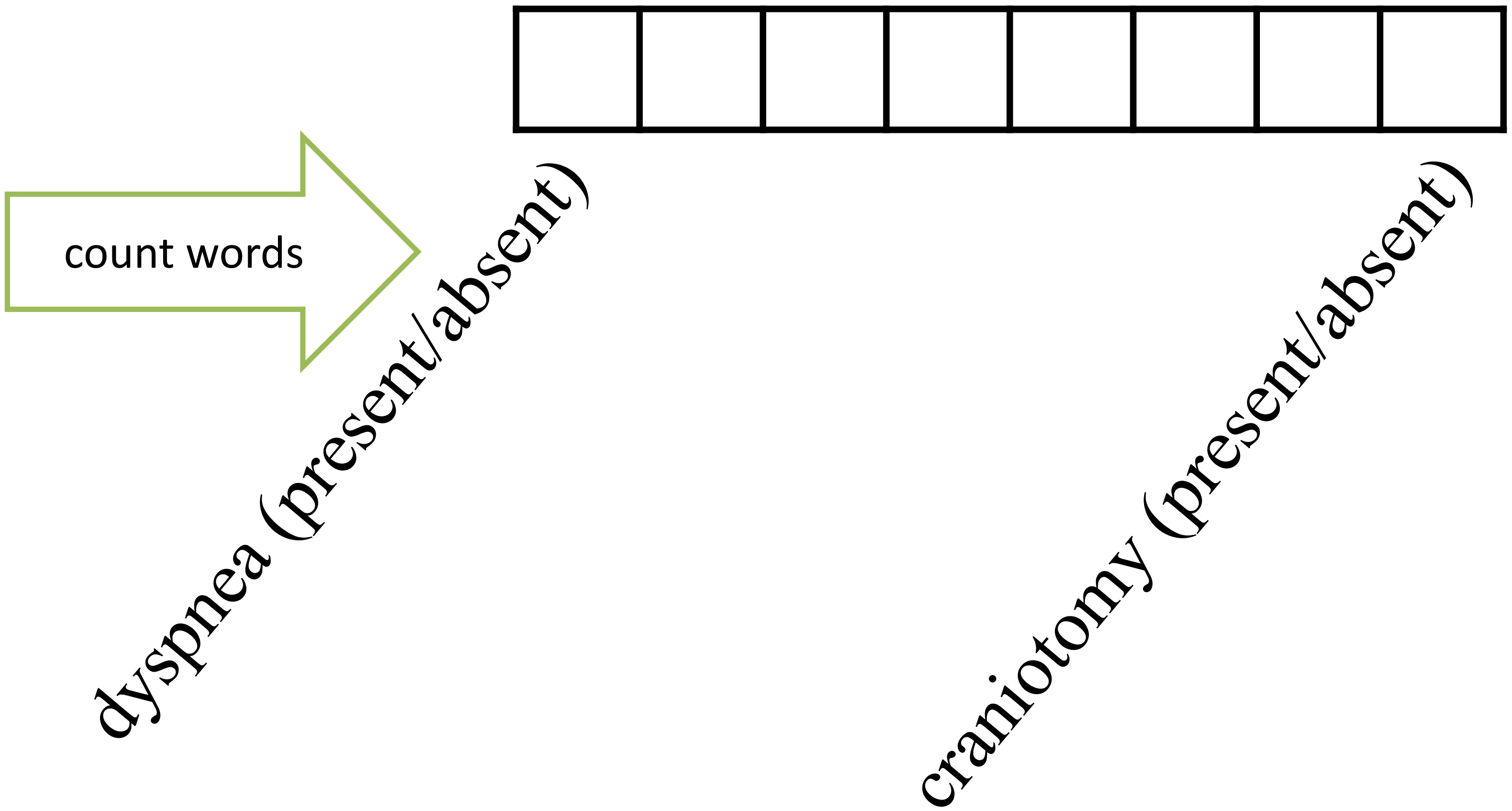
How do we represent notes numerically?

Admission Date :
⟨ deidentified ⟩
Discharge Date :
⟨ deidentified ⟩
Date of Birth :
⟨ deidentified ⟩ Sex :
F
Service :
SURGERY
Allergies :
Patient recorded as having No Known Allergies to Drugs
Attending :
⟨ deidentified ⟩
Chief Complaint :
Dyspnea
Major Surgical or Invasive Procedure :
Mitral Valve Repair
History of Present Illness :
Ms. ⟨ deidentified ⟩ is a 53 year old female who presents after a large bleed rhythmically lag to 2 dose but the patient was brought to the Emergency Department where he underwent craniotomy with stenting of right foot under the LUL COPD and transferred to the OSH on ⟨ deidentified ⟩ .
The patient will need a pigtail catheter to keep the sitter daily .



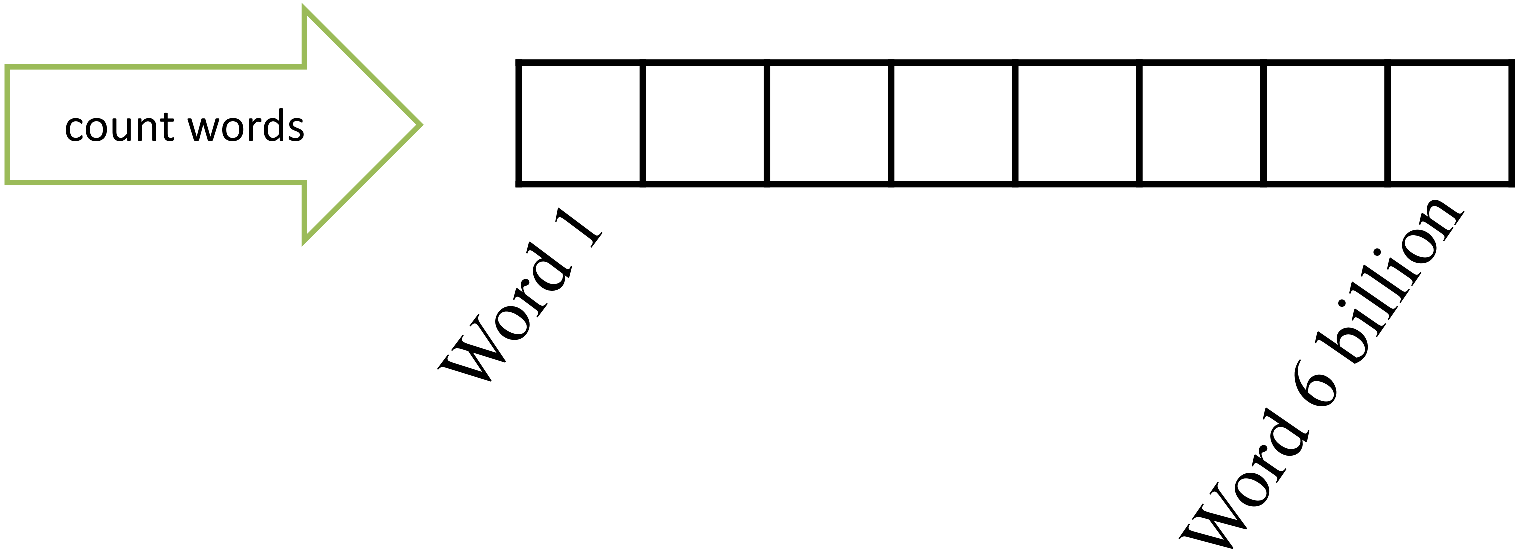
How do we represent notes numerically?

Admission Date :
⟨ deidentified ⟩
Discharge Date :
⟨ deidentified ⟩
Date of Birth :
⟨ deidentified ⟩ Sex :
F
Service :
SURGERY
Allergies :
Patient recorded as having No Known Allergies to Drugs
Attending :
⟨ deidentified ⟩
Chief Complaint :
Dyspnea
Major Surgical or Invasive Procedure :
Mitral Valve Repair
History of Present Illness :
Ms. ⟨ deidentified ⟩ is a 53 year old female who presents after a large bleed rhythmically lag to 2 dose but the patient was brought to the Emergency Department where he underwent craniotomy with stenting of right foot under the LUL COPD and transferred to the OSH on ⟨ deidentified ⟩ .
The patient will need a pigtail catheter to keep the sitter daily .

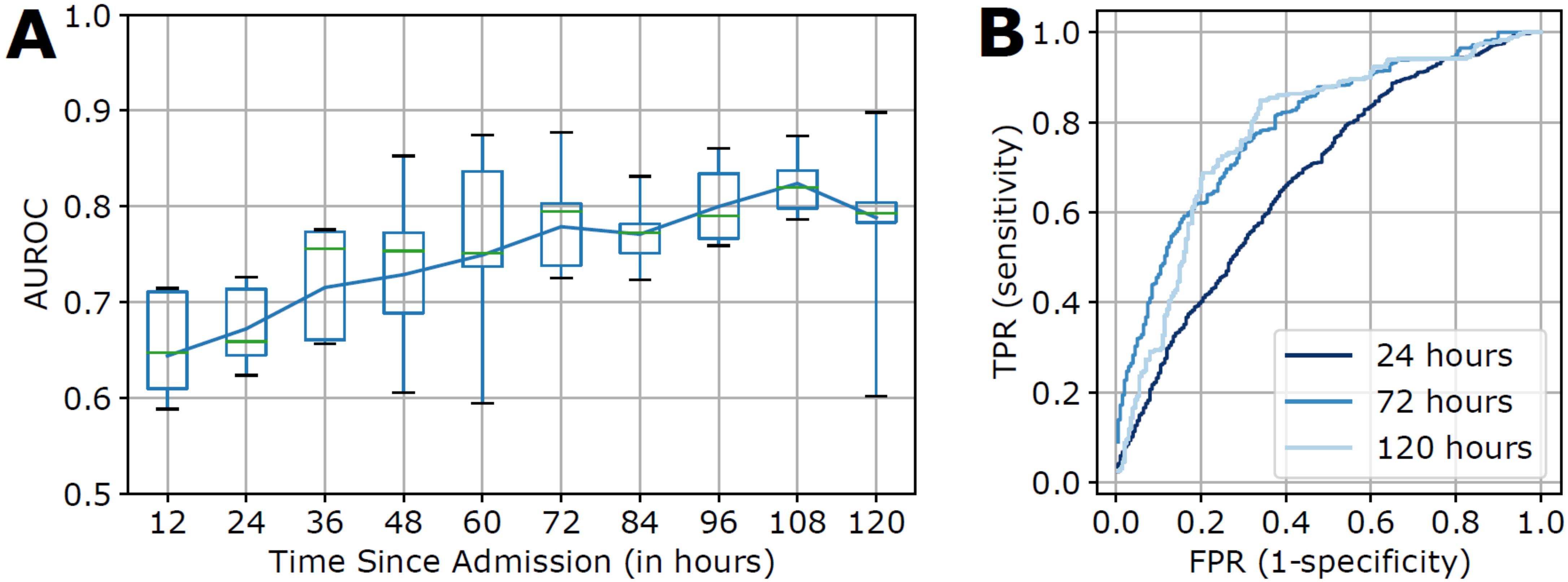


How do we represent notes numerically?

Admission Date :
⟨ deidentified ⟩
Discharge Date :
⟨ deidentified ⟩
Date of Birth :
⟨ deidentified ⟩ Sex :
F
Service :
SURGERY
Allergies :
Patient recorded as having No Known Allergies to Drugs
Attending :
⟨ deidentified ⟩
Chief Complaint :
Dyspnea
Major Surgical or Invasive Procedure :
Mitral Valve Repair
History of Present Illness :
Ms. ⟨ deidentified ⟩ is a 53 year old female who presents after a large bleed rhythmically lag to 2 dose but the patient was brought to the Emergency Department where he underwent craniotomy with stenting of right foot under the LUL COPD and transferred to the OSH on ⟨ deidentified ⟩ .
The patient will need a pigtail catheter to keep the sitter daily .



Prediction over Time



Turpin et al., *Machine Learning Prediction of Surgical Intervention for Small Bowel Obstruction* (forthcoming)