



# Journal Club

---

21 JULY 2023


PRESENTER: NAT TANGCHITNOB



# The paper

Article | [Open Access](#) | [Published: 08 May 2018](#)

## Scalable and accurate deep learning with electronic health records

[Alvin Rajkomar](#) , [Eyal Oren](#), [Kai Chen](#), [Andrew M. Dai](#), [Nissan Hajaj](#), [Michaela Hardt](#), [Peter J. Liu](#), [Xiaobing Liu](#), [Jake Marcus](#), [Mimi Sun](#), [Patrik Sundberg](#), [Hector Yee](#), [Kun Zhang](#), [Yi Zhang](#), [Gerardo Flores](#), [Gavin E. Duggan](#), [Jamie Irvine](#), [Quoc Le](#), [Kurt Litsch](#), [Alexander Mossin](#), [Justin Tansuwan](#), [De Wang](#), [James Wexler](#), [Jimbo Wilson](#), ... [Jeffrey Dean](#) [+ Show authors](#)

[npj Digital Medicine](#) **1**, Article number: 18 (2018) | [Cite this article](#)

**240k** Accesses | **1060** Citations | **2033** Altmetric | [Metrics](#)

### Authors and Affiliations

#### Google Inc, Mountain View, CA, USA

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Michael D. Howell, Claire Cui, Greg S. Corrado & Jeffrey Dean

#### University of California, San Francisco, San Francisco, CA, USA

Alvin Rajkomar, Dana Ludwig & Atul J. Butte

#### University of Chicago Medicine, Chicago, IL, USA

Samuel L. Volchenbom

#### Stanford University, Stanford, CA, USA

Nigam H. Shah

### Introduction

### Methods

- Datasets
- Data representation
- Outcome definition
- Algorithms
- Prediction time

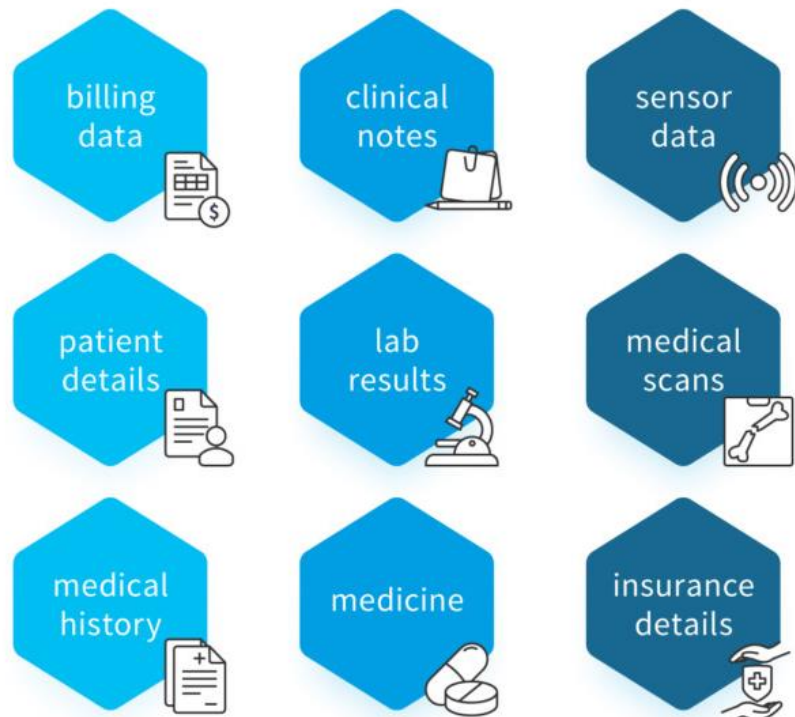
### Results

- Demographics
- Model performance

### Discussion

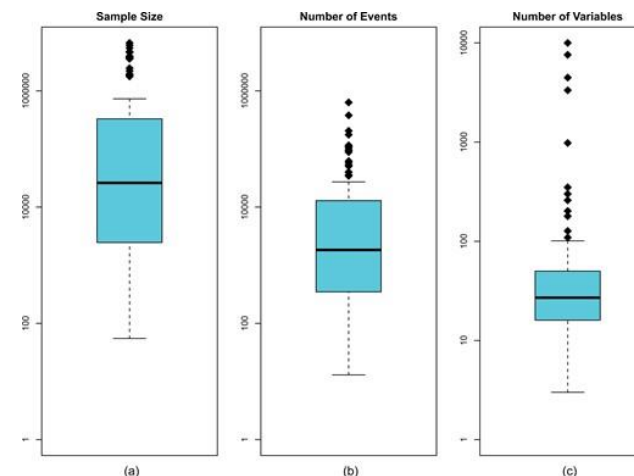
### Impact to society

# Electronic medical records (EHRs)



## Challenge of using EHR data

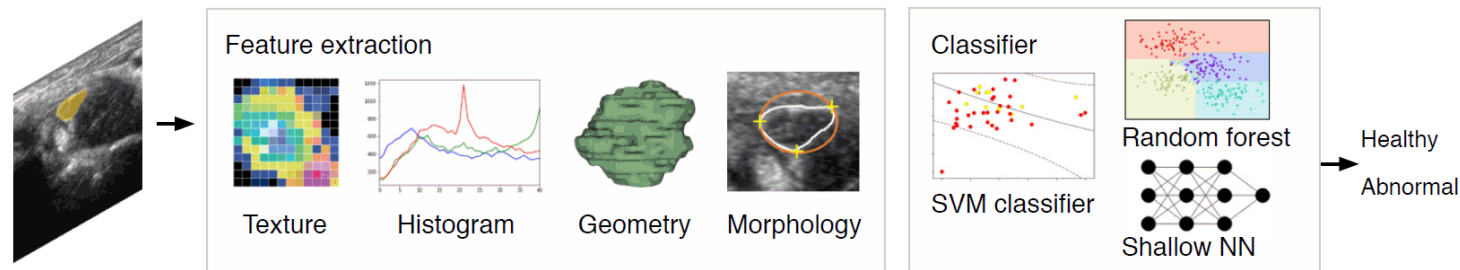
1. For traditional predictive modeling techniques, each model requires a custom dataset with specific variables.
2. The number of potential predictor variables in the electronic health record (EHR) may easily number in the thousands



A systematic review of prediction models using EHR data (n=107)  
Use a median of only 27 variables from data at a single center (Goldstein et al., 2017)

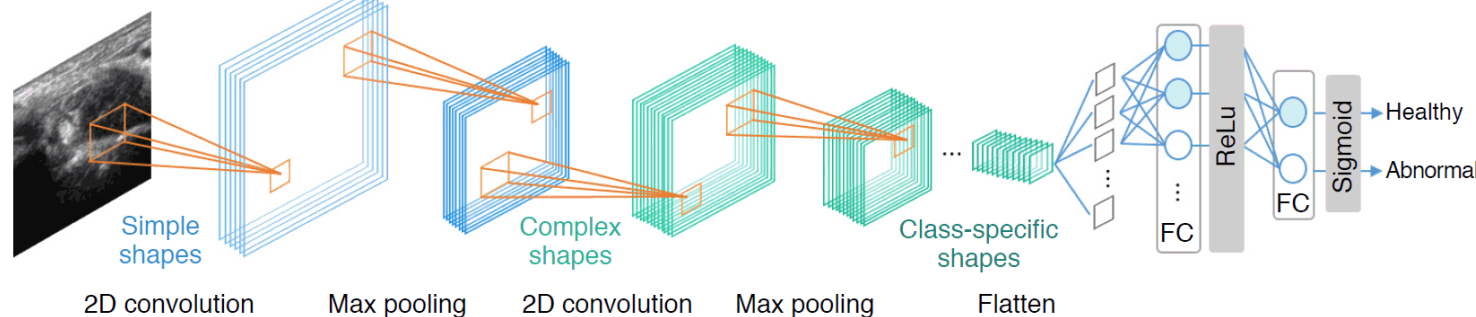
# Deep learning versus Traditional ML

A Machine learning (ML)



- 80% of the effort is
- Preprocessing
  - Merging
  - Customizing
  - Cleaning datasets

B Deep learning (DL)



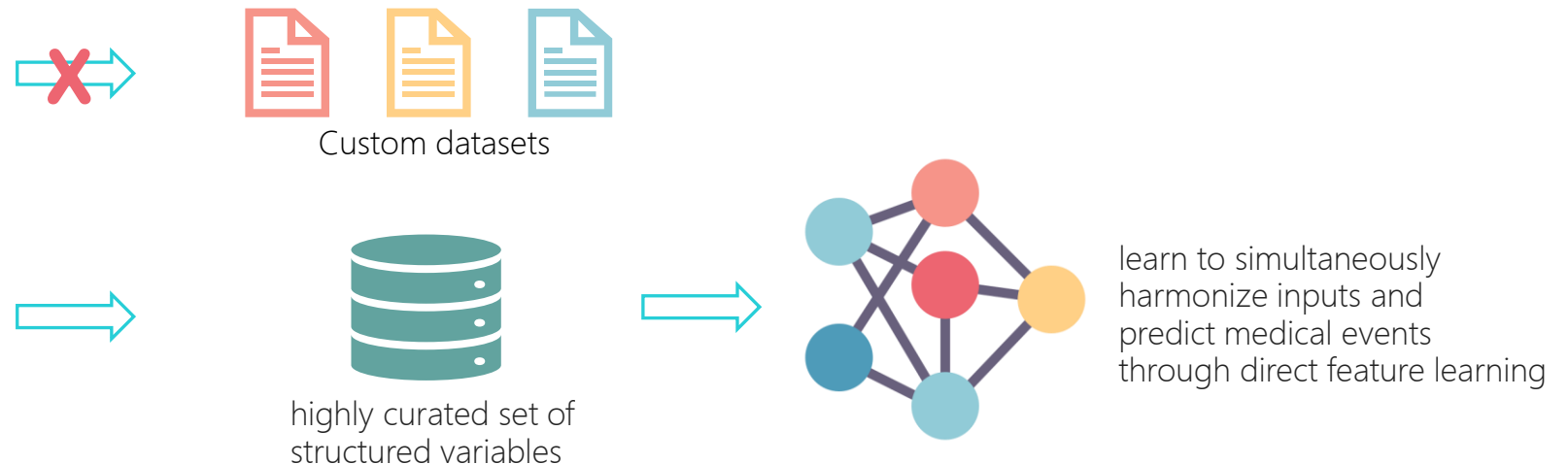
- Key advantage of Deep Learning
- No need to specify which potential variables to consider and in what combinations
  - Neural networks are able to learn representations of the key factors and interactions from the data itself

## Contribution of this study

1. Report a generic data processing pipeline that can take raw EHR data as input, and produce FHIR outputs without manual feature harmonization.
2. Based on data from two academic hospitals with a general patient population, we demonstrate the effectiveness of deep learning models in a wide variety of predictive problems and settings.

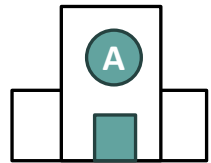
# Hypothesis

Deep learning approaches could incorporate the entire EHR, including free-text notes, to produce predictions for a wide range of clinical problems and outcomes that outperform state-of-the-art traditional predictive models.





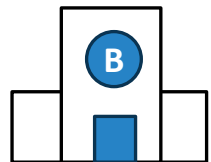
# Datasets & Data representations



## Hospital A

University of California, San Francisco (UCSF)

In-patient and Out-patient data (2012-2016)



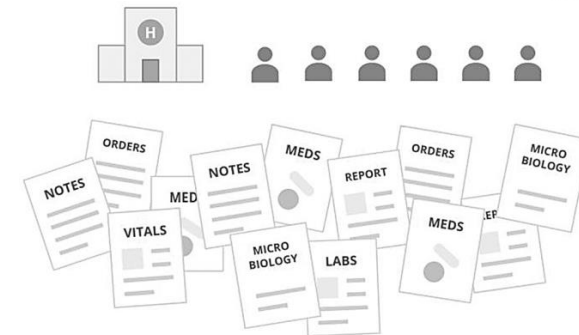
## Hospital B

University of Chicago Medicine (UCM)

In-patient and Out-patient data (2009-2016)

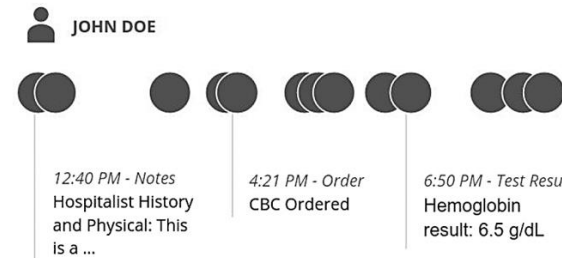
### Inclusion criteria for the study (In-patient Encounters)

1. Patients 18 years or older
2. Hospitalizations of 24 hours or longer
3. Encounter was confirmed as complete or noncancelled
4. Encounter had a start and end time
5. Encounter class was defined as inpatient as defined in dataset
6. Administrative encounters were excluded (No ICD-9 diagnosis)



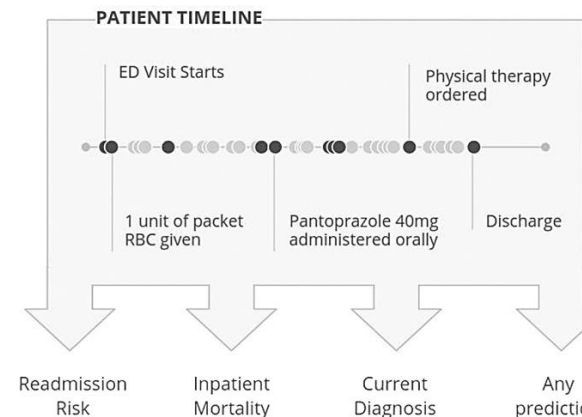
1

Health systems collect and store electronic health records in various formats in databases.



2

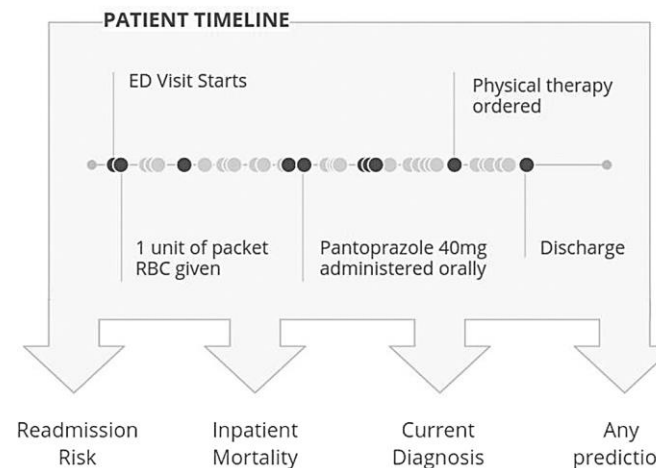
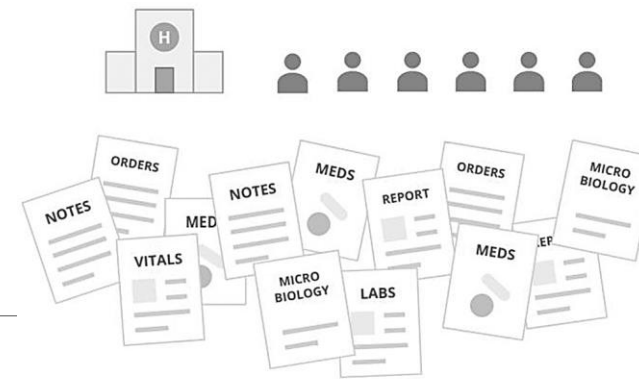
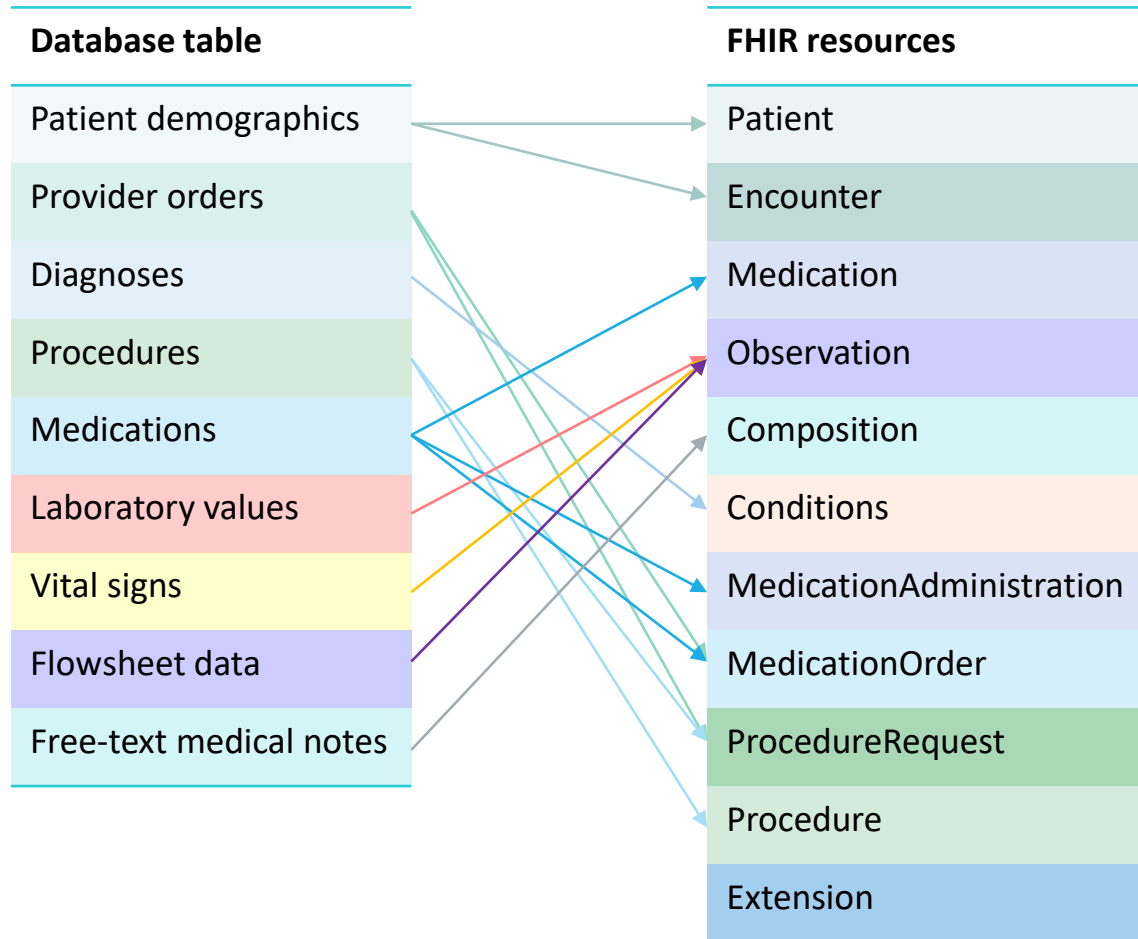
All available data for each patient is converted to events recorded in containers based on the Fast Healthcare Interoperability Resource (FHIR) specification.



3

The FHIR resources are placed in temporal order, depicting all events recorded in the EHR (i.e. timeline). The deep learning model uses this full history to make each prediction.

# Data representations



1 Health systems collect and store electronic health records in various formats in databases.

2 All available data for each patient is converted to events recorded in containers based on the Fast Healthcare Interoperability Resource (FHIR) specification.

3 The FHIR resources are placed in temporal order, depicting all events recorded in the EHR (i.e. timeline). The deep learning model uses this full history to make each prediction.





### Patient Timeline



### FHIR Resource

```

medication_order { contained { medication {
  code {
    text { value: "Zosyn" }
    coding {
      system { value: "RxNorm" }
      code { value: "1659133" } } } }
  ingredient { item_codeable_concept {
    text { value: "Piperacillin" }
    coding {
      system { value: "Hospital A. Ingredient Code" }
      code { value: "203134" } } } } }
  ingredient { item_codeable_concept {
    text { value: "Tazobactam" }
    coding {
      system { value: "Hospital A. Ingredient Code" }
      code { value: "221167" } } } } } }
  effective_period {
    start { value_us: 882518400000000 } } } }
procedure {
  code {
    text { value: "Ventilator Management" }
    coding {
      system { value: "CPT" }
      code { value: "94002" } } } }
  performed_date_time { value_us: 882518500000000 } }
  
```

### Feature Type and Token ID

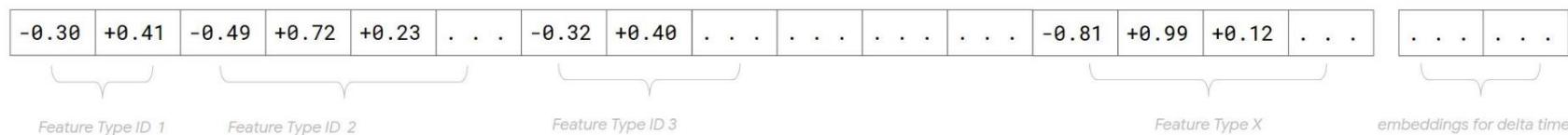
### Embedding

### Weight

Feature Type and Token ID	Embedding	Weight
1-< 17>	-0.30   +0.41	+0.20
2-< 35>	-0.49   +0.72   +0.23   . . .	+0.33
3-< 85>	-0.33   +0.39   . . .	+0.12
3-< 19>	-0.31   +0.41   . . .	+0.14
4-<702>	-0.33   +0.39   . . .	+0.12
4-<913>	-0.70   +0.88   -0.13   . . .	+0.31
5-<137>	-0.72   +0.83   -0.09   . . .	+0.41
5-<139>	-0.13   +0.41   +0.23	+0.89
6-< 21>	-0.78   +0.41   -0.98	+0.12
	-0.30   +0.41   +0.23   . . .	+0.33

Embeddings of same feature type ID are averaged using weights in RNN

### Input to Recurrent Neural Network (RNN) at a single timestep



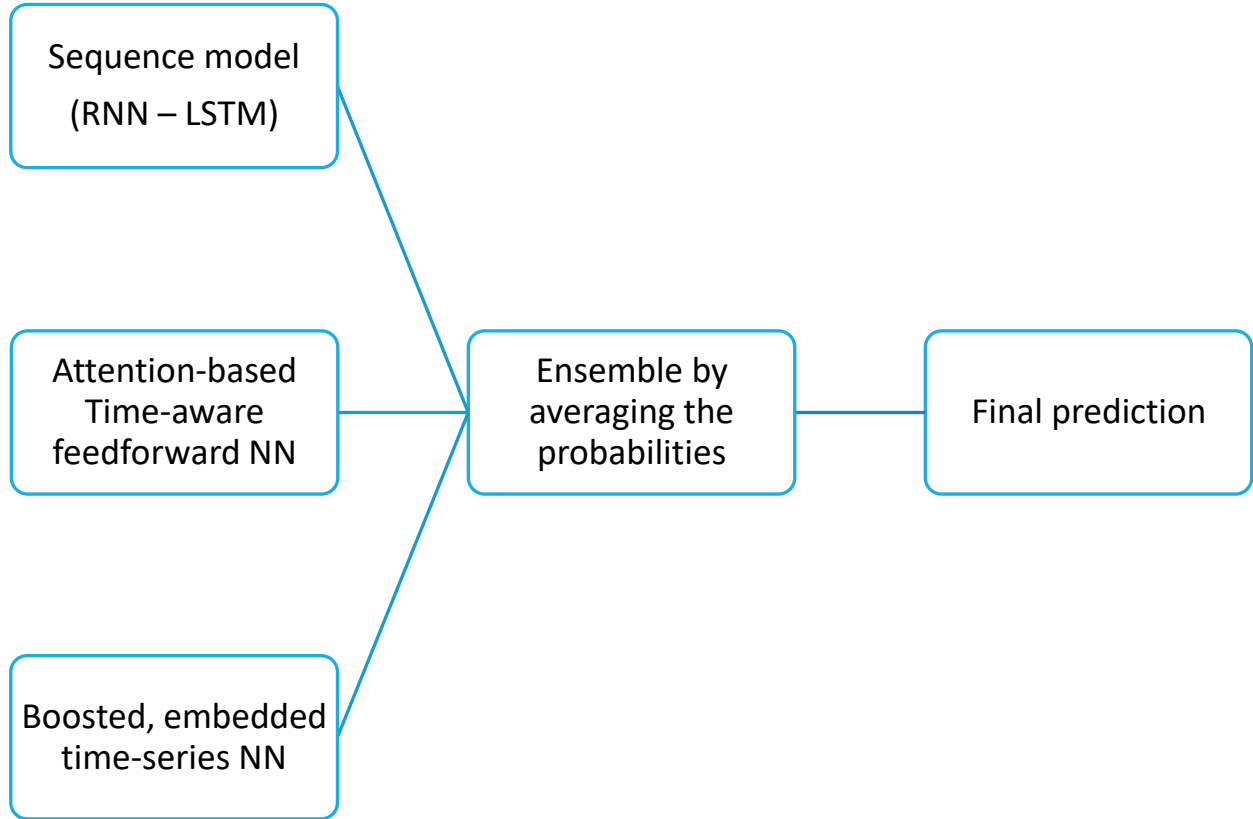
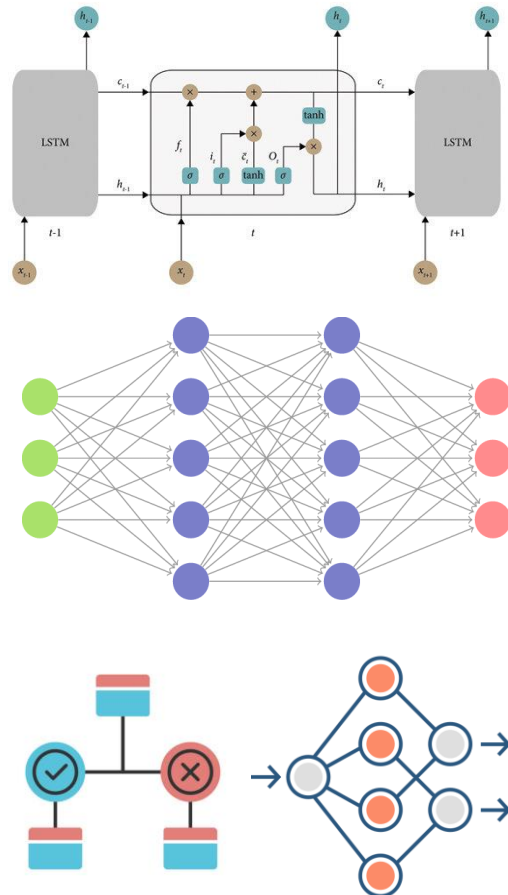




# Outcome definitions

Domains	Outcomes	Definitions	
Important clinical outcome	Inpatient mortality (Death)	- Discharge disposition of “expired”	Binary
Standard measure of quality of care	30-day unplanned readmission	- Admission within 30 days after discharge from an index hospitalization into the same institution - Exclude planned readmissions (e.g. chemotherapy)	
Measure of resource utilization	Long length of stay	- Length of stay $\geq 7$ days (75 <sup>th</sup> percentile of hospital stays)	
Measure of understanding patient’s problems	Diagnoses	- Primary and secondary ICD-9 billing diagnosis (14,025 codes)	Multilabel

# Algorithms



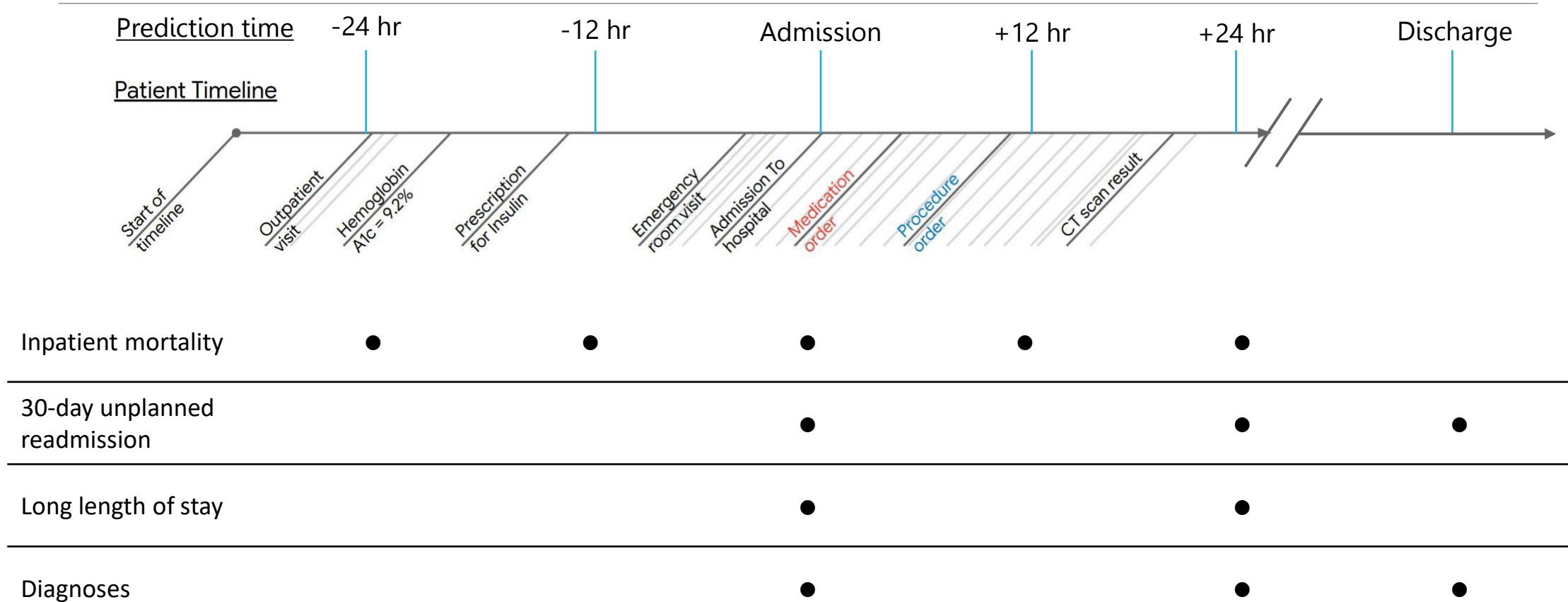


# Baseline models

Outcomes	Logistic models	Hand-engineered Features	
Inpatient mortality	augmented Early Warning Score (aEWS) (Smith et al., 2013)	SBP, HR, RR, Body Temp	
		24 common lab tests	
30-day unplanned readmission	Modified HOSPITAL (mHOSPITAL) (Donzé et al., 2013)	Na level	Hb level
		hospital service	
		occurrence of CPT codes	
		number of prior hospitalizations	
		length of the current hospitalization	
Long length of stay	Modified Liu (mLiu) (Liu et al., 2010)	age	gender
		hierarchical condition categories	
		admission source	
		hospital service	
		24 common lab tests	
Diagnoses	No baseline model		



# Prediction timing





# Results

216,221 hospitalizations

114,003 patients

- Train 80%

- Validate 10%

- Test 10%

4930 (2.3%) in-hospital death

27,918 (12.9%) unplanned 30-day readmissions

1-228 diagnoses/patient

**Table 1.** Characteristics of hospitalizations in training and test sets

	Training data (n = 194,470)		Test data (n = 21,751)	
	Hospital A (n = 85,522)	Hospital B (n = 108,948)	Hospital A (n = 9624)	Hospital B (n = 12,127)
<i>Demographics</i>				
Age, median (IQR) y	56 (29)	57 (29)	55 (29)	57 (30)
Female sex, no. (%)	46,848 (54.8%)	62,004 (56.9%)	5364 (55.7%)	6935 (57.2%)
<i>Disease cohort, no. (%)</i>				
Medical	46,579 (54.5%)	55,087 (50.6%)	5263 (54.7%)	6112 (50.4%)
Cardiovascular	4616 (5.4%)	6903 (6.3%)	528 (5.5%)	749 (6.2%)
Cardiopulmonary	3498 (4.1%)	9028 (8.3%)	388 (4.0%)	1102 (9.1%)
Neurology	6247 (7.3%)	6653 (6.1%)	697 (7.2%)	736 (6.1%)
Cancer	14,544 (17.0%)	19,328 (17.7%)	1617 (16.8%)	2087 (17.2%)
Psychiatry	788 (0.9%)	339 (0.3%)	64 (0.7%)	35 (0.3%)
Obstetrics and newborn	8997 (10.5%)	10,462 (9.6%)	1036 (10.8%)	1184 (9.8%)
Other	253 (0.3%)	1148 (1.1%)	31 (0.3%)	122 (1.0%)
<i>Previous hospitalizations, no. (%)</i>				
0 hospitalizations	54,954 (64.3%)	56,197 (51.6%)	6123 (63.6%)	6194 (51.1%)
≥1 and <2 hospitalizations	14,522 (17.0%)	19,807 (18.2%)	1620 (16.8%)	2175 (17.9%)
≥2 and <6 hospitalizations	12,591 (14.7%)	24,009 (22.0%)	1412 (14.7%)	2638 (21.8%)
≥6 hospitalizations	3455 (4.0%)	8935 (8.2%)	469 (4.9%)	1120 (9.2%)
<i>Discharge location no. (%)</i>				
Home	70,040 (81.9%)	91,273 (83.8%)	7938 (82.5%)	10,109 (83.4%)
Skilled nursing facility	6601 (7.7%)	5594 (5.1%)	720 (7.5%)	622 (5.1%)
Rehabilitation	2666 (3.1%)	5136 (4.7%)	312 (3.2%)	649 (5.4%)
Another healthcare facility	2189 (2.6%)	2052 (1.9%)	243 (2.5%)	220 (1.8%)
Expired	1816 (2.1%)	2679 (2.5%)	170 (1.8%)	265 (2.2%)
Other	2210 (2.6%)	2214 (2.0%)	241 (2.5%)	262 (2.2%)
<i>Primary outcomes</i>				
In-hospital deaths, no. (%)	1816 (2.1%)	2679 (2.5%)	170 (1.8%)	265 (2.2%)
30-day readmissions, no. (%)	9136 (10.7%)	15,932 (14.6%)	1013 (10.5%)	1837 (15.1%)
Hospital stays at least 7 days, no. (%)	20,411 (23.9%)	26,109 (24.0%)	2145 (22.3%)	2931 (24.2%)
No. of ICD-9 diagnoses, median (IQR)	12 (16)	10 (10)	12 (16)	10 (10)

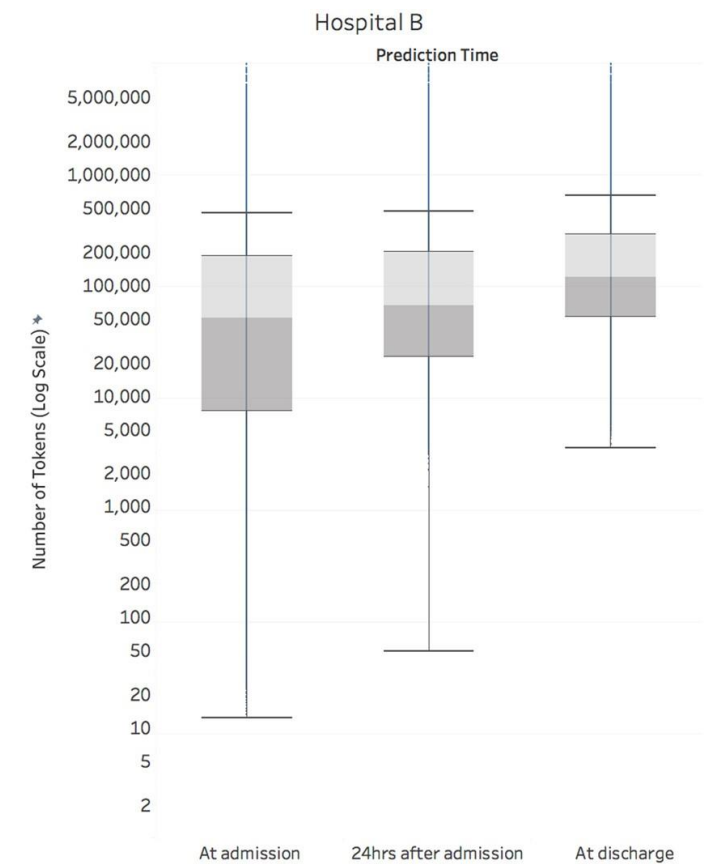
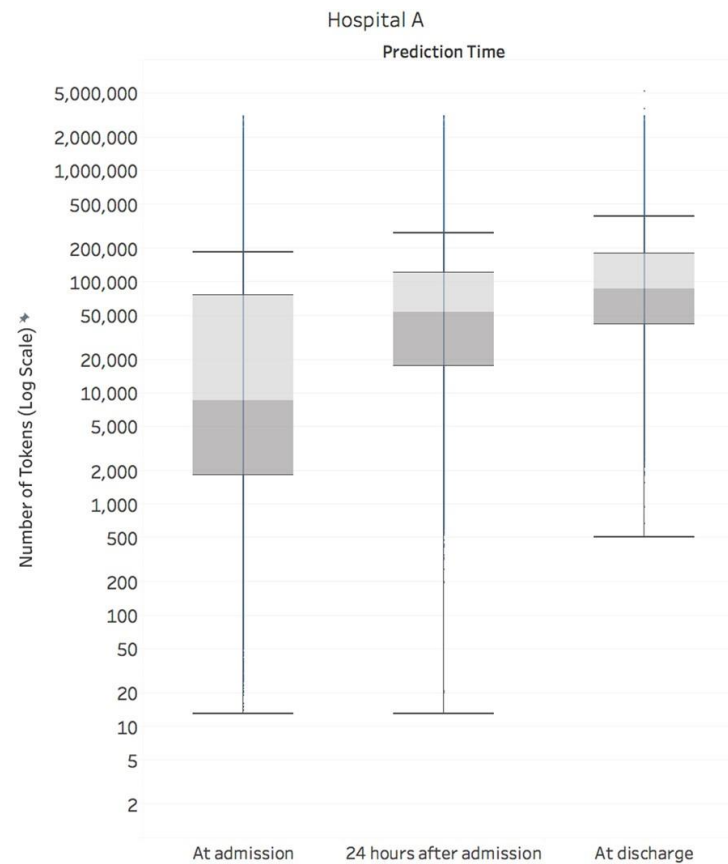


# Amount of data in the EHR (log scale)

Average tokens

137,882 (At admission)

216,744 (At discharge)





**Table 2.** Prediction accuracy of each task made at different time points

	Hospital A	Hospital B
<i>Inpatient mortality, AUROC<sup>a</sup> (95% CI)</i>		
24 h before admission	0.87 (0.85–0.89)	0.81 (0.79–0.83)
At admission	0.90 (0.88–0.92)	0.90 (0.86–0.91)
24 h after admission	<b>0.95 (0.94–0.96)</b>	<b>0.93 (0.92–0.94)</b>
Baseline (aEWS <sup>b</sup> ) at 24 h after admission	0.85 (0.81–0.89)	0.86 (0.83–0.88)
<i>30-day readmission, AUROC (95% CI)</i>		
At admission	0.73 (0.71–0.74)	0.72 (0.71–0.73)
At 24 h after admission	0.74 (0.72–0.75)	0.73 (0.72–0.74)
At discharge	<b>0.77 (0.75–0.78)</b>	<b>0.76 (0.75–0.77)</b>
Baseline (mHOSPITAL <sup>c</sup> ) at discharge	0.70 (0.68–0.72)	0.68 (0.67–0.69)
<i>Length of stay at least 7 days, AUROC (95% CI)</i>		
At admission	0.81 (0.80–0.82)	0.80 (0.80–0.81)
At 24 h after admission	<b>0.86 (0.86–0.87)</b>	<b>0.85 (0.85–0.86)</b>
Baseline (Liu <sup>d</sup> ) at 24 h after admission	0.76 (0.75–0.77)	0.74 (0.73–0.75)
<i>Discharge diagnoses (weighted AUROC)</i>		
At admission	0.87	0.86
At 24 h after admission	0.89	0.88
At discharge	<b>0.90</b>	<b>0.90</b>

<sup>a</sup>Area under the receiver operator curve

<sup>b</sup>Augmented Early Warning System score

<sup>c</sup>Modified HOSPITAL score for readmission

<sup>d</sup>Modified Liu score for long length of stay

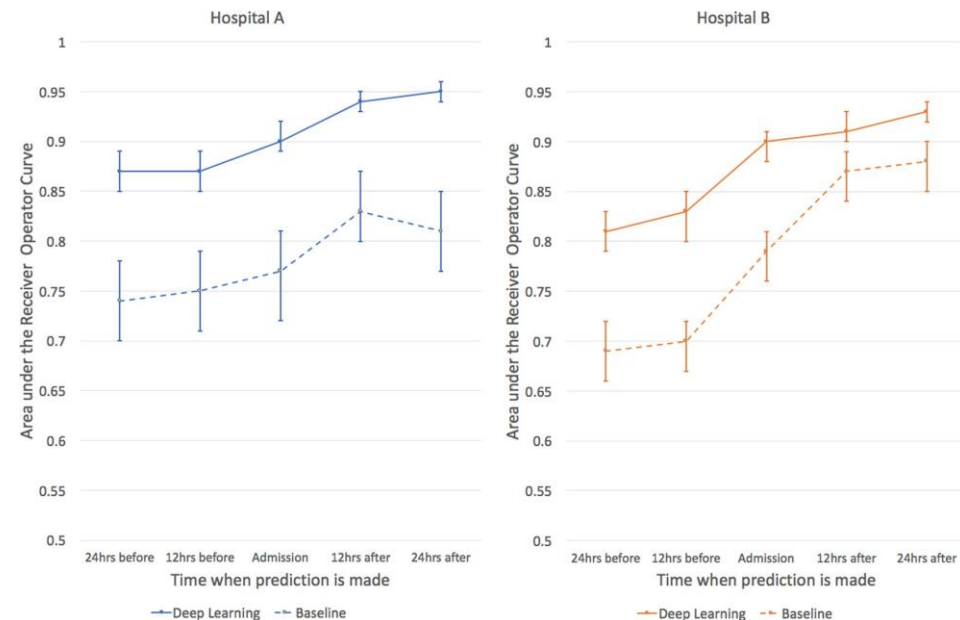
The bold values indicate the highest area-under-the-receiver-operator-curve for each prediction task

# Model performance

At 24 h, the work-up-to-detection ratio (number needed to evaluate) of our model compared to the aEWS for predicting patient mortality

- 7.4 vs 14.3 (Hospital A)

- 8.0 vs 15.4 (Hospital B)





# Model performance

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

	Hospital A	Hospital B
<b>Inpatient Mortality, AUROC<sup>1</sup>(95% CI)</b>		
Deep learning 24 hours after admission	<b>0.95</b> (0.94-0.96)	<b>0.93</b> (0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS <sup>2</sup> ) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
<b>30-day Readmission, AUROC (95% CI)</b>		
Deep learning at discharge	<b>0.77</b> (0.75-0.78)	<b>0.76</b> (0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL <sup>3</sup> ) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
<b>Length of Stay at least 7 days AUROC (95% CI)</b>		
Deep learning 24 hours after admission	<b>0.86</b> (0.86-0.87)	<b>0.85</b> (0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu <sup>4</sup> ) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

<sup>1</sup> Area under the receiver operator curve

<sup>2</sup> Augmented early warning score

<sup>3</sup> Modified HOSPITAL score

<sup>4</sup> Modified Liu score

## Baseline

- Using hand-engineered features according to literature review

## Full feature simple

- Logistic regression model trained with these features using Adam optimizer and early-stopping as regularization

- All available predictor variables, ignoring temporal order

## Full feature enhanced

- The features were bucketized into five time-buckets, representing intervals of less than 1 day, 1 week, 1 month, 1 year, or greater than 1 year.

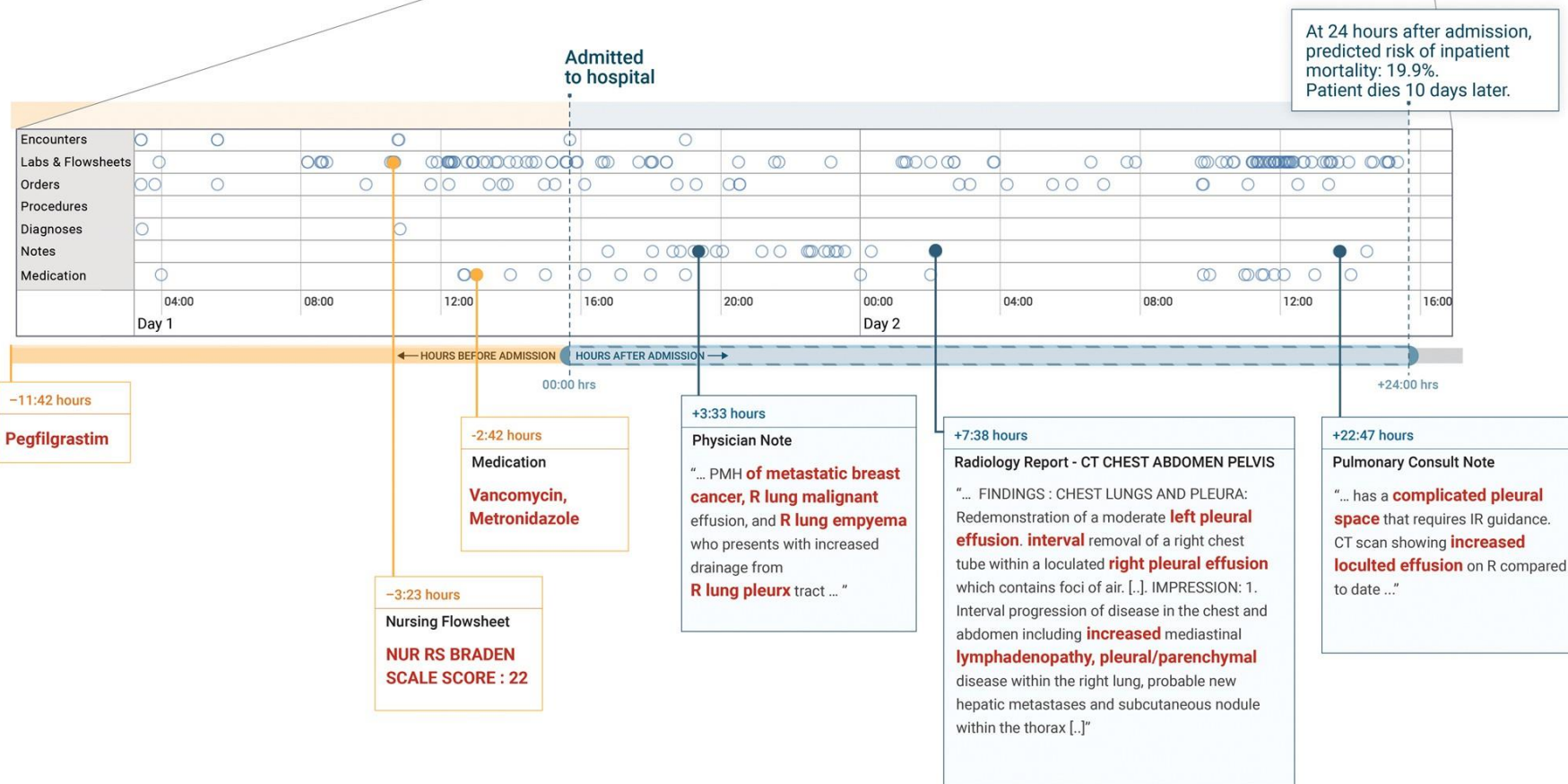
# Case study of model interpretation



Metastatic breast cancer with malignant pleural effusions and empyema

Risk of death prediction  
 By DL (19.9%)  
 By aEWS (9.3%)  
 175,639 data points (tokens)

## Patient Timeline





# Discussion

---

- Deep learning models performed better than traditional predictive models.
- Deep learning models achieved accurate predictions earlier than traditional models.
- The approach incorporated the entire electronic health record (EHR), including free-text notes, for predictions.
- The models outperformed existing EHR models in predicting mortality, unexpected readmission, and increased length of stay.
- No hand-selection of important variables but allowed the model to identify relevant data for each prediction.



# Discussion

---

- The study was retrospective and prospective trials are needed to demonstrate the improvement of care through accurate predictions.
- Further research is required to determine how models trained at one site can be best applied to different sites.
- The prediction of a patient's ICD-9 diagnoses was challenging but demonstrated the potential for aiding decision support and clinical trial recruitment.
- Further research is needed to explore the applicability and clinical utility of the approach and other methods for interpreting deep learning models.



# Impact to society

---

**Bloomberg**

## Google Is Training Machines to Predict When a Patient Will Die

AI advances by the 'Medical Brain' team could help the internet giant finally break into the health-care business

MEDTECH

## Google AI predicts hospital inpatient death risks with 95% accuracy

By **Conor Hale** · Jun 20, 2018 10:20am



GOOGLE · Published June 19, 2018 9:58am EDT

## Google AI can predict when you'll die with 95 percent accuracy, researchers say







# References

---

Donzé, J., Aujesky, D., Williams, D., & Schnipper, J. L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA internal medicine*, 173(8), 632-638.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1), 198.

Liu, V., Kipnis, P., Gould, M. K., & Escobar, G. J. (2010). Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Medical care*, 48(8), 739-744.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., . . . Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1), 18. <https://doi.org/10.1038/s41746-018-0029-1>

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4), 465-470.



Thank you for your attention