ARTICLE #1

Journal Article Summary Sheet Spring 2025

Use this Journal Article Summary Sheet when you are assigned to read a journal article.

THIS MUST BE AN EXPERIMENTAL ARTICLE - NO REVIEWS

Do not copy/paste any text of the article into this section - this includes "almost" copying and pasting, where some words are changed/rearranged. This will result in the assignment being returned to you with no credit. Each section will be graded on clarity, completeness and the student's ability to explain.

Read the article in this order: Abstract, Introduction, Discussion, Methods, Results

Citation/Article Information
Paste the APA citation to the article here: Mataraso, S. J., Espinosa, C. A., Seong, D., Reincke, S. M., Berson, E., Reiss, J. D., ... & Aghaeepour, N. (2025). A machine learning approach to leveraging electronic health records for enhanced omics analysis. *Nature Machine Intelligence*, *7*(2), 293-306.
What main area of science research is this article from? Machine Learning
What specific area of science is the article from? Machine Learning in health care

SCIENTIST INFORMATION
What is the first Author's name and what institution are they affiliated with? Samson J. Mataraso, Stanford University
What is the last Author's name and what institution are they affiliated with? Nima Aghaeepour, Stanford University
How many times has this article been cited? 9
What topics has the author's most recent five articles focused on? Machine Learning and bioinformatics
Where is the scientist's lab located/how would you be able to work with them? I go to California many summers anyways to visit family.

TITLE INFORMATION - Define all of the scientific terms in the article title and tell how they connect to the main idea of the study
Omics - the characterization and quantification of entire sets of biological molecules

Review of Literature: List ten key facts from the introduction of this article that were discovered in previous research. For each fact, give the author and year. No copying/pasting directly from the source.	
Fact	**Source (Author, Year)**

1. New omics technologies can measure many molecules cheaply, but small study sizes make results hard to repeat.	Schüssler-Fiorenza Rose et al., 2019
2. While statistics can help reduce false positives, machine learning has fewer tools, so some new methods use known important features to guide the model and prevent overfitting.	Ruiz et al., 2023
3. Another method is transfer learning, where a model is trained on a large dataset first and then applied to a smaller, related dataset.	Jiang et al., 2020
4. Modern deep learning has been added to traditional models like the Cox model to better analyze time to event data, even when some patient data is incomplete.	Katzman et al., 2016
5. While past methods mainly used omics or metadata, this study adds electronic health records to improve analysis, taking advantage of growing access to public and private EHR databases.	Goldberger et al., 2000
6. Multimodal machine learning mixes different data types by joining features early, combining processed data in the middle, or merging predictions at the end.	Steyaert et al., 2023
7. Some frameworks can mix these methods for a better end product.	Ding et al., 2022
8. EHR data is easier to get for many people, but combining it with omics data is difficult because early and intermediate fusion need full data from both types, and late fusion can't capture how the data types interact well.	Guarrasia et al., 2024
9. The scientists used COMET to predict labor timing in over 30,000 pregnant people from Stanford's EHR data, and for 61 of them, they measured 1,317 proteins from blood samples taken late in pregnancy.	Stelzer et al., 2021
10. The scientists used t-SNE to visualize the combined data by projecting relationships into two dimensions, grouping features with similar patterns.	Li et al., 2003

Vocabulary: List ten important scientific terms from the introduction to the article and provide definitions for them. Give the source where you got the definition (name of website is ok).	
Vocabulary term	**Definition/Source**
1. Cohort	**A group of people who all share one or more characteristics. (Science Direct)**
2. Genomics	**The branch of molecular biology concerned with the structure, function, evolution, and mapping of genomes. (National Human Genome Research Institute)**
3. Proteomics	**The large scale study of proteins, particularly with regard to their functions and structures. (Science Direct)**
4. Metabolomics	**The scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates, and products of metabolism. (National institutes of health)**
5. Clinical Decision Support System	**Health information technology systems designed to assist healthcare providers in making clinical decisions. (National institutes of health)**
6. Embedding	**The representation of data in a lower dimensional space, often used to capture the semantic meaning of the data. (GeeksforGeeks)**
7. Transfer Learning	**A machine learning technique where a model developed for a particular task is reused as the starting point for a model on a second task, allowing knowledge gained from one task to be applied to another (Nature)**
8. Multimodal Learning	**The process of integrating and analyzing data from multiple sources. (GeeksforGeeks)**
9. Early Vs Late fusion	**Combining the sources for a multimodal model before vs after processing. (GeeksforGeeks)**
10. Representation Learning	**Enabling a model to choose the most important features to focus on. (Science Direct)**

Gap in Literature

Identify the gap in literature (What they are trying to find out) They are trying to create a cheaper and more effective model for omics studies using integration of transfer learning.

Variables
Identify the Independent Variable in the study. The model?
Is this variable Numeric or Categorical? Categorical
Identify the Dependent Variable in the study. The accuracy/efficacy
Is this variable Numeric or Categorical? categorical

Methods - Describe each method from the study. *<u>This includes equations.</u>*
For this section, we want you to answer these questions:

- **What is this method used to measure/discover?**
- **What did they do?**
- **Why is it being used in this article?**

If you are having a hard time figuring out what the method is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). <u>Copy the table below for each method used.</u>

Some sample prompts:
"Please explain how the research technique _____ is used in science in easily understood terms"
"Please explain how the technique _____ is used in the article titled _____. "

Method #1 Data Sets	
What is Method #1 used for in science? Datasets are used to provide real world data that supports model training, testing, and validation in scientific studies.	**Source: medium**
Why did they use it in this study? Using these datasets allowed the scientists to study both	**Source (if any):**

a focused clinical group and a large diverse population. **What did they do?** They used two real world datasets, one focused on pregnancy and the other from a large general population, to supply proteomic and health data for developing and evaluating their predictive models. **What did they measure?** They measured 1,305 proteins in the Stanford group and about 2,894 proteins per person in the UK Biobank	

Method #2 Extraction of EHR data and pretraining cohort	
What is Method #2 used for in science? EHR data extraction is used to gather medical history and health information from electronic records for use in research and model training.	**Source: National Institutes of Health**
Why did they use it in this study? Using EHR data gave the scientists detailed health information before labor or diagnosis, which helped improve the accuracy of their models. **What did they do?** They pulled EHR data from specific tables like measurements, observations, and drug records. For pregnant patients, they collected data from the start of pregnancy to the time of sampling. For the pretraining group, they simulated sampling dates and gathered earlier health records. **What did they measure?** They extracted health records from the data set.	**Source (if any):**

Method #3 Embedding Process	
What is Method #3 used for science? The embedding process is used to turn data into numerical vectors making them usable by machine learning models.	**Source: science direct**
Why did they use it in this study? This helped the scientists turn messy EHR data into a clean and consistent format that could be used as input	**Source (if any):**

for deep learning models. **What did they do?** They grouped EHR codes by patient and day, treated each day's codes like words in a sentence, and used word2vec to create a 400 number vector for each code **What did they measure?** They created embeddings that represent each day's EHR data for every patient.	

Method #4 COMET deep learning architecture	
What is Method # used for in science? A deep learning model is used to build models that can learn patterns from complex data to make predictions or classifications.	**Source: GeeksforGeeks**
Why did they use it in this study? They used this method to combine EHR and omics data in a single model that could work for both predicting outcomes like time until labor or classifying outcomes like cancer death, showing that the approach works across different tasks. **What did they do?** They built a model called COMET with separate parts for EHR data, omics data, and a layer that combines both. **What did they measure?** They used EHR and omics data to predict time until labor and risk of death from cancer.	**Source (if any):**

Method #5 COMET hyperparameter details	
What is Method #5 used for in science? Hyperparameter tuning is used to find the best model settings to improve performance and prevent overfitting.	**Source: GeeksforGeeks**
Why did they use it in this study? Using this method helped the scientists get the most accurate and stable results from their deep learning models across different datasets and tasks. **What did they do?**	**Source (if any):**

They used threefold cross validation and grid search to test different combinations of settings, chose the best ones, and applied those settings to all later experiments. They repeated each experiment 25 times with different data splits. **What did they measure?** They worked with EHR and omics data and used model loss and prediction accuracy across different data splits.	

Method #6 Transformer-based architecture	
What is Method #6 used for in science? Transformer based architecture is used to process sequences of data by learning relationships between elements using attention mechanisms.	**Source: GeeksforGeeks**
Why did they use it in this study? They used a transformer to better model complex patterns in EHR data. **What did they do?** They changed the COMET model to use a transformer instead of an RNN for reading EHR records. They turned the EHR data into tokens, added time information, and used the transformer's output to represent each patient. **What did they measure?** They used the EHR data as input and created a detailed patient summary from it.	**Source (if any):**

Method #7 Ridge Regression Baseline	
What is Method # used for in science? Ridge regression is a type of linear model that adds a penalty to prevent overfitting by keeping the model's weights small.	**Source:**
Why did they use it in this study? They used ridge regression as a simple baseline to compare against their deep learning models, and added a version that could include knowledge from pretraining to see if it improved results. **What did they do?**	**Source (if any):**

They cleaned and normalized the data, one hot encoded EHR features, and used cross validation to test different settings. They tried versions with and without prior knowledge from pretraining and picked the best model based on validation performance. **What did they measure?** They used EHR and omics features and looked at how well the ridge model could make predictions.	

Method #8 Logistic regression	
What is Method # used for in science? Logistic regression is a method used to predict binary outcomes based on input features.	**Source: GeeksforGeeks**
Why did they use it in this study? They used it as a simple baseline to compare against more complex models and tested whether adding prior knowledge from pretraining could improve its predictions. **What did they do?** They cleaned the data and tested different model settings using cross-validation. They tried versions with and without pretrained priors and chose the best based on how well the model predicted outcomes. **What did they measure?** They used EHR and omics data to predict binary outcomes and measured how well the model performed.	**Source (if any):**

Method #9 External Validation	
What is Method #9 used for in science? External validation is used to test whether a model's findings hold true in new, independent datasets.	**Source: GeeksforGeeks**
Why did they use it in this study? They used it to increase the reliability of their model by proving it identifies the same protein as an indicator repeatedly. **What did they do?**	**Source (if any):**

They used external validation to check if the proteins identified by the COMET model were also linked to labor timing or cancer mortality in other patient groups. **What did they measure?** They used proteomics data from outside sources and calculated the correlation between selected proteins and the outcome.	

Method #10 Intermediate-node predictions	
What is Method # used for in science? Intermediate node predictions use internal parts of a model to understand what the model has learned from different types of input data.	**Source: Science Direct**
Why did they use it in this study? They used this method to separate the influence of EHR data, omics data, and their combination, so they could see how each part contributed to the final prediction. **What did they do?** They used this method to see how much each part of the data helped the model make its final prediction. **What did they measure?** They looked at the model's internal outputs to study how the EHR data, the omics data, and their combination each affected the prediction.	**Source (if any):**

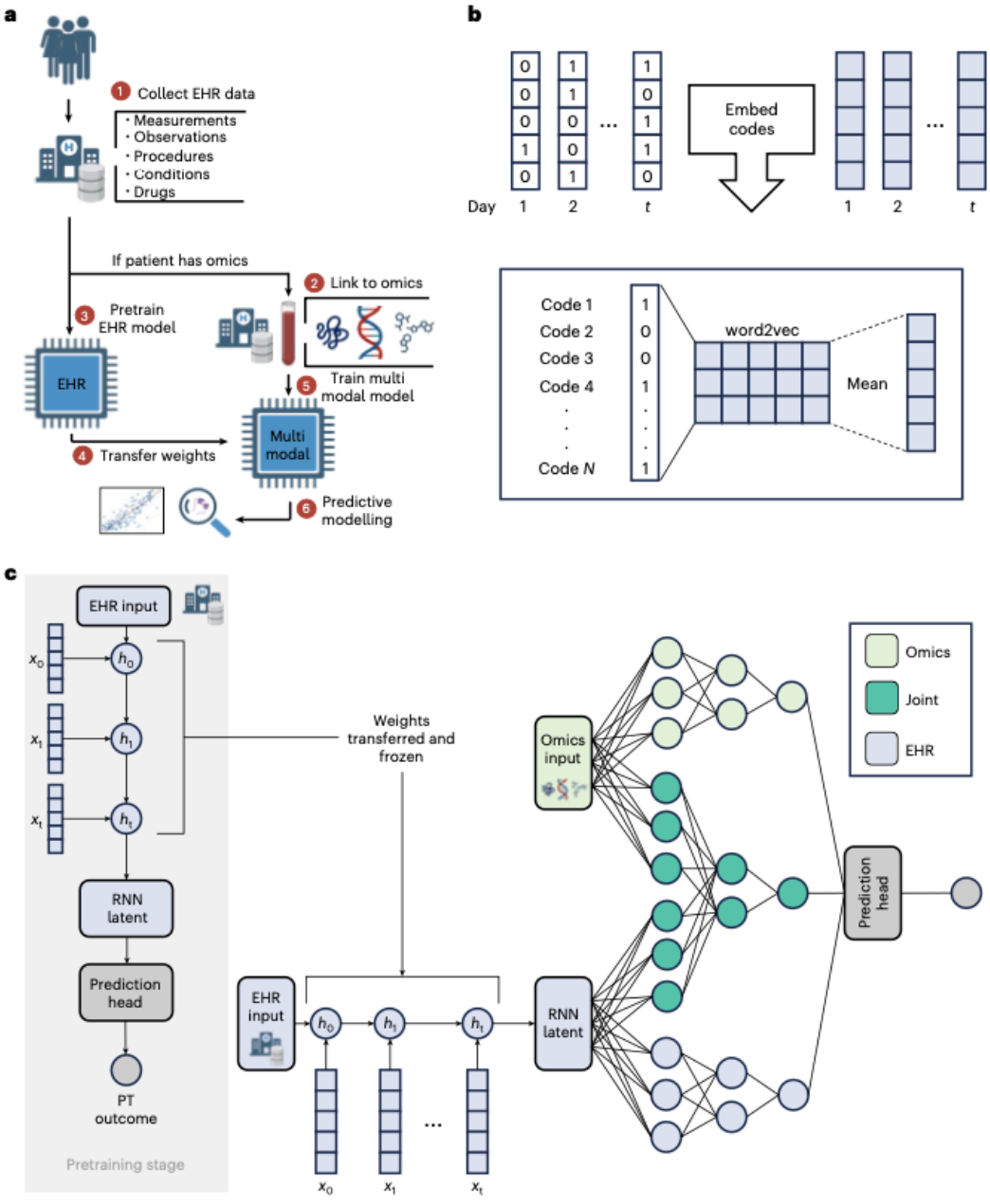
Method #11 Parameter space visualization	
What is Method # used for in science? Parameter space visualization is used to understand how a model behaves by comparing its outputs instead of its internal settings.	**Source: National institutes of Health**
Why did they use it in this study? They used it to track how the model changes during training and to compare different parts of the model, like the EHR, omics, and joint sections, in a way that focuses on function rather than raw parameters. **What did they do?** At each training step, they ran all the data through the	**Source (if any):**

model and collected the outputs from the intermediate nodes and final prediction. **What did they measure?** They used the model's prediction outputs to study how different parts of the model contributed.	

Results - Describe all of the key results from the study. For each section or figure, tell what was discovered in your own words. PASTE FIGURES BELOW WITH YOUR WORK. If you are having a hard time figuring out what the results mean, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting).

"Please explain Figure _____ in the article _____ in easily understood terms"
Copy the table below for each section of the results or figure.

Results section name or figure #1 Paste results figure here:



What does this section or figure tell us?

Panel A:

COMET first trains a model using only EHR data. Then it uses that trained model to help build a new one that combines EHR and omics data for better predictions.

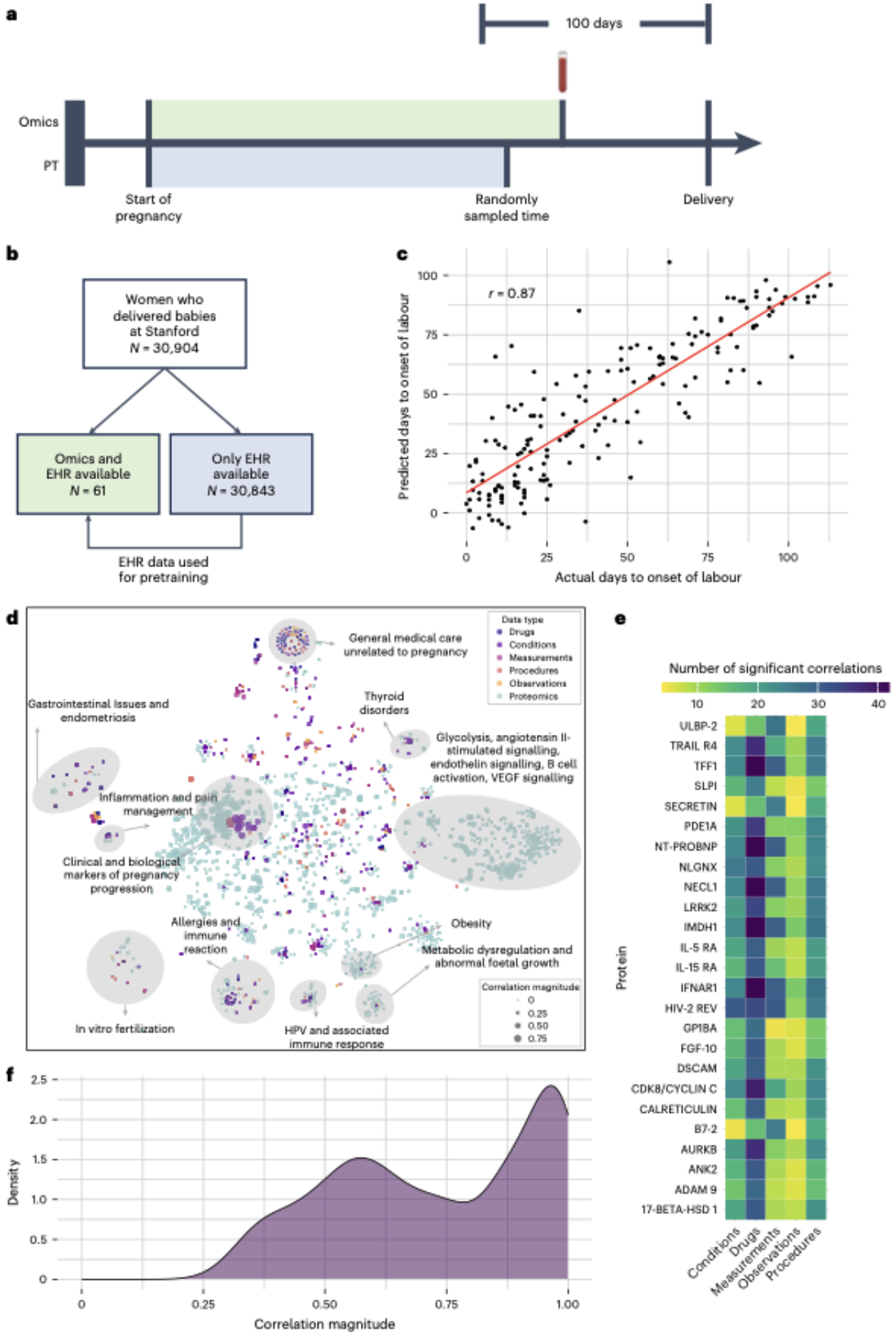
Panel B:

EHR records are turned into numbers using

Source:

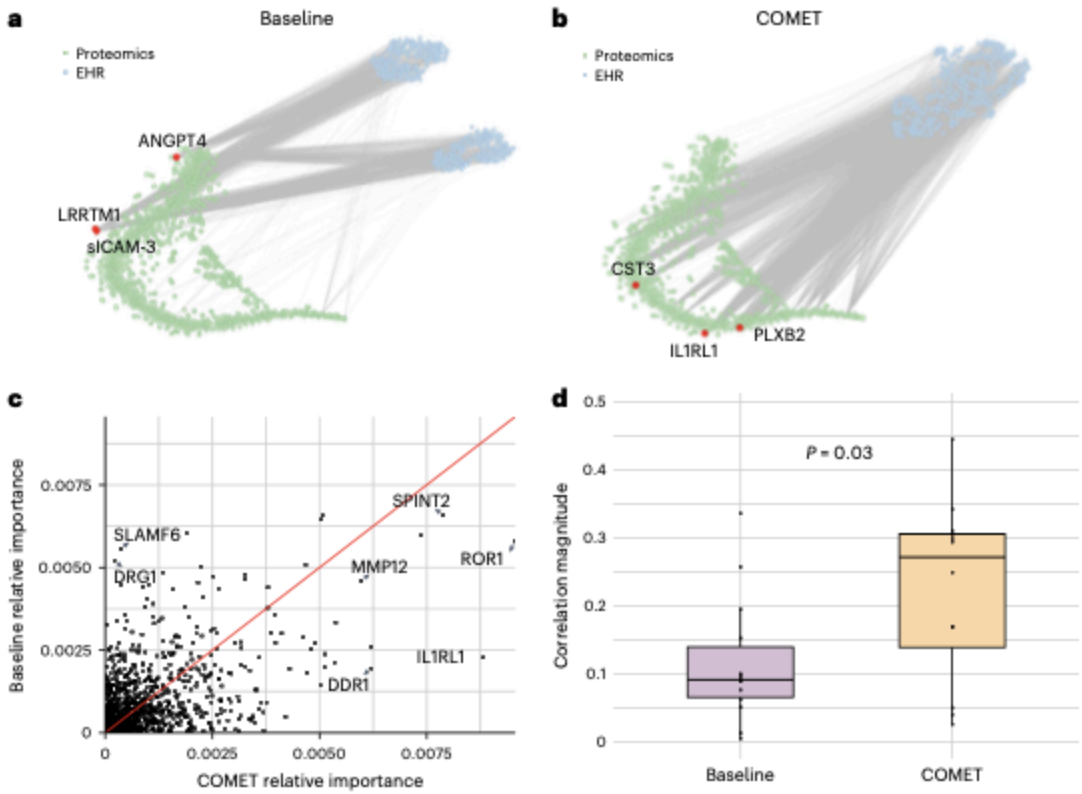
word2vec. The model averages these daily to understand the patient over time. Panel C: An RNN learns patterns from EHR data. Its knowledge is reused in a bigger model that also takes omics data to make stronger predictions. In its totality, this figure shows us the end result of their testing. This is a visual representation of how the model is created and how it actually works.	

Results section name or figure #2 Paste results figure here:


a
100 days
Omics
PT
Start of pregnancy
Randomly sampled time
Delivery
b
Women who delivered babies at Stanford N = 30,904
Omics and EHR available N = 61
Only EHR available N = 30,843
EHR data used for pretraining
c
r = 0.87
Predicted days to onset of labour
Actual days to onset of labour
0
25
50
75
100
d
Data type
Drugs
Conditions
Measurements
Procedures
Observations
Proteomics
General medical care unrelated to pregnancy
Thyroid disorders
Gastrointestinal issues and endometriosis
Glycolysis, angiotensin II-stimulated signalling, endothelin signalling, B cell activation, VEGF signalling
Inflammation and pain management
Clinical and biological markers of pregnancy progression
Allergies and immune reaction
Obesity
Metabolic dysregulation and abnormal foetal growth
In vitro fertilization
HPV and associated immune response
Correlation magnitude
0
0.25
0.50
0.75
e
Number of significant correlations
10
20
30
40
Protein
ULBP-2
TRAIL R4
TFF1
SLPI
SECRETIN
PDE1A
NT-PROBNP
NLGNX
NECL1
LRRK2
IMDH1
IL-5 RA
IL-15 RA
IFNAR1
HIV-2 REV
GP1BA
FGF-10
DSCAM
CDK8/CYCLIN C
CALRETICULIN
B7-2
AURKB
ANK2
ADAM 9
17-BETA-HSD 1
Conditions
Drugs
Measurements
Observations
Procedures
f
Density
0
0.5
1.0
1.5
2.0
2.5
0
0.25
0.50
0.75
1.00
Correlation magnitude

What does this section or figure tell us?	**Source:**
Panel A: EHR data was only used from early pregnancy up to a certain endpoint and the model predicted from there. Panel B: Only 61 of the 31k patients had omics and EHR data, while the rest were only EHR. Panel C: COMET made very accurate predictions for time to labour, showing it works well even with a small omics cohort. Panel D: EHR and protein features are clustered together by the model in ways that reflect real biological systems involved in pregnancy. This supports the credibility of the model Panel E: This panel shows significant correlation between some EHR data and proteins. This supports the model's credibility. Panel F: Many EHR features didn’t strongly link to proteins, showing that proteomics adds new, useful information. Overall, this figure supports the reliability of COMET by showing that it can not only make good predictions, but it also uses real connections between EHR data and proteins that are known.	

Results section name or figure #3 Paste results figure here:



What does this section or figure tell us?

Panel A vs B:

This is a visual comparison between a baseline model (A model which does not use the large EHR data base to pretrain and transfer learn) and COMET. It is clear that COMET is much better at linking EHR data and omics data.

Panel C:

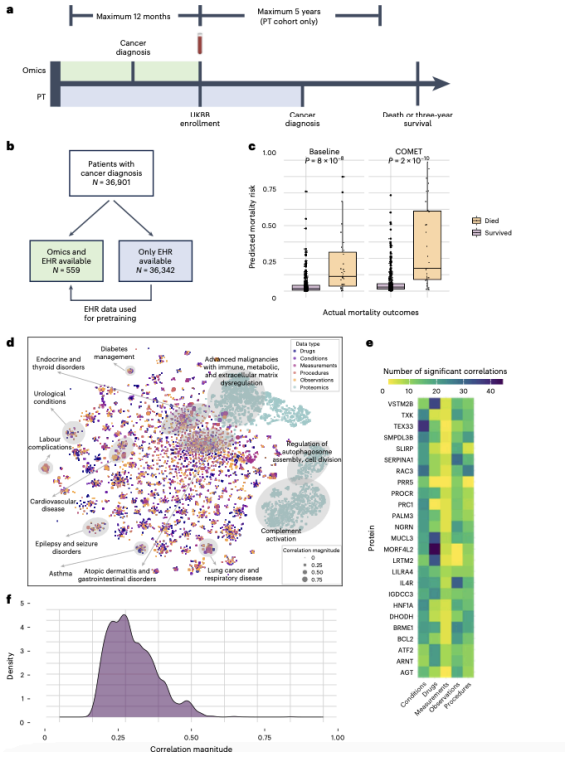
This shows us the importance of different proteins to COMET vs a standard model. COMET finds more and stronger importance then the standard model.

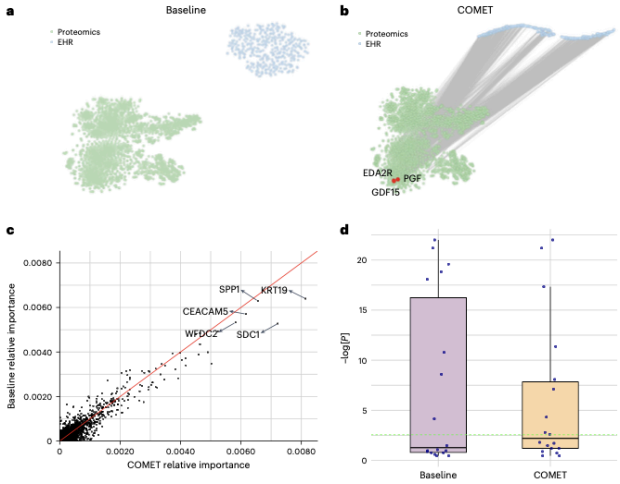
Panel D:

A comparison using the known (external to the training data) times of labor to compare the real correlation between proteins and the models predicted. COMET has significantly better predictors.

Source:

Results section name or figure #4 and #5 Paste results figure here:





What does this section or figure tell us?

These figures are pretty much the same as #2 and #3 but for mortality, so I'm just going to go over changes rather than redescribe everything.

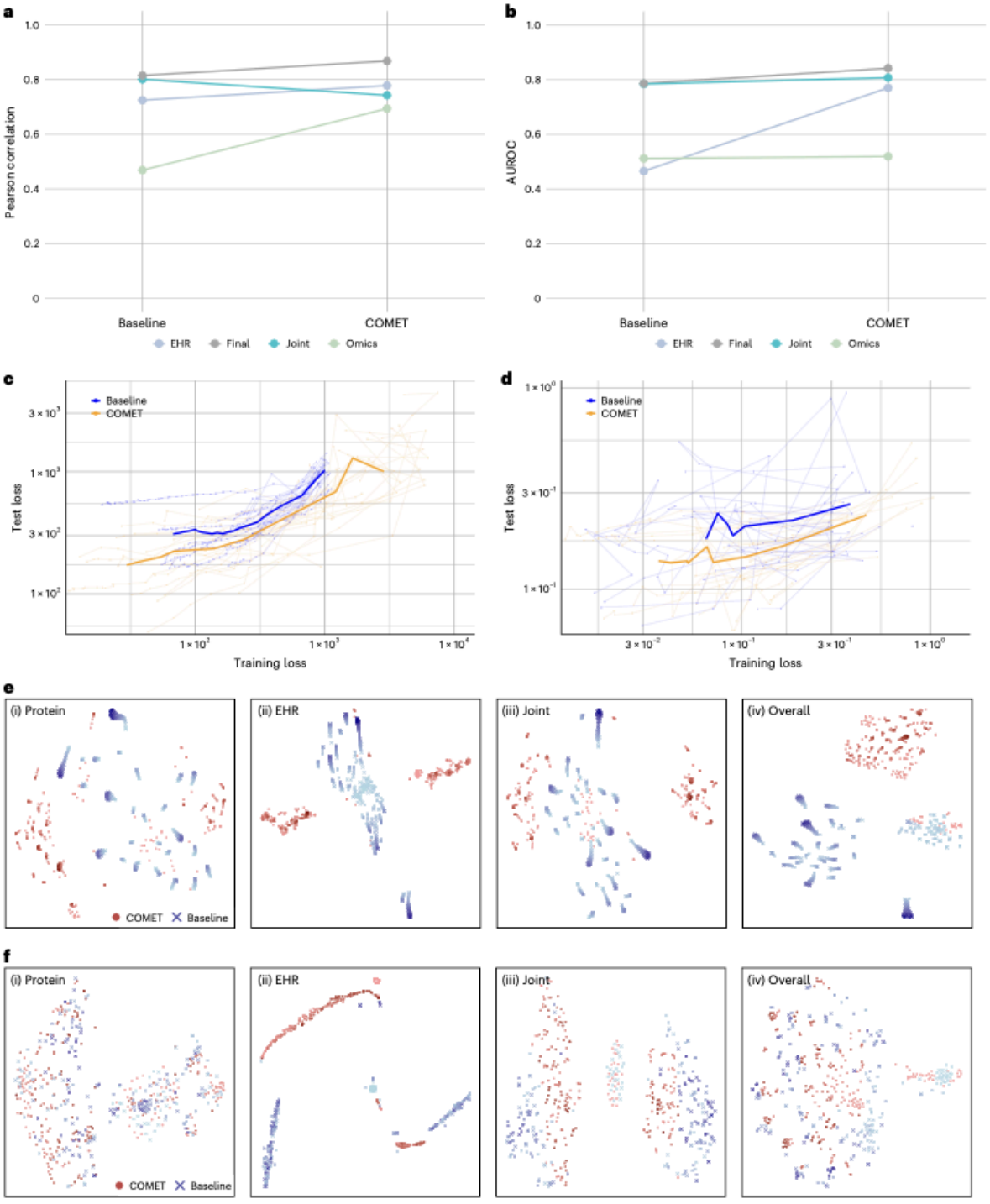
The cancer mortality model had a larger proportion of omics to EHR data, this could very likely lead to a better performance with this task.

This time the model seems to be better than the baseline by an even wider margin, supporting our earlier observation.

Format of panel D in second figure is changed to better reflect the difference between COMET and the baseline

Source:

Results section name or figure #6 Paste results figure here:



What does this section or figure tell us?	**Source:**
Panels A and B: These panels show the intermediate nodes in COMET vs baseline. In both cancer mortality and time to labor the COMET model out performs the	

baseline. Panels C and D: These panels generally show the effectiveness of training with new data. In both cancer mortality and time to labor, COMET has less overfitting and adapts better to the data then the baseline. Panels E and F: These panels show that COMET explores different and better areas of the parameter space than the baseline. In both cancer mortality and labor tasks, this helps the model learn patterns that are more generalizable and biologically accurate.	

Data Analysis - Describe one data analysis technique used in this article. If you are having a hard time figuring out what a data analysis technique or concept is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). Copy the table below for each part of the data analysis.

Some good prompts for this might be:
"Please explain how the data analysis technique _____ is used in science in easily understood terms"
"Please explain how _____ is used in the article titled _____. "

Data Analysis concept - Logistic Regression	
What is this method used for in science (types of variables)? It's used to predict the probability of a 'yes/no' or 'True/False' outcome using any number of categorical or numeric inputs	**Source:**
Why did they use it in this study? To compare whether or not the COMET worked better than the baseline model.	**Source:**

Discussion/Significance

Explain why this work is important/explain the specific contribution this article makes to the scientific literature The model they created is a provenly cheaper, more accurate, and more efficient model than the standard model. This could prove to be extremely useful in omics analysis.
Question for further research #1 with Independent and Dependent Variable How does the amount of available EHR data impact the performance of the model? I = EHR data D = performance
Question for further research #2 with Independent and Dependent Variable How does the number of EHR features included affect model accuracy in predicting omics traits? I = Number of EHR features D = Model accuracy
Question for further research #3 with Independent and Dependent Variable How does using different types of machine learning models affect disease prediction? I = Type of machine learning model D = Accuracy of disease prediction

ARTICLE #2

Journal Article Summary Sheet Spring 2025

Use this Journal Article Summary Sheet when you are assigned to read a journal article.

THIS MUST BE AN EXPERIMENTAL ARTICLE - NO REVIEWS

Do not copy/paste any text of the article into this section - this includes “almost” copying and pasting, where some words are changed/rearranged. This will result in the assignment being returned to you with no credit. Each section will be graded on clarity, completeness and the student’s ability to explain.

Read the article in this order: Abstract, Introduction, Discussion, Methods, Results

Citation/Article Information
Paste the APA citation to the article here: Ivanov, O., Molander, K., Dunne, R., Liu, S., Brecher, D., Masek, K., ... & Reilly, C. (2022). Detection of sepsis during emergency department triage using machine learning. *arXiv preprint arXiv:2204.07657*.
What main area of science research is this article from? Machine Learning
What specific area of science is the article from? Applications of ML models in healthcare.

SCIENTIST INFORMATION
What is the first Author's name and what institution are they affiliated with? Oleksandr Ivanov, Medinition inc
What is the last Author's name and what institution are they affiliated with? Christian Reilly, Medinition inc
How many times has this article been cited? 2
What topics has the author's most recent five articles focused on? ML and Healthcare
Where is the scientist's lab located/how would you be able to work with them? California

TITLE INFORMATION - Define all of the scientific terms in the article title and tell how they connect to the main idea of the study
Sepsis - When the bodies response to an infection goes out of control and attacks the body.

Review of Literature: List ten key facts from the introduction of this article that were discovered in previous research. For each fact, give the author and year. No copying/pasting directly from the source.

Fact	**Source (Author, Year)**
1. **Sepsis is a dangerous condition where the body's response to an infection becomes unbalanced and starts to damage its own organs.**	**(Singer et al. 2016)**
2. **Sepsis is a leading cause of death and illness worldwide**	**(Angus et al. 2001)**
3. **32 Million people a year get sepsis and 5 million die from it.**	**(Fleischmann, 2016)**
4. **In the U.S., over 1.7 million people get sepsis each year, and nearly 270,000 of them die from it.**	**(CDC, 2021)**
5. **In 2017, the U.S. spent more than $38 billion treating**	**(Liang, 2020)**

sepsis, making it the costliest medical condition to treat.	
6. Even survivors of sepsis often have long term health consequences.	**(Iwashyna et al. 2010)**
7. Even with improvements in medical care, the number of sepsis cases has not gone down.	**(Iwashyna et al. 2012)**
8. Sepsis remains the primary cause of death by infection	**(Singer et al. 2016)**
9. Sepsis is hard to treat because it's difficult to detect early. The symptoms are often mild and vary from person to person at first, but the condition can quickly become life threatening.	**(Vincent, 2016, de Grooth et al. 2018)**
10. Even a short delay of a few hours in treating sepsis can raise the risk of death.	**(Seymour et al. 2017)**

Vocabulary: List ten important scientific terms from the introduction to the article and provide definitions for them. Give the source where you got the definition (name of website is ok).

Vocabulary term	Definition/Source
1. Sepsis	**A dangerous condition where the body reacts too strongly to an infection and starts to harm itself. (National Health Institutes)**
2. Triage	**The process of quickly checking patients in the emergency room to decide who needs help first. (National Health Institutes)**
3. Systemic Inflammatory Response Syndrome (SIRS)	**The body's response to an infection, often causing fever, and other symptoms . (NIH)**
4. Quick Sequential Organ Failure Assessment Score (qSOFA)	**A methodology used to determine how critical a sepsis patient is. (NIH)**
5. XGBoost	**An optimization method that boosts the efficiency of ML algorithms. (Nvidia)**
6. Area Under the Curve (AUC)	**A metric to evaluate how well a model works used for models with binary outputs. (GeeksforGeeks)**

7. logistic regression	**The process by which a computer uses data to predict binary variables. (GeeksforGeeks)**
8. Sensitivity	**The true positive rate, or how often the machine accurately detects a given variable. (GeeksforGeeks)**
9. Specificity	**The true negative rate, or how often the model accurately doesn't predict the event. (GeeksforGeeks)**
10. PHI	**Protected Health information**

Gap in Literature
Identify the gap in literature (What they are trying to find out) They want to make a model that can be used for sepsis triage analysis.

Variables
Identify the Independent Variable in the study The model
Is this variable Numeric or Categorical? categorical
Identify the Dependent Variable in the study Accuracy of prediction
Is this variable Numeric or Categorical? Categorical

Methods - Describe each method from the study. *<u>This includes equations.</u>*
For this section, we want you to answer these questions:

- **What is this method used to measure/discover?**
- **What did they do?**
- **Why is it being used in this article?**

If you are having a hard time figuring out what the method is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). <u>Copy the table below for each method used.</u>

Some sample prompts:

“Please explain how the research technique _____ is used in science in easily understood terms”
“Please explain how the technique _____ is used in the article titled _____. “

Method #1 Data Collection	
What is Method #1 used for in science? Data collection is used to gather information needed to analyze and train machine learning models.	**Source: N/a**
Why did they use it in this study? They needed patient data from different hospitals to train and test a model that can detect sepsis during emergency department triage. **What did they do?** They got approval to use patient data, removed all personal information to protect privacy, and combined data from multiple hospitals into one consistent format while keeping track of hospital differences. **What did they measure?** They worked with patient records from several hospitals, and prepared the data to be used later in the study.	**Source (if any):**

Method #2 Study Sites and Demographics	
What is Method #2 used for in science? Analyzing study sites and demographics helps researchers understand where the data comes from and what kind of population is included in the study.	**Source: NIH**
Why did they use it in this study? They wanted to focus on adult patients in emergency departments and ensure the data was clean and relevant for detecting sepsis using machine learning. **What did they do?** They collected data from 16 hospitals across California, Oregon, and Hawaii. **What did they measure?** They reported the number of records, number of sepsis cases, and demographic details such as age, and also the	**Source (if any):**

rates of sepsis, severe sepsis, and septic shock.	

Method #3 Sepsis Diagnosis Definition	
What is Method #3 used for in science? Defining and labeling medical conditions helps researchers train and evaluate models by knowing which cases are positive or negative for a disease.	**Source: NIH**
Why did they use it in this study? They needed to label who had sepsis so the model could learn to detect it from triage data. **What did they do?** They used doctor's notes written within 24 hours of arrival and applied a standard definition of sepsis to decide who had it. **What did they measure?** They measured whether a patient had sepsis, severe sepsis, or septic shock, based on what was written in the free text diagnoses.	**Source (if any):**

Method #4 Sepsis screening protocol	
What is Method #4 used for in science? Rule based screening methods help quickly identify patients who might have a condition.	**Source: NIH**
Why did they use it in this study? They wanted to compare how well a standard, commonly used rule based method performs against their machine learning model. **What did they do?** They applied the standard SIRS protocol based on vital signs and signs of infection to patients at triage. **What did they measure?** They measured whether patients met the SIRS criteria plus signs of infection during triage, and compared those results to the KATE Sepsis model.	**Source (if any):**

Method #5 Features overview

What is Method #5 used for in science? Using features means selecting important information from the data to help the machine learning model make predictions.	**Source: GeeksforGeeks**
Why did they use it in this study? They used features from patient records to help the model predict who might have sepsis during emergency department triage. **What did they do?** They used different types of data from the electronic health record such as vital signs, age, sex, and medical history. They cleaned the data by removing impossible values and only used information available at the time of triage. **What did they measure?** They measured and used these features to train the machine learning model to predict sepsis based on data available right at triage.	**Source (if any):**

Method #6 Clinical concepts extraction algorithm	
What is Method #6 used for in science? Clinical concept extraction is used to pull out important medical terms from free text in health records so models can understand and use this information.	**Source: NIH**
Why did they use it in this study? They needed to turn free text notes from doctors and nurses into structured features to help the machine learning model better understand each patient. **What did they do?** They created a step by step algorithm to process raw text, break it into phrases, and match the phrases to known medical terms using a large medical database called UMLS. These terms were then added as features. **What did they measure?** They collected and counted unique clinical terms from different sections of each patient's record, such as medical history or complaints.	**Source (if any):**

Method #7 Machine learning algorithms	
What is Method #7 used for in science? Machine learning algorithms find patterns in data to make predictions or classifications.	**Source:**
Why did they use it in this study? They used XGBoost and logistic regression to handle large clinical datasets and improve sepsis prediction. **What did they do?** They trained an XGBoost model and used its output as input for logistic regression, with balanced class weights and fivefold cross validation. **What did they measure?** They measured how well the model predicted sepsis using patient records.	**Source (if any):**

Method #8 Software	
What is Method #8 used for in science? Software tools are used to run and evaluate machine learning models and data processing steps.	**Source: GeeksforGeeks**
Why did they use it in this study? They used different tools to handle text extraction, model training, and evaluation. **What did they do?** They built the concept extraction tool in Java and developed the machine learning pipeline in Python. **What did they measure?** They used the software to analyze model performance and run statistical tests.	**Source (if any):**

Method #9 Data Availability	
What is Method #9 used for in science? Data sharing allows others to access and verify research findings.	**Source: NIH**
Why did they use it in this study? They had to limit access to the full data because of agreements with hospitals.	**Source (if any):**

What did they do? They kept the full dataset private but offered a small sample upon request. **What did they measure?** N/a.	

Results - Describe all of the key results from the study. For each section or figure, tell what was discovered in your own words. PASTE FIGURES BELOW WITH YOUR WORK. If you are having a hard time figuring out what the results mean, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting).

"Please explain Figure _____ in the article _____ in easily understood terms"
Copy the table below for each section of the results or figure.

Results section Primary results	
What does this section or figure tell us? KATE Sepsis greatly outperforms standard screening, with much higher AUC and sensitivity across all sepsis categories. It detects most of the cases standard screening finds, but the reverse is not true. It benefits from using a wider range of patient features beyond basic vital signs. This suggests a significant advantage in early and accurate detection.	**Source:**

Results section Compare of KATE Sepsis and standard screening	
What does this section or figure tell us? KATE Sepsis matches standard screening in specificity but is much more sensitive. Standard screening misses many cases because it relies only on abnormal vital signs, which most sepsis patients don't have at triage. Many sepsis patients present with normal SIRS vital signs (see	**Source:**

figure #2). This explains why standard screening underperforms compared to KATE Sepsis.	

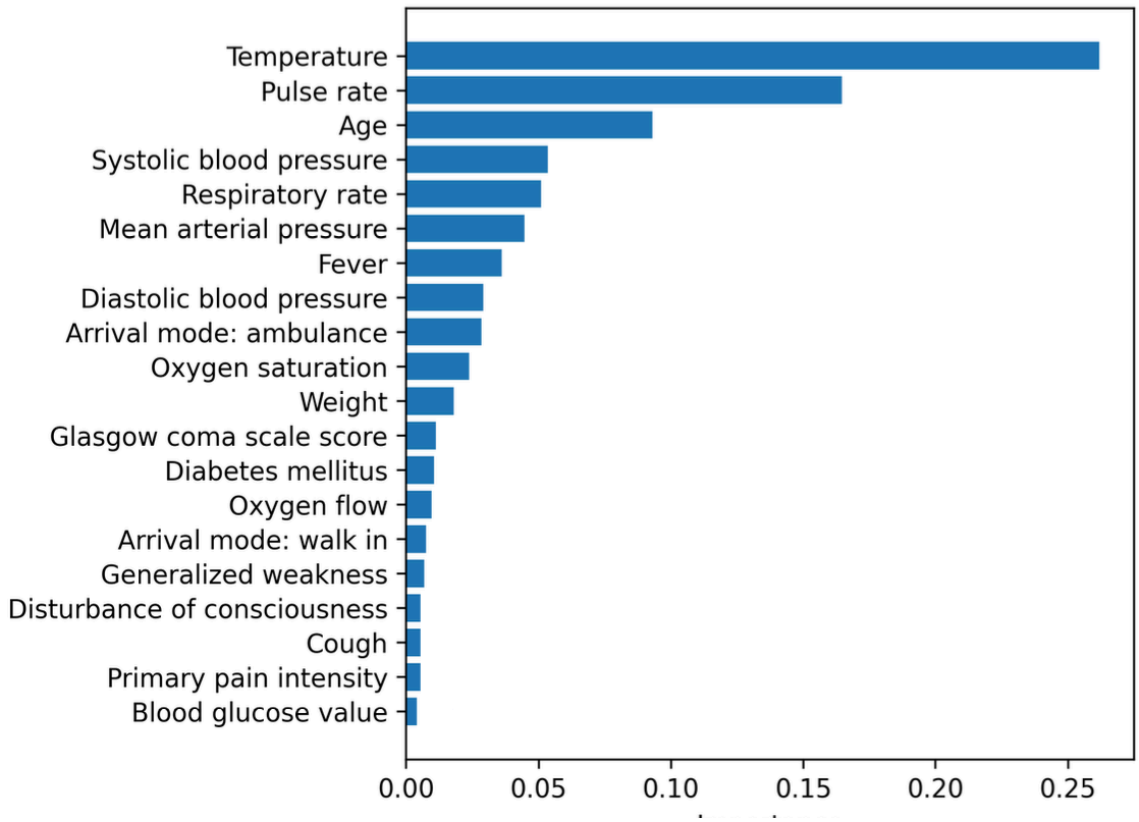
Table #3 Paste results table here:

Table 3. Performance metrics and 95% confidence intervals (in parentheses) for the adult population from 16 study hospital sites (Feb 2015 to Jul 2021) of KATE Sepsis and standard screening evaluated 512,949 medical records with 9,257 sepsis diagnoses using 5-fold cross validation.

Group	AUC	Sensitivity	Specificity	F1-score	Accuracy	Precision
KATE Sepsis	0.9423 (0.9401 - 0.9441)	0.7109 (0.7012 - 0.7198)	0.9481 (0.9475 - 0.9487)	0.31365 (0.3068 - 0.3193)	0.94385 (0.9432 - 0.9444)	0.20121 (0.1961 - 0.2055)
Standard screening	0.6826 (0.6774 - 0.6878)	0.408 (0.3971 - 0.4186)	0.9572 (0.9568 - 0.9578)	0.21844 (0.2128 - 0.2247)	0.94731 (0.9468 - 0.9479)	0.14915 (0.1448 - 0.1539)

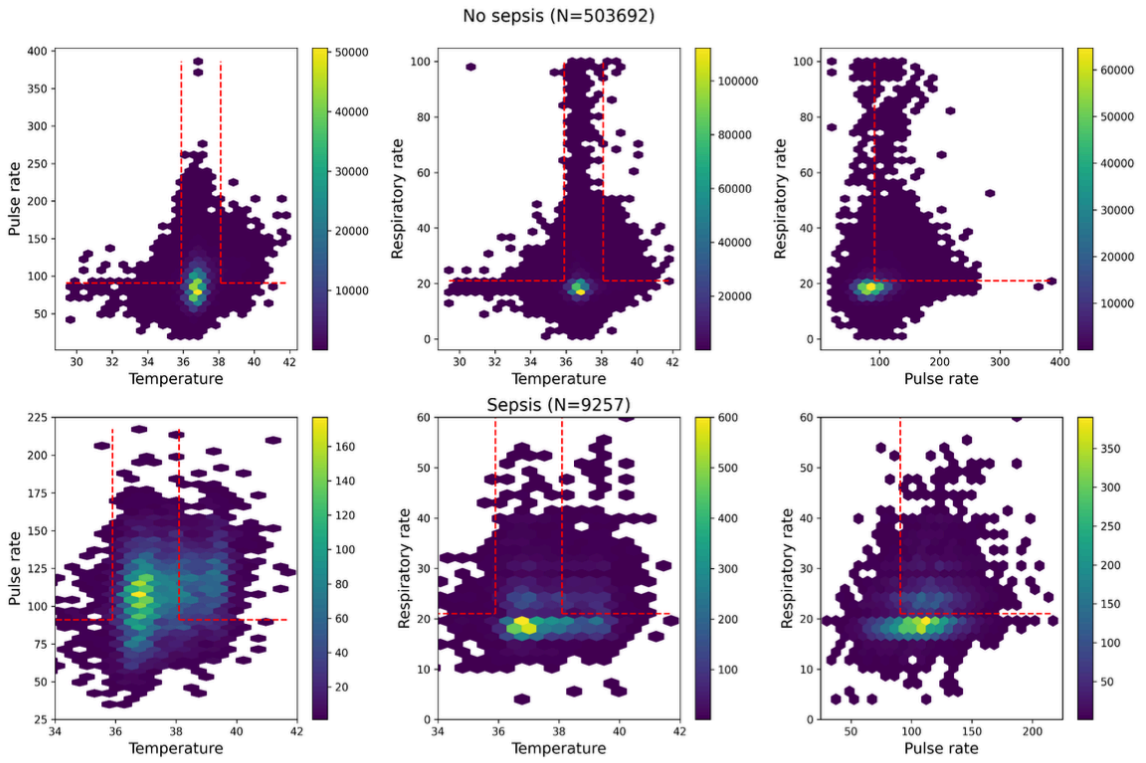
What does this section or figure tell us? This table shows us that over the 500,000+ test cases their model has continually proven to catch almost twice as many true positives while keeping the false positives the same.	**Source:**

figure #1 Paste results figure here:



What does this section or figure tell us?	**Source:**
This figure shows how the newly created KATE sepsis model values different features. Temperature being the most important factor, but with pulse rate not far behind.	

Results section name or figure # 2 Paste results figure here:



What does this section or figure tell us?	**Source:**
These heatmaps compare vital signs at triage for patients without sepsis and with sepsis. The red boxes show the SIRS criteria ranges. Most sepsis patients' vital signs fall outside the SIRS triggering area. Overall, this figure shows how lacking current triage technology is and why their model is such a breakthrough.	

Data Analysis - Describe one data analysis technique used in this article. If you are having a hard time figuring out what a data analysis technique or concept is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). Copy the table below for each part of the data analysis.

Some good prompts for this might be:
"Please explain how the data analysis technique _____ is used in science in easily understood terms"
"Please explain how _____ is used in the article titled _____. "

Data Analysis concept XGBoost	
What is this method used for in science (types of variables)? XGBoost is a machine learning method that builds many decision trees to predict outcomes. It works with numbers and categories and is good at handling complex healthcare data.	**Source:**
Why did they use it in this study? They used XGBoost to predict sepsis early at emergency triage by analyzing patient vital signs and other info quickly and accurately.	**Source:**

Discussion/Significance

Explain why this work is important/explain the specific contribution this article makes to the scientific literature The model they created could theoretically help save lives, it showed significantly more accuracy then the widely used SIRS method.
Question for further research #1 with Independent and Dependent Variable Does adding more clinical text data improve sepsis prediction accuracy? IV: Amount of data DV: Accuracy
Question for further research #2 with Independent and Dependent Variable Can machine learning models trained on sepsis triage data predict sepsis outcomes like length of stay or mortality? IV: ML model DV: prediction
Question for further research #3 with Independent and Dependent Variable Can use of the KATE Sepsis model in sepsis triage improve clinical outcomes? IV: Real world testing Vs. Simulated DV: Effectiveness

ARTICLE #3

Journal Article Summary Sheet Spring 2025

Use this Journal Article Summary Sheet when you are assigned to read a journal article.

THIS MUST BE AN EXPERIMENTAL ARTICLE - NO REVIEWS

Do not copy/paste any text of the article into this section - this includes "almost" copying and pasting, where some words are changed/rearranged. This will result in the assignment being returned to you with no credit. Each section will be graded on clarity, completeness and the student's ability to explain.

Read the article in this order: Abstract, Introduction, Discussion, Methods, Results

Citation/Article Information
Paste the APA citation to the article here: Almeida, T., Moreno, P., & Barata, C. (2025). Prediction of 30-day hospital readmission with clinical notes and EHR information. *arXiv preprint arXiv:2503.23050*.
What main area of science research is this article from? Machine Learning
What specific area of science is the article from? Machine Learning in healthcare

SCIENTIST INFORMATION
What is the first Author's name and what institution are they affiliated with? Tiago A. Almeida, University of Sao Carlos
What is the last Author's name and what institution are they affiliated with? Catarina Barata, Institute for Systems and Robotics
How many times has this article been cited? none
What topics has the author's most recent five articles focused on?
Where is the scientist's lab located/how would you be able to work with them?

TITLE INFORMATION - Define all of the scientific terms in the article title and tell how they connect to the main idea of the study

Review of Literature: List ten key facts from the introduction of this article that were discovered in previous research. For each fact, give the author and year. No copying/pasting directly from the source.	
Fact	**Source (Author, Year)**
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Vocabulary: List ten important scientific terms from the introduction to the article and provide definitions for them. Give the source where you got the definition (name of website is ok).	
Vocabulary term	**Definition/Source**
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

9.	
10.	

Gap in Literature
Identify the gap in literature (What they are trying to find out)

Variables
Identify the Independent Variable in the study
Is this variable Numeric or Categorical?
Identify the Dependent Variable in the study
Is this variable Numeric or Categorical?

Methods - Describe each method from the study. *<u>This includes equations.</u>*
For this section, we want you to answer these questions:

- **What is this method used to measure/discover?**
- **What did they do?**
- **Why is it being used in this article?**

If you are having a hard time figuring out what the method is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). <u>Copy the table below for each method used.</u>

Some sample prompts:
"Please explain how the research technique _____ is used in science in easily understood terms"
"Please explain how the technique _____ is used in the article titled _____. "

Method #1 Name	
What is Method #1 used for in science?	**Source:**
Why did they use it in this study? **What did they do?** **What did they measure?**	**Source (if any):**

Results - Describe all of the key results from the study. For each section or figure, tell what was discovered in your own words. PASTE FIGURES BELOW WITH YOUR WORK. If you are having a hard time figuring out what the results mean, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting).

"Please explain Figure _____ in the article _____ in easily understood terms"
Copy the table below for each section of the results or figure.

Results section name or figure # _____ Paste results figure here:	
What does this section or figure tell us?	**Source:**

Data Analysis - Describe one data analysis technique used in this article. If you are having a hard time figuring out what a data analysis technique or concept is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). Copy the table below for each part of the data analysis.

Some good prompts for this might be:
"Please explain how the data analysis technique _____ is used in science in easily understood terms"
"Please explain how _____ is used in the article titled _____. "

Data Analysis concept Name	
What is this method used for in science (types of variables)?	**Source:**

Why did they use it in this study?	Source:

Discussion/Significance

Explain why this work is important/explain the specific contribution this article makes to the scientific literature
Question for further research #1 with Independent and Dependent Variable
Question for further research #2 with Independent and Dependent Variable
Question for further research #3 with Independent and Dependent Variable

ARTICLE #4

Journal Article Summary Sheet Spring 2025

Use this Journal Article Summary Sheet when you are assigned to read a journal article.

THIS MUST BE AN EXPERIMENTAL ARTICLE - NO REVIEWS

Do not copy/paste any text of the article into this section - this includes "almost" copying and pasting, where some words are changed/rearranged. This will result in the assignment being returned to you with no credit. Each section will be graded on clarity, completeness and the student's ability to explain.

Read the article in this order: Abstract, Introduction, Discussion, Methods, Results

Citation/Article Information
Paste the APA citation to the article here:

What main area of science research is this article from?
What specific area of science is the article from?

SCIENTIST INFORMATION
What is the first Author's name and what institution are they affiliated with?
What is the last Author's name and what institution are they affiliated with?
How many times has this article been cited?
What topics has the author's most recent five articles focused on?
Where is the scientist's lab located/how would you be able to work with them?

TITLE INFORMATION - Define all of the scientific terms in the article title and tell how they connect to the main idea of the study

Review of Literature: List ten key facts from the introduction of this article that were discovered in previous research. For each fact, give the author and year. No copying/pasting directly from the source.	
Fact	**Source (Author, Year)**
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Vocabulary: List ten important scientific terms from the introduction to the article and provide definitions for them. Give the source where you got the definition (name of website is ok).	
Vocabulary term	**Definition/Source**
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Gap in Literature
Identify the gap in literature (What they are trying to find out)

Variables
Identify the Independent Variable in the study
Is this variable Numeric or Categorical?
Identify the Dependent Variable in the study
Is this variable Numeric or Categorical?

Methods - Describe each method from the study. *<u>This includes equations.</u>*
For this section, we want you to answer these questions:

- **What is this method used to measure/discover?**
- **What did they do?**
- **Why is it being used in this article?**

If you are having a hard time figuring out what the method is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). <u>Copy the table below for each method used.</u>

Some sample prompts:
"Please explain how the research technique _____ is used in science in easily understood terms"
"Please explain how the technique _____ is used in the article titled _____. "

Method #1 Name	
What is Method #1 used for in science?	**Source:**
Why did they use it in this study? **What did they do?** **What did they measure?**	**Source (if any):**

Results - Describe all of the key results from the study. For each section or figure, tell what was discovered in your own words. PASTE FIGURES BELOW WITH YOUR WORK. If you are having a hard time figuring out what the results mean, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting).

"Please explain Figure _____ in the article _____ in easily understood terms"
Copy the table below for each section of the results or figure.

Results section name or figure # _____ Paste results figure here:	
What does this section or figure tell us?	**Source:**

Data Analysis - Describe one data analysis technique used in this article. If you are having a hard time figuring out what a data analysis technique or concept is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). Copy the table below for each part of the data analysis.

Some good prompts for this might be:
"Please explain how the data analysis technique _____ is used in science in easily understood terms"
"Please explain how _____ is used in the article titled _____. "

Data Analysis concept Name	
What is this method used for in science (types of variables)?	**Source:**
Why did they use it in this study?	**Source:**

Discussion/Significance

Explain why this work is important/explain the specific contribution this article makes to the scientific literature
Question for further research #1 with Independent and Dependent Variable
Question for further research #2 with Independent and Dependent Variable
Question for further research #3 with Independent and Dependent Variable

ARTICLE #5

Journal Article Summary Sheet Spring 2025

Use this Journal Article Summary Sheet when you are assigned to read a journal article.

THIS MUST BE AN EXPERIMENTAL ARTICLE - NO REVIEWS

Do not copy/paste any text of the article into this section - this includes "almost" copying and pasting, where some words are changed/rearranged. This will result in the assignment being returned to you with no credit. Each section will be graded on clarity, completeness and the student's ability to explain.

Read the article in this order: Abstract, Introduction, Discussion, Methods, Results

Citation/Article Information
Paste the APA citation to the article here:
What main area of science research is this article from?
What specific area of science is the article from?

SCIENTIST INFORMATION
What is the first Author's name and what institution are they affiliated with?
What is the last Author's name and what institution are they affiliated with?
How many times has this article been cited?
What topics has the author's most recent five articles focused on?
Where is the scientist's lab located/how would you be able to work with them?

TITLE INFORMATION - Define all of the scientific terms in the article title and tell how they connect to the main idea of the study

Review of Literature: List ten key facts from the introduction of this article that were discovered in previous research. For each fact, give the author and year. No copying/pasting directly from the source.	
Fact	**Source (Author, Year)**
1.	

2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Vocabulary: List ten important scientific terms from the introduction to the article and provide definitions for them. Give the source where you got the definition (name of website is ok).	
Vocabulary term	**Definition/Source**
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Gap in Literature
Identify the gap in literature (What they are trying to find out)

Variables
Identify the Independent Variable in the study
Is this variable Numeric or Categorical?
Identify the Dependent Variable in the study
Is this variable Numeric or Categorical?

Methods - Describe each method from the study. *<u>This includes equations.</u>*
For this section, we want you to answer these questions:

- **What is this method used to measure/discover?**
- **What did they do?**
- **Why is it being used in this article?**

If you are having a hard time figuring out what the method is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). <u>Copy the table below for each method used.</u>

Some sample prompts:
"Please explain how the research technique _____ is used in science in easily understood terms"
"Please explain how the technique _____ is used in the article titled _____. "

Method #1 Name	
What is Method #1 used for in science?	**Source:**
Why did they use it in this study? **What did they do?** **What did they measure?**	**Source (if any):**

Results - Describe all of the key results from the study. For each section or figure, tell what was discovered in your own words. PASTE FIGURES BELOW WITH YOUR WORK. If you are having a hard time figuring out what the results mean, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting).

"Please explain Figure _____ in the article _____ in easily understood terms"
Copy the table below for each section of the results or figure.

Results section name or figure # _____ Paste results figure here:	
What does this section or figure tell us?	**Source:**

Data Analysis - Describe one data analysis technique used in this article. If you are having a hard time figuring out what a data analysis technique or concept is or why it was used in the study, you may do online research. List your source. You may use AI to figure this out. If you do this, write down the prompt you entered into AI under "source" and make sure to paraphrase/ write your answer in your own words (no copying and pasting). Copy the table below for each part of the data analysis.

Some good prompts for this might be:
"Please explain how the data analysis technique _____ is used in science in easily understood terms"
"Please explain how _____ is used in the article titled _____. "

Data Analysis concept Name	
What is this method used for in science (types of variables)?	**Source:**
Why did they use it in this study?	**Source:**

Discussion/Significance

Explain why this work is important/explain the specific contribution this article makes to the scientific literature
Question for further research #1 with Independent and Dependent Variable
Question for further research #2 with Independent and Dependent Variable
Question for further research #3 with Independent and Dependent Variable