# Time-dependent profiling of distinct stages prior to breast cancer onset using free-text diagnosis names

Benjamin Holmes, Foad Green, Joshua Loving, Amol Bhalla
Syapse, San Francisco, CA

Syapse®
ISPOR 2023
*May 7-10, 2023*
*Abstract Number: RWD14*

## BACKGROUND

Early identification of disease in a patient's journey is of critical importance in patient care.[1] However, the complex, multidimensional and frequently incomplete nature of patient data make this difficult.[2] While well-structured patient data can leave gaps in understanding of patient conditions, there is another rich source of information: free-text about the patient contained in Electronic Health Records. Machine Learning (ML) methods have promising applications in this space, but can suffer from low explainability, expert requirements for feature engineering, and prohibitively long training requirements.[3] We present a novel interpretable pipeline to predict relative time to diagnosis in breast cancer (BC) patients using natural language vectors produced by sentence transformers on patients' ICD-10-CM diagnosis descriptions, and which required minimal feature engineering.

## METHODS

- 12,360 patients with a BC diagnosis, stratified by age, race, and sex, were pulled from the Syapse Learning Health Network about cancer patients from 2013-2022. All ICD-10 diagnosis descriptions from these patients' EHRs were vectorized using Biomed-RoBERTa, a BERT sentence transformer.[4]
- Vectors were clustered with K-means, with a silhouette score used for optimal cluster number. The centroids of these clusters were again clustered, and all the diagnoses from these centroid-clusters were evaluated by clinical experts for coherence and utility.
- Patients were randomly assigned one of two categories, 30 days or 150 days before breast cancer diagnosis. Medical history after this date was blanked. In both populations, one year of each patient's medical history was considered.
- A patient's diagnosis in the visits during this time was transformed with the sentence transformer into vector space, and this position, its' proximity to the diagnosis of breast cancer, and the distance that the patient had moved in this vector space in the previous 1, 3, and 5 visits was used to create an XGBoost model to determine class membership, with a 75/25 training/test split.

## RESULTS

- 9,915 individual diagnoses were extracted from the patient files. Silhouette scores determined that the optimal cluster number from these diagnoses was 563. The centroids of each of these clusters was taken, and further grouped into an even 100 clusters.
- When the centroid-clusters were evaluated manually, they were found to group the diagnoses into clear topics. All patients' BC diagnoses came from a single centroid cluster, and none had diagnoses from this cluster before BC diagnosis date.
- Features were kept intentionally simple:distance from the terminal (BC) cluster, and distance the patient's diagnosis had moved in the last 1, 3, and 5 visits. These features were a simple preview of the patient's path through the diagnosis space and towards the terminal cluster.
- The XGBoost model trained using patient position in this vector space achieved an F1 score of 0.73, with 75% accuracy.
- Two versions of the XGBoost model were attempted, one which took as input the vectors of the specific diagnoses a patient received at each visit, and another which instead used the centroid of the cluster that each diagnosis belonged to. If these two models gave equivalent results across the patient group, it was an indication that the clusters were tightly-grouped: that the centroids of clusters were good approximations for the spatial coordinates of any member of their cluster. Indeed, there was no significant change in model accuracy between these two feature sets, indicating that the clusters were tightly-grouped.

## CONCLUSIONS

- We demonstrate that we can discover important information about a patient using a sentence transformer based ML pipeline. In particular, we predicted time to diagnosis via classification.
- The unsupervised clustering technique applied to vectorized text descriptions of diagnoses produced useful groupings of the diagnoses, providing additional interpretability, particularly for clinically focused users.
- Of special interest was the generalizability of this method: given nothing but diagnosis names, this system produced clinically-useful clusters and found a signal separating two patient groups at different stages before BC diagnosis.
- Further work will incorporate refinements to the models in this framework as well as incorporation of text data from different sources.

**FIGURE 1** Creation of patient groups split by XGBoost

Patients were split into one of two groups: in one, their medical history 30 days before initial diagnosis was excluded, and records up to one year before that date were considered. Inthe other group, the exclusion period was extended to 150 days, again with a one-year examination period

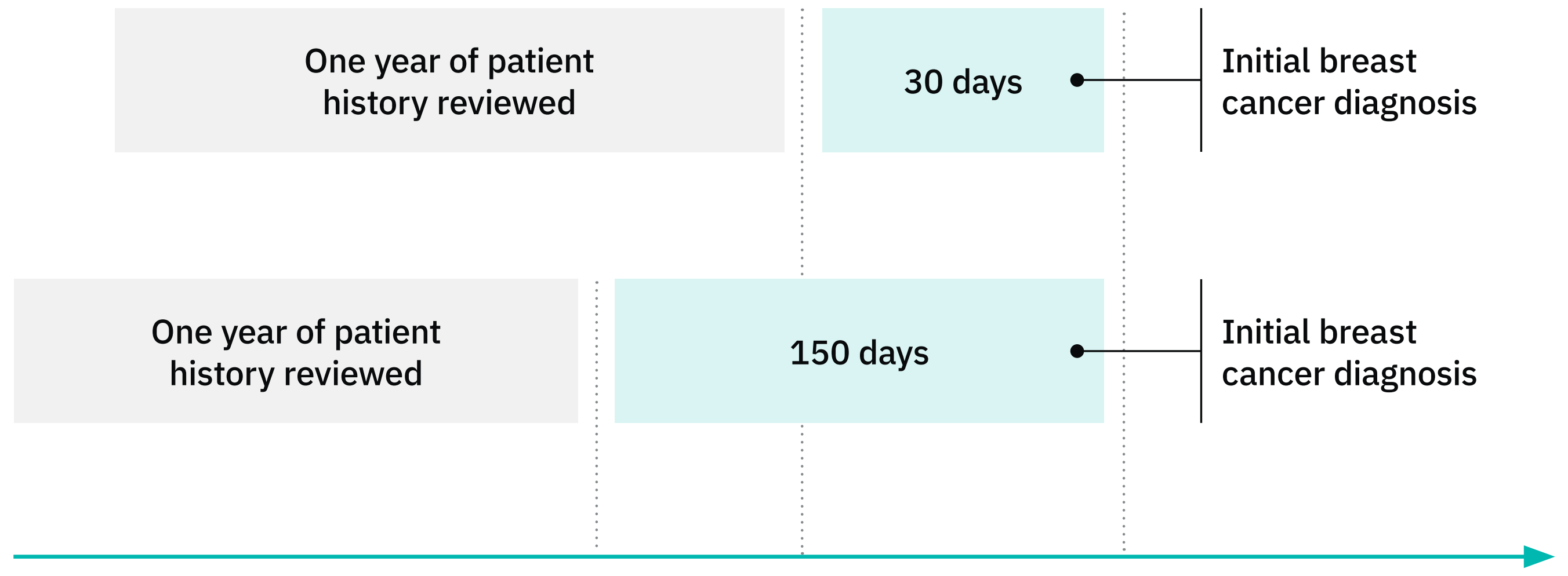

One year of patient history reviewed
30 days
Initial breast cancer diagnosis
One year of patient history reviewed
150 days
Initial breast cancer diagnosis

**FIGURE 3** ROC curve on binary classification XGB Model

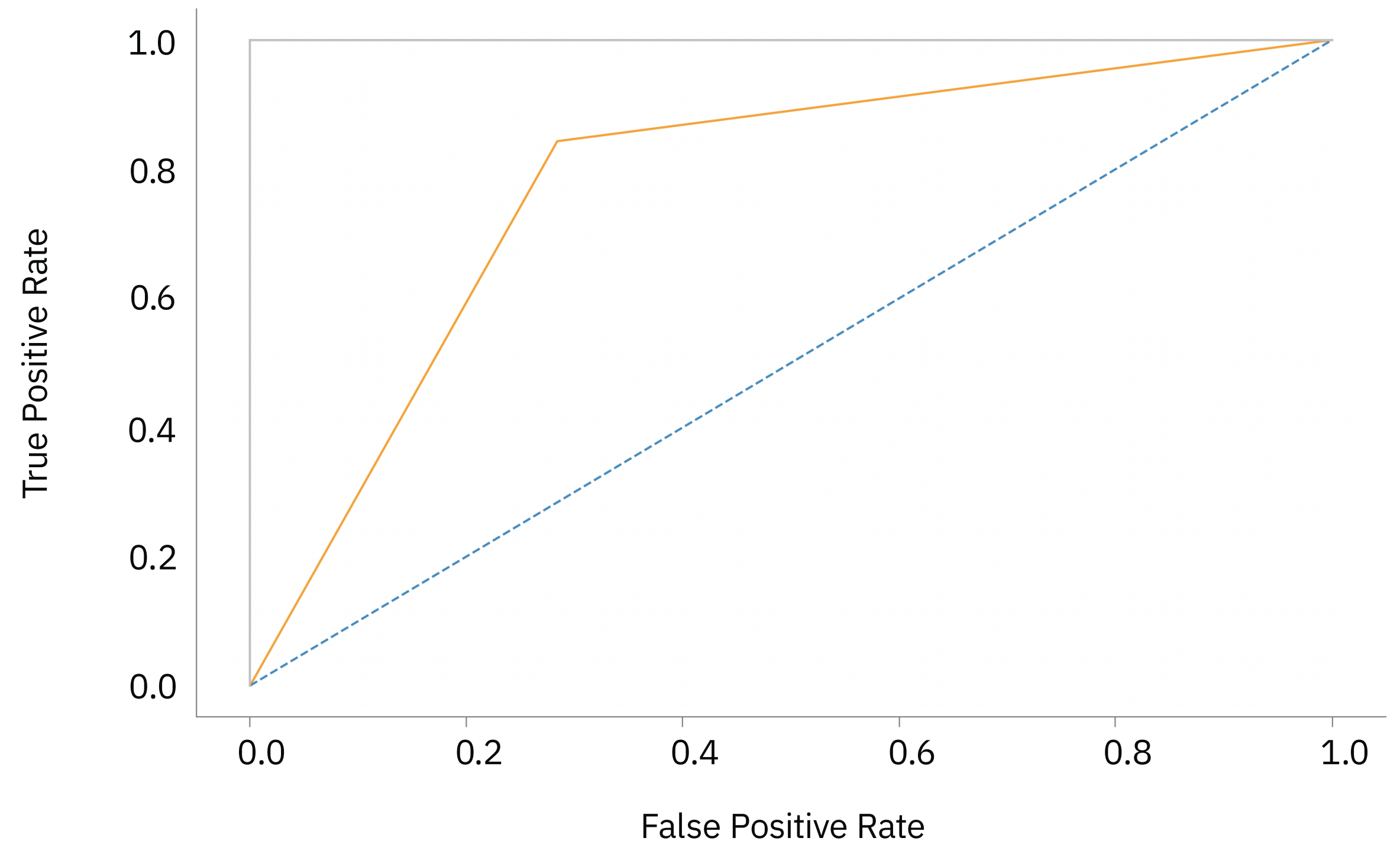
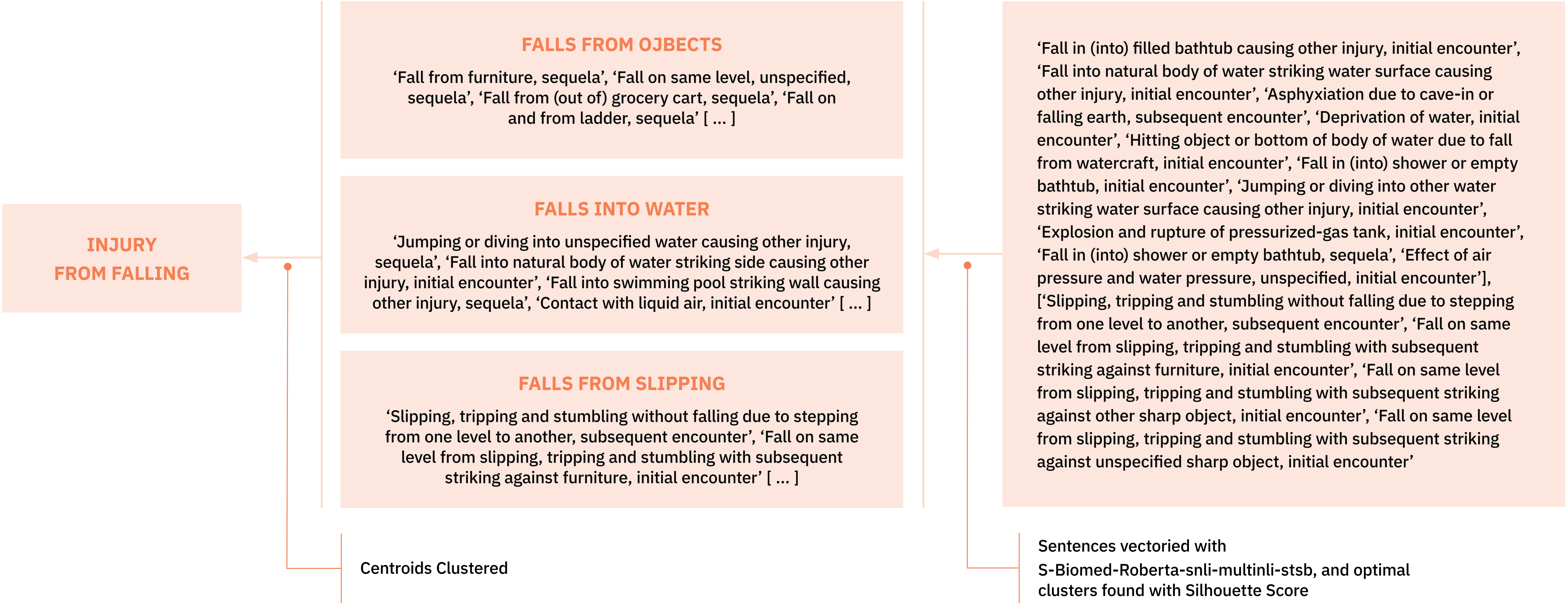


**FIGURE 2** Cluster Formation from free-text



INJURY FROM FALLING

FALLS FROM OJBECTS

'Fall from furniture, sequela', 'Fall on same level, unspecified, sequela', 'Fall from (out of) grocery cart, sequela', 'Fall on and from ladder, sequela' [ ... ]

FALLS INTO WATER

'Jumping or diving into unspecified water causing other injury, sequela', 'Fall into natural body of water striking side causing other injury, initial encounter', 'Fall into swimming pool striking wall causing other injury, sequela', 'Contact with liquid air, initial encounter' [ ... ]

FALLS FROM SLIPPING

'Slipping, tripping and stumbling without falling due to stepping from one level to another, subsequent encounter', 'Fall on same level from slipping, tripping and stumbling with subsequent striking against furniture, initial encounter' [ ... ]

'Fall in (into) filled bathtub causing other injury, initial encounter', 'Fall into natural body of water striking water surface causing other injury, initial encounter', 'Asphyxiation due to cave-in or falling earth, subsequent encounter', 'Deprivation of water, initial encounter', 'Hitting object or bottom of body of water due to fall from watercraft, initial encounter', 'Fall in (into) shower or empty bathtub, initial encounter', 'Jumping or diving into other water striking water surface causing other injury, initial encounter', 'Explosion and rupture of pressurized-gas tank, initial encounter', 'Fall in (into) shower or empty bathtub, sequela', 'Effect of air pressure and water pressure, unspecified, initial encounter'], ['Slipping, tripping and stumbling without falling due to stepping from one level to another, subsequent encounter', 'Fall on same level from slipping, tripping and stumbling with subsequent striking against furniture, initial encounter', 'Fall on same level from slipping, tripping and stumbling with subsequent striking against other sharp object, initial encounter', 'Fall on same level from slipping, tripping and stumbling with subsequent striking against unspecified sharp object, initial encounter'

Centroids Clustered

Sentences vectoried with S-Biomed-Roberta-snli-multinli-stsb, and optimal clusters found with Silhouette Score

## REFERENCES

1. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. Nat Rev Cancer. 2003;3(4):243-252. doi:10.1038/nrc1041
2. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinform. 2010;2010:1-5. Published 2010 Mar 1.
3. Pawar, Urja, Donna O'Shea, Susan Rea and Ruairi O'Reilly. "Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain." Irish Conference on Artificial Intelligence and Cognitive Science (2020).
4. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Published online August 27, 2019. doi:10.48550/arXiv.1908.10084