

# IA4Elderly : Exploitation des données du SNDS pour la représentation des parcours de soins des personnes âgées atteintes d'un cancer.

Marie Guyomard, Louis Tassy et Raquel Ureña

Journée Scientifique de l'Institut Laënnec  
15 Mai 2025

Inserm



IRD Institut de Recherche  
pour le Développement  
FRANCE



Aix-Marseille  
université  
Socialement engagée

SESSTIM, Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université,  
Marseille, France

<https://sesstim.univ-amu.fr/>

# Table of Contents

IA4Elderly

SNDS

Simulated data

pySNDS

Representation Learning

Future Works



## Cancer: A Disease of the Aged



## Cancer: A Disease of the Aged



## Cancer: A Disease of the Aged



### Objective

Development of a clinical decision support algorithm to personalize therapeutical pathways of these patients.

# Table of Contents

IA4Elderly

SNDS

Simulated data

pySNDS

Representation Learning

Future Works



## SNDS

### French inter-scheme consumption datamart (DCIR)

**Socio-demographic data**

Age, sex, commune of residence, insurance scheme, date of birth and death

**Medico-administrative data**

Long-term diseases, work-related accidents or diseases, disabilities

**Ambulatory care data**

Dates, medical or paramedical visits, claims information (drugs and medical devices, biology, imagery)

### French national hospital discharge database (PMSI)

**Administrative data**

Dates of admission and discharge, hospital unit

**Medical data**

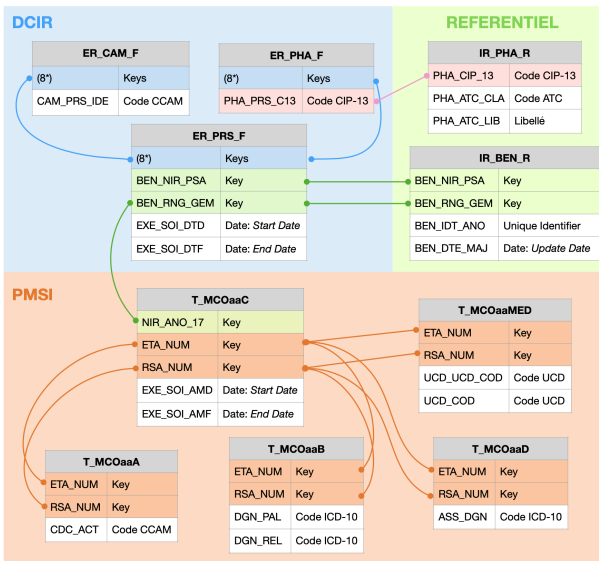
Main, related and associated diagnosis (ICD-10)

**Hospital care data**

Dates, high cost drugs and procedures, related costs

### French national causes-of-death register

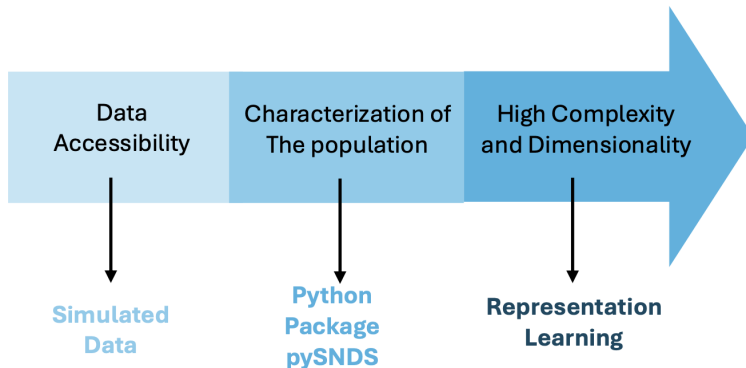




## Several Challenges



## Several Challenges



# Table of Contents

IA4Elderly

SNDS

Simulated data

pySNDS

Representation Learning


Future Works



## Breast Cancer Simulated Data

Collaboration with Thomas Guyet (INRIA, Lyon)

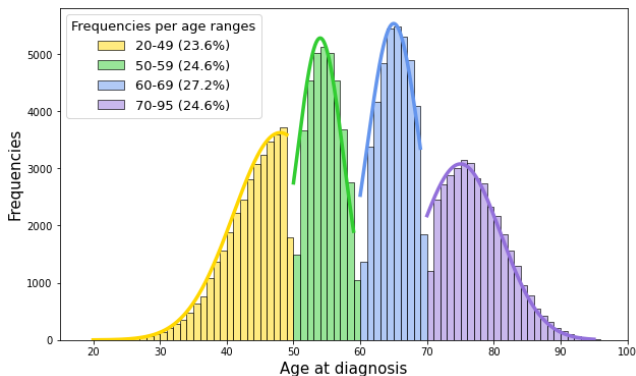
### Methodology

- SNDS synthetic data generator:  SNDSGenerator
- Compliant with privacy : statistical model generator
- Based on open source litterature

*“The French Early Breast Cancer Cohort (FRESH): A Resource for Breast Cancer Research and Evaluations of Oncology Practices Based on the French National Healthcare System Database (SNDS), Dumas et al. (2022)”.*

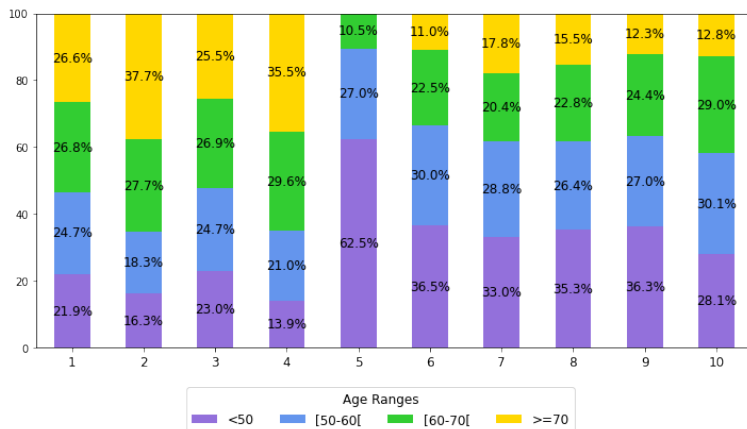
## Data

- 150,000 female breast cancer patients



## Data

- 150,000 female breast cancer patients
- 10 most common therapeutic pathways



## Contributions

- Open-source dataset compliant with reality
  1. Distributions of the population
  2. Complex architecture of SNDS

*“A Realistic Open-Source SNDS-Structured Simulated Database for Privacy-Compliant Analysis of Female” (en soumission)*

## Contributions

- Open-source dataset compliant with reality
  1. Distributions of the population
  2. Complex architecture of SNDS
- Uses
  1. Educational tool
  2. Testing algorithms

*“A Realistic Open-Source SNDS-Structured Simulated Database for Privacy-Compliant Analysis of Female” (en soumission)*

## Contributions

- Open-source dataset compliant with reality
  1. Distributions of the population
  2. Complex architecture of SNDS
- Uses
  1. Educational tool
  2. Testing algorithms
- Codes provided to facilitate reproducibility and simulation of other data sets

*“A Realistic Open-Source SNDS-Structured Simulated Database for Privacy-Compliant Analysis of Female” (en soumission)*

# Table of Contents

IA4Elderly

SNDS

Simulated data

pySNDS

Representation Learning

Future Works



## pySNDS

Comprehensive understanding of the study population:

1. **Automated Population Identification** – It navigates the SNDS structure to identify specific populations and their characteristics
2. **Detection of Targeted Medical Events** – It identifies the occurrence of specific medical events within the SNDS for a given population and determines their occurrence dates, including the first appearance.

+ Tools to characterize the breast cancer population

*“pySNDS” (en soumission)*

# Table of Contents

IA4Elderly

SNDS

Simulated data

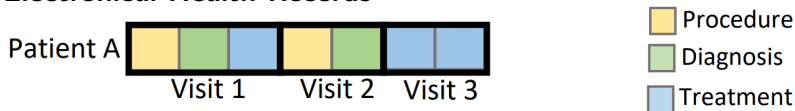
pySNDS

Representation Learning

Future Works

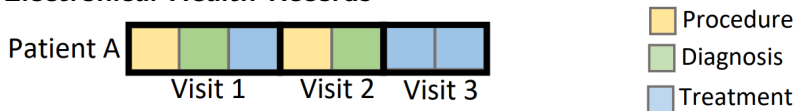


## Electronical Health Records



*Figure: An example of EHR.*

## Electronical Health Records

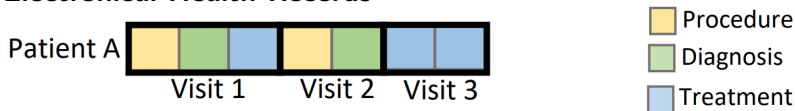


*Figure: An example of EHR.*

### Challenges

- Temporal Dynamic: temporal dependencies;

## Electronical Health Records

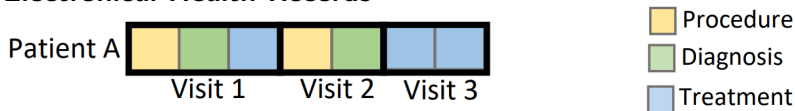


*Figure: An example of EHR.*

### Challenges

- Temporal Dynamic: temporal dependencies;
- Multi-modality: a single visit contains multiple medical codes;

## Electronical Health Records



*Figure: An example of EHR.*

### Challenges

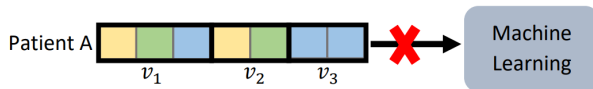
- Temporal Dynamic: temporal dependencies;
- Multi-modality: a single visit contains multiple medical codes;
- Unstructured data;
- Highly dimensional: thousands of unique medical codes.

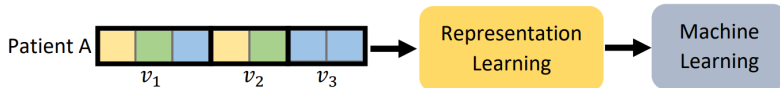
## VICAN Database

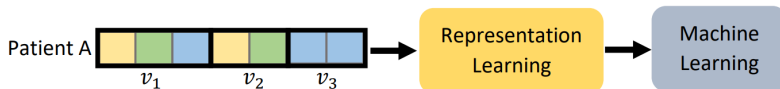
- VICAN study [\[Bouhnik, 2015\]](#)
- Female patients with Breast Cancer
- 1,304,361 events, 6111 patients (213 visits in average)
- 3407 unique medical codes

## VICAN Database

- VICAN study [Bouhnik, 2015]
- Female patients with Breast Cancer
- 1,304,361 events, 6111 patients (213 visits in average)
- 3407 unique medical codes







### Definition (Representation Learning Task)

*Patient Representation Learning task involves extracting meaningful information from the dense mathematical representation of a patient within an embedding space or latent space.*

$$f_C : \mathbb{R}^L \rightarrow \mathbb{R}^m. \quad (1)$$

[Si, 2021], [Shickel, 2017]

### 3 main Deep Learning strategies

- Natural Language Processing [Y. Choi, 2016], [E. Choi, 2016a-d]
- Autoencoders [Miotto, 2016], [Landi, 2020], [Baytas, 2017]
- Transformers [Li, 2020], [Rasmy, 2021]

### 3 main Deep Learning strategies

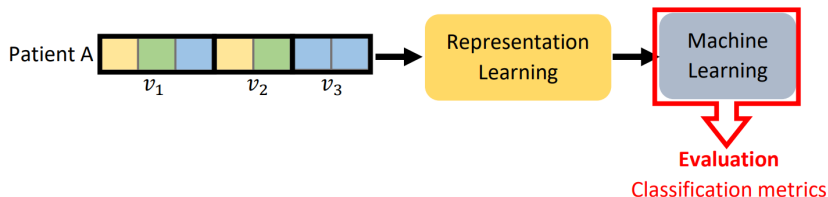
- Natural Language Processing [Y. Choi, 2016], [E. Choi, 2016a-d]
- Autoencoders [Miotto, 2016], [Landi, 2020], [Baytas, 2017]
- Transformers [Li, 2020], [Rasmy, 2021]

### 3 types of representation

- Medical Codes [Y. Choi, 2016], [E. Choi, 2016a,b,d], [Li, 2020], [Rasmy, 2021]
- Visit [E. Choi, 2016b-d], [Rasmy, 2021]
- Patient [E. Choi, 2016a], [Miotto, 2016], [Landi, 2020], [Baytas, 2017]

## 🌟 Evaluation Method

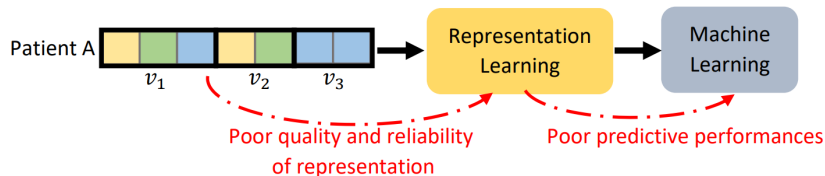
**Quality** and **Reliability** are assessed through the performance resulting from the prediction task fitted on the embedding space by the mean of **classification metrics mostly**.



[Choi, 2016c], [Choi, 2016d], [Miotto, 2016]

## 🌟 Evaluation Method

**Quality** and **Reliability** are assessed through the performance resulting from the prediction task fitted on the embedding space by the mean of **classification metrics mostly**.



[Choi, 2016c], [Choi, 2016d], [Miotto, 2016]

- Validation of state of the art Representation Learning tools
  - ▶ Quantify their accuracies
  - ▶ Analyse their reliability

- Validation of state of the art Representation Learning tools
  - ▶ Quantify their accuracies
  - ▶ Analyse their reliability

## 1. Fit general latent spaces (unsupervised tools)

Strategy / Types	NLP	Autoencoder	Transformer
Medical code	<b>Skip-Gram</b> [Y.Choi, 2016], [E.choi, 2016a] [E.Choi, 2016d]	-	<i>Out of scope</i>
Visit	<b>Med2Vec</b> [E.Choi, 2016b], [E.choi, 2016c]	-	<i>Supervised Tools</i>
Patient	-	<b>Deep Patient</b> [Miotto, 2016]	

- Validation of state of the art Representation Learning tools
  - ▶ Quantify their accuracies
  - ▶ Analyse their reliability

## 1. Fit general latent spaces (unsupervised tools)

Strategy / Types	NLP	Autoencoder	Transformer
Medical code	<b>Skip-Gram</b> [Y.Choi, 2016], [E.choi, 2016a] [E.Choi, 2016d]	-	<i>Out of scope</i>
Visit	<b>Med2Vec</b> [E.Choi, 2016b], [E.choi, 2016c]	-	<i>Supervised Tools</i>
Patient	-	<b>Deep Patient</b> [Miotto, 2016]	

## 2. Clustering task

## Résultats

- Assessing the quality of RL tools only on empirical metrics is not sufficient;
- Unsupervised study: methods with **higher value of silhouette score does not necessarily align with patients' clinical reality**;
- Need of evaluation metrics assessing both the performance and the consistency of patient RL tools.

*"Encoding breast cancer patients' medical pathways from reimbursement data using representation learning: a benchmark for clustering tasks", IEEE 37th International Symposium on Computer-Based Medical Systems, 2024*

*"Representation Learning pour la codification des parcours thérapeutiques de patientes atteintes de cancer du sein à partir de données de remboursement : un benchmark pour des tâches de clustering", Atelier IACD - EGC, 2025*

# Table of Contents

IA4Elderly

SNDS

Simulated data

pySNDS

Representation Learning

Future Works



## Future Works

- Mini-RHU *Granted*
- ERC *Interview step*

## Future Works

- Mini-RHU *Granted*
- ERC *Interview step*

### 1. Theoretical

- ▶ Develop an empirical metric to evaluate both performance and reliability of RL tools
- ▶ Develop an intrinsically interpretable RL tool

## Future Works

- Mini-RHU *Granted*
- ERC *Interview step*

### 1. Theoretical

- ▶ Develop an empirical metric to evaluate both performance and reliability of RL tools
- ▶ Develop an intrinsically interpretable RL tool

### 2. Pratical

- ▶ HDH access
- ▶ Application of all developed tools on elderly population with cancer

# References

- Y. Si, and al. **Deep representation learning of patient data from electronic health records (ehr): A systematic review.** *Journal of biomedical informatics*, vol. 115, p. 103671. (2021)
- B. Shickel and al. **Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis.** *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604. (2017)
- Y. Choi, and al. **Learning low-dimensional representations of medical concepts.** *AMIA Summits on Translational Science Proceedings*, p. 41, (2016)
- E. Choi, and al. **Medical concept representation learning from electronic health records and its application on heart failure prediction.** *arXiv preprint arXiv:1602.03686*, (2016a)
- E. Choi, and al. **Multi-layer representation learning for medical concepts,** *in proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1495–1504. (2016b)
- E. Choi, and al. **Doctor ai: Predicting clinical events via recurrent neural networks.** *Machine learning for healthcare conference.* PMLR, pp. 301–318. (2016c)
- E. Choi, and al. **Retain: An interpretable predictive model for healthcare using reverse time attention mechanism.** *Advances in neural information processing systems*, vol. 29. (2016d)
- R. Miotto, and al. **Deep patient: an unsupervised representation to predict the future of patients from the electronic health records.** *Scientific reports*, vol. 6, no. 1, pp. 1–10. (2016)
- I. Landi et al. **Deep representation learning of electronic health records to unlock patient stratification at scale.** *NPJ digital medicine*, vol. 3, no. 1, p. 96. (2020)
- I. M. Baytas et al. **Patient subtyping via time-aware lstm networks.** *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74. (2017)
- Y. Li et al. **Behrt: transformer for electronic health records.** *Scientific reports*, vol. 10, no. 1, p. 7155. (2020)
- L. Rasmy et al. **Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction.** *NPJ digital medicine*, vol. 4, no. 1, p. 86. (2021)

# Table of Contents

## Appendix

Skip-Gram

Med2Vec

Deep Patient

Evaluation of Patient Representations

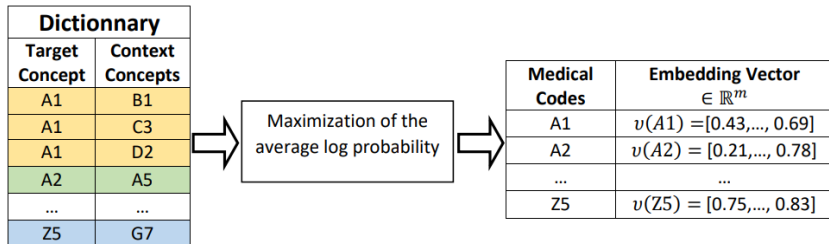
Experimental Settings

Results - Performance

Results - Clinical Reliability

## Skip-Gram

- Natural Language Processing
- Medical Code Representation [Y.Choi, 2016]

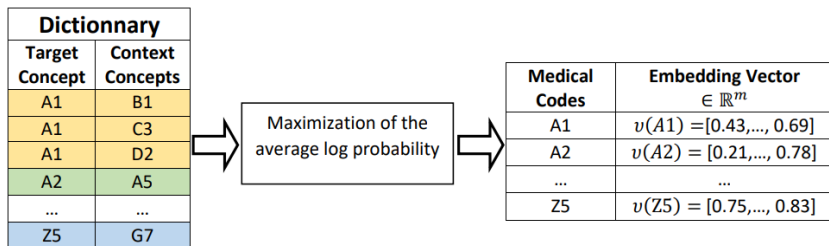


*Schema of Skip-Gram.*

*\*Theoretical information are provided in Appendix.*

## Skip-Gram

- Natural Language Processing
- Medical Code Representation [Y.Choi, 2016]

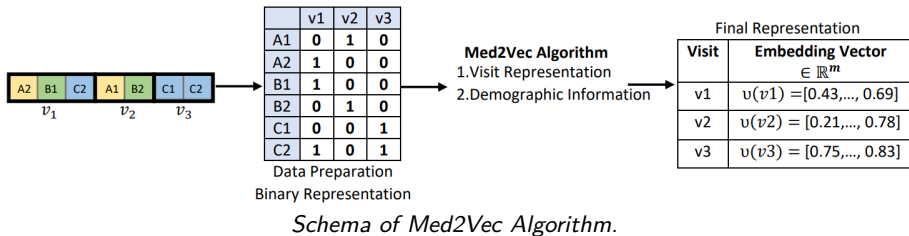


*Schema of Skip-Gram.*

- Patient Representation: sum all the medical codes' embedded vectors appearing for a patient [E.Choi, 2016a].

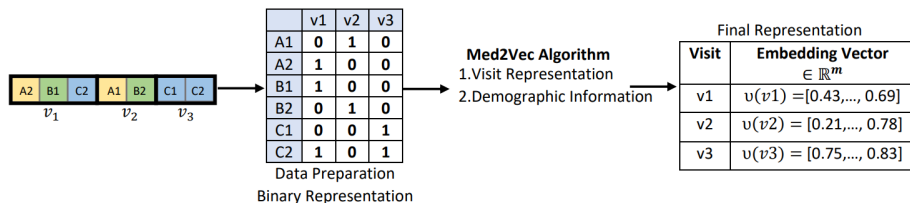
*\*Theoretical information are provided in Appendix.*

- Multi-Layer Perceptron x Natural Language Processing
- Visit Representation [E.Choi, 2016b]



*\*Theoretical information are provided in Appendix.*

- Multi-Layer Perceptron x Natural Language Processing
- Visit Representation [E.Choi, 2016b]

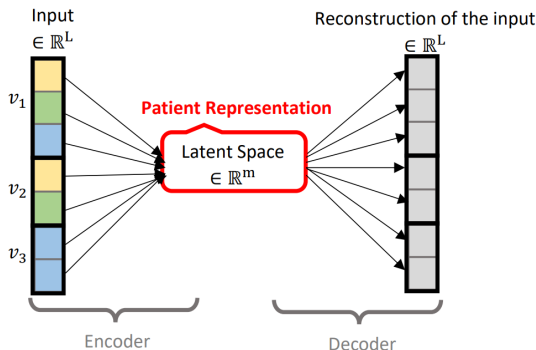


*Schema of Med2Vec Algorithm.*

- Patient Representation: sum all the visit representations.

*\*Theoretical information are provided in Appendix.*

- Denoising Stacked Autoencoder
- Patient Representation [Miotto, 2016b]



*Schema of an Autoencoder.*

*\*Theoretical information are provided in Appendix.*

# Evaluation of Patient Representations

---

## Clustering

- Clustering Methods
  1. K-means
  2. Gaussian Mixture Model
- Performance:
  1. Metric: silhouette score and Davies-Bouldin index
  2. Visualization: PCA and t-SNE
- Reliability: Chi-squared test on the clusters

### Data

- VICAN study [[Bouhnik, 2015](#)]
- Female patients with Breast Cancer
- 1,304,361 events, 6111 patients (213 visits in average)
- 3407 unique medical codes

### Data

- VICAN study [[Bouhnik, 2015](#)]
- Female patients with Breast Cancer
- 1,304,361 events, 6111 patients (**213 visits in average**)
- **3407 unique medical codes**

**Need of Representation Learning Tools !**

## Learning

1. Representation Learning
  - ▶ Gridsearch of the hyperparameters
  - ▶ Training of the hyperparameters

## Learning

### 1. Representation Learning

- ▶ Gridsearch of the hyperparameters
- ▶ Training of the hyperparameters

### 2. Clustering Task

- ▶ Gridsearch of the optimal number of clusters
  - ▶ 10-folds CV
  - ▶ Maximization of the silhouette score on validation sample
- ▶ Training of the clusters
  - ▶ 10-folds CV

Results - Performance

---

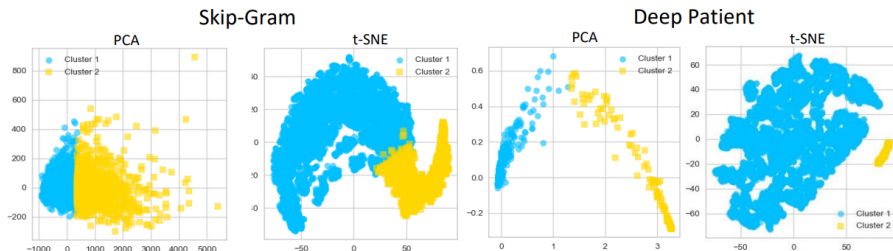
	Training Sample		Validation Sample	
	Silhouette Score ↑	Davies- Bouldin ind. ↓	Silhouette Score ↑	Davies- Bouldin ind. ↓
<b>Skip-Gram</b>	0.6 (0.005)	0.34 (0.005)	0.6 (0.006)	0.344 (0.02)
<b>Med2Vec</b>	0.55 (0.004)	0.3 (0)	0.54 (0.006)	0.31 (0.005)
<b>Deep Patient</b>	0.98 (0)	0.13 (0.005)	0.98 (0.002)	0.13 (0.007)

*Average metrics (std) over the 10-folds for the k-means clustering task.*

## Results - Performance

	Training Sample		Validation Sample	
	Silhouette Score $\uparrow$	Davies-Bouldin ind. $\downarrow$	Silhouette Score $\uparrow$	Davies-Bouldin ind. $\downarrow$
<b>Skip-Gram</b>	0.6 (0.005)	0.34 (0.005)	0.6 (0.006)	0.344 (0.02)
<b>Med2Vec</b>	0.55 (0.004)	0.3 (0)	0.54 (0.006)	0.31 (0.005)
<b>Deep Patient</b>	0.98 (0)	0.13 (0.005)	0.98 (0.002)	0.13 (0.007)

*Average metrics (std) over the 10-folds for the k-means clustering task.*



## Results - Clinical Reliability

	Skip-Gram	Med2Vec	Deep Patient
Partial Mastectomy	<0.05 (0)	0.07 (0.04)	<0.05 (0.02)
Mastectomy	<0.05 (0)	0.37 (0.13)	<0.05 (0.01)
Axillary Surgery	<0.05 (0)	<0.05 (0)	0.7 (0.23)
Chemotherapy Y/N	<0.05 (0)	<0.05 (0)	0.5 (0.27)
Chemotherapy Setting	<0.05 (0)	<0.05 (0)	<0.05 (0.03)
Chemotherapy Regimen	<0.05 (0)	<0.05 (0)	0.1 (0.22)
Targeted Therapy Y/N	0.87 (0.12)	<0.05 (0)	0.6 (0.31)
Targeted Therapy Setting	0.7 (0.01)	<0.05 (0)	0.7 (0.2)
Targeted therapy Regimen	0.34 (0.12)	<0.05 (0)	0.6 (0.31)
Radiotherapy Y/N	<0.05 (0.03)	<0.05 (0)	0.4 (0.23)
Radiotherapy Setting	<0.05 (0.21)	<0.05 (0)	<0.05 (0)
Endocrine Therapy Y/N	<0.05 (0.01)	<0.05 (0)	0.2 (0.2)
Endocrine Therapy Setting	<0.05 (0.03)	<0.05 (0)	<0.05 (0)
Endocrine Therapy Regimen	<0.05 (0)	<0.05 (0)	<0.05 (0)
BC Sub Type	<0.05 (0)	<0.05 (0)	0.2 (0.12)
Nodal status	<0.05 (0.01)	<0.05 (0)	0.06 (0.07)
Metastatic	<0.05 (0)	<0.05 (0)	<0.05 (0)

*Average (std) of Chi-squared test p-values between the k-means clusters and the BC characteristics obtained on 5 random sub samples.*