Évaluation de la qualité et de la fiabilité de méthodes de *Representation Learning* appliquées aux données du SNDS pour le cancer du sein

M. Guyomard, AD. Bouhnik, L. Tassy, R. Urena

Séminaire Interne, SESSTIM, 7 Février 2025

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

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

Table of Contents

SESSTIM

SNDS

- Administrative DataBase
- 60 million people
- 3 main data sources :
  - Health insurance expenditure (*Système National d'Information InterRégime de l'Assurance Maladie*, CNAM)
  - Healthcare establishments (*Programme de Médicalisation des Systèmes d'Information*, ATIH)
  - Medical causes of death (CépiDc, INSERM)

SNDS
Système national des données de santé

Electronical Health Records

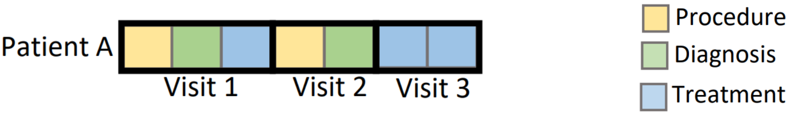


Figure: An example of EHR.

Electronical Health Records

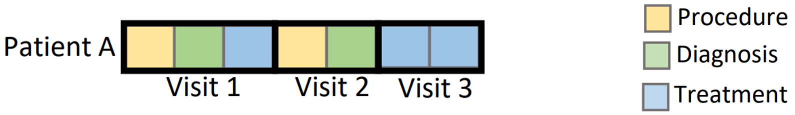


Figure: An example of EHR.

Challenges

- Temporal Dynamic: temporal dependencies;

Electronical Health Records

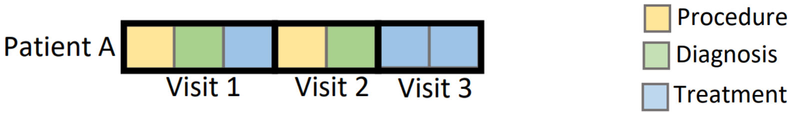


Figure: An example of EHR.

Challenges

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

Electronical Health Records

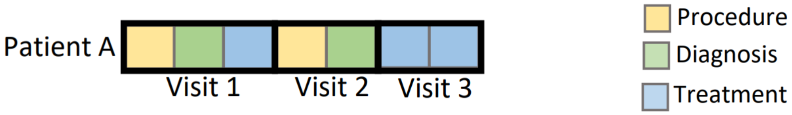
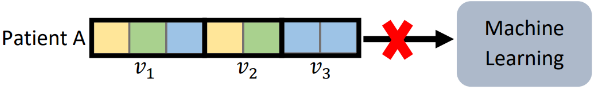


Figure: An example of EHR.

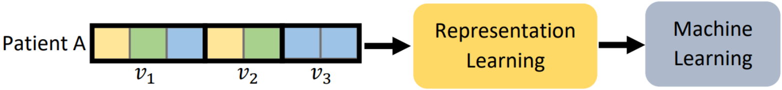
Challenges

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

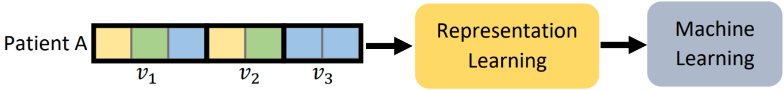
Representation Learning



Representation Learning



Representation Learning



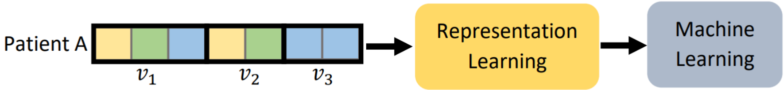
Definition (Representation Learning Task)

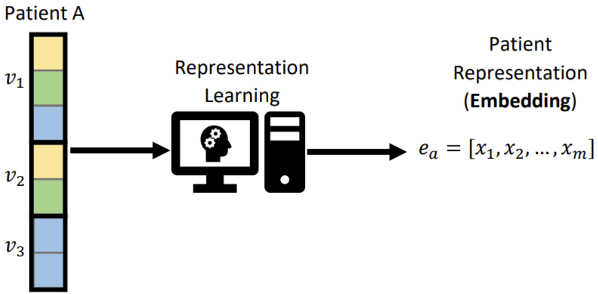
Patient Representation Learning task involves extracting meaningful information from the dense mathematical representation of a patient within an embedding space or latent space.

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

[Si, 2021], [Shickel, 2017]

Representation Learning





Representation Learning

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]

Representation Learning

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]

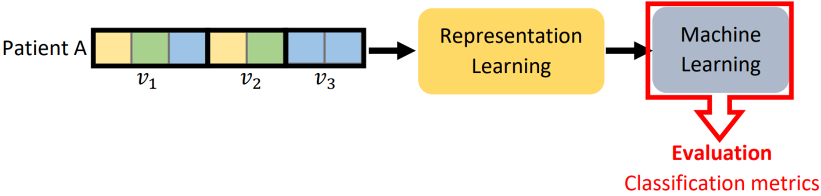
3 types of representation

- 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]

Representation Learning

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**.

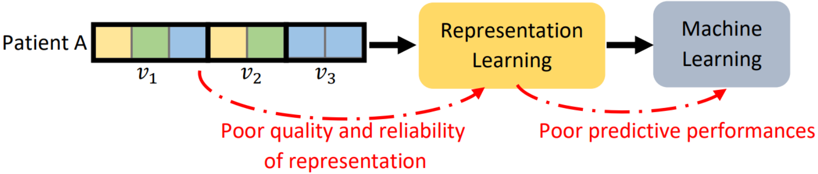


[Choi, 2016c], [Choi, 2016d], [Miotto, 2016]

Representation Learning

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]

Objectives

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

Objectives

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

1. Fit general latent spaces (unsupervised tools)
   We aim at learning **general embedding**, not specific to a prediction task.

Objectives

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

1. Fit general latent spaces (unsupervised tools)
   We aim at learning **general embedding**, not specific to a prediction task.

2. Clustering task

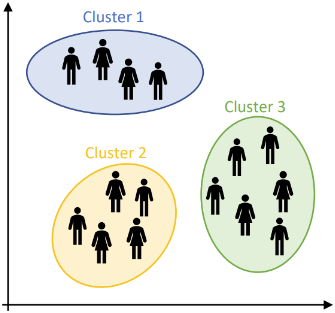
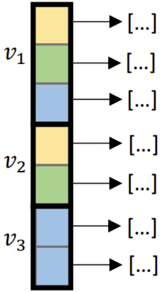


Table of Contents

Representation Learning

Skip-Gram

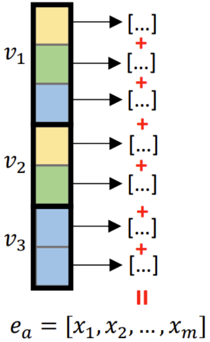
Medical Codes Representations [Y.Choi, 2016]



Representation Learning

Skip-Gram

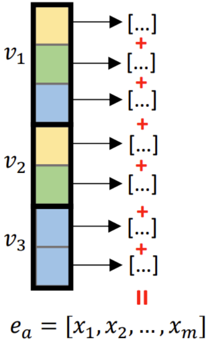
Medical Codes Representations [Y.Choi, 2016]



Representation Learning

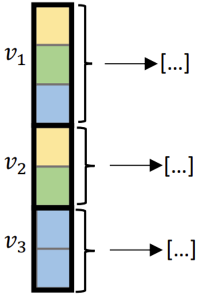
Skip-Gram

Medical Codes Representations [Y.Choi, 2016]



Med2Vec

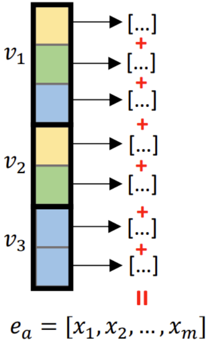
Visits Representations [E.Choi, 2016b]



Representation Learning

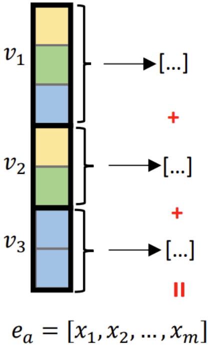
Skip-Gram

Medical Codes Representations [Y.Choi, 2016]



Med2Vec

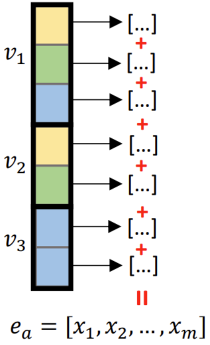
Visits Representations [E.Choi, 2016b]



Representation Learning

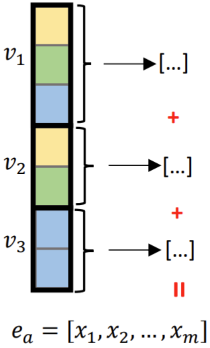
Skip-Gram

Medical Codes Representations
[Y.Choi, 2016]



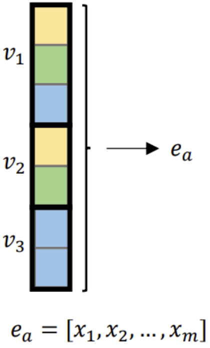
Med2Vec

Visits Representations
[E.Choi, 2016b]



Deep Patient

Patients Representations
[Miotto, 2016b]



Evaluation of Patient Representations

Clustering : K-means

- Performance:
  1. Empirical Metric:
     - Silhouette score (↑) : cohesion of a cluster and its separation from other clusters
     - Davies-Bouldin index (↓) : cluster separability and compactness

Evaluation of Patient Representations

Clustering : K-means

- Performance:
  1. Empirical Metric:
     - Silhouette score (↑) : cohesion of a cluster and its separation from other clusters
     - Davies-Bouldin index (↓) : cluster separability and compactness
  2. Visualization:
     - PCA
     - t-SNE

Evaluation of Patient Representations

Clustering : K-means

- Performance:
  1. Empirical Metric:
     - Silhouette score (↑) : cohesion of a cluster and its separation from other clusters
     - Davies-Bouldin index (↓) : cluster separability and compactness
  2. Visualization:
     - PCA
     - t-SNE
- Reliability: Chi-squared test on the clusters ($p < 0.05$)

Data

Database

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

Data

Database

- 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 !

Data

Characterization of the population

French Early Breast Cancer Cohort (FRESH) methodology [Dumas, 2022]

1. Cancer
   - Sub-type (Luminal, TNBC, HER)
   - Nodal Status (+/−)

2. Treatments

	Setting	Regimen
Surgery		Partial Mastectomy / Mastectomy
Chemotherapy	Neo / adj	Anthracyclines, Docetaxel, Paclitaxel, etc
Targeted Therapy	Neo / adj	Trastuzumab only, Pertuzumab +/ trastuzumab
Radiotherapy	Neo / adj	-
Endoctrine Therapy	Neo / Adj	Aromatase Inhibitor, Tamoxifen, agonists and combinations

Table of Contents

Performance

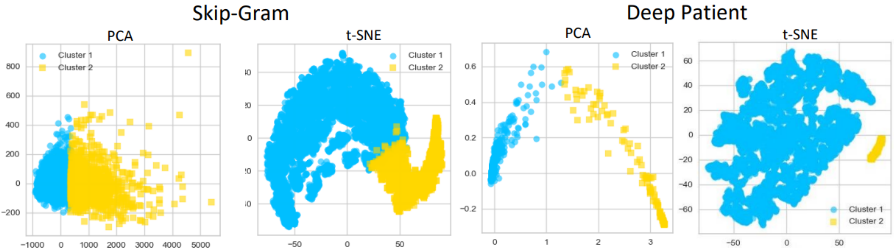
	Training Sample		Validation Sample	
	Silhouette Score ↑	Davies-Bouldin ind. ↓	Silhouette Score ↑	Davies-Bouldin ind. ↓
Skip-Gram	0.6 (0.005)	0.34 (0.005)	0.6 (0.006)	0.344 (0.02)
Med2Vec	0.55 (0.004)	0.3 (0)	0.54 (0.006)	0.31 (0.005)
Deep Patient	0.98 (0)	0.13 (0.005)	0.98 (0.002)	0.13 (0.007)

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

Performance

	Training Sample		Validation Sample	
	Silhouette Score ↑	Davies-Bouldin ind. ↓	Silhouette Score ↑	Davies-Bouldin ind. ↓
Skip-Gram	0.6 (0.005)	0.34 (0.005)	0.6 (0.006)	0.344 (0.02)
Med2Vec	0.55 (0.004)	0.3 (0)	0.54 (0.006)	0.31 (0.005)
Deep Patient	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.



Visualization through PCA and t-SNE of the k-means clusters.

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.*

Table of Contents

SESTIM

Conclusion

- 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.

Conclusion

- 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.

Future works

1. Develop an empirical metric to evaluate both performance and reliability of RL tools;
2. Develop an intrinsically interpretable RL tool, based on graphs representation learning tools.

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)

AD Bouhnik, and al. **The labour market, psychosocial outcomes and health conditions in cancer survivors: protocol for a nationwide longitudinal survey 2 and 5 years after cancer diagnosis (the vican survey)**. *BMJ open*, vol. 5, no. 3. (2015)

References (2)

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

EHR

Notation

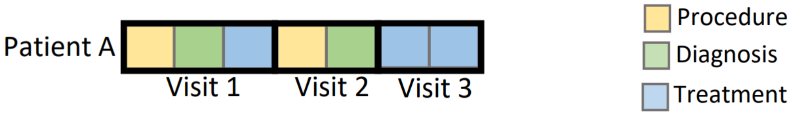


Figure: An example of EHR.

EHR

Notation

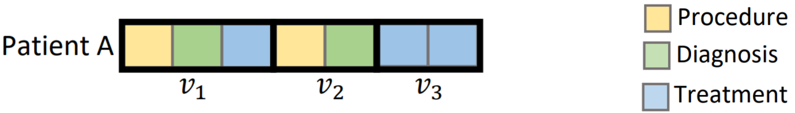


Figure: An example of EHR.

- $V = \{v_1, \ldots, v_n\}$;

$n = 3$

EHR

Notation

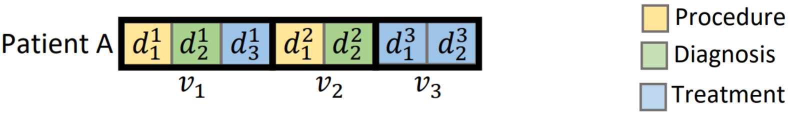


Figure: An example of EHR.

- $V = \{v_1, \ldots, v_n\}$;
- j-th visit: $v_j = \{d_1^j, d_2^j, \ldots, d_{k_j}^j\}$;

$n = 3$

$k_1 = 3,\ k_2 = k_3 = 2$

EHR

Notation

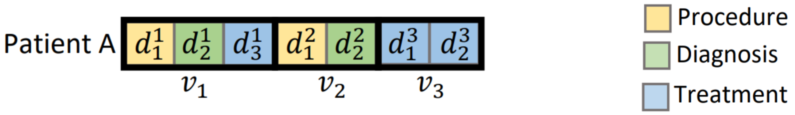


Figure: An example of EHR.

- $V = \{v_1, \dots, v_n\}$;
- j-th visit: $v_j = \{d_1^j, d_2^j, \dots, d_{k_j}^j\}$;
- $v_j \subseteq \mathcal{C}, \quad \mathcal{C} = \{c_1, \dots, c_{|\mathcal{C}|}\}$;

$n = 3$
$k_1 = 3,\ k_2 = k_3 = 2$

EHR

Notation

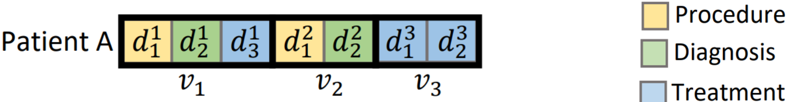


Figure: An example of EHR.

- $V = \{v_1, \ldots, v_n\}$;
- j-th visit: $v_j = \{d_1^j, d_2^j, \ldots, d_{k_j}^j\}$;
- $v_j \subseteq \mathcal{C}, \quad \mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$;
- $L = \sum_{t=1}^{n} |v_t|$.

$n = 3$

$k_1 = 3,\ k_2 = k_3 = 2$

$L = 7$

Skip-Gram Algorithm

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

Dictionnary	
Target Concept	**Context Concepts**
A1	B1
A1	C3
A1	D2
A2	A5
...	...
Z5	G7

⇒ Maximization of the average log probability ⇒

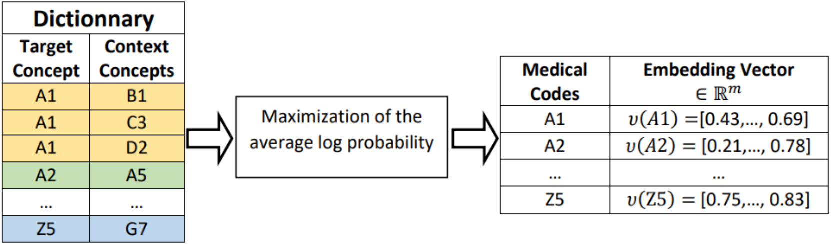
Medical Codes	**Embedding Vector $\in \mathbb{R}^m$**
A1	$\upsilon(A1)$ =[0.43,..., 0.69]
A2	$\upsilon(A2)$ = [0.21,..., 0.78]
...	...
Z5	$\upsilon(Z5)$ = [0.75,..., 0.83]

Schema of Skip-Gram.

**Theoretical information are provided in Appendix.*

Skip-Gram Algorithm

- 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.*

Skip-Gram Algorithm

[Y.Choi, 2016]

- Medical representation: $\nu(c)$

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{-w\leq j\leq w, j\neq 0}\log p(c_{t+j}|c_t), \tag{1}$$

with w representing the size of the context window and

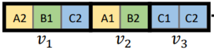
$$p(c_{t+j}|c_t) = \frac{\exp(\nu(c_{t+j})^T\nu(c_t))}{\sum_{c=1}^{|\mathcal{C}|}\exp(\nu(c)^T\nu(c_t))}. \tag{2}$$

- Patient representation [E.Choi, 2016a]

$$e^{SG} = \sum_{t=1}^{n}\sum_{j=1}^{k_t}\nu(d_j^t) \in \mathbb{R}^m. \tag{3}$$

Med2Vec Algorithm

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



	v1	v2	v3
A1	**0**	**1**	**0**
A2	**1**	**0**	**0**
B1	**1**	**0**	**0**
B2	**0**	**1**	**0**
C1	**0**	**0**	**1**
C2	**1**	**0**	**1**

Data Preparation
Binary Representation

Med2Vec Algorithm
1.Visit Representation
2.Demographic Information

Final Representation

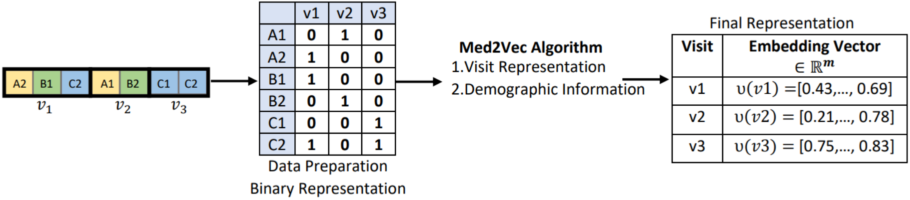
Visit	**Embedding Vector** $\in \mathbb{R}^m$
v1	$\upsilon(v1)$ =[0.43,..., 0.69]
v2	$\upsilon(v2)$ = [0.21,..., 0.78]
v3	$\upsilon(v3)$ = [0.75,..., 0.83]

Schema of Med2Vec Algorithm.

**Theoretical information are provided in Appendix.*

Med2Vec Algorithm

- 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.*

Med2Vec Algorithm

[E.Choi, 2016b]

- Visit representation
  1. Intermediate visit representation given a visit $\bar{v}_t \in \{0,1\}^{|\mathcal{C}|}$

$$u_t = \phi(W_c \bar{v}_t + b_c) \in \mathbb{R}^{m'}, \quad (4)$$

with $\phi(x) = \max\{0, x\}$, $W_c \in \mathbb{R}^{m' \times |\mathcal{C}|}$ and $b_c \in \mathbb{R}^{m'}$.

  2. Concatenation with demographic information $d_t \in \mathbb{R}^d$

$$\nu_t = \phi(W_v[u_t, d_t] + b_v) \in \mathbb{R}^m, \quad (5)$$

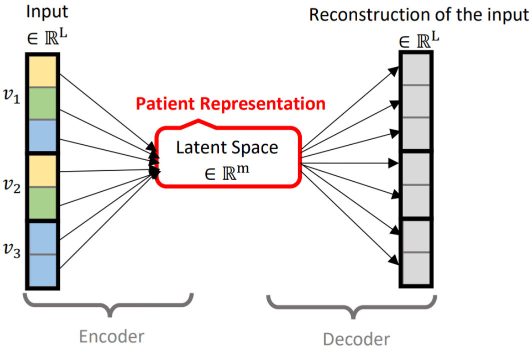
with $W_v \in \mathbb{R}^{m \times (m' \times d)}$ and $b_v \in \mathbb{R}^m$.

- Patient representation

$$e^{Med} = \sum_{t=1}^{n} \nu_t \in \mathbb{R}^m. \quad (6)$$

Deep Patient Algorithm

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



Schema of an Autoencoder.

**Theoretical information are provided in Appendix.*

Deep Patient Algorithm

[Miotto, 2016b]

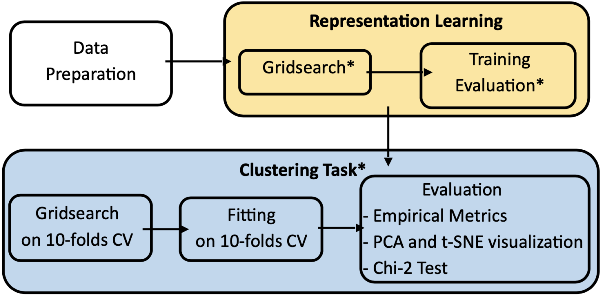
- Patient representation
- Denoising Stacked Autoencoder

1. Masking Noise algorithm on the input $\tilde{V} \in \mathbb{R}^L$.
2. Encoder
$$y = f_\theta(\tilde{V}) = s(W\tilde{V} + b), \quad (7)$$
with $s(\cdot)$ a non-linear transformation, $W \in \mathbb{R}^{m \times L}$ and $b \in \mathbb{R}^m$.
3. Decoder
$$z = g_{\theta'}(y) = s(W'y + b'), \quad (8)$$
with $W' \in \mathbb{R}^{L \times m}$ and $b' \in \mathbb{R}^m$.

Data

- VICAN study [Bouhnik, 2015], a national survey on French cancer survivors
- Inclusion Criteria of patients : (i) Female, (ii) diagnosed with Breast Cancer, (iii) who have reached the age of majority and (iv) have undergone surgery
- Exclusion criteria of patients : affected by another form of cancer
- **1,304,361 events**, **6111 patients** with an average of 213 visits (min 4, max 1111)
- 3,407 medical codes at first
  - 2447 diagnosis (ICD-10 Classification)
  - 1977 procedures (Anatomical Therapeutic Chemical, ATC)
  - 1043 medications (Classification Commune des Actes Médicaux, CCAM)
- Grouping of the medical codes based on their hierarchical structure [Y.Choi, 2016], [E.Choi, 2016a]
  - 2 digits
  - It remains **3,407 unique medical codes**

Experimental Settings



Experimental settings. **The complementary tools provided on Github.*

Experimental Settings

Learning

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

Experimental Settings

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

Experimental Settings

	Epoch	Learning Rate	Tested Parameters
Skip-Gram	40	1e-3	Window Size: $\{\mathbf{5}, 10\}$ # False neighbors: $\{\mathbf{5}, 10\}$ Embedding Dim: $\{10, 20, \mathbf{50}, 100\}$
Med2Vec	5	1e-6	Temporary Dim: $\{20, \mathbf{50}, 100\}$ Final Dim: $\{\mathbf{20}, 50, 100\}$ Window Size: $\{\mathbf{1}, 3, 5\}$
Deep Patient	100	1e-3	Embedding Dim: $\{10, \mathbf{20}, 50, 100\}$ # Layers: $\{1, \mathbf{3}, 5\}$ Corruption Rate: $\{\mathbf{0.01}, 0.05, 0.01\}$

Settings for the Gridsearch step, optimal parameters are in bold.

Performance results

	Training Sample		Validation Sample	
	Silhouette Score	Davies-Bouldin ind.	Silhouette Score	Davies-Bouldin ind.
K-means				
SG	0.6 (0.005)	0.34 (0.005)	0.6 (0.006)	0.344 (0.02)
M2V	0.55 (0.004)	0.3 (0)	0.54 (0.006)	0.31 (0.005)
DP	0.98 (0)	0.13 (0.005)	0.98 (0 002)	0.13 (0.007)
Gaussian Mixture Model				
SG	0.37 (0.01)	0.52 (0.008)	0.35 (0.01)	0.52 (0.01)
M2V	0.06 (0.06)	1.1 (0.4)	0.3 (0.09)	0.8 (0.2)
DP	0.9 (0)	0.62 (0.01)	0.9 (0.005)	0.6 (0.09)

Average (standard deviation) results obtained on clustering over the 10−*folds CV, for Skip-Gram (SG), Med2Vec (M2V) and Deep Patient (DP) algorithms.*