Modélisation et analyse des parcours de soins

T. Guyet
thomas.guyet@inria.fr

Inria – AIstroSight – (Chaire AIRACLES)

EINS – 2024

Outline

1. Parcours de soins
   - Parcours de soins en santé publique
   - Analyser les parcours de soins
   - Représentation des trajectoires de soins (cas d'étude AP-HP)
   - A web semantic approach
2. Décrire les trajectoires de soins
   - Problématique générale
   - Catégorisation d'exemples
   - Métriques et clustering de séquences
   - Exemple OPTISOINS
3. Abstraction de trajectoires
4. Conclusion

Parcours de soins en santé publique

Parcours de soins

⇒ objet de réflexion et d'analyse pour les différents acteurs impliqués dans l'organisation et l'adaptation du système de santé

Notion liée à l'organisation des soins

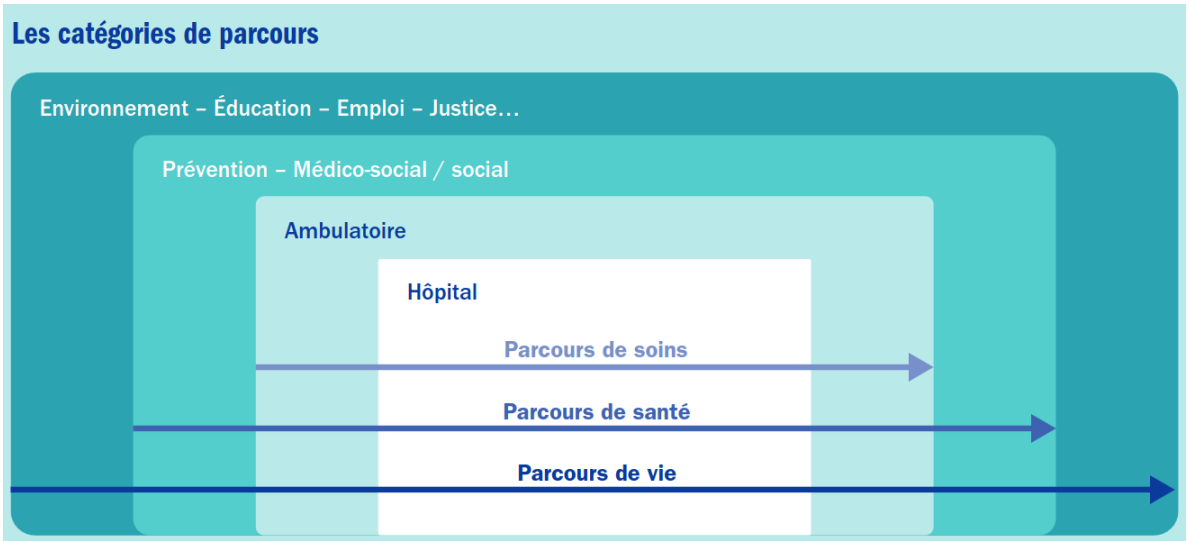
- en France: introduction du "parcours de soins coordonnés" confié au médecin traitant par la loi de 2004
- notion reprise par la Haute Autorité de santé (HAS) pour la prise en charge des maladies chroniques

Concrétisé par des données observationnelles de soins

- acquisition et historisation en routine de données de soins
  → données longitudinales

Parcours de soins ... et autres variantes

- "parcours de santé": proposé par le HCAAM[1] pour élargir le périmètre pris en compte, de façon à assurer la globalité à la fois soignante et sociale
- "parcours de vie": intégrer des évènements de vie (hors santé) dans la prise en charge : famille et entourage, scolarisation, prévention de la désinsertion professionnelle, réinsertion, logement ...



Source: "Parcours de santé : enjeux et perspectives", adsp n° 88

Parcours de soins ... et autres variantes

- "parcours de santé": proposé par le HCAAM[1] pour élargir le périmètre pris en compte, de façon à assurer la globalité à la fois soignante et sociale
- "parcours de vie": intégrer des évènements de vie (hors santé) dans la prise en charge : famille et entourage, scolarisation, prévention de la désinsertion professionnelle, réinsertion, logement ...

Idées principales

- le soin doit être vu comme une **prise en charge globale** d'un patient
- le **système de soins** doit être **organisé** pour favoriser cette prise en charge globale

- "parcours organisationnel": suivi attendu pour la **qualité des soins**
- "trajectoire de soins" : ce qui est réellement observé
- "guidelines" / "bonnes pratiques" : issue de l'épidémiologie

Vers la "médecine de parcours"

- Mise en place de la "médecine de parcours" (Rapport Cordier, 2013)
- Développer une meilleure coordination des interventions professionnelles, fondées sur de bonnes pratiques
- Objectif : délivrer "*les bons soins par les bons professionnels dans les bonnes structures, au bon moment*"

Différents leviers

- la coordination des soins (médecin traitant)
- l'organisation des structures de soins (locale et nationale)
- la mise en place de parcours recommandés
- ...

Différentes vues du parcours (projet SafePaw)

- vue **Patient** (*parcours de vie*)
  - prismes:
    - de l'efficacité du traitement
    - de la facilité à suivre le traitement
  - importance des contraintes d'accès aux soignants ou aux équipements
- vue **Régulateur** (*parcours de soin*)
  - prisme médico-économique: soins du plus grand nombre pour un coût raisonnable pour la société
  - utilisation optimale des ressources de soins
  - besoin "guidelines" évalués medico-économiquement
- vue **Soignant**
  - prisme de la coordination

- perception du parcours différent selon les acteurs
- objectifs d'amélioration différents
- angles d'analyse potentiellement différents ou biaisés

Différentes vues du parcours: exemples d'objectifs

- vue **Patient**
  - améliorer l'accompagnement des soins (compréhension des parcours, coordination des soins,
  - améliorer l'accompagnement social (accès aux aides)
  - personnalisation des traitements et implication du patient
  - aide aux aidants
- vue **Régulateur**
  - égalité d'accès aux soins sur le territoire
  - amélioration médico-administrative: traiter avec un coût minimal pour la société (inclus la prévention)
  - visualiser l'ensemble des activités de manière unifiée
  - évaluer les coûts de prise en charge
- vue **Soignant**
  - donner un cadre à la prise en charge (*guidelines*)
  - faciliter et fluidifier les échanges entre soignants, entre patient et soignants
  - accompagnement thérapeutique/diagnostic : améliorer l'observance, réduire le non-recours au soin

Analyser les parcours

Analyser les parcours

- décrire
- comprendre
- améliorer

les prises en charge des patients / l'organisation des soins.

Multitude d'approches ...

- approches transversales (par pathologie) vs verticales
- centré sur les perceptions humaines et sociales
  - en science sociale
  - en science du management et des organisations
- centré sur les aspects médico-économiques
  - en médecine, épidémiologie
  - ... dont celles menées sur données médico-administratives

Analyser les trajectoires à partir de bases médico-administratives

Opportunité

Les bases de données hospitalières et médico-administratives offrent une vue sur les enchainements de soins par patient (ou par soignant)

Analyser les trajectoires = exploiter des données patients longitudinales pour

- décrire
- comprendre
- améliorer

Analyser les trajectoires à partir de base médico-administratives

Motivations de l'approche par les données

- Décrire les soins délivrés sous l'angle de la prise en charge
  - Déterminer quels sont les parcours de soins effectifs ?
  - Comprendre l'organisation effective des soins
  - Identifier des leviers d'amélioration des parcours ou des organisations
- Évaluer des parcours en vie réelle
  - Mener des études épidémiologiques sur les prises en charge
- Déterminer des parcours de soin "optimaux" (vers des "bonnes pratiques")
  - Passage de la T2A à la tarification de la prise en charge

Dans la suite, on s'intéresse à ...

$\rightarrow$ l'analyse de parcours au travers de l'analyse de données médico-administratives

$\Rightarrow$ l'apport des méthodes d'analyse de données dans ces questions

Outline

1. Parcours de soins
   - Parcours de soins en santé publique
   - Analyser les parcours de soins
   - Représentation des trajectoires de soins (cas d'étude AP-HP)
   - A web semantic approach
2. Décrire les trajectoires de soins
   - Problématique générale
   - Catégorisation d'exemples
   - Métriques et clustering de séquences
   - Exemple OPTISOINS
3. Abstraction de trajectoires
4. Conclusion

Parcours et trajectoires : clarification

Trois notions distinctes dans le contexte de l'analyse de parcours

1. **Trajectoire de soins**
2. **Trajectoire observée**/décrite
3. **Parcours de soins**

Parcours et trajectoires : clarification

Trois notions distinctes dans le contexte de l'analyse de parcours

1. **Trajectoire de soins**
   - ensemble des évènements de soins d'une personne
   - il s'agit de la donnée "réelle" : uniquement théorique
2. **Trajectoire observée**/décrite
   - *description* d'une trajectoire
   - les informations quelle contient sont partielles
   - offre un point de vue particulier sur la trajectoire

   → *données* longitudinales d'**un patient particulier**
3. **Parcours de soins**
   - description d'une abstraction d'un enchainement des évènements de soins
   - correspond à un **type de patients**
   - par exemple un *guidelines*

Parcours et trajectoires : clarification

Trois notions distinctes dans le contexte de l'analyse de parcours

1. **Trajectoire de soins**
   - ensemble des évènements de soins d'une personne
   - il s'agit de la donnée "réelle" : uniquement théorique
2. **Trajectoire observée**/décrite
   - *description* d'une trajectoire
   - les informations quelle contient sont partielles
   - offre un point de vue particulier sur la trajectoire

   → *données* longitudinales d'**un patient particulier**
3. **Parcours de soins**
   - description d'une abstraction d'un enchainement des évènements de soins
   - correspond à un **type de patients**
   - par exemple un *guidelines*

Trajectoire = Trajectoire observée

Si dans la suite on parle de trajectoire, il ne faut pas perdre de vue qu'il s'agit systématiquement d'une trajectoire observée

Différentes notions : exemple

Un homme de 35 ans a été admis à l'hôpital pour un gonflement périorbitaire, une rougeur et une douleur le 24 mai 2014. On lui a alors diagnostiqué une cellulite périorbitaire. Il a été traité avec de la clindamycine intraveineuse et de la ciprofloxacine, ce qui a réduit la rougeur et le gonflement de l'orbite. Cependant, le deuxième jour suivant le traitement antibiotique, il a développé des nausées et des douleurs abdominales dans le quadrant supérieur droit, et ses tests de la fonction hépatique ont commencé à augmenter. Un diagnostic de lésion hépatique idiosyncrasique d'origine médicamenteuse a été posé.

Trajectoire observée (point de vue medico-administratif)

- Hospitalisation avec le code CIM H050 au 24 mai (Inflammation aiguë de l'orbite [Cellulite orbite])
- Delivrance de J01FF01 et J01MA02 (codes ATC) du 24 mai au 26 mai
- Test NABM 0516 (ALAT,TGP) le 24 mai
- Test NABM 0516 (ALAT,TGP) le 25 mai
- Test NABM 0516 (ALAT,TGP) le 26 mai
- CIM S361 (Lésion traumatique du foie et de la vésicule biliaire) le 26 mai

Différentes notions : exemple

Un homme de 35 ans a été admis à l'hôpital pour un gonflement périorbitaire, une rougeur et une douleur le 24 mai 2014. On lui a alors diagnostiqué une cellulite périorbitaire. Il a été traité avec de la clindamycine intraveineuse et de la ciprofloxacine, ce qui a réduit la rougeur et le gonflement de l'orbite. Cependant, le deuxième jour suivant le traitement antibiotique, il a développé des nausées et des douleurs abdominales dans le quadrant supérieur droit, et ses tests de la fonction hépatique ont commencé à augmenter. Un diagnostic de lésion hépatique idiosyncrasique d'origine médicamenteuse a été posé.

Parcours de traitement de l'inflammation aiguë de l'orbite (*fictif*)

- Diagnostique H050 à J_0
- Démarrage immédiat d'un traitement antibiotique par Fluoroquinolones (J01MA) de J_0 à $J + 7$, dosage entre 20 et 30 mg/j IV
- Suivi de la fonction hépatique pendant toutes la durée du traitement jusqu'à 1 semaine après

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins

2. Catégorisation des trajectoires
3. Abstraction des trajectoires en parcours
4. Interpréter les parcours

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins
   - → définir un modèle de représentation des trajectoires
   - → représenter l'information contenue dans une base de données par des trajectoires
2. Catégorisation des trajectoires
3. Abstraction des trajectoires en parcours
4. Interpréter les parcours

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins
   - → définir un modèle de représentation des trajectoires
   - → représenter l'information contenue dans une base de données par des trajectoires
2. Catégorisation des trajectoires
3. Abstraction des trajectoires en parcours
4. Interpréter les parcours

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins
   - → définir un modèle de représentation des trajectoires
   - → représenter l'information contenue dans une base de données par des trajectoires
2. Catégorisation des trajectoires
3. Abstraction des trajectoires en parcours
4. Interpréter les parcours

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins
   - → définir un modèle de représentation des trajectoires
   - → représenter l'information contenue dans une base de données par des trajectoires
2. Catégorisation des trajectoires
3. Abstraction des trajectoires en parcours
4. Interpréter les parcours

Étapes similaires à celle d'une étude épidémiologiques, mais:

- on part de données déjà disponibles
- on utilise des représentations des patients qui capturent la notion de trajectoire

Outline

1. Parcours de soins
   - Parcours de soins en santé publique
   - Analyser les parcours de soins
   - Représentation des trajectoires de soins (cas d'étude AP-HP)
   - A web semantic approach
2. Décrire les trajectoires de soins
   - Problématique générale
   - Catégorisation d'exemples
   - Métriques et clustering de séquences
   - Exemple OPTISOINS
3. Abstraction de trajectoires
4. Conclusion

Cas d'étude : cancers broncho-pulmonaires

Utilisation d'un cas d'étude

- Objectif: illustrer les différentes étapes et les difficultés rencontrées lors de la réalisation d'un étude d'analyse de parcours de soins
- Contexte:
  - projet OPTISOINS
  - données hospitalières (APHP)
- limites:
  - probablement d'autres difficultés non-rencontrées dans ce cas
  - dépendant de cette tâche et source de données

Cas d'étude : cancers broncho-pulmonaires

Cancers broncho-pulmonaires

- 50000 nouveaux cas/an
- Traitement curatif : chirurgie
- Disparité observée entre services : p.ex. taux d'hospitalisation en réanimation
- Le délai de prise en charge entre la suspicion de cancer broncho-pulmonaire et l'intervention chirurgicale est un facteur pronostique majeur

- Question : existe-t-il des parcours plus favorables pour le pronostique du patient?
  - → quelle vue ???
  - → "des parcours": il n'y a probablement pas un parcours unique, dépend de:
    - typologie de patients
    - environnement local (accès aux soins)

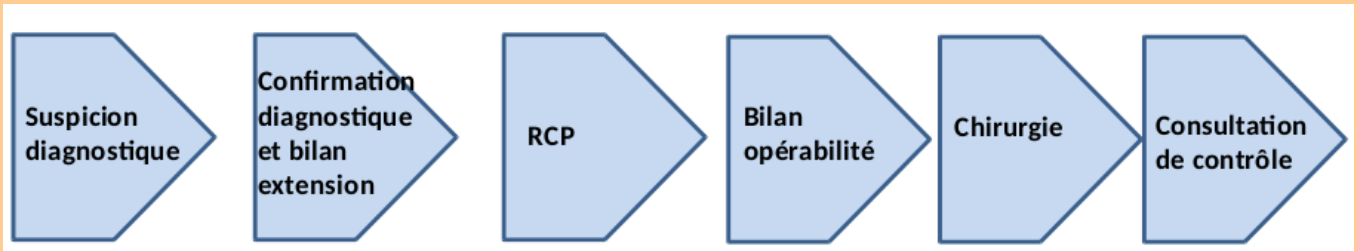
Cas d'étude : cancers broncho-pulmonaires

Cancers broncho-pulmonaires

- 50000 nouveaux cas/an
- Traitement curatif : chirurgie
- Disparité observée entre services : p.ex. taux d'hospitalisation en réanimation
- Le délai de prise en charge entre la suspicion de cancer broncho-pulmonaire et l'intervention chirurgicale est un facteur pronostique majeur

- Question : existe-t-il des parcours plus favorables pour le pronostique du patient?

Parcours typique



Représentation des trajectoires de soins

Problématiques

- Capturer l'information utile pour décrire ce qu'on souhaite observer dans une trajectoire
- Couvrir la population de manière large et diverse

Défi spécifique à l'utilisation des BDMA

- Identifier des sources de données: e.g. données hospitalières, SNDS, cohorte, etc.
- Extraire l'information utile à partir des données
  - les données ne sont pas forcément facilement exploitables : données textuelles
  - les données sont secondaires : saut sémantique

$\Rightarrow$ On fait au mieux avec ce qu'on a!

Représentation des trajectoires de soins (*OPTISOINS*)

- Base de données: Utilisation des données de l'EDS de l'AP-HP
  - → La base de données choisie est un "filtre" d'observation (e.g. différent entre SNDS / EDS)
- Reconstruction des trajectoires de soins
  - → identification d'une cohorte de patients répondant aux critères de l'étude
  - → identification des évènements et variables d'intérêt
  - → structurer les informations

Représentation des trajectoires de soins (*OPTISOINS*) I

Cohorte de patients

- Critères de l'étude: traitement du cancer bronco-pulmonaire
- Prisme de l'hopital: identifier les exérèses pulmonaires
- Prisme de la base de données: repérage par code CIM dans les actes médicaux
  - Lobectomie : GFFA004, GFFA006, GFFA008, GFFA009, GFFA010, GFFA013, GFFA015, GFFA016, GFFA018, GFFA019, GFFA022, GFFA023, GFFA026, GFFA027, GFFA030, GFFA031, GFFA033, GFFA034
  - Pneumonectomie : GFFA001, GFFA002, GFFA007, GFFA011, GFFA012, GFFA024, GFFA025, GFFA028
  - Segmentectomie : GFFA029
  - Wedge : GFFA017, GFFA021, GFFC002

Représentation des trajectoires de soins (*OPTISOINS*) II

Repérages de variables/évènements d'intérêt

- Acte chirurgical: idem!
- Tenue d'une réunion RCP et sa date ?
  - Pas d'acte médical associé $\Rightarrow$ pas de codage
  - ... mais des comptes-rendus informatisés $\Rightarrow$ analyse des compte-rendus
    - codification des comptes-rendus: CR:CR-IMAGE/ CR:CR-ACTE / CR:CR-DIAG
    - analyse du contenu textuel

Analyse des comptes-rendus textuels

$\rightarrow$ Essentiel pour les données hospilatières

- la plupart du temps: identification par mots clés
- vers l'utilisation des outils de TAL/NLP

Représentation des trajectoires de soins (*OPTISOINS*) III

Liste des types d'évènements recherchés

- exerèses
- traitements: chimio, immuno, radio, etc.
- RCP
- anapath
- scanner thoracique
- fibro broncho
- consultations (pre-post-op)
- passage en réa
- rehospit
- hdj

Liste de variables socio-démographiques recherchées

- Age
- Genre
- Comorbidités: Cancer, BPCO, cardiopathie Ischémique, diabète, HTA, Insuffisance respiratoire, Insuffisance renale
- IMC
- Tabac (PA)

Représentation des trajectoires de soins (*OPTISOINS*) IV

⇒ construction d'une table de patients + table évènementielle

	patient_num	encounter_num	start_date	location_cd	uf_name	hopital	concept_cd	event_name
0	[illegible]	[illegible]	2018-01-02	UFR:021037	021037 : CCH CHIR THORACIQUE (date de fin = 01...	Cochin	[CCAM:GFFA017]	Exerese pulmonaire ref
1	[illegible]	[illegible]	2018-01-02	UFR:011580	011580 : BCH CHIR VASC THOR	Bichat	[CCAM:GFFA013]	Exerese pulmonaire ref
2	[illegible]	[illegible]	2018-01-02	UFR:021037	021037 : CCH CHIR THORACIQUE (date de fin = 01...	Cochin	[CCAM:GFFA017, CCAM:GFFA029]	Exerese pulmonaire ref
3	[illegible]	[illegible]	2018-01-03	UFR:087144	087144 : TNN CHIR THOR VASC	Tenon	[CCAM:GFFA009]	Exerese pulmonaire ref
4	[illegible]	[illegible]	2018-01-03	UFR:011580	011580 : BCH CHIR VASC THOR	Bichat	[CCAM:GFFA009]	Exerese pulmonaire ref
5	[illegible]	[illegible]	2018-01-03	UFR:095063	095063 : AVC UF URG.THOR.VASCU.	Avicenne	[CCAM:GFFC002]	Exerese pulmonaire ref
6	[illegible]	[illegible]	2018-01-03	UFR:011580	011580 : BCH CHIR VASC THOR	Bichat	[CCAM:GFFA021]	Exerese pulmonaire ref
7	[illegible]	[illegible]	2018-01-03	UFR:021037	021037 : CCH CHIR THORACIQUE (date de fin = 01...	Cochin	[CCAM:GFFA009]	Exerese pulmonaire ref

Repérage d'évènements dans des documents médicaux textuels

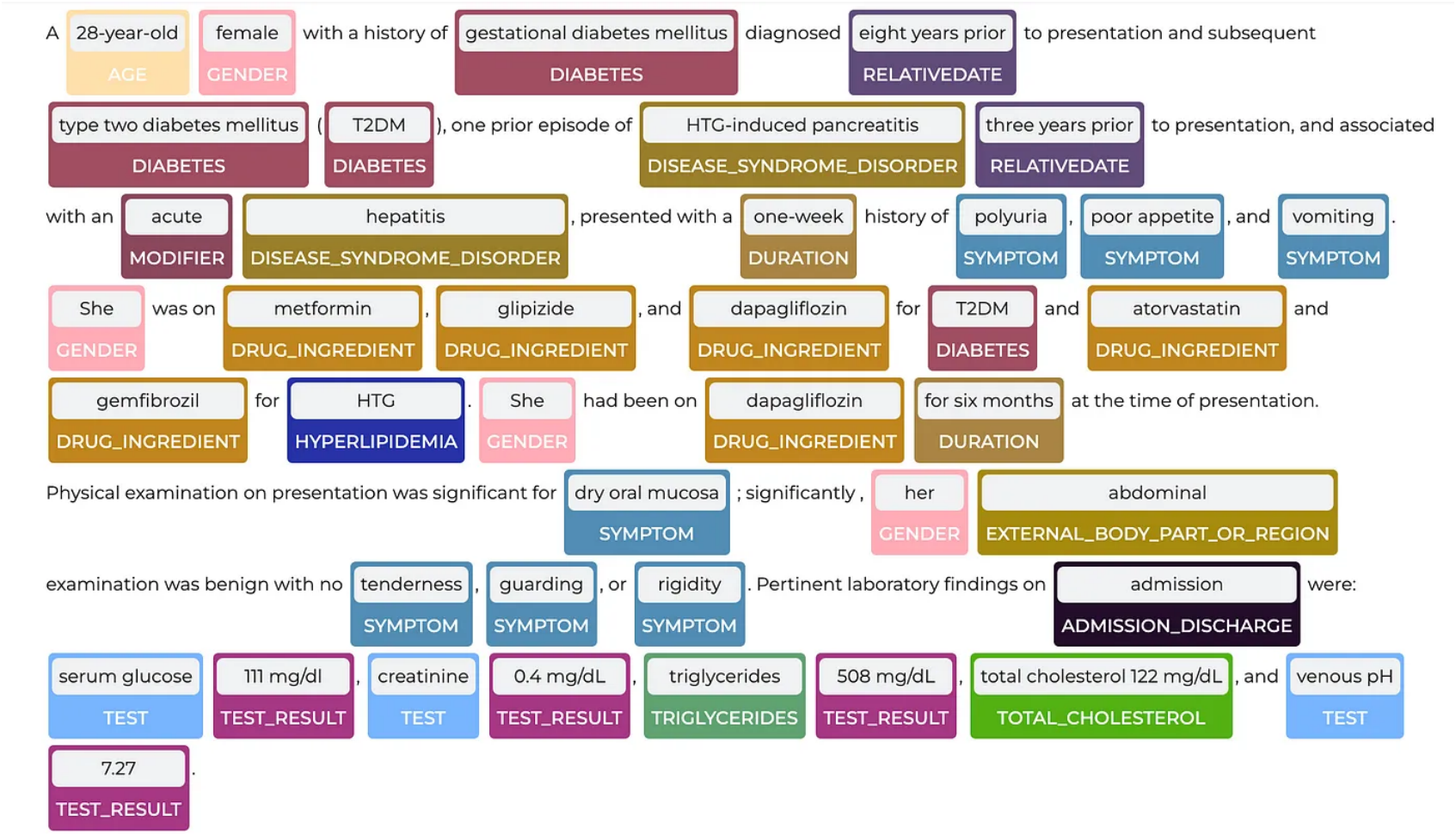
Exemples de besoin

- extraction d'une date
- extraction d'un stage de cancer
- extraction d'un nombre de paquets jours
- filtrage des CR d'anapath

Méthodologies

1. Méthodes à base de *reg exp* (expressions régulières)
   `"[ée]x[ée]r[éèe]se|pulm|poumon|bronch|lobe|segment|wedge|thora|pleural|pl[eè]vre"`
2. Méthodes à base de modèles de langue: NER (Named Entity Recognition)
   → identification de concepts par un modèle appris sur des données

Named Entity Recognition in Healthcare



Source: https://www.johnsnowlabs.com/in-depth-comparison-of-spark-nlp-for-healthcare-and-chatgpt-on-clinical-named-entity-recognition/.

NER avancés

Négations

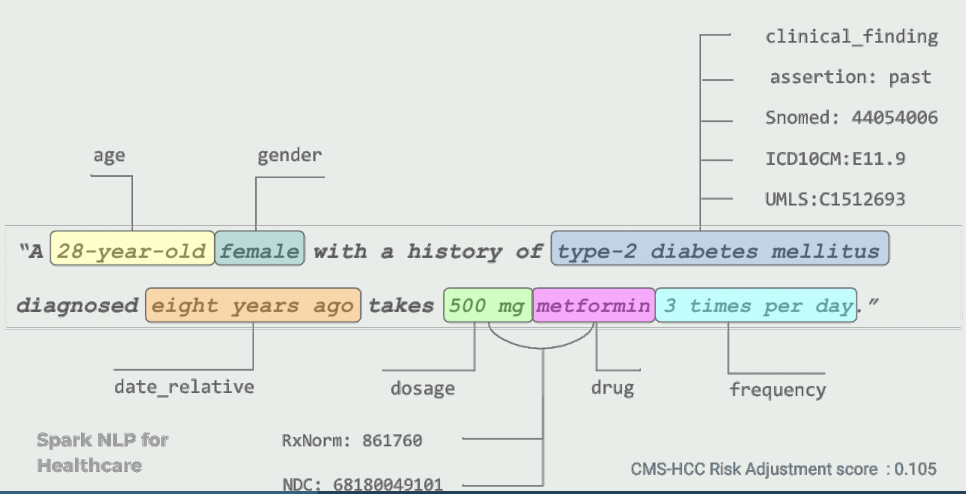
→ les négations sont très fréquemment utilisées dans les documents médicaux

- `Absence d'EP`
- `Pas de perte de poids ni d'épisodes de douleurs à l'estomac`

→ Problème de détection des négations et des portées de la négation

Ontologies

- Détection d'entités qui sont des concepts dans des ontologies médicales



Représentation des trajectoires de soins

Difficultés et limites

- Les difficultés et les limites proviennent du fait
  - d'une utilisation secondaire de données
  - d'un écart sémantique entre les données et les notions de la trajectoire
- Techniquement: manipulation de nombreux outils combinés ... principalement requêtes SQL et NLP
- Nécessite une très bonne connaissance de la base (et de son contenu) pour trouver les indices identifiant les évènements médicaux d'intérêt

Il est fortement recommandé de:

1. Évaluer la qualité des méthodologies d'identification d'évènements
2. Documenter la manière et la logique de repérage des évènements

⇒ **cette étape est fastidieuse, longue mais déterminante pour l'analyse**

Outline

1. Parcours de soins
   - Parcours de soins en santé publique
   - Analyser les parcours de soins
   - Représentation des trajectoires de soins (cas d'étude AP-HP)
   - A web semantic approach
2. Décrire les trajectoires de soins
   - Problématique générale
   - Catégorisation d'exemples
   - Métriques et clustering de séquences
   - Exemple OPTISOINS
3. Abstraction de trajectoires
4. Conclusion

Utilisation d'ontologies pour faciliter la reconstruction des trajectoires : exemple d'utilisation du SNDS

- un des soucis majeurs est l'écart sémantique
- les méthodes de raisonnement automatique sur les connaissances permettent de combler ces écarts (cf. raisonnement ontologique)

Intuition

- Une exérèse pulmonaire est repérée par une requêtes SQL complexe (cf plus haut)
- Les outils du raisonnement ontologiques permettent de définir des concepts de haut niveau plus richement et intuitivement que le SQL
  - $\rightarrow$ définir le concept d'exérèse pulmonaire comme une équivalence de notion plus complexes
- Il existe de plus de nombreuses ontologies médicales

Alternative du web sémantique

Semantic web

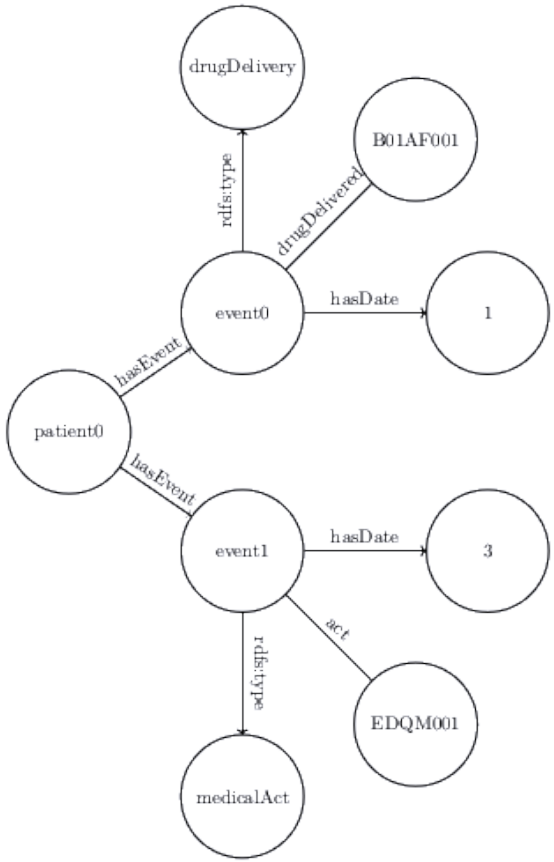
- Adapté pour représenter des données et raisonner sur des connaissances formalisées
- language de requête expressif (SPARQL)
- Peu adapté pour manipuler des données et des requêtes temporelles (*nativement*) [ZWLC19]

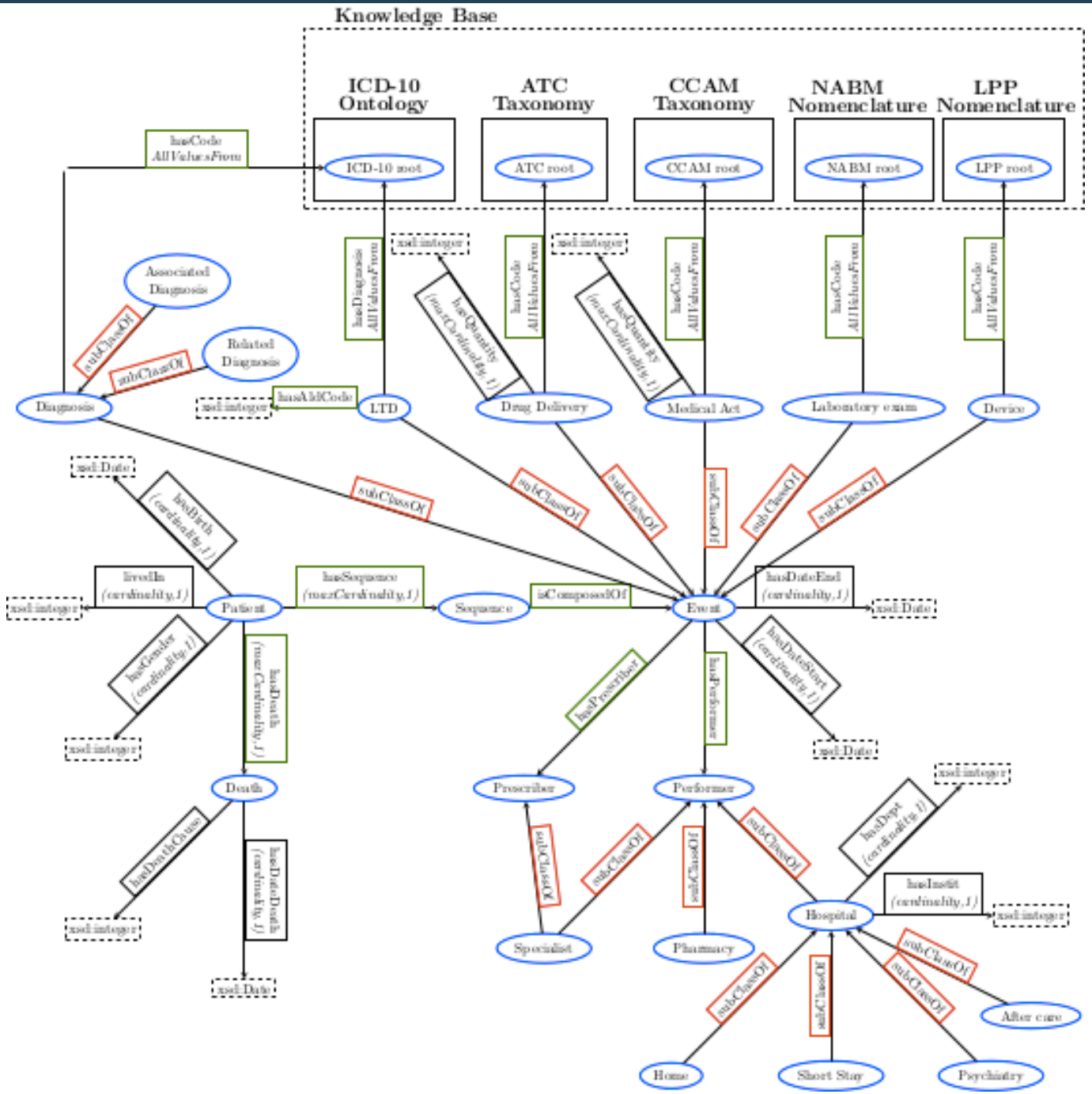
Interrogation de trajectoire par des parcours (J. Bakalara [BGD+19])

- Objectif:
- Représentation des trajectoires de soins en RDF
- Parcours = requête SPARQL
- Requêtage efficace pour l'identification des patients adhérents au parcours

Représentation RDF des trajectoires de soins

- Modèle de représentation des trajectoires de soins en RDF
  - Adaptation du travail de Y. Rivault [RDLM19]
  - Dérive de la notion de séquence d'événements
    - un type
    - une date (nombre entier)
  - Modélisation plus complète des notions du SNDS
- Outil de transformation des données SNDS en format RDF
  - Expérimentées sur données synthétiques et réelles




Knowledge Base
ICD-10 Ontology
ATC Taxonomy
CCAM Taxonomy
NABM Nomenclature
LPP Nomenclature
ICD-10 root
ATC root
CCAM root
NABM root
LPP root
hasCode AllValuesFrom
Associated Diagnosis
Related Diagnosis
subClassOf
Diagnosis
hasAltCode
LTD
Drug Delivery
Medical Act
Laboratory exam
Device
Patient
Sequence
Event
Death
Prescriber
Performer
Specialist
Pharmacy
Hospital
Home
Short Stay
Psychiatry
After care
xsd:integer
xsd:Date

Wrap up

À ce stade de notre étude ... on a vu:

- différentes notions : parcours/trajectoires, de vie/de soins/...,
- l'identification des évènements médicaux dans une grande masse de données
  - requêtes parfois complexes (*temporelles*)
  - utilisation d'outils d'extraction d'information dans des textes
- la notion d'**écart sémantique** entre la donnée brut et l'information recherchée
  - → l'utilisation d'ontologie peut aider à combler cet écart.
- la construction d'une table d'évènements: on dispose d'un tableau de données décrivant, par patient et par dates, des évènements de santé.

⇒ Construction de trajectoires à partir de données médico-administratives

⏩ Représentations informatiques

Outline

1 Parcours de soins
- Parcours de soins en santé publique
- Analyser les parcours de soins
- Représentation des trajectoires de soins (cas d'étude AP-HP)
- A web semantic approach

2 Décrire les trajectoires de soins
- Problématique générale
- Catégorisation d'exemples
- Métriques et clustering de séquences
- Exemple OPTISOINS

3 Abstraction de trajectoires

4 Conclusion

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins
   - → définir un modèle de représentation des trajectoires
   - → représenter l'information contenue dans une base de données par des trajectoires
2. Catégorisation des trajectoires
   - → créer des groupes de trajectoires qui sont similaires
3. Abstraction des trajectoires en parcours
4. Interpréter les parcours

Décrire l'ensemble des trajectoires de soins

Objectif : proposer une compréhension d'un ensemble de trajectoires de soin

- existe-t-il des groupes de trajectoires similaires ?
- qu'on en commun des groupes de trajectoires ?
- en quoi un trajectoires est-il proche d'un groupe de parcours ?

Deux types de questions méthodologiques

- comment regrouper des trajectoires "similaires"? (*clustering*)
- comment abstraire des trajectoires ? (*vers des parcours ?*)

Dans la suite : point de vue "longitudinal"

On cherche à conserver la dimension longitudinale des trajectoires/parcours

$\rightarrow$ conserver le séquencement des évènements

$\rightarrow$ conserver les dates des évènements

$\Rightarrow$ les techniques de clustering usuelles doivent être adaptées

Outline

1 Parcours de soins
- Parcours de soins en santé publique
- Analyser les parcours de soins
- Représentation des trajectoires de soins (cas d'étude AP-HP)
- A web semantic approach

2 Décrire les trajectoires de soins
- Problématique générale
- Catégorisation d'exemples
- Métriques et clustering de séquences
- Exemple OPTISOINS

3 Abstraction de trajectoires

4 Conclusion

Clustering examples

Clustering

- Overall objective of clustering: Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n examples, their clustering aims to identify groups/clusters of *similar* examples that are *dissimilar* to the other groups/clusters.
- Classical approaches:
  - K-Means
  - Hierarchical clustering
  - Self Organizing Maps
  - and others: density based clustering (DB-SCAN), EM algorithm, Affinity propagation, ...

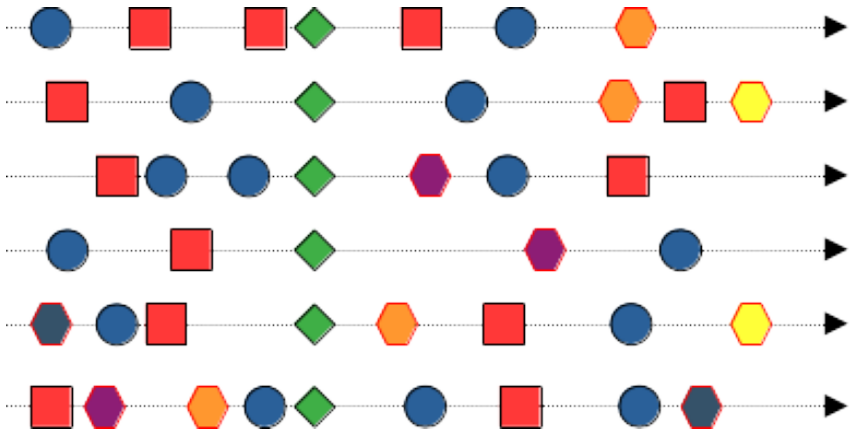
Two important choices

- similarity measure between examples
- aggregation function: how to create a representative

$\Rightarrow$ contributes to the semantics of the clusters

Comparaison de séquences temporelles : sémantique(s)

Comment catégoriser ces séquences ?

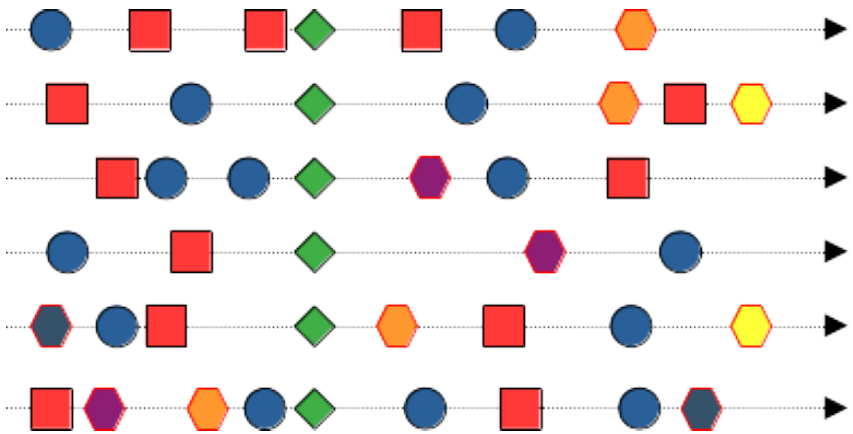


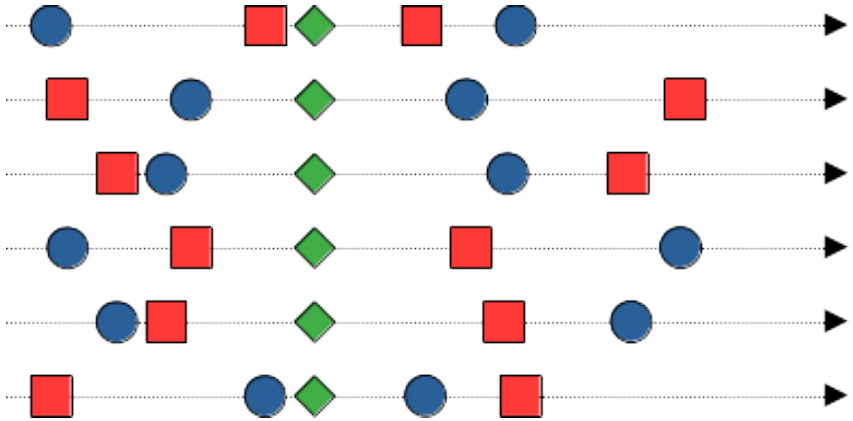
⇒ Quel métrique peut capturer ces notions ?

Comparaison de séquences temporelles : sémantique(s)

Comment catégoriser ces séquences ?

- être robuste à des évènements manquants/additionnels





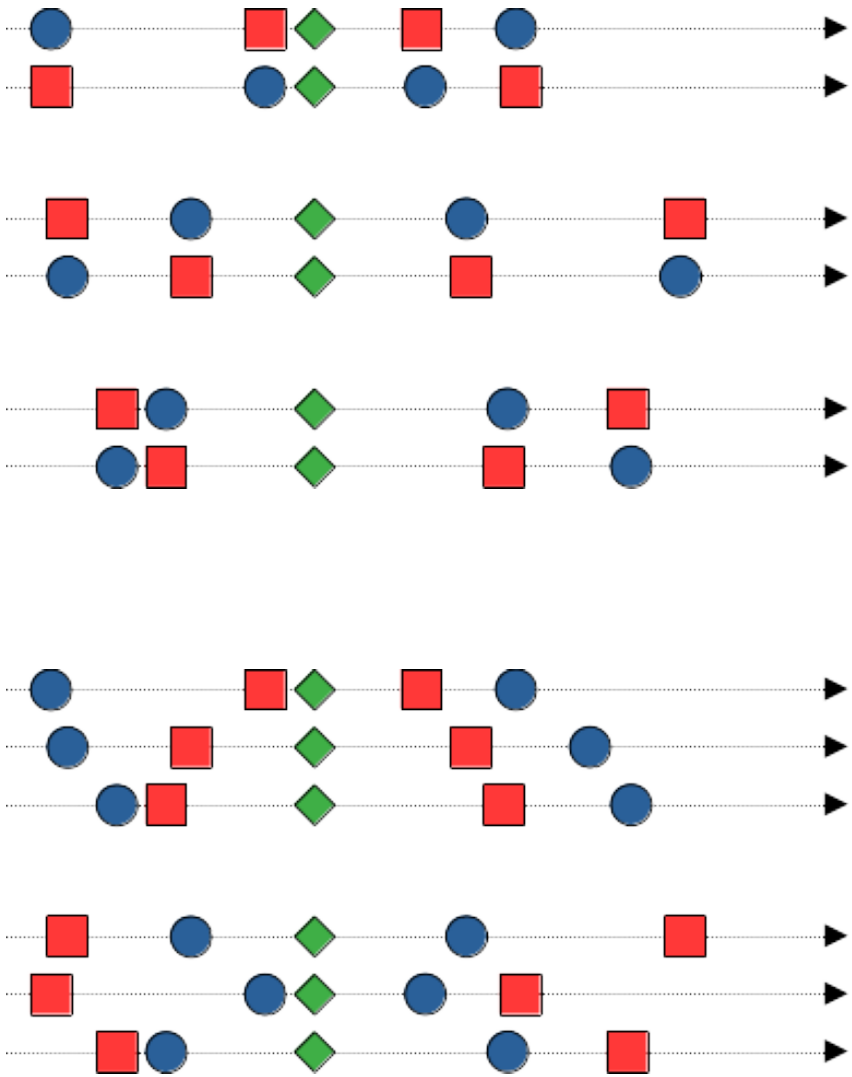
$\Rightarrow$ Quel métrique peut capturer ces notions ?

Comparaison de séquences temporelles : sémantique(s)

Comment catégoriser ces séquences ?

- être robuste à des évènements manquants/additionnels
- deux dimensions: **temporelle** ou **symbolique**

⇒ Quel métrique peut capturer ces notions ?



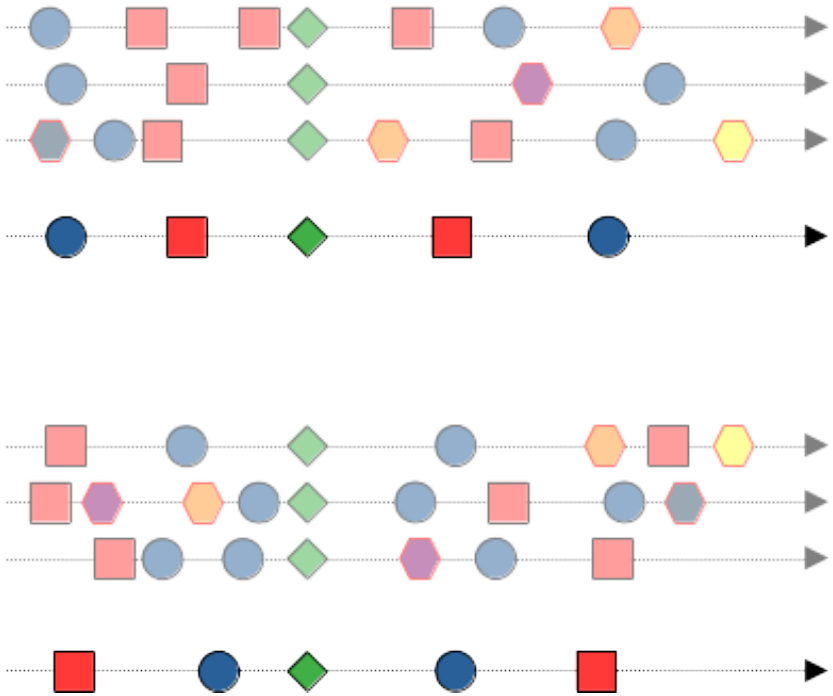
Comparaison de séquences temporelles : sémantique(s)

Comment catégoriser ces séquences ?

- être robuste à des évènements manquants/additionnels
- deux dimensions: **temporelle** ou **symbolique**

Et engendrer une séquence moyenne ...

$\Rightarrow$ Quel métrique peut capturer ces notions ?



Outline

1 Parcours de soins
- Parcours de soins en santé publique
- Analyser les parcours de soins
- Représentation des trajectoires de soins (cas d'étude AP-HP)
- A web semantic approach

2 Décrire les trajectoires de soins
- Problématique générale
- Catégorisation d'exemples
- Métriques et clustering de séquences
- Exemple OPTISOINS

3 Abstraction de trajectoires

4 Conclusion

Métriques entre séquences temporelles

Similarités

- Comparaison comme vecteurs : distance euclidienne
- Distances d'éditions : Levenshtein, etc.
- Comparaison flexible dans le temps : DTW, drop-DTW, LCSS, ...
- Comparaison lfexibles differentiables : soft-DTW, divergence

⏩ Métriques

Clustering

- Catégorisation basées sur les métriques :
  - → CAH, K-Means, etc.
- Autres approches
  - TriClustering
  - Informationnel
  - Kernel based
  - ...

⏩ Clustering avancés

Outils pour mener une analyse de clusters

TraMiner [GR13]

- State Sequences Analysis: représentation d'une séquence comme une suite contigue d'états
- collection de métriques et de méthodes
- outils de visualisation

Better to come ...

Outline

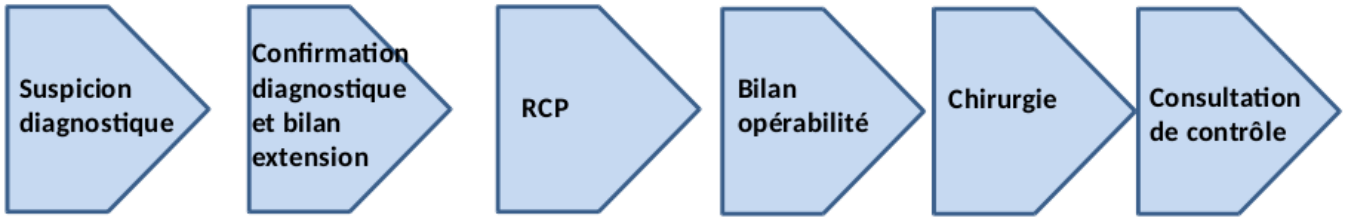
1 Parcours de soins
- Parcours de soins en santé publique
- Analyser les parcours de soins
- Représentation des trajectoires de soins (cas d'étude AP-HP)
- A web semantic approach

2 Décrire les trajectoires de soins
- Problématique générale
- Catégorisation d'exemples
- Métriques et clustering de séquences
- Exemple OPTISOINS

3 Abstraction de trajectoires

4 Conclusion

Clustering of OPTISOINS care trajectories



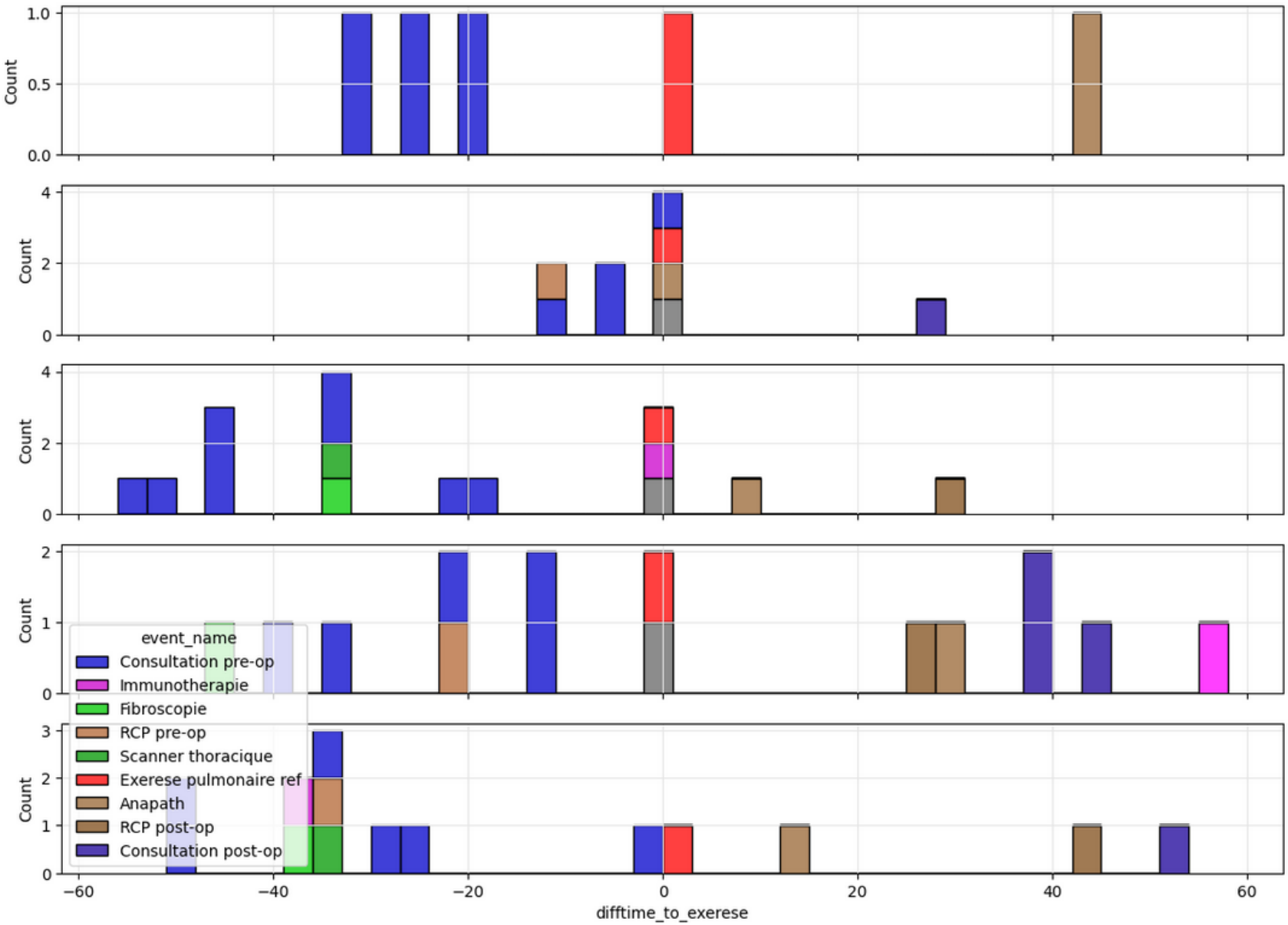
Care trajectories preparation

- collection of the dates for these events for 3311 patients
- index date: exeresis
- lot of missing events in trajectories
  - → motivation for using a method robust to missing events

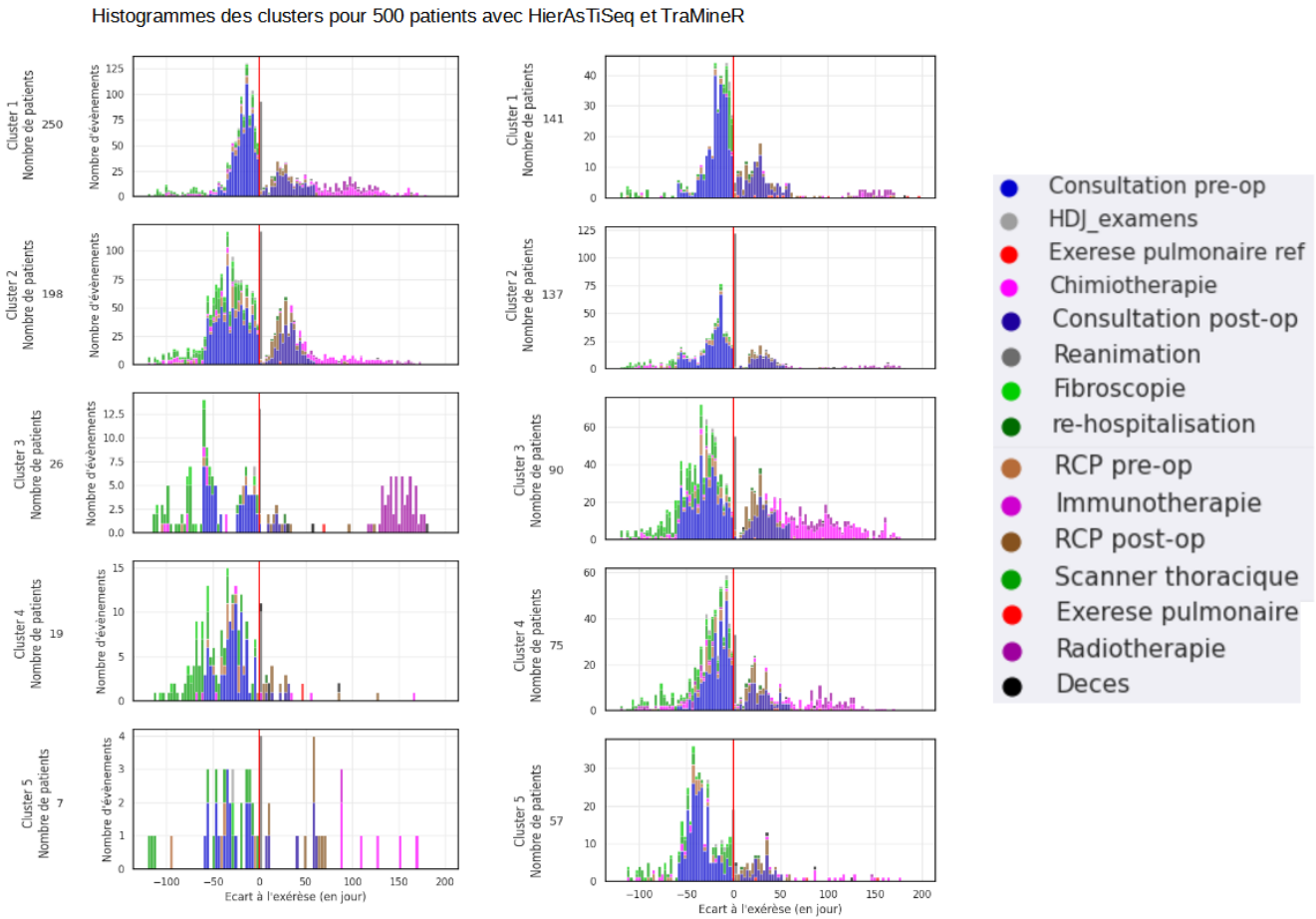
Clustering of care trajectories

- 5 clusters

Quelques exemples de trajectoires

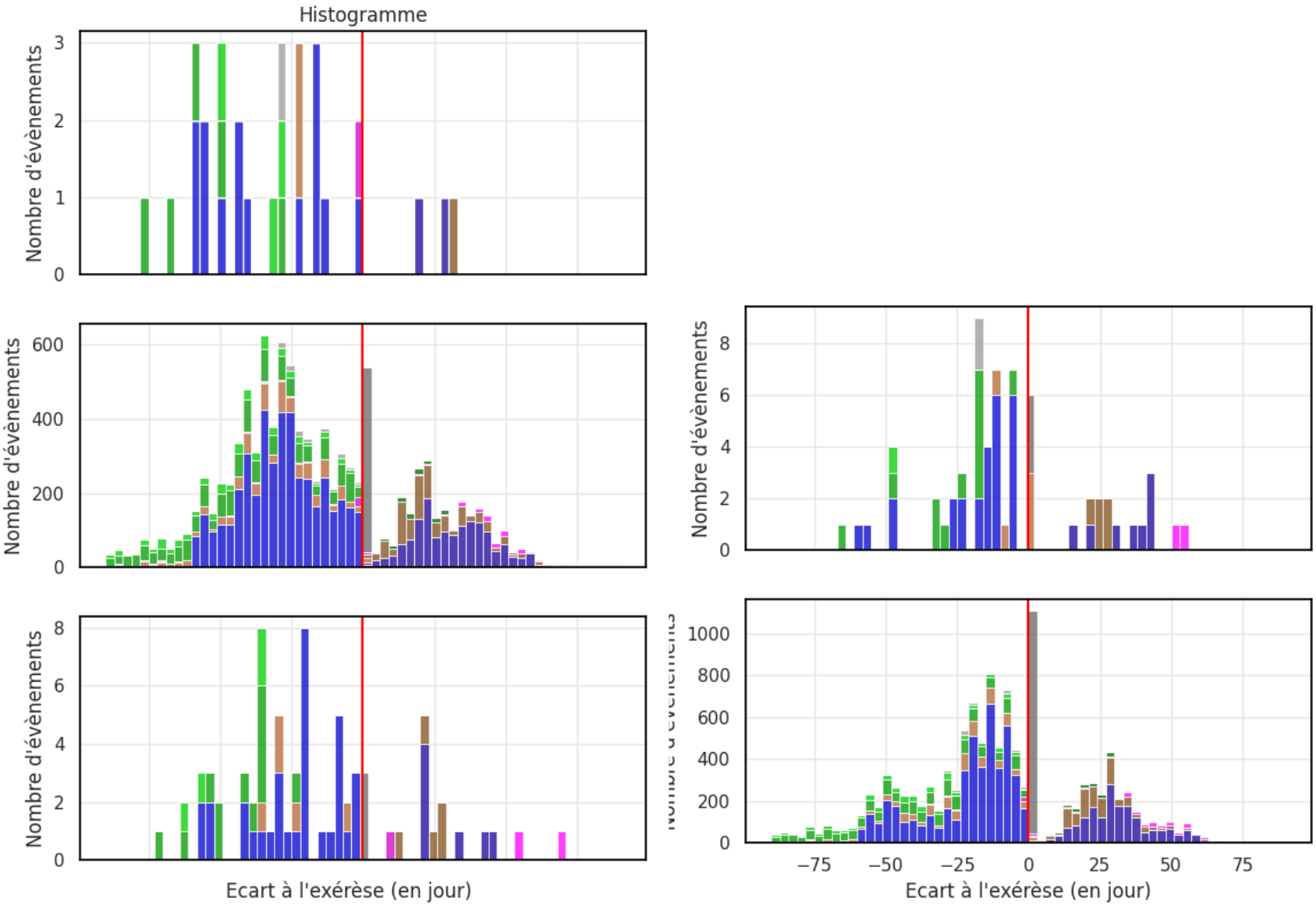


Comparaison HierAsTiSeq et TraMineR sur OPTISOINS (500 patients)

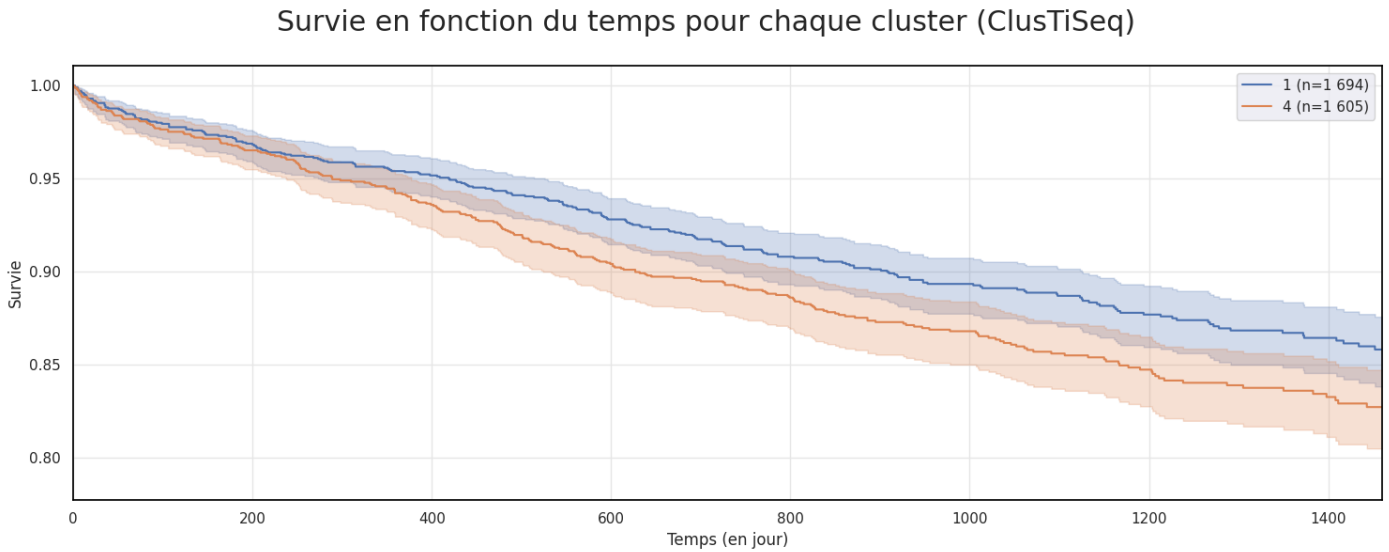


→ soucis de passage “à l’échelle” avec TraMineR

Résultats KMeans OPTISOINS (3311 patients)



Résultats KMeans OPTISOINS (3311 patients)



→ Légère difference de survie entre les deux groupes les plus volumineux

→ Pas d'identification de typologie de patients spécifiques

⇛ Importance du parcours de prise en charge

Wrap up

À ce stade de notre étude ... on a vu:

- utilisation de méthode de clustering de trajectoires similaires
- importance du choix de la métrique dans des algorithmes de clustering
  - → choix algorithmique et **sémantique**
- nombreuses alternatives de métriques sur des trajectoires de soins
  - → une meilleure métrique ??

⇒ Construction de groupes de trajectoires similaires

Wrap up

Ce qu'on n'a pas vu:

- Méthodes statistiques pour la modélisation de séries temporelles
  - clustering par approche fonctionnelle
  - etc.

Analyser les parcours de soins : méthodologie générale

1. Représentation des trajectoires de soins
   - → définir un modèle de représentation des trajectoires
   - → représenter l'information contenue dans une base de données par des trajectoires
2. Catégorisation des trajectoires
   - → créer des groupes de trajectoires qui sont similaires
3. Abstraction des trajectoires en parcours
   - → définir un modèle de représentation des parcours
   - → généraliser un ensemble de trajectoires sous forme d'un ou plusieurs parcours
4. Interpréter les parcours

Abstraction de trajectoires vers des parcours

→ Objectif : proposer une vue synthétique d'un ensemble de trajectoires

- pour généraliser
- pour résumer
- pour définir un comportement moyen

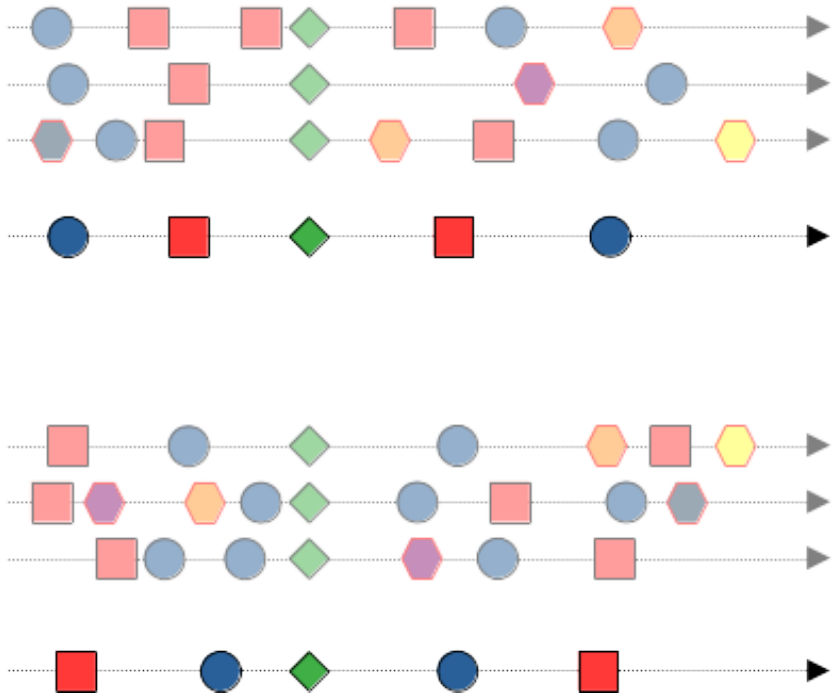
- Remarques :
  - l'information attendue est possiblement plus riche que la seule moyenne
  - pas de contrainte sur la forme de ma généralisation (en particulier pas forcément identique à une trajectoire)
  - doit être en cohérence avec la manière de regrouper des séquences
  - fonctionnalité généralement souhaitée : trajectory matching

→ vers la notion de parcours comme généralisation de trajecoires)

Exemple OPTISOINS

- Comment décrire les parcours des clusters

Timed sequence abstraction: average trajectories

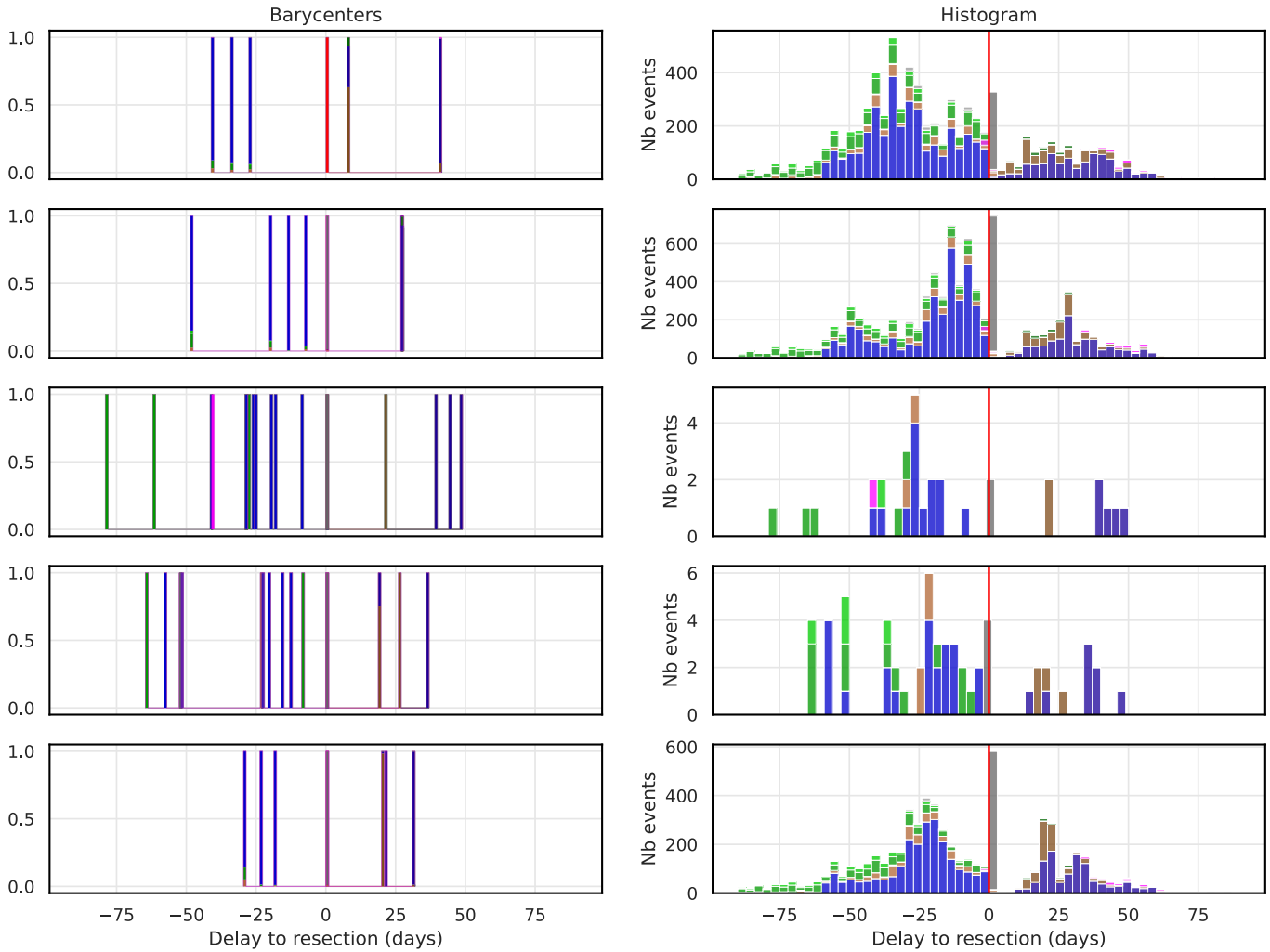


Pathway as an average trajectory

- Intuitive way to summarize a group of trajectories
- Not always so simple

- Clustering techniques: require a distances and the computation of representatives (barycenters)
- Create an **average** of the timed sequences
- The *representatives are of the same nature than objects*

OPTISOINS: moyennage par drop-DTW

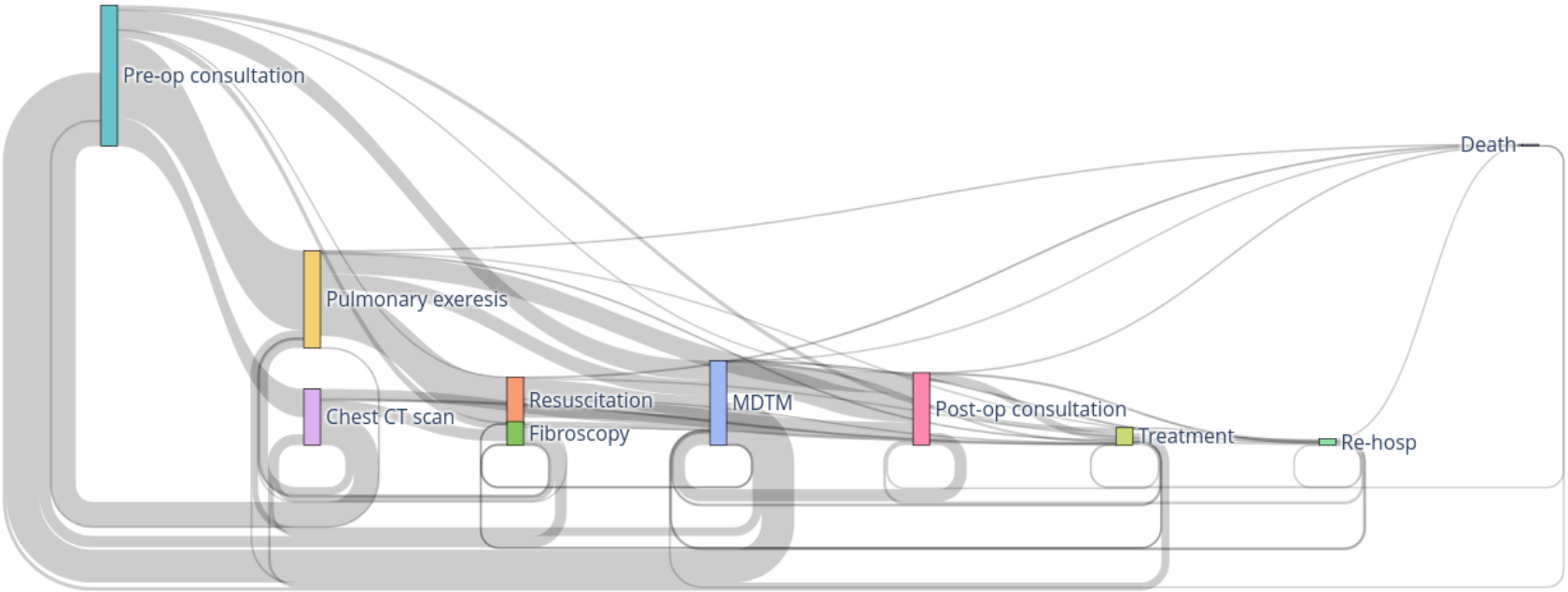


Timed sequence abstraction

Different ways to make temporal abstraction

- Induce a representative of a collection of timed sequences
- Use a single object built from time sequences (or not!)
- The domain of object maybe very different from the domain of timed sequences
- Two different objectives:
  - summarize the timed sequences: create a condensed representation of the whole sequences (**union**)
  - essentialize the timed sequences: capture the maximal commonalities in sequences (**intersection**)

Analyse visuelle de la cohorte

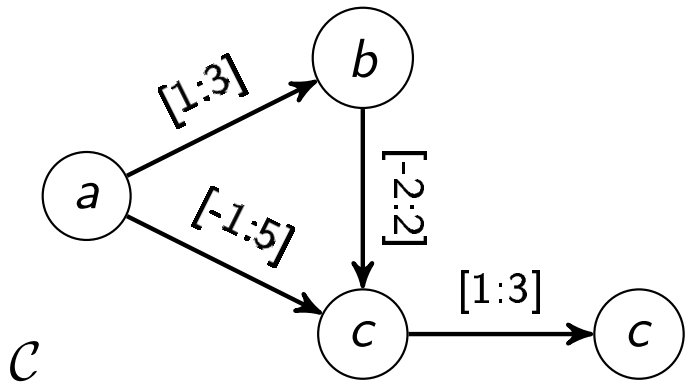


Example of temporal model: chronicles

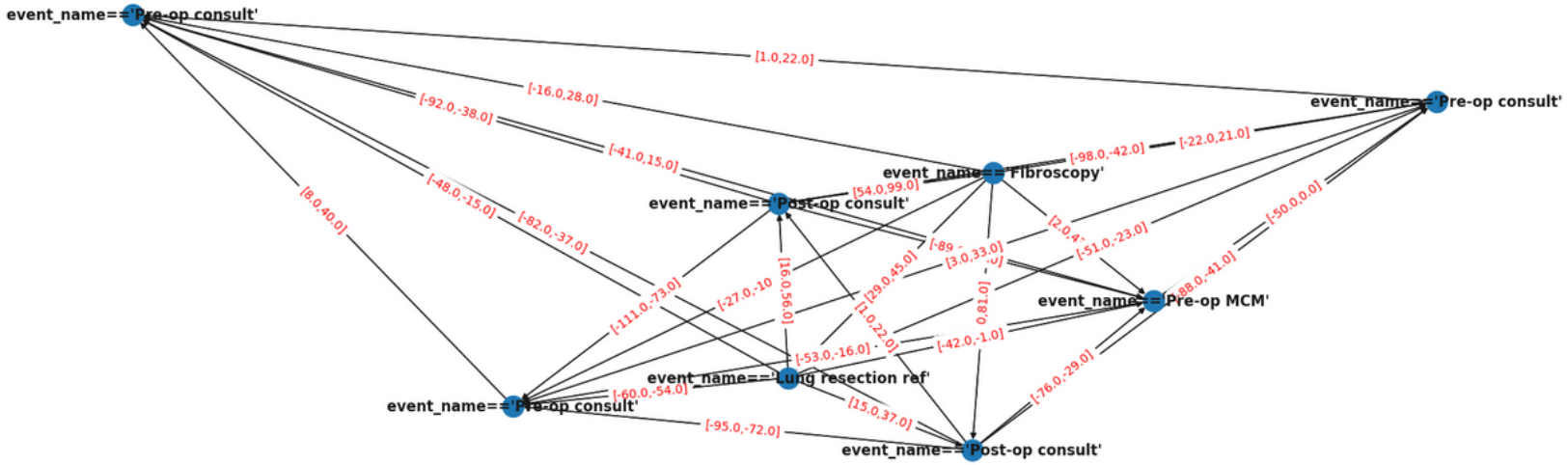
Chronicles [GM22, BG23]

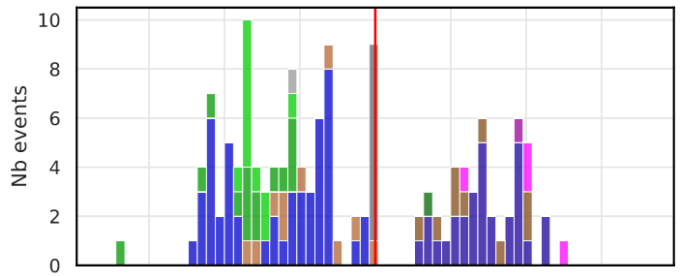
- Temporal model
  - multiset of events
  - flexible temporal constraints
- Interesting to represent a collection of sequences
  - → More expressive than a simple "average" sequence
  - → Topological properties to define clean notions of union/intersection

⏩ Autres méthodes d'abstraction



Example of temporal model: chronicles (OPTISOINS)





Wrap up

À ce stade de notre étude ... on a vu:

- le clustering de trajectoires similaires s'accompagne de la génération d'une représentation abstraite des clusters
  - pour faciliter leur interprétation
  - pour des raisons algorithmiques
- abstration comme généralisation (union) vs comme résumé (intersection)
- la création d'une "séquence moyenne"
- une abstraction peut prendre des formes variées
  - résumé visuel
  - abstraction avec des formalismes plus ou moins compexes

Conclusion générale: THM

Modélisation/représentation longitudinale des trajectoires ou parcours

- notions de parcours/trajectoires, de vie/de santé/de soins
- différents points de vue : patients, soignants, régulateurs
- reconstruction des trajectoires à partir des données de santé

Intérêt des ontologies: aider à combler les écarts sémantiques

- proposer une représentation des données plus intelligible
- structurer une information
- permettre de représenter des parcours

Description des parcours : clustering et abstraction

- Proposition d'une vue "informatique" de clustering de trajectoires
- Clustering de trajectoires : importance sémantique de la métrique
  → à choisir avec soin !
- Abstraction: construction de *représentants* d'un groupe = un parcours

Questions ?

References I

Alexis Bondu, Marc Boullé, and Benoît Grossin.

Saxo: An optimized data-driven symbolic representation of time series.
In The 2013 international joint conference on neural networks (IJCNN), pages 1–9. IEEE, 2013.

Philippe Besnard and Thomas Guyet.

Chronicles.
Springer Briefs, 2023.

Johanne Bakalara, Thomas Guyet, Olivier Dameron, Emmanuel Oger, and André Happe.

Temporal models of care sequences for the exploration of medico-administrative data.
In Workshop IA&Santé, PFIA, 2019.

Adeline Bailly, Simon Malinowski, Romain Tavenard, Thomas Guyet, and Laetitia Chapel.

Bag-of-temporal-sift-words for time series classification.
In ECML/PKDD workshop on advanced analytics and learning on temporal data, 2015.

Mathieu Blondel, Arthur Mensch, and Jean-Philippe Vert.

Differentiable divergences between time series.
In Arindam Banerjee and Kenji Fukumizu, editors, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 3853–3861. PMLR, 13–15 Apr 2021.

Debarun Bhattacharjya, Dharmashankar Subramanian, and Tian Gao.

Proximal graphical event models.
Advances in Neural Information Processing Systems, 31, 2018.

Marco Cuturi and Mathieu Blondel.

Soft-dtw: a differentiable loss function for time-series.
In International conference on machine learning, pages 894–903. PMLR, 2017.

References II

Yuanzhe Chen, Panpan Xu, and Liu Ren.

Sequence synopsis: Optimize visual summary of temporal event data.
IEEE transactions on visualization and computer graphics, 24(1):45–55, 2017.

Mikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan Jepson.

Drop-dtw: Aligning common signal between sequences while dropping outliers.
Advances in Neural Information Processing Systems, 34:13782–13793, 2021.

Elias Egho, Chedy Raïssi, Toon Calders, Nicolas Jay, and Amedeo Napoli.

On measuring similarity for sequences of itemsets.
Data Mining and Knowledge Discovery, 29(3):732–764, 2015.

Dominique Gay, Romain Guigourès, Marc Boullé, and Fabrice Clérot.

TESS: Temporal event sequence summarization.
In 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 1–10, 2015.

Thomas Guyet and Nicolas Markey.

Logical Forms of Chronicles.
In Alexander Artikis, Roberto Posenato, and Stefano Tonetta, editors, International Symposium on Temporal Representation and Reasoning (TIME), volume 247 of Leibniz International Proceedings in Informatics (LIPIcs), pages 7:1–7:15, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Enoal Gesny, Pierre Pinson, and Thomas Guyet.

Catégorisation de séquences temporelles – application à l'analyse de parcours de soins.
In Proceedings of EGC, pages 119–130, 2024.

References III

Alexis Gabadinho and Gilbert Ritschard.

Searching for typical life trajectories applied to childbirth histories.
Gendered life courses–Between individualization and standardization. A European approach applied to Switzerland, pages 287–312, 2013.

Rui Henriques and Sara C. Madeira.

Triclustering algorithms for three-dimensional data analysis: A comprehensive survey.
ACM Comput. Surv., 51(5), 2018.

Etienne Audureau Hana Sebia, Thomas Guyet.

Une extension de la décomposition tensorielle au phénotypage temporel.
In Actes de la conférence Extraction et Gestion de Connaissances, pages 1–12, 2023.

François Petitjean, Alain Ketterlin, and Pierre Gançarski.

A global averaging method for dynamic time warping, with applications to clustering.
Pattern recognition, 44(3):678–693, 2011.

Kishan Rama, Helena Canhão, Alexandra M Carvalho, and Susana Vinga.

Aliclu-temporal sequence alignment for clustering longitudinal clinical data.
BMC medical informatics and decision making, 19(1):1–11, 2019.

Yann Rivault, Olivier Dameron, and Nolwenn Le Meur.

queryMed: Semantic web functions for linking pharmacological and medical knowledge to data.
Bioinformatics, 2019.

Stan Salvador and Philip Chan.

Toward accurate dynamic time warping in linear time and space.
Intelligent Data Analysis, 11(5):561–580, 2007.

References IV

Bing Su and Gang Hua.

Order-preserving wasserstein distance for sequence matching.
In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1049–1057, 2017.

Sven Schneider, Maria Maximova, and Holger Giese.

Probabilistic metric temporal graph logic.
In International Conference on Graph Transformation, pages 58–76. Springer, 2022.

Ahmed Shifaz, Charlotte Pelletier, Francois Petitjean, and Geoffrey I Webb.

Elastic similarity measures for multivariate time series classification.
arXiv preprint arXiv:2102.10231, 2021.

Tara Safavi, Chandra Sripada, and Danai Koutra.

Scalable hashing-based network discovery.
In 2017 IEEE International Conference on Data Mining (ICDM), pages 405–414. IEEE, 2017.

Nikolaj Tatti and Jilles Vreeken.

The long and the short of it: summarising event sequences with serial episodes.
In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 462–470, 2012.

Xiufan Yu, Karthikeyan Shanmugam, Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, and Lingzhou Xue.

Hawkesian graphical event models.
In Manfred Jaeger and Thomas Dyhre Nielsen, editors, Proceedings of the 10th International Conference on Probabilistic Graphical Models, volume 138 of Proceedings of Machine Learning Research, pages 569–580. PMLR, 2020.

References V

Feng Zhou and Fernando Torre.

Canonical time warping for alignment of human behavior.
Advances in neural information processing systems, 22, 2009.

Fu Zhang, Ke Wang, Zhiyin Li, and Jingwei Cheng.

Temporal data representation and querying based on RDF.
IEEE Access, 7:85000–85023, 2019.

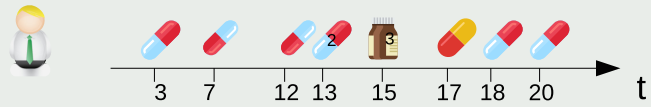
Représentation des trajectoires de soins : complément

Fondement d'informatique

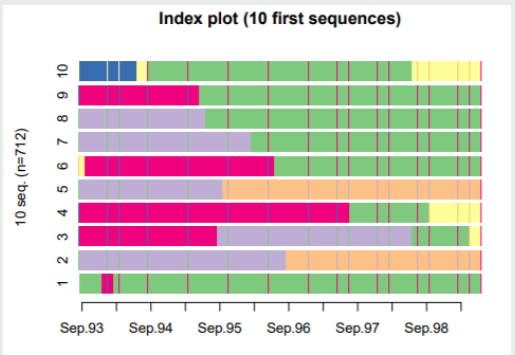
⇒ le **choix d'une représentation** informatique des données impacte les capacités d'analyse

Considération sur le choix de la représentation des évènements

- approche basée "évènements"
  - une trajectoire est un ensemble d'évènements
  - les évènements sont généralements ponctuels



- approche basée "état"
  - une trajectoire est une suite d'état
  - les évènements sont généralements ponctuels

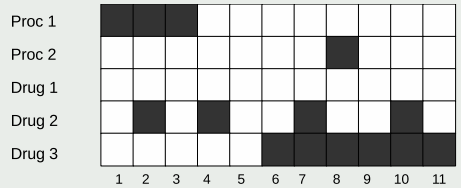


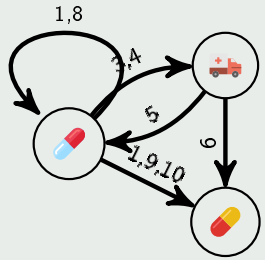
Alternative representations of timed sequences

Some alternative representations

$\langle$(💊, 2)(💊, 3)(🚑, 6)(💊, 11), (💊, 12)$\rangle$

- 💊 $\xrightarrow{1}$ 💊 $\xrightarrow{3}$ 🚑 $\xrightarrow{5}$ 🌡 $\xrightarrow{1}$ 💊
- binary matrix representation
- transition graph representations
- enhance the description of events with supplementary nodes: exemple with RDF [BGD+19]





The representation of the data impacts the choices of the methods and the results

Notion de date index

Problématique : alignement des dates

- la date est rarement utile pour les trajectoires
- difficile de mettre en relation des dates très différentes

→ problème d'alignement entre les trajectoires

Date index

- Date d'un évènement qu'on retrouve dans une trajectoire de manière unique et non-ambigue
- Les évènements de la trajectoire sont positionnés par rapport à cette date index
- Les trajectoires d'une cohorte peuvent être alignées sur cette date pour faciliter les comparaisons

Timed sequence similarities

Let $\boldsymbol{x} = \langle (x_1, t_1), \ldots, (x_n, t_n) \rangle$ and $\boldsymbol{y} = \langle (y_1, t_1), \ldots, (y_m, t_m) \rangle$ be two sequences

- $\cdot -_{\mathcal{E}} \cdot : \mathcal{E} \times \mathcal{E} \mapsto \mathbb{R}$ denotes the dissimilarity between two elements of $\mathcal{E}$.
- $sim(\boldsymbol{x}, \boldsymbol{y})$ denotes the similarity measure between two sequences
- $d(\boldsymbol{x}, \boldsymbol{y})$ denotes the dissimilarity measure between two sequences

How to choose a similarity?

- **semantically sound** with the problem/the intuition: the value is high when examples are intuitively the same
- **computationally efficient**: it does not take to much time to compute
- **suitable for algorithm**: the similarity properties required by algorithms are satisfied

Classical vectorial dissimilarity

pre-condition

- sequence with a regular time-sampling
- same length ($n = m$)

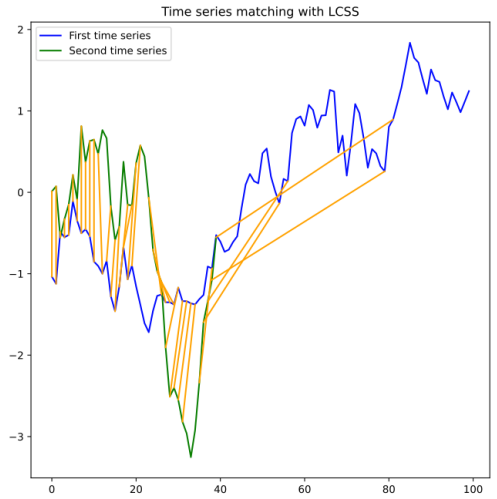
- Euclidean $sim_{ED}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} ||x_i -_{\mathcal{E}} y_i||^2}$
- Minkowski ($p \geq 1$) $sim_{mink}(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} ||x_i -_{\mathcal{E}} y_i||^p\right)^{\frac{1}{p}}$
- Manhatan $sim_{manh}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} |x_i -_{\mathcal{E}} y_i|$
- infinite $sim_{inf}(\boldsymbol{x}, \boldsymbol{y}) = \max_{i=1}^{n} |x_i -_{\mathcal{E}} y_i|$

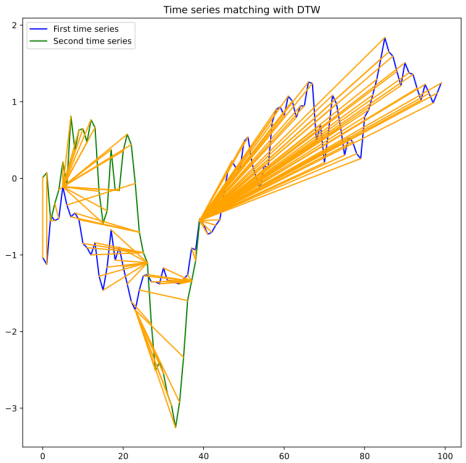
Time-elastic similarities

pre-condition

- sequence with a regular time-sampling
- ~~same length ($n = m$)~~

- Dynamic Time Warping
- Canonical Time Warping [ZT09]
- Longuest Common Subsequences (LCSS)
- Recent review of elastic similarity measures [SPPW21]

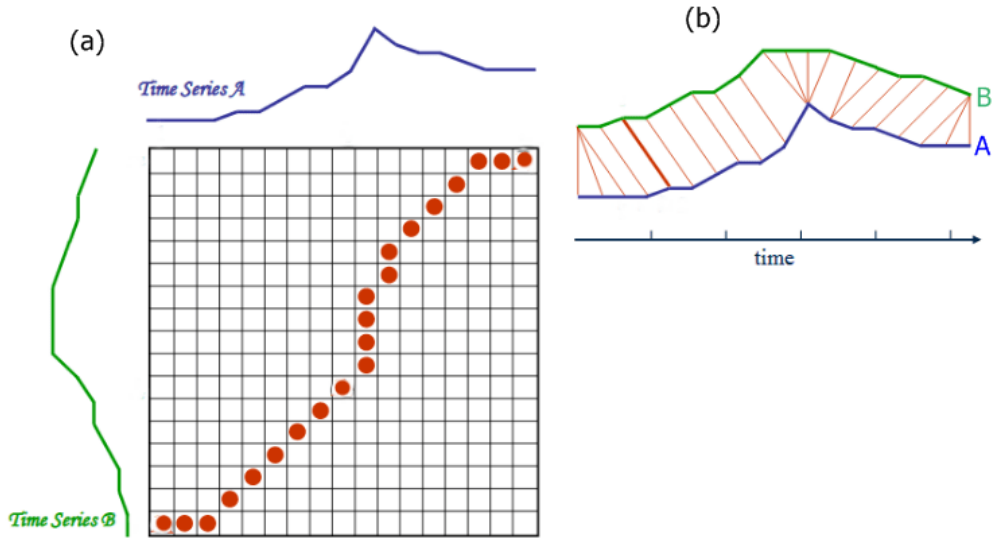




These similarity measures are based on dynamic programming

Time-elastic similarities: DTW

$$dtw(\boldsymbol{x}, \boldsymbol{y}) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} ||x_i -_{\mathcal{E}} y_i||^2}$$



Alternatives

- Sakoe-Chiba constraints: constraint on the "leash" length
- FastDTW: approximate DTW, linear time [SC07]
- soft-DWT: differentiable approximated DTW [CB17]
- soft-DTW divergence [BMV21]
- Piecewise Dynamic Time Warping
- ... Gromov-DTW, DTW with Global Invariances

DTW is not a distance

DTW is not a distance

- positive: $dtw(y, z) > 0$
- symmetric: $dtw(y, z) = dtw(z, y)$
- the triangular inegality is not satisfied:
  $dtw(x, y) + dtw(y, z) \not\leq dtw(x, z)$

DTW with KMeans?

- K-Means require a distance to ensure the algorithm convergence

$\rightarrow$ DBA algorithm [PKG11] is an alternative for KMeans with DTW

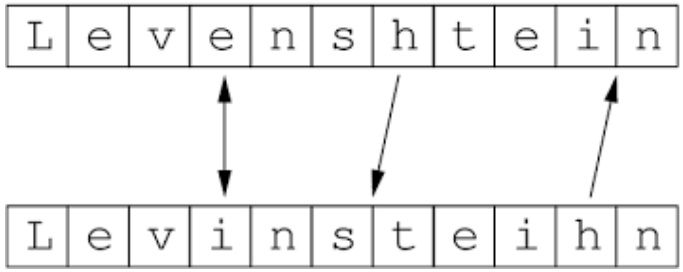
- Uses the Fréchet means

$$barycenter = \min_{\mu} \sum_{x \in \mathcal{D}} DTW(\mu, x)^2$$

- Time-consuming computation

DTW for symbolic sequences: Levenshtein distance and other edit distances

The Levenshtein distance between two strings a, b is given by:



$$lev(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ lev(tail(a), tail(b)) & \text{if } a[0] = b[0], \\ 1 + \min \begin{pmatrix} lev(tail(a), b) \\ lev(a, tail(b)) \\ lev(tail(a), tail(b)) \end{pmatrix} & \text{otherwise,} \end{cases}$$

Other edit distances

- different weights for letter replacement, insertion, deletion
- Smith & Waterman
- see string similarity metric for more ...

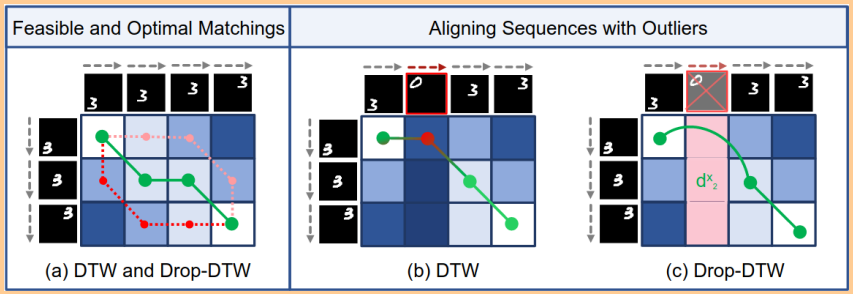
DTW for symbolic sequences

Symbolic version of DTW: Needleman-Wunsch algorithm

- Same principle: find the best mapping between two sequences, with three operations
  - *Match*: The two letters at the current index are the same
  - *Mismatch*: The two letters at the current index are different
  - *Indel* (Insertion or Deletion): One letter aligning to a gap in the other string
- Difference scoring system, the classical one is *Match* : $+1$, *Mismatch* : -1, *Indel* : -1

Drop-DTW [DHD+21]

- discard outliers from the computation



HierASTiSeq: métrique pour séquences temporelles [GPG24]

Métrique : Dynamic Time Warping

$$DTW(X, Y) = \min_{M \in \mathcal{M}} \langle M, C \rangle$$

où

- $M \in \{0,1\}^{K \times N}$, $M_{i,j} = 1$ si x_i associé à y_j, et 0 sinon
- $C_{i,j} = d_\Sigma(x_i, y_j)$: coût d'associer x_i à y_j

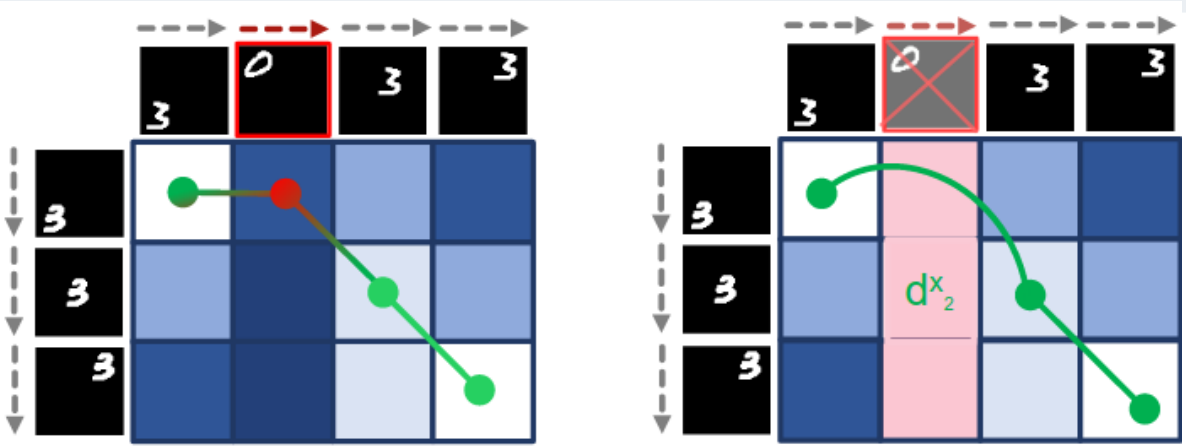
HierASTiSeq: métrique pour séquences temporelles [GPG24]

Métrique : Drop-Dynamic Time Warping [DHD+21]

$$DropDTW(X, Y) = \min_{M \in \mathcal{M}} \langle M, C \rangle + \delta.(P_c(M) + P_r(M))$$

où

- $M \in \{0,1\}^{K \times N}$, $M_{i,j} = 1$ si x_i associé à y_j, et 0 sinon
- $C_{i,j} = d_\Sigma(x_i, y_j)$: coût d'associer x_i à y_j
- $P_c(M)/P_r(M)$: nombre de colonnes/lignes vides de M ($M_{:,j} = 0$)
- δ : drop-cost



HierASTiSeq: métrique pour séquences temporelles [GPG24]

Métrique : Drop-Dynamic Time Warping [DHD+21]

$$DTW(X, Y) = \min_{M \in \mathcal{M}} \langle M, C \rangle$$

où

- $M \in \{0,1\}^{K \times N}$, $M_{i,j} = 1$ si x_i associé à y_j, et 0 sinon
- $C_{i,j} = d_\Sigma(x_i, y_j)$: coût d'associer x_i à y_j
- $P_c(M)/P_r(M)$: nombre de colonnes/lignes vides de M ($M_{:,j} = 0$)
- δ : drop-cost

Sémantique intéressante pour les parcours de soins

- Gère des séquences de différentes longueurs
- Flexibilité temporelle
- Considère les dates et les types d'évènements (à travers d_Σ)

$\rightarrow$ **ne permet pas de moyenner des séquences**

Subsequence based similarities

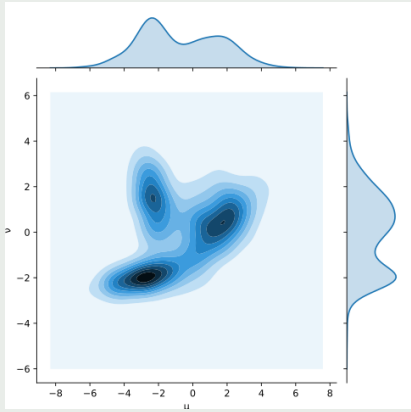
- The similarity presented above are "global" similarities
- In some cases (eg. long care pathways),

- ABC [SSK17]
- Bag-of-words [BMT+15]
  - → vector representation of sequence with the numbers of occurrences of a collection of
- Counting similar subsequences [ERC+15]
  - → theoretical results about the possibility to efficiently count the number of occurrences of subsequence in sequences of itemsets
- Genetic sequence alignments based on short reads ...

Wasserstein distance (intuitions)

Transportation plan

- $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$ and $\boldsymbol{y} = (y_1, \ldots, y_M) \in \mathbb{R}^M$ two vectors
- Cost of the transportation plan $d_\Gamma(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \Gamma(x_i, y_j) \boldsymbol{C}(x_i, y_j)$



where

- $\Gamma \in \mathbb{R}_+^{N \times M}$ is a transportation plan with constrained marginals $\Gamma \times 1_M = \boldsymbol{x}$ and $1_N \times \Gamma = \boldsymbol{y}$ (move all the mass)
- $\boldsymbol{C} : \mathbb{R}^N \times \mathbb{R}^M \mapsto \mathbb{R}_+$ is a cost function

Wasserstein distance

$\rightarrow$ there are infinite number of transportation plans Γ

- Wasserstein distance is the minimal cost of a transportation plan

$$d_{Wass}(\boldsymbol{X}, \boldsymbol{y}) = \min_{\Gamma} d_\Gamma(\boldsymbol{x}, \boldsymbol{y})$$

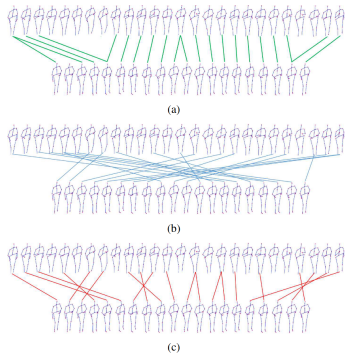
Wasserstein distance for sequences

Sinkhorn distance

- Wasserstein is too computational
- Sinkhorn distance: adaptation of Wasserstein but easy to compute

Order preserving Wasserstein distance [SH17]

- Wasserstein distance with additional cost on the alignement with "too far" elements of the sequence
- The alignments *does not preserve the order*!
- Efficient computation
- Top: DTW, middle: Wasserstein, bottom: OPW



Advanced clustering methods

Advanced clustering methods denotes techniques that are more than the adaptation of a classical clustering technique for a temporal dissimilarity.

Kernel based distances and clustering (1/2)

Kernel

- Define a "comparison function": $K : X \times X \mapsto \mathbb{R}$
- Constraint: symmetric, positive, definite function
- Distance induced by Kernels

$$d_K(x_1, x_2)^2 = K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2)$$

Kernel trick

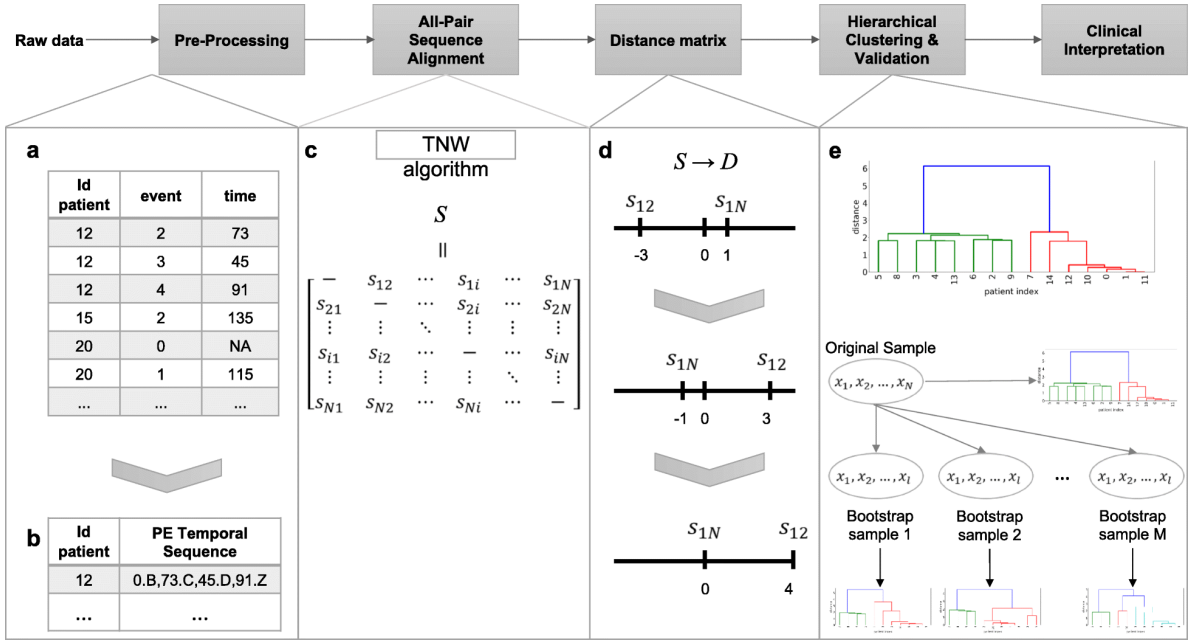
Any algorithm to process finite-dimensional vectors that can be expressed only in terms of pairwise inner products can be applied to potentially infinite-dimensional vectors in the feature space of a positive definite kernel by replacing each inner product evaluation by a kernel evaluation.

$\rightarrow$ Kernel K-Means ...

AliClu [RCCV19]

AliClu: algorithm for clustering longitudinal clinical data

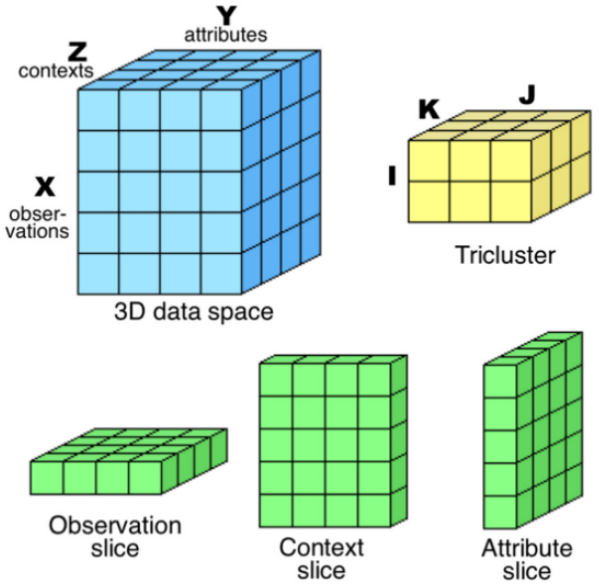
- Distance: Temporal Needelman-Wunch
- Agglomerative hierarchical clustering
  - Linkage criterion?
- Stepwise algorithm



Triclustering

Triclustering [HM18]

- Clustering technique for three-dimensional data (e.g. (*patient*, *event*, *time*))
- Identify *coherent* 3D sub-matrices in such large data structure
  - → subgroup of patients, sharing the same events at the same time instant



Different tasks / resolution techniques

- tasks: different "merit functions", different properties
- resolution techniques: greedy, stochastic, optimal, ...

→ TESS [GGBC15], SAXO [BBG13], SWoTTeD [HS23], ...

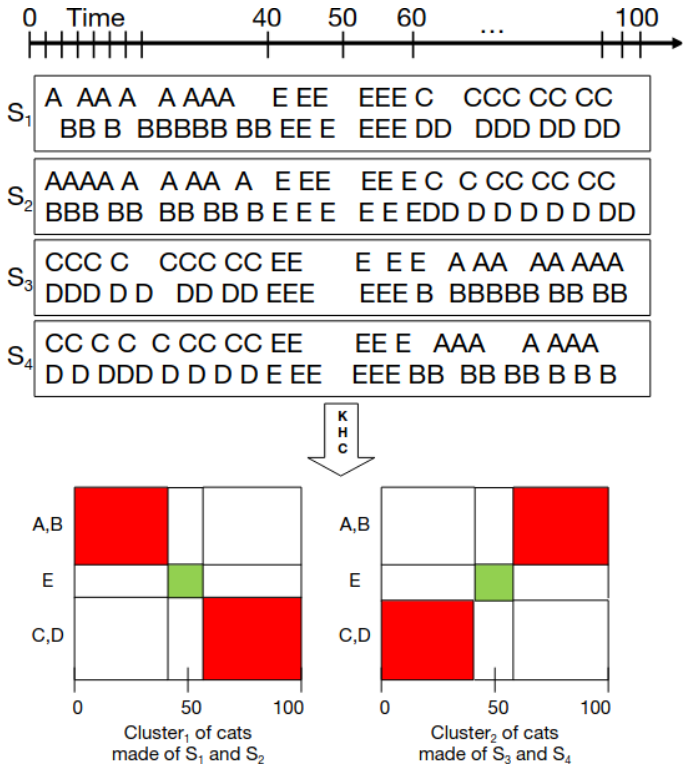
Triclustering: TESS [GGBC15]

TESS: Temporal event sequence summarization

- Grid model of the data
- Evaluation of a decomposition based on MDL principle

$$cost(M) = L(M) + L(D|M)$$

- Greedy algorithm to find the best grid
  1. start with the finest grained data grid
  2. evaluate all merges between clusters of sequence ids, clusters of events and adjacent time intervals
  3. perform the best merge (decreasing cost)
- Provide visualizations with Mutual Information and Contrast



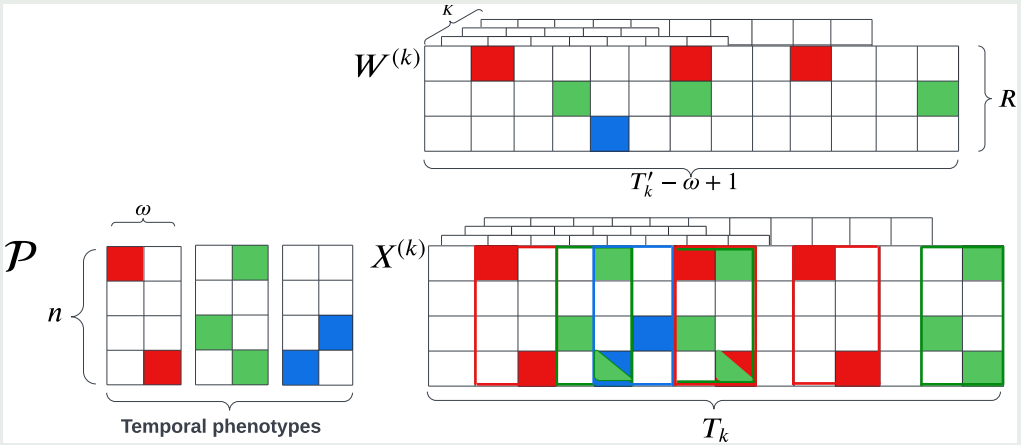
Triclustering: Tensor decomposition

Tensor decomposition

- Generic technique for the reduction of a 3D-tensor into 1D-tensors (generalisation of the SVD/ACP decomposition)
  - A practical use case: $\mathcal{X} = \mathcal{P} \otimes \mathcal{W} \otimes \mathcal{K}$ where
    - $\mathcal{P} \in \mathbb{R}^{R \times n}$ describes the R latent behaviors (phenotypes)
    - $\mathcal{W} \in \mathbb{R}^{R \times m}$ describes the R typical care pathways
    - $\mathcal{K} \in \mathbb{R}^{R \times m}$ describes the R typical **groups of patients**
- Several solutions for care-pathways: PARAFAC2, CNTF, LogPar, ...

SwoTTeD: an internal solution [HS23]

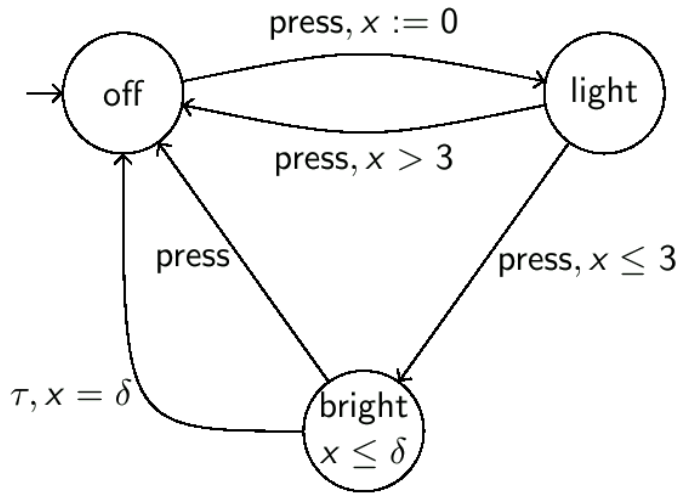
- based on the decomposition with *temporal phenotypes*



Formal language models

Formal languages

- Temporal logic: language to represent and manipulate
  - Linear Temporal Logic (LTL), Metric Temporal Logic (MTL), ...
  - Automata, Buchi automata, Petri Nets, Timed Automata, ...
- Advanced techniques for model inference



$$\varphi = \Diamond(B \wedge x_2.\Diamond(A \wedge x_2 \leq 4 \wedge x_1.\Diamond(C \wedge x_2 \leq 7 \wedge x_1 > 1 \wedge x_3.\Diamond(C \wedge x_3 \geq 4))))$$

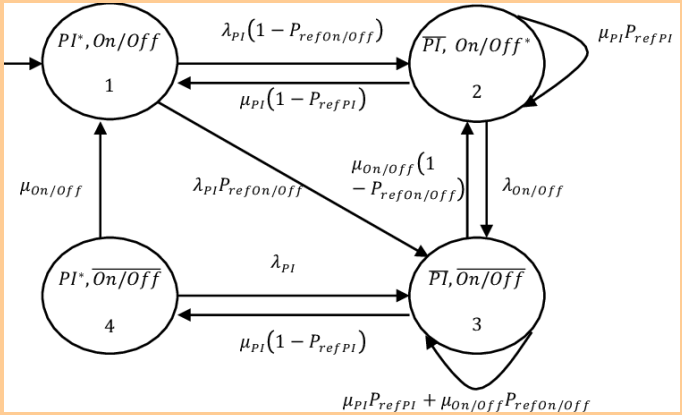
$\rightarrow$ rigid models for real data ...

Probabilistic graphical models

Probabilistic models

- (Hidden) Markov models
- Graphical Event Models [BSG18, YSB+20]

→ machine learning techniques for inference of these probabilistic models from timed sequences



Probabilistic formal models

- Probabilistic timed automata
- Probabilistic Metric Temporal Graphs [SMG22]

→ inference for these models??

Pattern mining based approaches

Pattern mining

- Algorithmic approach to identify interesting patterns occurring in a (large) set of structured objects
- Structured objects: itemsets, sequences, graphs
- Wide range of pattern languages (including chronicles)

$\Rightarrow$ a dataset of sequences is abstracted by its set of patterns

Summarizing Event Sequences with Serial Episodes [TV12]

- MDL based approach