HE
VA

Re-identifying cancer treatment lines in real-world data

an innovative sequence alignment algorithm

MYLORD

Martin PRODEL & Marie LAURENT

IA for Health, November 10th, 2020

Objective

Case study in hematology: Multiple Myeloma (MM)

HEVA

MM disease history

succession of remission and relapse, which form **treatment line**

How many patients under each line?

MM treatment

treatment line are treated by **protocols**

In which line is a protocol given?

MM care evolution

personalized medicine thanks to fast evolving **recommendations**

How does the patient management change year after year?

Automatically describe the treatment lines received by MM patients in France from the SNDS

MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

Why do we need AI?

To learn the matches between medical theory and Real-World Data, despite the variations

HEVA

Medical theory

VTD cycle

D1 – D4 – D8 – D11

D1

D1

Complex protocols

≠

Patient reality

One part of a patient's treatment sequence

D1 – D4 – D8 – D13

D1 - D13

D1 - D13

Deviations from theoretical protocols

≠

Real-World Data

SNDS: French claim databases

68 million people

Completeness of reimbursement

Unnamed injections

For drugs, only date of sale is available

MYLORD

Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

Why do we need AI?

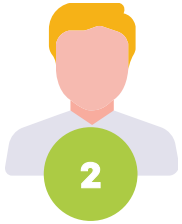
Because of the number of patients and of protocols

Identify theoretical protocols

Numerous protocols

41 cycles

VS.



Cohort selection (2014 - 2017)

Thousands of patients

17 442 patients

MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

HEVA

HEVA

ATLAS

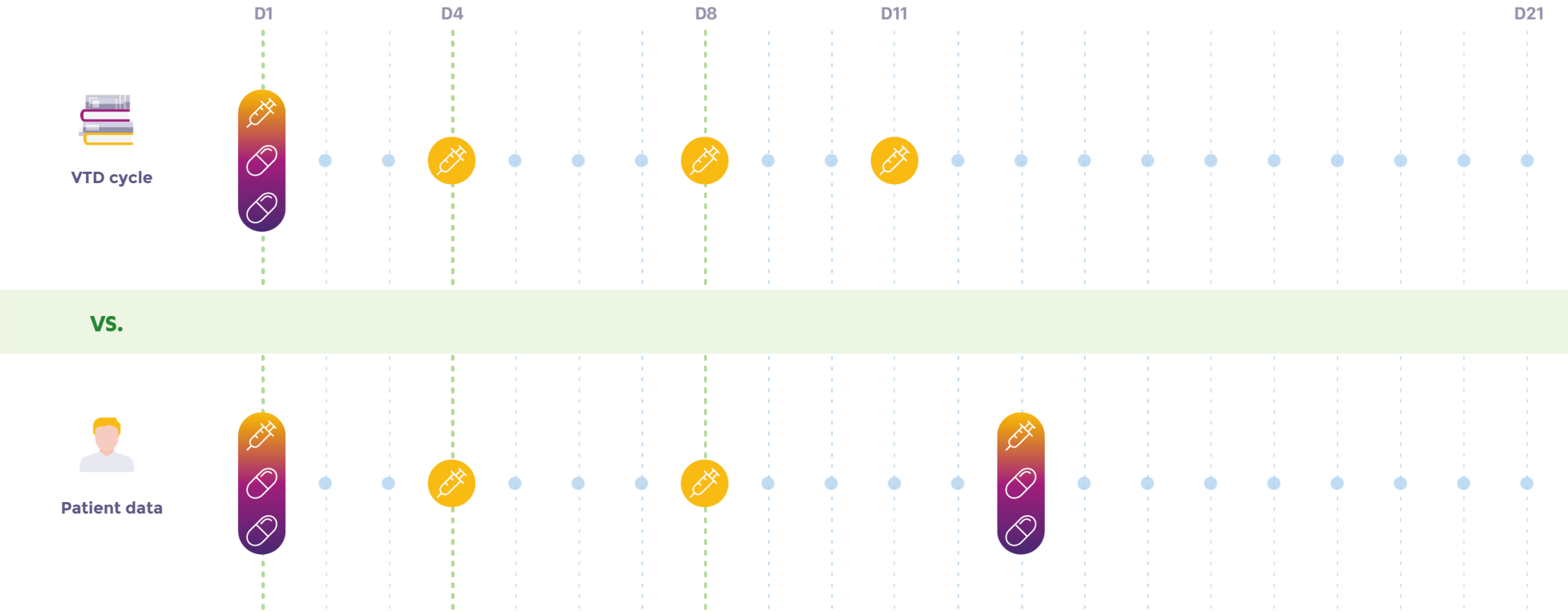
Analysis of Treatment Lines using Alignment of Sequences

By HEVA

1 cycle (1) vs. (2) 1 part of the patient's sequence

(3) (4)

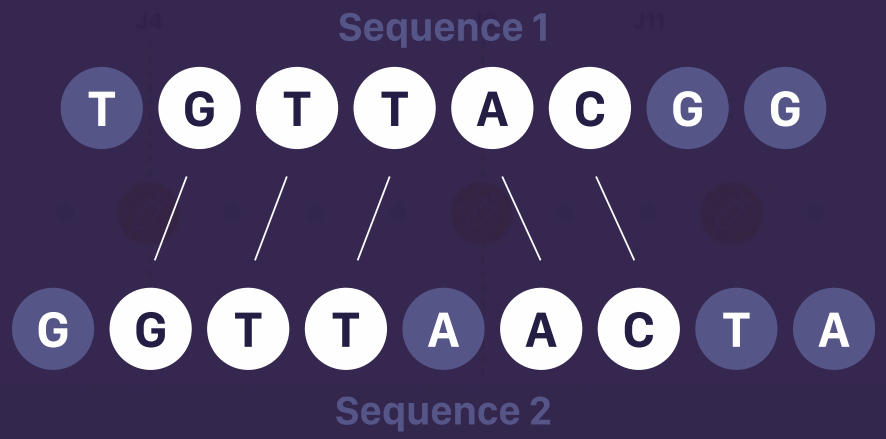
Alignment and score computation



Velcade + Thalidomide + Déxaméthasone Velcade

Method

Which sections are the most similar?

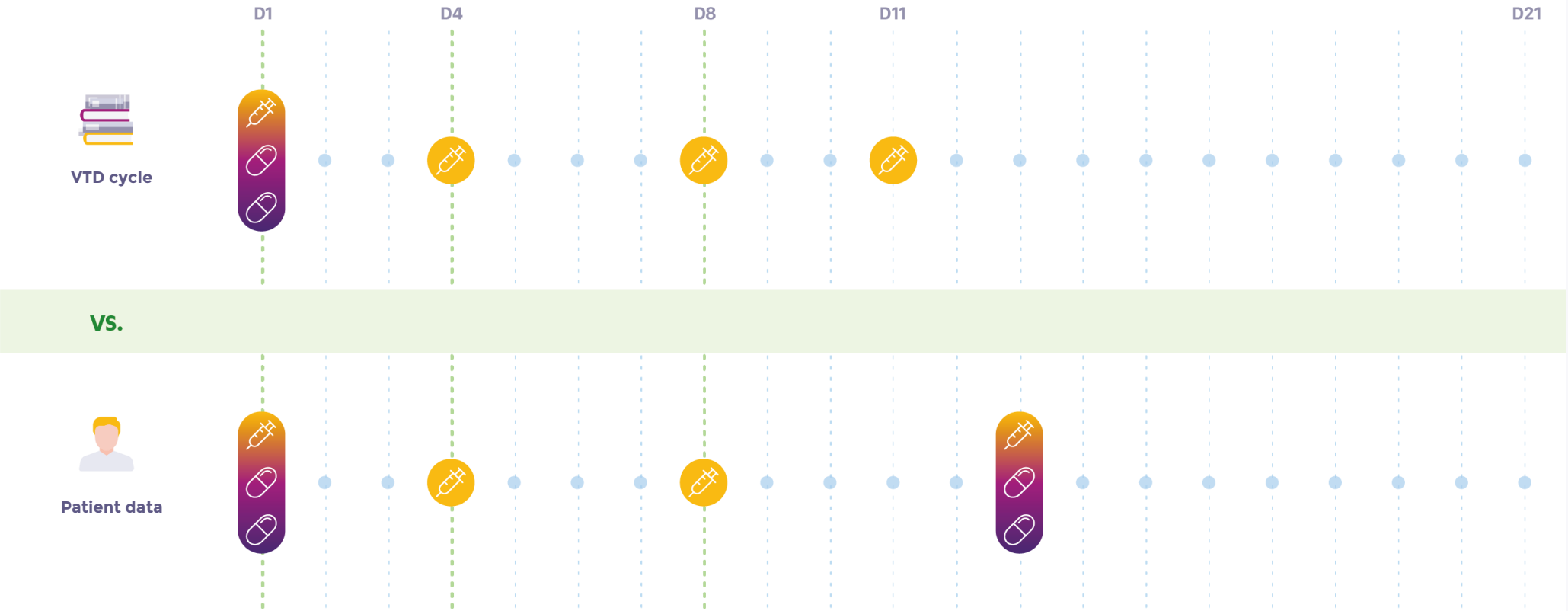


Adaptation of the Smith Waterman algorithm,
used for DNA sequences alignment

1 cycle (1) vs. (2) 1 part of the patient's sequence

(3) (4)

Alignment and score computation

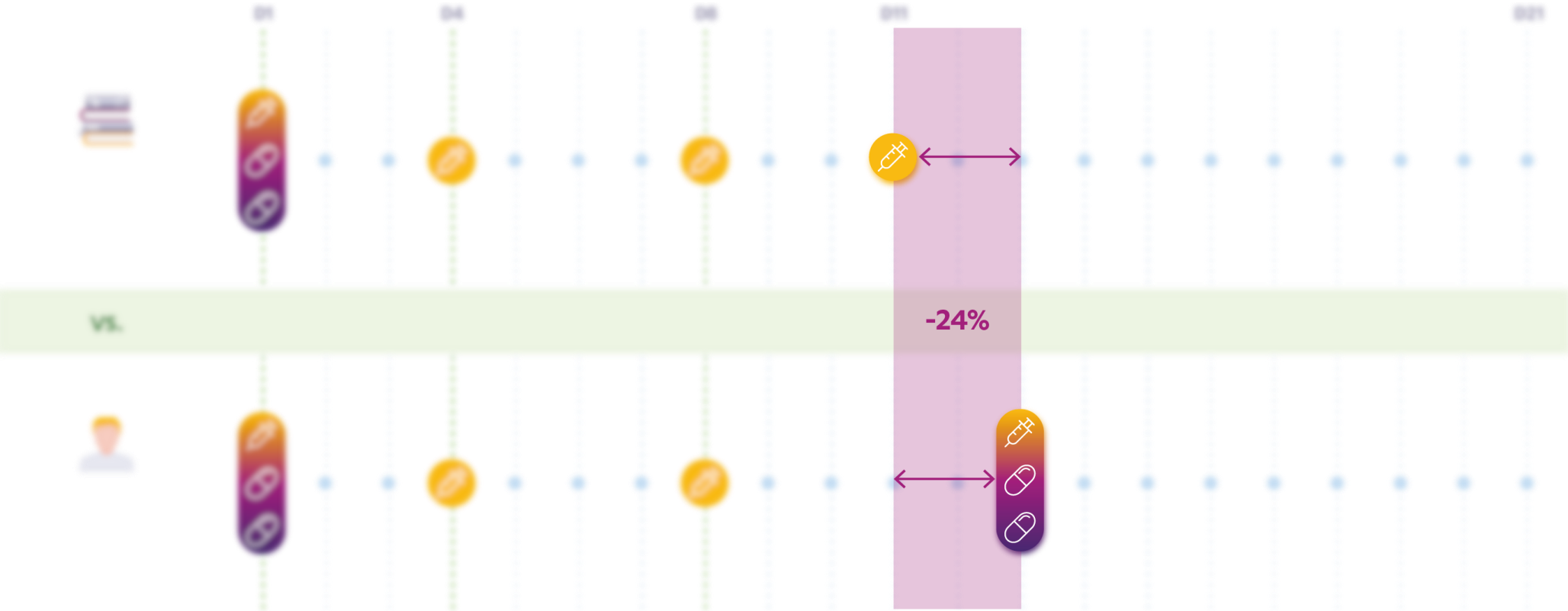


Velcade + Thalidomide + Déxaméthasone Velcade

1 cycle (1) vs. (2) 1 part of the patient's sequence

3 4

Alignment and score computation

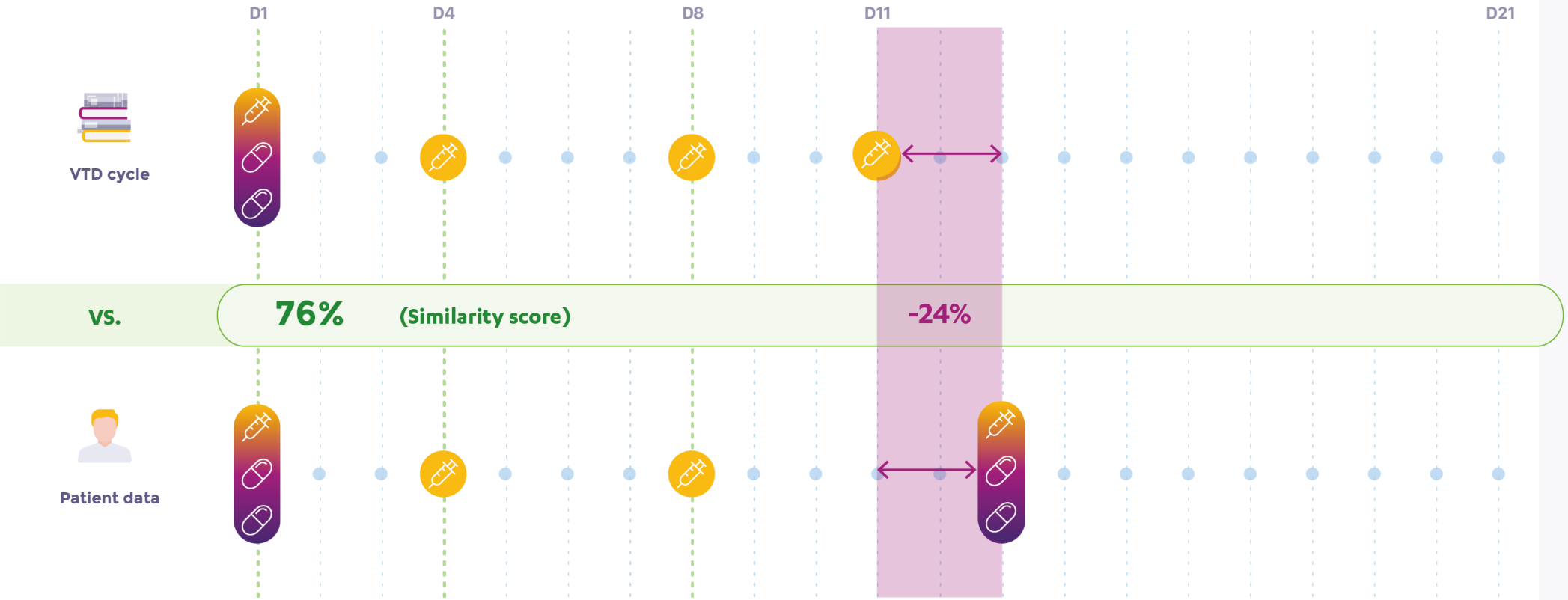


Velcade + Thalidomide + Déxaméthasone Velcade

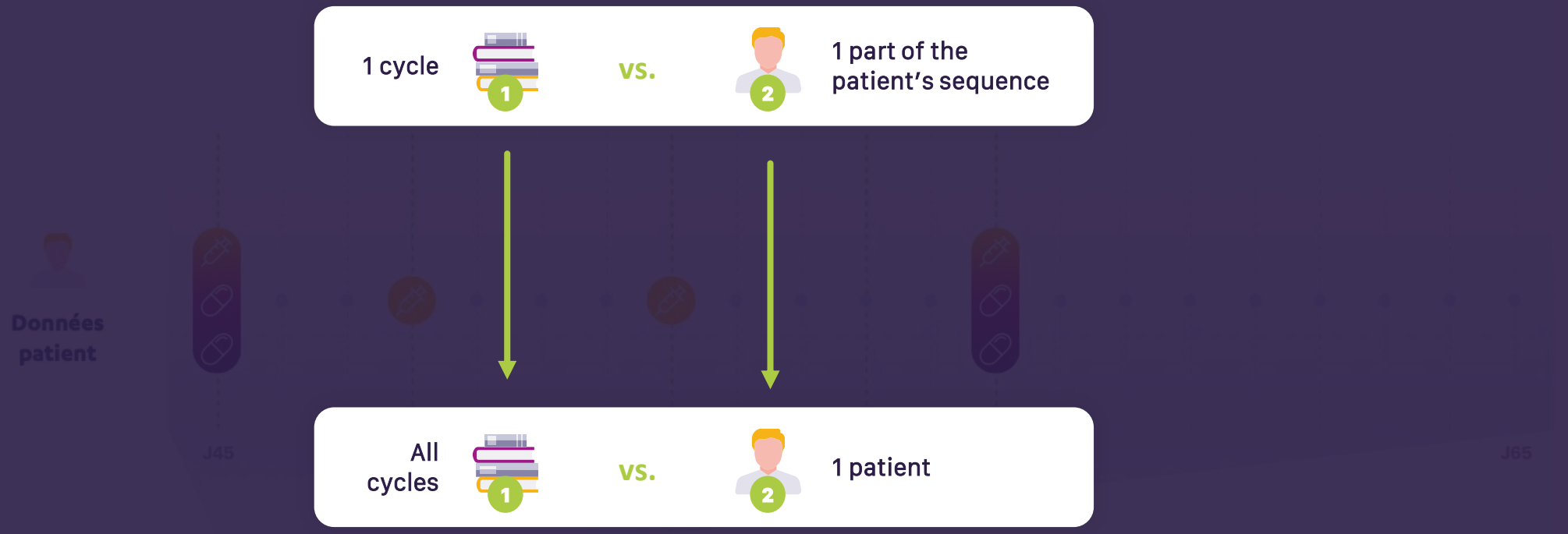
1 cycle (1) vs. (2) 1 part of the patient's sequence

(3) (4)

Alignment and score computation



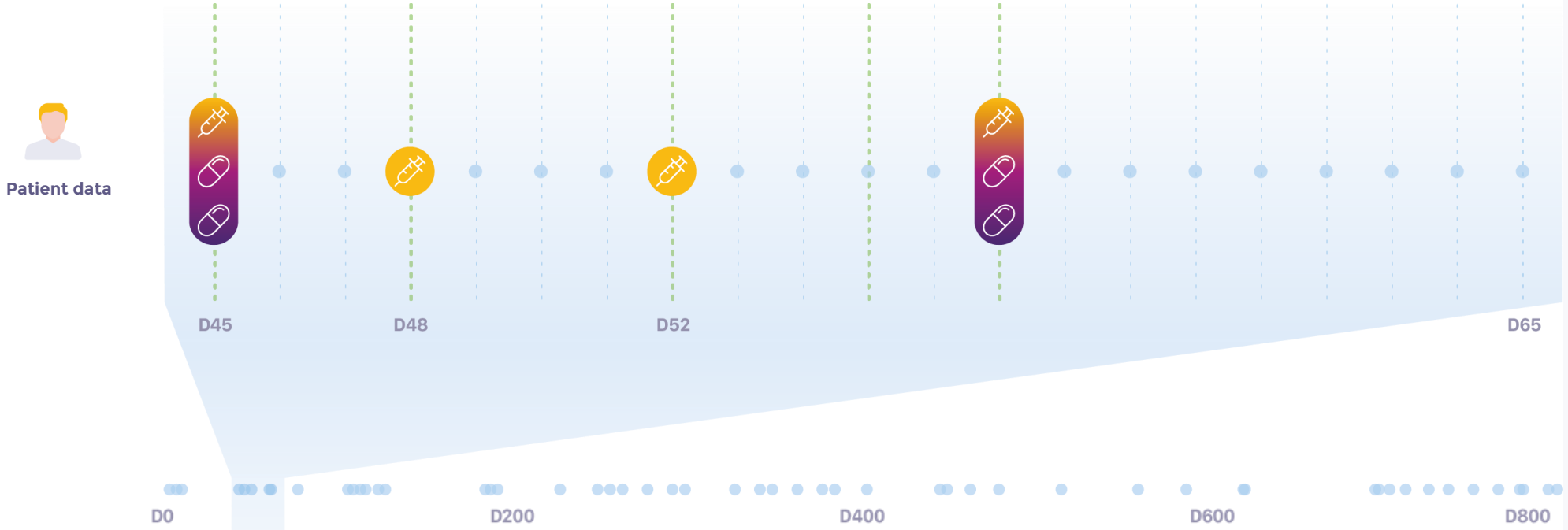
Velcade + Thalidomide + Déxaméthasone Velcade


1 cycle
1
vs.
2
1 part of the patient's sequence
All cycles
1
vs.
2
1 patient

All cycles 1 vs. 2 1 patient

HEVA

Scan the entire patient's history (800 days)



Analyze of the repetition of cycles over time
→ Reconstruction of complete treatment lines

MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

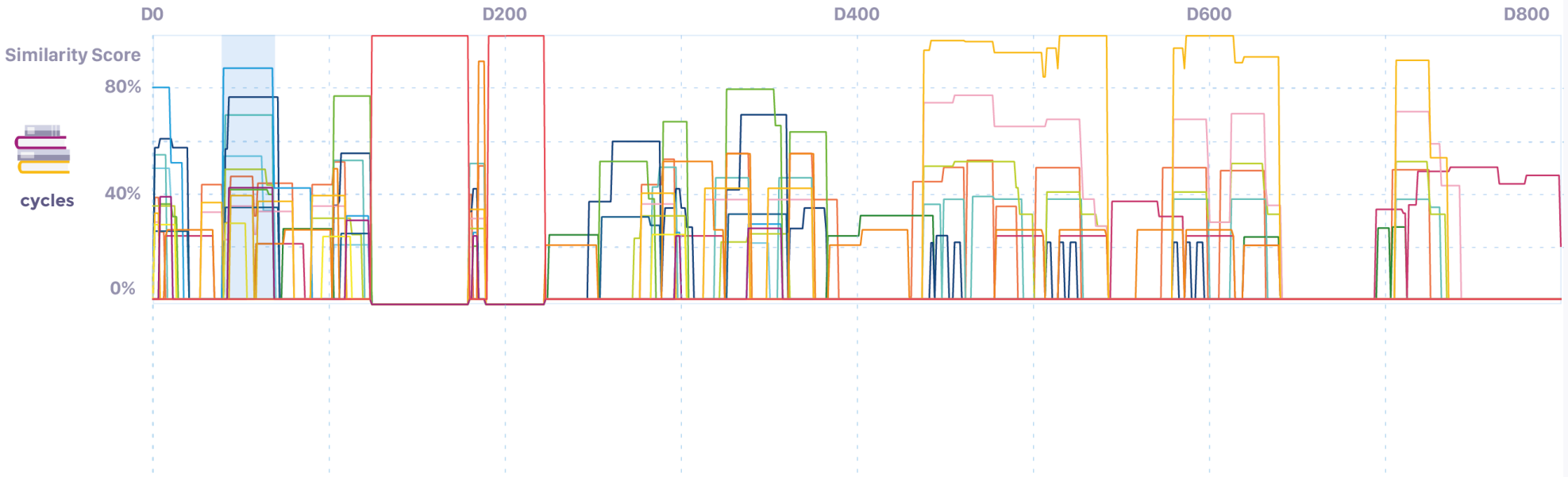
All cycles (1) vs. (2) 1 patient

HEVA

5

Display all the similarity scores between a patient and each of the theoretical cycles

- VTD
- Pom Dex
- Dara Dex
- Dara Vd
- VTD Dara
- Dara Rd
- Pom Dex Dara
- Pom Vel Dex
- MPV
- VRD
- BVD
- Rd
- MP
- V
- Pano Vd
- ASCT



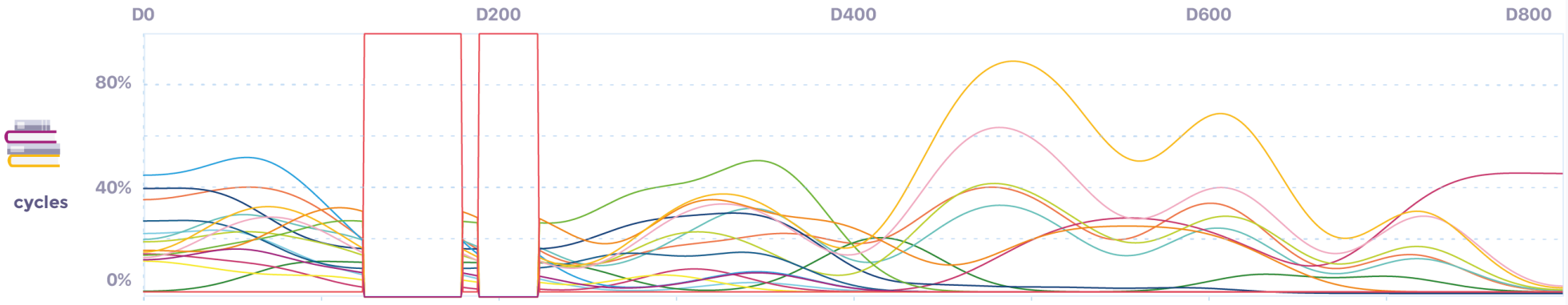
MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

HEVA

6

Smooth the scores over time

- VTD
- Pom Dex
- Dara Dex
- Dara Vd
- VTD Dara
- Dara Rd
- Pom Dex Dara
- Pom Vel Dex
- MPV
- VRD
- BVD
- Rd
- MP
- V
- Pano Vd
- ASCT

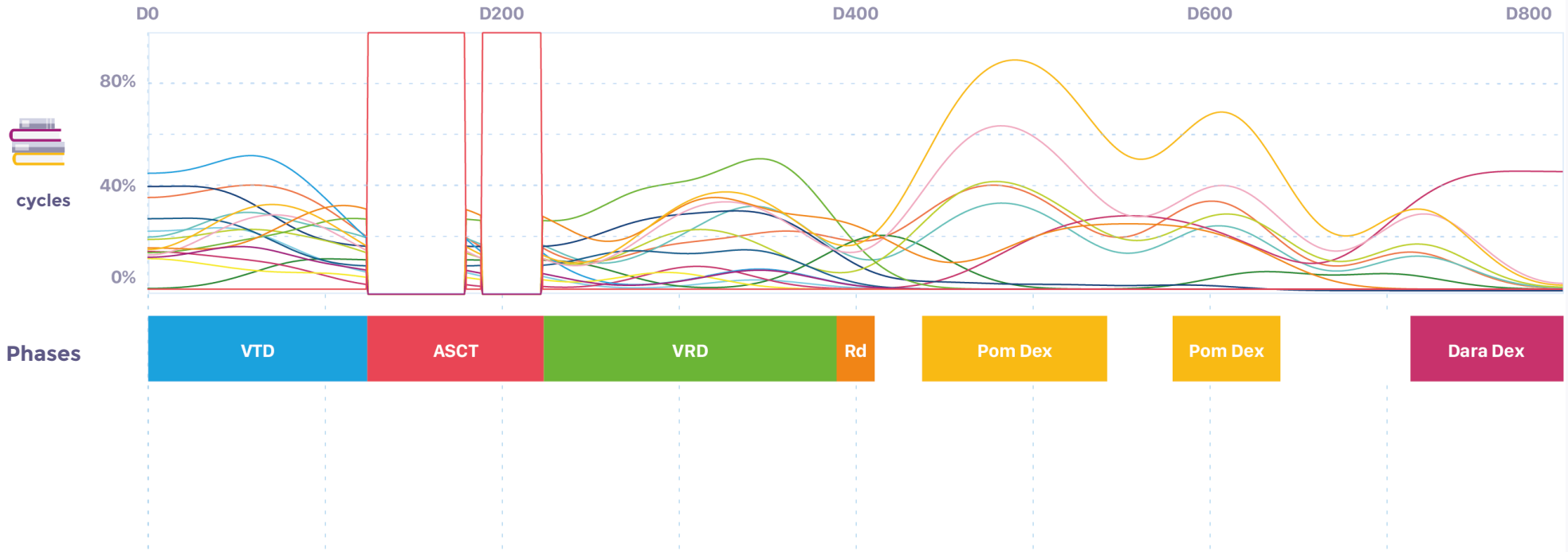


MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

6

Smooth the scores over time
and identify the best phase at each time of the follow-up

- VTD
- Pom Dex
- Dara Dex
- Dara Vd
- VTD Dara
- Dara Rd
- Pom Dex Dara
- Pom Vel Dex
- MPV
- VRD
- BVD
- Rd
- MP
- V
- Pano Vd
- ASCT



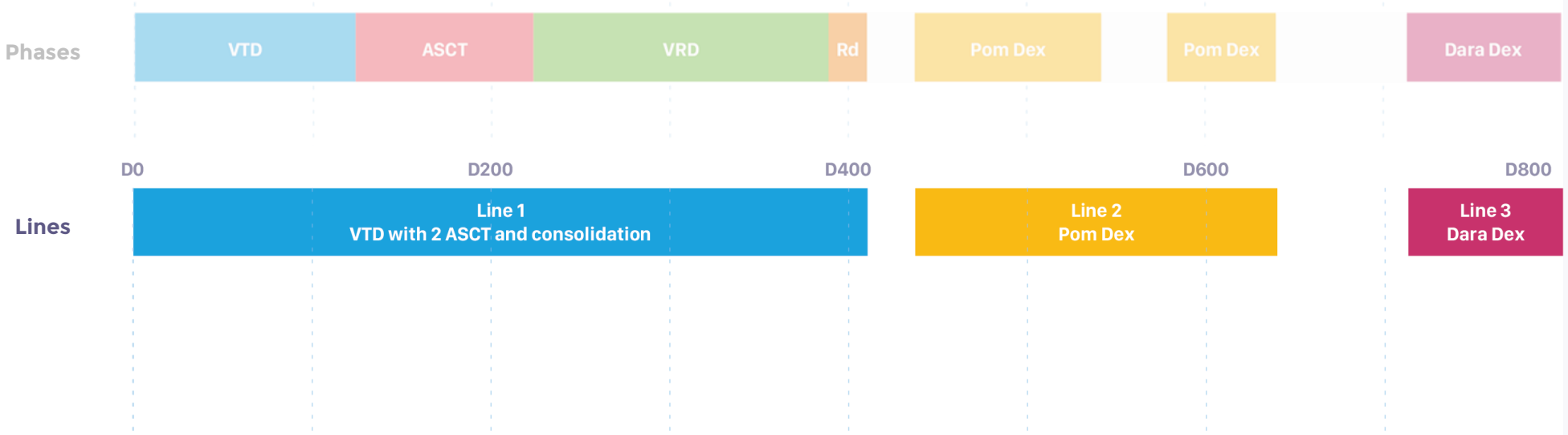
MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

HEVA

7

Gather phases into medical lines

- VTD
- Pom Dex
- Dara Dex
- Dara Vd
- VTD Dara
- Dara Rd
- Pom Dex Dara
- Pom Vel Dex
- MPV
- VRD
- BVD
- Rd
- MP
- V
- Pano Vd
- ASCT



MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

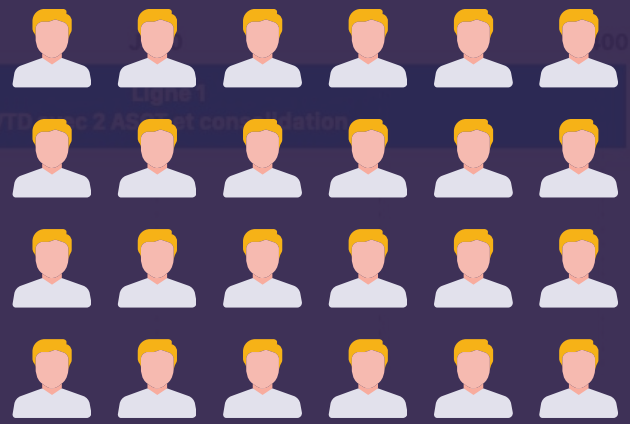
7

Résultats lignes pour le patient

Go from each individual patient result

to

smart data-visualizations of the entire cohort



Cohort-level results

Empower dataviz'

HEVA

Restitution

Tables

Histograms

Sunburst

Kaplan-Meier

Confrontation with the clinical knowledge

to improve the algorithm

Counts
of patients,
by lines,
by protocols,
by year

Duration
of lines,
of inter-lines

MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

Results

Re-identification power of the algorithm

ATLAS

1st execution

77 %
of the patients

Have all their lines identified

2 cycles identified a posteriori were added

Improvement of the parameterization

Improvement of medical understanding

ATLAS

2nd execution

92 %
of the patients

Have all their lines identified

MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

Advantages of the AI approach

HEVA

Flexibility and performant on RWD

Custom-made for each disease

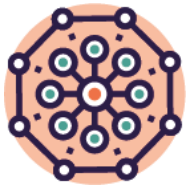
Confidence scores allow to work with several iterations, to obtain high-quality results

MYLORD
Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

AI in healthcare

HEVA

Therapeutic innovations arise from new treatment sequences



Pharmaceutical companies and public institutions face the challenge to find such sequences

In our experience, algorithms and data have a role to play

MYLORD

Multiple MYeloma: an epidemioLOgical study using SNIIRAM Database

HE
VA

Thank you for your attention

MYLORD

Scientific article on the ATLAS algorithm

ATLAS: a Robust Algorithm for Temporal Sequence Alignment of Treatment Lines using Claim Databases

Martin Prodel
dept. of Data Analytics
HEVA
Lyon, France
mprodel@hevaweb.com

Ludovic Lamarsalle
dept. of Data Analytics
HEVA
Lyon, France
llamarsalle@hevaweb.com

Vincent Augusto
Mines Saint-Étienne, Univ. Clermont Auvergne
CNRS, UMR 6158 LIMOS
Center for Healthcare Engineering
F-42023 Saint-Étienne, France
augusto@emse.fr

***Abstract*—Comparison of cancer treatment protocols against patient history is important to assess the efficiency of chemotherapy protocols. However, manual identification of protocols is not possible when considering a large population of patients. This paper proposes a new method called ATLAS (Analysis of Treatment Lines using Alignment of Sequences) to tackle the problem related to protocol identification using claim databases as data input. The proposed algorithm is an extension of the Smith-Waterman algorithm and allows to compare any patient's medical history against a list of theoretical protocols taking into account temporal information in sequences as well as missing data. Numerical experiments show that the proposed method could identify the right protocol in over 95% of the cases for realistically noised sequences (15,000 generated patients aligned with 15 protocols). The method is meant to be used as a decision aid tool for practitioners.**

***Index Terms*—sequence alignment, temporal Smith-Waterman, claim data, healthcare, oncology, treatment lines**

I. Introduction

A. Context

Chemotherapy is a type of cancer treatment that uses one or more anti-cancer drugs as part of a standardized chemotherapy regimen. Chemotherapy may be given with a curative intent, or it may aim to prolong life or to reduce symptoms (palliative chemotherapy). A *regimen* is a systematic plan of treatment administration designed to improve the health of the patient. The regimens recommended by the national guidelines are called *protocols*. These protocols are evaluated through clinical trials, but also a posteriori using historical patient data.

To assess the efficiency of chemotherapy regimens, epidemiologists need to identify the actual treatments received by a cohort of patients in large databases [1]–[4]. It helps improve existing protocols and thus the quality of care. Protocol identification is possible thanks to the increasing availability of medical and claim data in recent years. Still, such identification is tedious because drug protocols are complex and numerous. Scientific challenges related to protocol identification are: co-occurrence of multi-drug treatments, data quality, temporal nature of the problem.

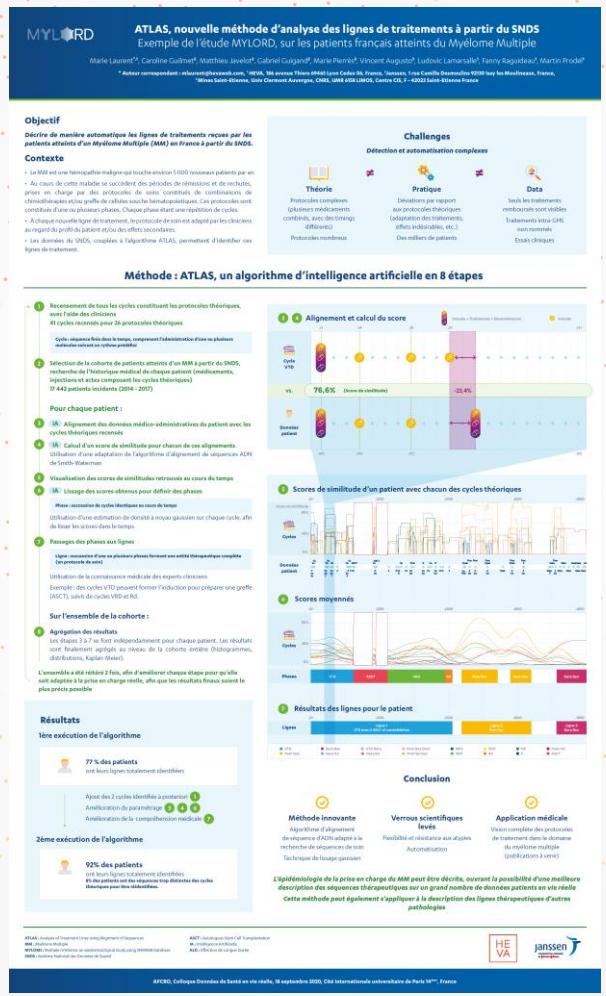
978-1-7281-1462-0/19/$31.00 ©2019 IEEE

Claim databases, Electronic Health Records (EHR) or Electronic Health Data (EHD), contain the information found on medical bills or claims, and are collected by insurers (public or private). Although claim databases are collected for reimbursement purposes, they contain relevant information on patients' medical condition and on delivered care. The increasing availability of claim data seems promising to lead epidemiological studies on how well treatment protocols are followed [5]. However, using data to identify which protocol a patient has followed is not straightforward. First, the exact administration date of some drugs might be unknown, only the delivery date in a pharmacy is known. The correct matching of such events to their theoretical administrations is uncertain. The second hindrance is missing data. Databases sporadically lack the information about a drug given to a patient, even if a chemotherapy session is registered. Those omitted administration blur the identification of the protocol, for which all the drugs are supposedly given at precise dates.

Existing works on protocol identification from claim databases mainly rely on sequence alignment algorithms. Those algorithms were made popular and widespread in the field of bioinformatics [6]. Sequence alignment is used to determine similar regions between two strings of DNA also named *sequences*. When aligning two sequences, one is the sequence in which we search similarities and the other is the "reference" sequence. Two classical and proven algorithms for sequence alignments are the Needleman-Wunch [7] and Smith-Waterman [8] algorithms. Both are based on dynamic programming and use a scoring matrix. Each value of this matrix represents the similarity between two elements, one from the sequence to align and the other from the reference sequence. The higher the score, the more the elements are similar.

These alignment methods are meant to identify shared patterns in two sequences, but they do not consider time between elements of a sequence, nor the uncertain position of such elements (e.g. uncertain administration date of drugs).

Poster (in French) on the ATLAS algorithm



MYLORD

ATLAS, nouvelle méthode d'analyse des lignes de traitements à partir du SNDS
Exemple de l'étude MYLORD, sur les patients français atteints du Myélome Multiple

Marie Laurent[1,*], Caroline Guillmet[2], Matthieu Javelot[2], Gabriel Guigan[2], Marie Pierrée[2], Vincent Augusto[3], Ludovic Lamarsalle[1], Fanny Raguideau[1], Martin Prodel[1]

[*] Auteur correspondant : mlaurent@hevaweb.com, [1]HEVA, 9B avenue Thiers 69461 Lyon Cedex 06, France, [2]Janssen, 1 rue Camille Desmoulins 92130 Issy les Moulineaux, France, [3]Mines Saint-Étienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Centre CIS, F – 42023 Saint-Étienne France

Objectif

Décrire de manière automatique les lignes de traitements reçues par les patients atteints d'un Myélome Multiple (MM) en France à partir du SNDS.

Contexte

- Le MM est une hémopathie maligne qui touche environ 5 000 nouveaux patients par an.
- Au cours de cette maladie se succèdent des périodes de rémissions et de rechutes, prises en charge par des protocoles de soins constitués de combinaisons de chimiothérapies et/ou greffe de cellules souche hématopoïétiques. Ces protocoles sont constitués d'une ou plusieurs phases. Chaque phase étant une répétition de cycles.
- À chaque nouvelle ligne de traitement, le protocole de soin est adapté par les cliniciens au regard du profil du patient et/ou des effets secondaires.
- Les données du SNDS, couplées à l'algorithme ATLAS, permettent d'identifier ces lignes de traitement.

Challenges

Détection et automatisation complexes

- **Théorie** – Protocoles complexes (plusieurs médicaments combinés, avec des timings différents) – Protocoles nombreux
- **Pratique** – Déviations par rapport aux protocoles théoriques (adaptation des traitements, effets indésirables, etc.) – Des milliers de patients
- **Data** – Seuls les traitements remboursés sont visibles – Traitements intra-GHS non nommés – Essais cliniques

Méthode : ATLAS, un algorithme d'intelligence artificielle en 8 étapes

1. Recensement de tous les cycles constituant les protocoles théoriques, avec l'aide des cliniciens
   41 cycles recensés pour 26 protocoles théoriques
   Cycle : séquence fixe dans le temps, composant l'administration d'une ou plusieurs molécules suivant un rythme prédéfini
2. Sélection de la cohorte de patients atteints d'un MM à partir du SNDS, recherche de l'historique médical de chaque patient (médicaments, injections et actes composant les cycles théoriques)
   17 442 patients incidents (2014 - 2017)

Pour chaque patient :

3. IA : Alignement des données médico-administratives du patient avec les cycles théoriques recensés
4. IA : Calcul d'un score de similitude pour chacun de ces alignements
   Utilisation d'une adaptation de l'algorithme d'alignement de séquences ADN de Smith-Waterman
5. Visualisation des scores de similitudes retrouvés au cours du temps
6. IA : Lissage des scores obtenus pour définir des phases
   Phase : succession de cycles identiques au cours de temps
   Utilisation d'une estimation de densité à noyau gaussien sur chaque cycle, afin de lisser les scores dans le temps
7. Passages des phases aux lignes
   Ligne : succession d'une ou plusieurs phases formant une entité thérapeutique complète (un protocole de soin)
   Utilisation de la connaissance médicale des experts cliniciens
   Exemple : des cycles VTD peuvent former l'induction pour préparer une greffe (ASCT), suivis de cycles VRD et Rd.

Sur l'ensemble de la cohorte :

8. Agrégation des résultats
   Les étapes 3 à 7 se font indépendamment pour chaque patient. Les résultats sont finalement agrégés au niveau de la cohorte entière (histogrammes, distributions, Kaplan-Meier).

L'ensemble a été réitéré 2 fois, afin d'améliorer chaque étape pour qu'elle soit adaptée à la prise en charge réelle, afin que les résultats finaux soient le plus précis possible

Résultats

1ère exécution de l'algorithme

77 % des patients ont leurs lignes totalement identifiées

Ajout des 2 cycles identifiés à posteriori
Amélioration du paramétrage
Amélioration de la compréhension médicale

2ème exécution de l'algorithme

92% des patients ont leurs lignes totalement identifiées
8% des patients ont des séquences trop distinctes des cycles théoriques pour être réidentifiés.

Alignement et calcul du score

Cycle VTD – Données patient – 76,6% (Score de similitude) vs. -23,4%

Scores de similitude d'un patient avec chacun des cycles théoriques

Cycles – Données patient

Scores moyennés

Cycles – Phases

Résultats des lignes pour le patient

Lignes

Conclusion

- **Méthode innovante** – Algorithme d'alignement de séquence d'ADN adapté à la recherche de séquences de soin – Technique de lissage gaussien
- **Verrous scientifiques levés** – Flexibilité et résistance aux atypies – Automatisation
- **Application médicale** – Vision complète des protocoles de traitement dans le domaine du myélome multiple (publications à venir)

L'épidémiologie de la prise en charge du MM peut être décrite, ouvrant la possibilité d'une meilleure description des séquences thérapeutiques sur un grand nombre de données patients en vie réelle

Cette méthode peut également s'appliquer à la description des lignes thérapeutiques d'autres pathologies

ATLAS : Analysis of Treatment Lines using Alignment of Sequences
MM : Myélome Multiple
MYLORD : Analyse rétrospective sur epidemiological study using MMRRR Database
SNDS : Système National des Données de Santé
ASCT : Autologous Stem Cell Transplantation
IA : Intelligence Artificielle
AVC : Affection du système vasculaire

HEVA – janssen

AFCRO, Colloque Données de Santé en vie réelle, 18 septembre 2020, Cité internationale universitaire de Paris 14ème, France

HEVA

Our website

Our videos