# Reconstruction of patient medical history from medico-administrative databases

Antoine Poirot-Bourdain

CEREMADE : Université Paris Dauphine PSL, CNRS
HeKA : INRIA, INSERM, Université Paris Cité

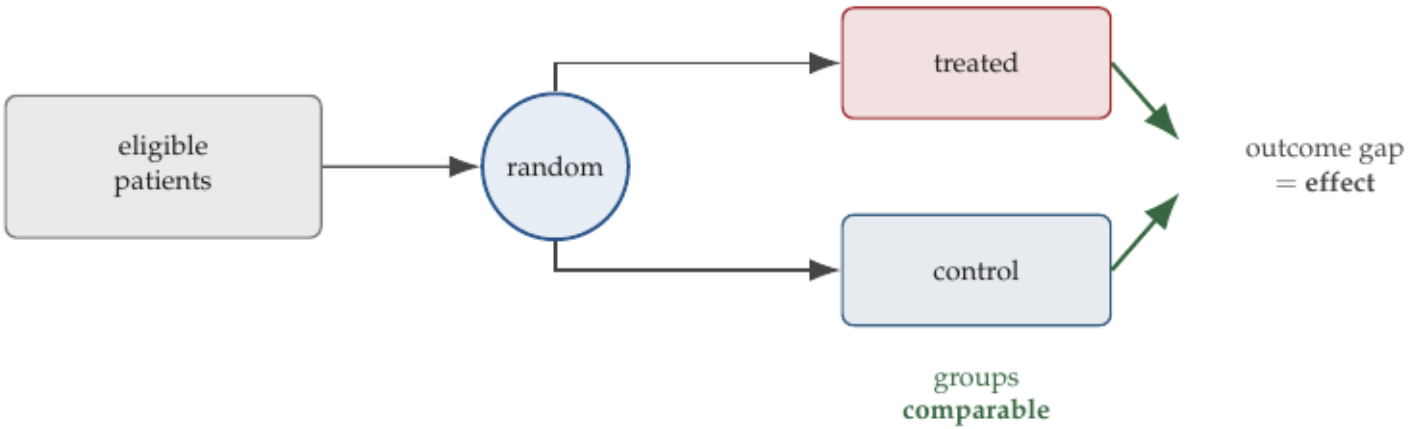
10/06/2026

# Why use medico-administrative databases ?

# Clinical trials: the gold standard

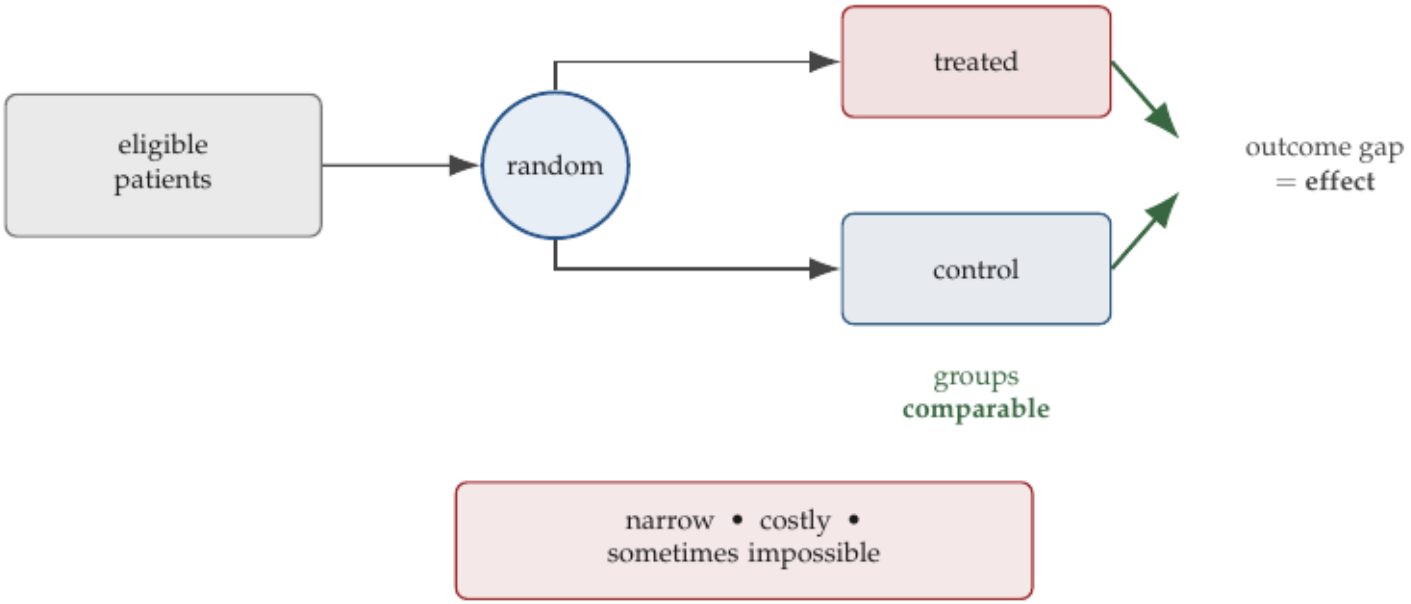

*Randomisation is what guarantees an unbiased effect.*

# Clinical trials: the gold standard



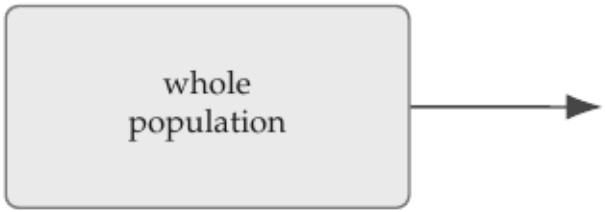

*Randomisation is what guarantees an unbiased effect.*

# Clinical trials: the gold standard



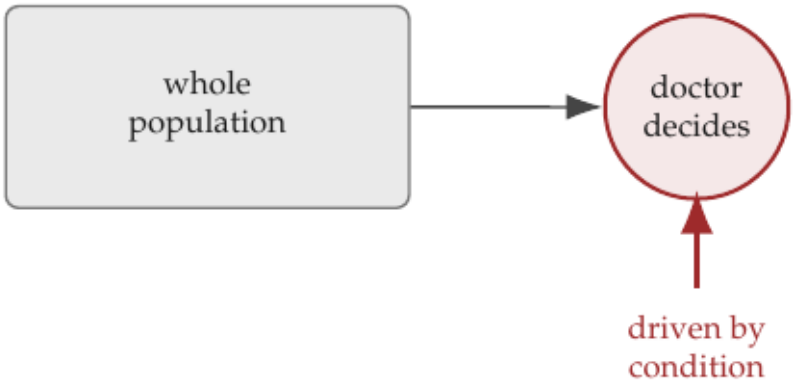

*Randomisation is what guarantees an unbiased effect.*

# Real wold data



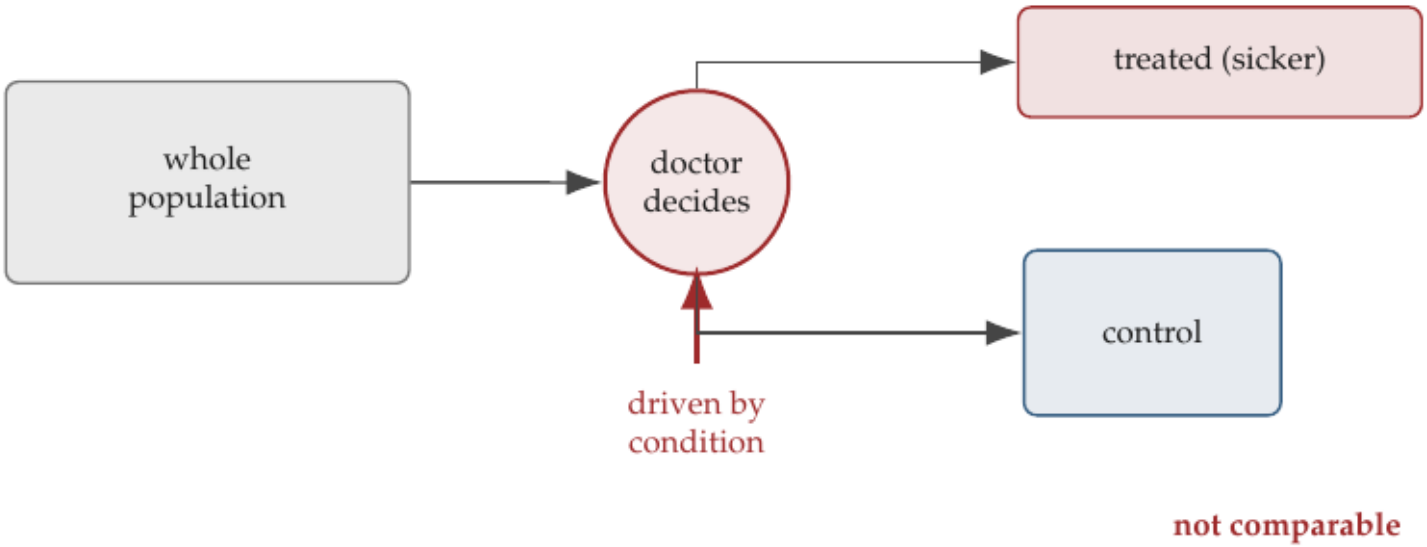

*The link condition → treatment is exactly what randomisation removes.*

# Real world data




*The link condition → treatment is exactly what randomisation removes.*

# Real wolrd data




*The link condition → treatment is exactly what randomisation removes.*

# Clincal trials vs real world data

### Randomised trial

+ unbiased *by design*
+ the causal gold standard
− small, selected, short
− not always feasible

### Real-world data

+ large, representative
+ long follow-up, always there
− groups not comparable
− confounded

# Adverse drug reactions

**During clinical trials**

- Only short term adverse drug reactions
- On a small subset of patients
- Without other drugs

**Pharmacovigilance**

- Adverse drug reactions are detected by the general practitioner.
- They are poorly detected.
- They are responsible for 8% of all hospitalizations.

# The SNDS

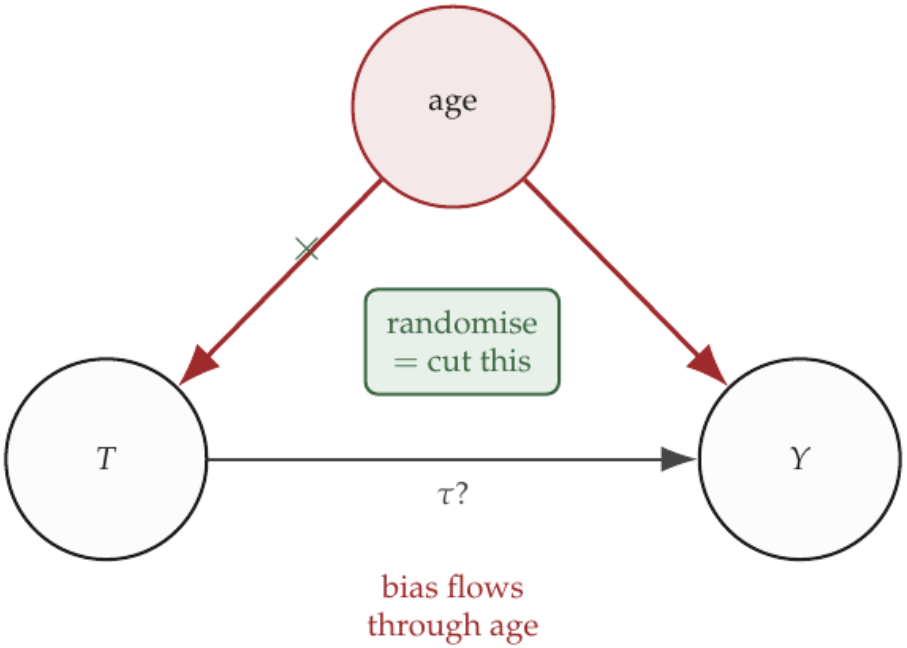
**A massive medicaldatabase**

- The medical history of 67 millions of patients
- For more than ten years
- Drug reimbursement and hospital stays

**A claim database**

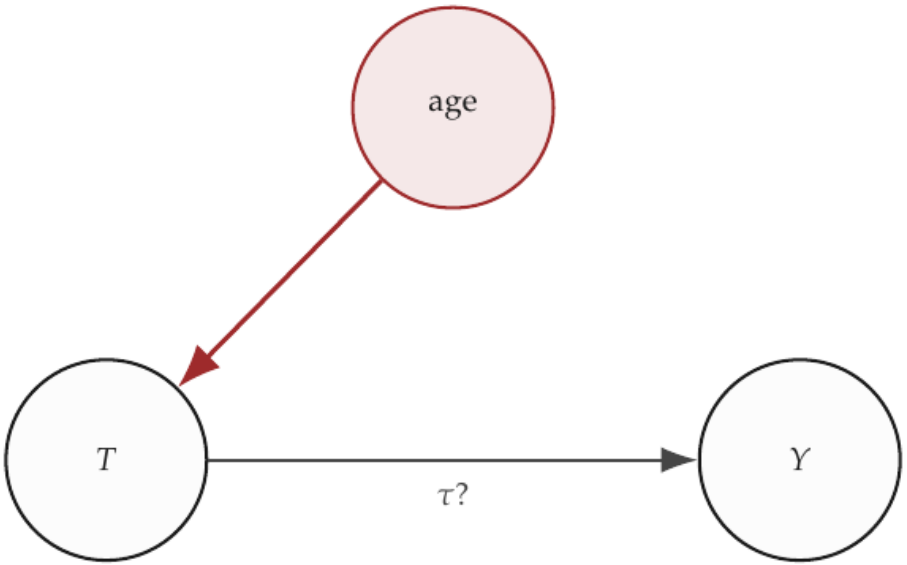
- The data was originally for administrative purpuse
- It has built-in biais
- It has blind spots

# The Causality problem



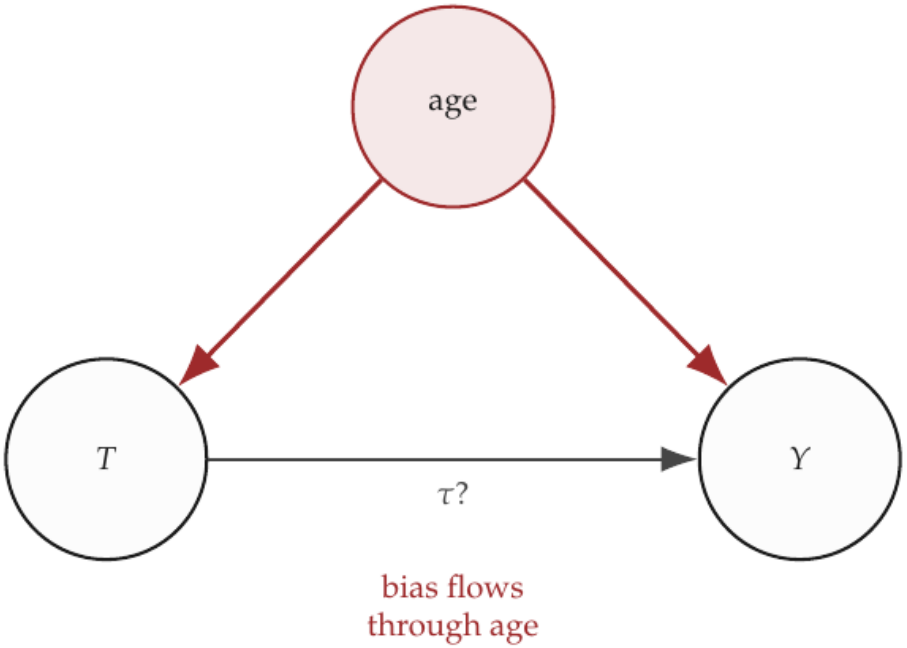

*If older patients get the drug and also have worse outcome due to their age, the drug can look harmful even when it helps.*

# The causality problem



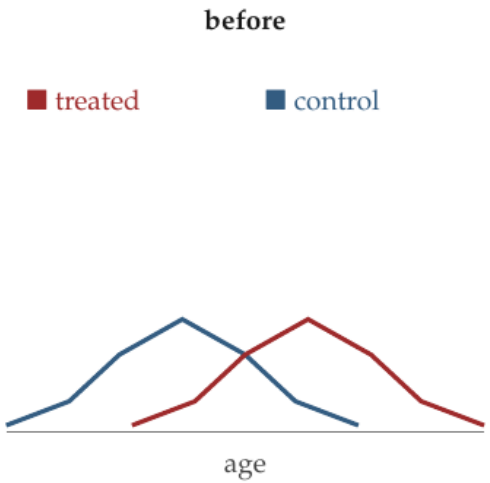

*If older patients get the drug and also have worse outcome due to their age, the drug can look harmful even when it helps.*

# The Causality problem



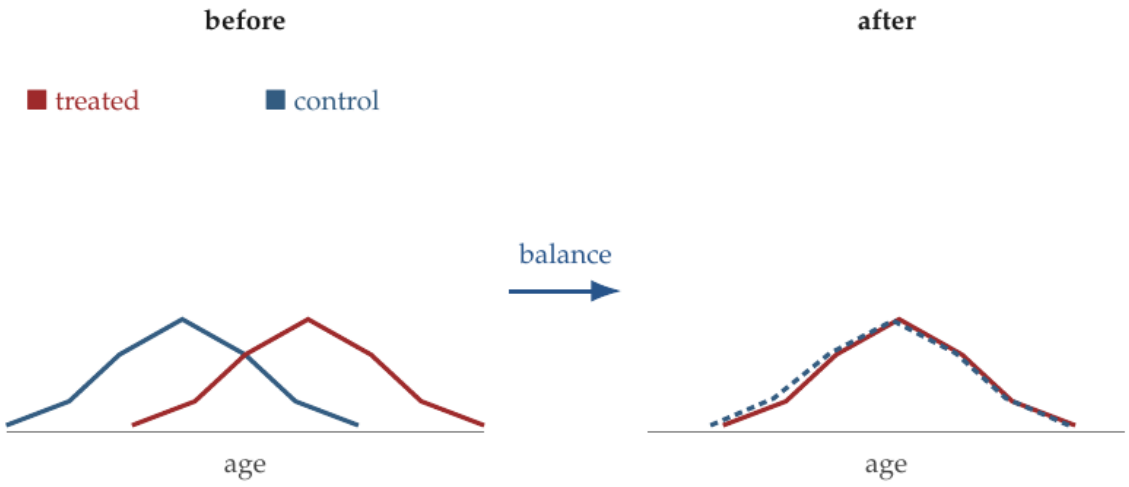

*If older patients get the drug and also have worse outcome due to their age, the drug can look harmful even when it helps.*

# Removing biais: balance the covaraites distributions



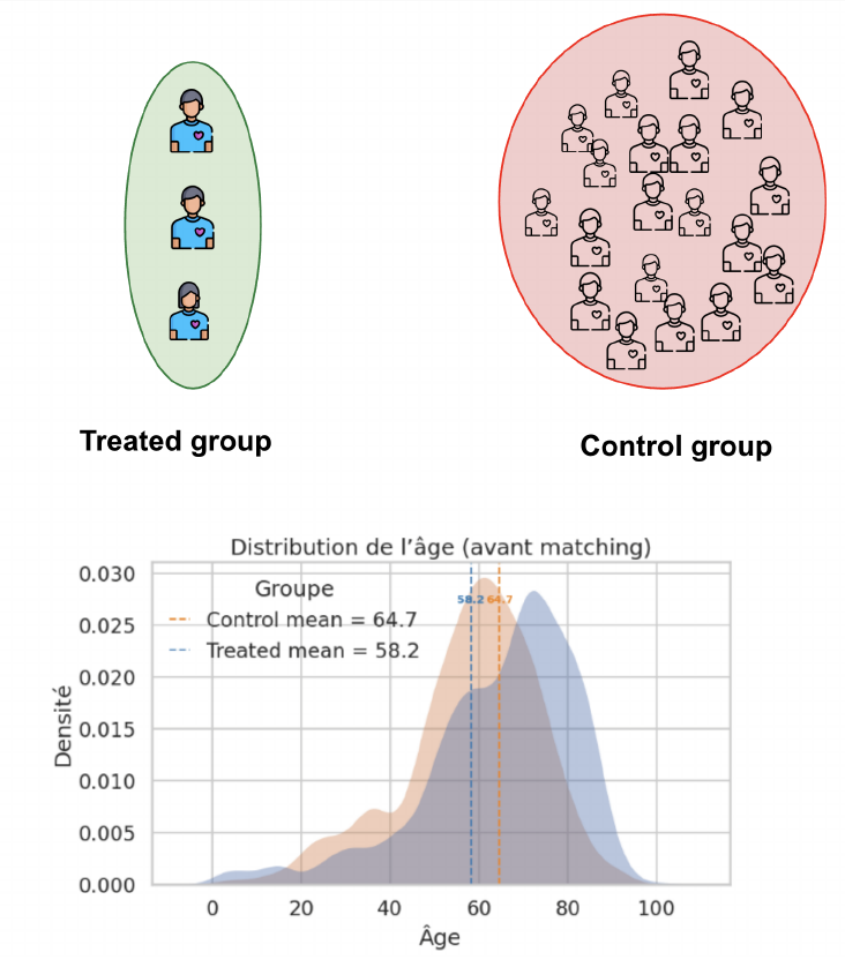

*The control age-density is reshaped until it matches the treated one.*

# Removing biais: balance the covaraites distributions




*The control age-density is reshaped until it matches the treated one.*

# Matching


Treated group
Control group
Distribution de l'âge (avant matching)
Groupe
Control mean = 64.7
Treated mean = 58.2
58.2
64.7
Densité
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Âge

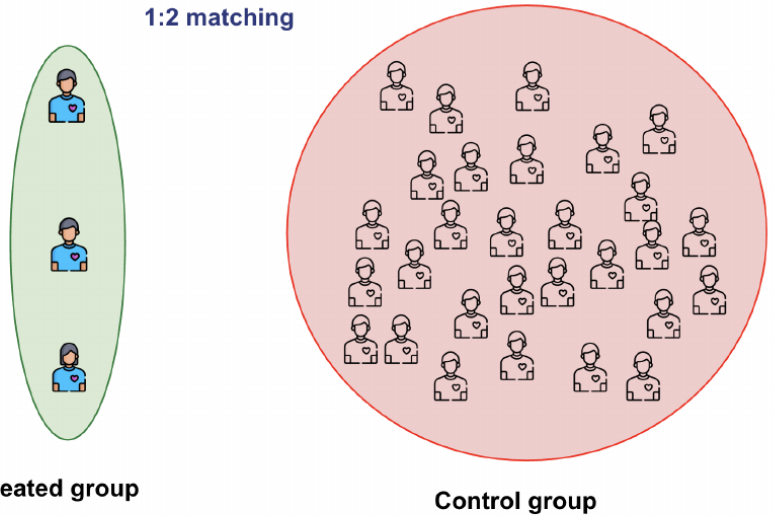
# Matching

Problem : the groups are **not exchangeable** due to **lack of randomization**

Solution : balance the groups by matching similar patients together

**Matching procedure**
(Distance-based matching)

> Select the patients from the control group that are **"similar"** to the ones in the treatment group



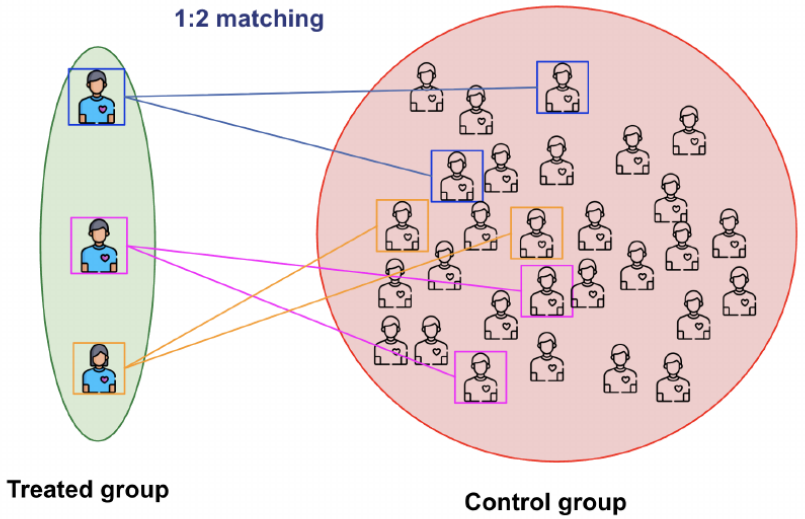
# Matching

Problem : the groups are **not exchangeable** due to **lack of randomization**

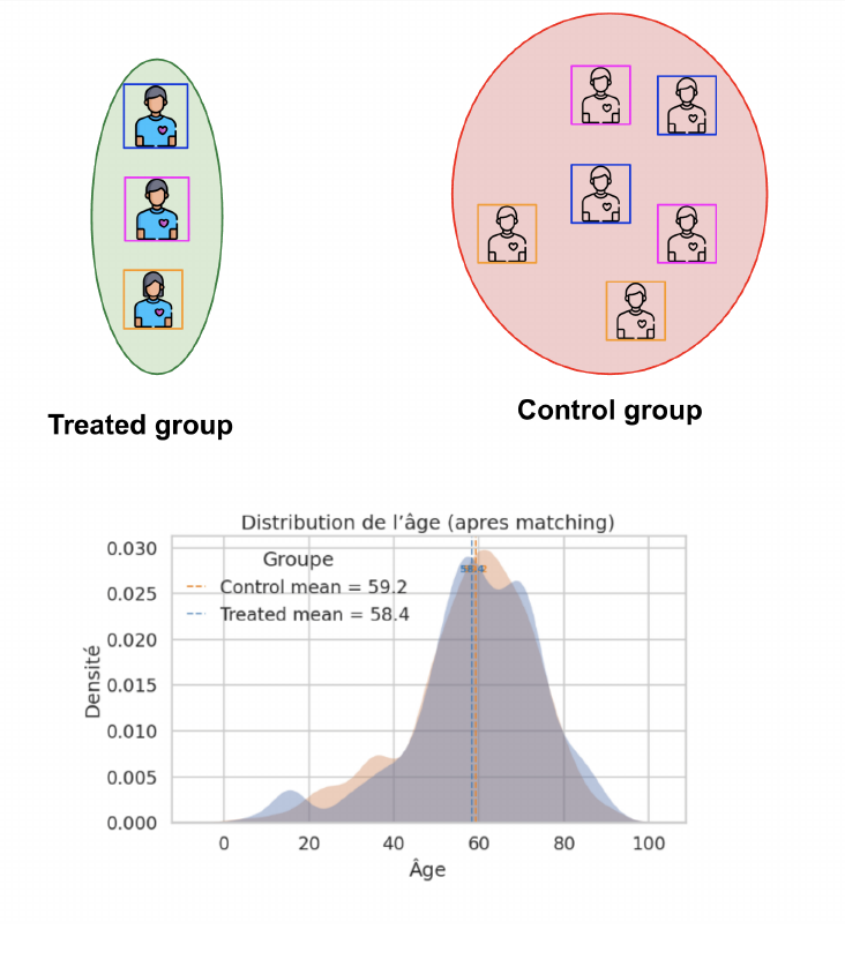
Solution : balance the groups by matching similar patients together

**Matching procedure**
(Distance-based matching)

> Select the patients from the control group that are **"similar"** to the ones in the treatment group



# Matching

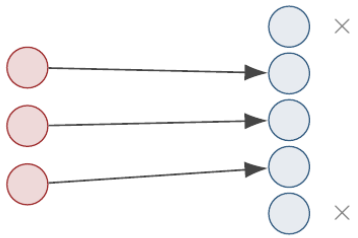

Treated group
Control group
Distribution de l'âge (apres matching)
Groupe
Control mean = 59.2
Treated mean = 58.4
Densité
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Âge

# Matching or weighting methodology

**Matching**

pair each treated $i$ with the control $m(i)$ nearest under a distance $d$

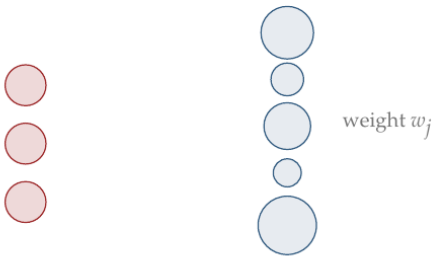
$$\min_{m} \sum_{i:\, T_i=1} d\left(X_i,\, X_{m(i)}\right)$$



**Weighting**

keep every control; reweight them so their distribution matches the treated $P_1$ under the *same* $d$

$$\min_{w \geq 0} \mathcal{D}_d\left(P_1,\, \textstyle\sum_{j:T_j=0} w_j\, \delta_{X_j}\right)$$




Both rest on one **patient distance** $d$ ($P_1$ the treated distribution): matching pairs on $d$; weighting matches the whole *distribution* under a discrepancy $\mathcal{D}_d$ built from $d$ (energy distance / maximum mean discrepancy). Euclidean $d$ on the means recovers $\left\|\bar{X}_1 - \sum_j w_j X_j\right\|$ – the open question is *which* $d$.

# Example of covariates for distance

- Age
- Sex
- Charlson score (comorbidity index)

# Adverse drug reaction analysis on the SNDS

# The SNDS

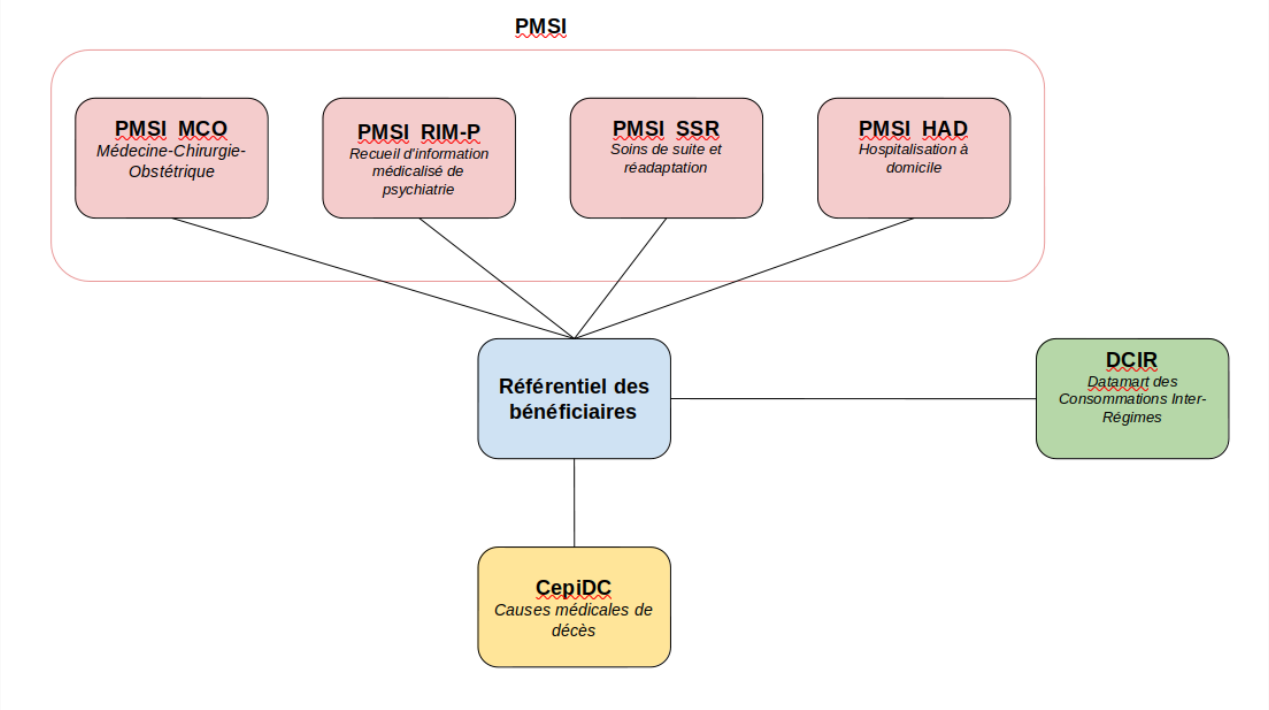
**A massive medical database**

- The medical history of 67 millions of patients
- For more than ten years
- Drug reimbursement and hospital stays

**A claim database**

- The data was originally for administrative purpuse
- It has built-in biais
- It has blind spots

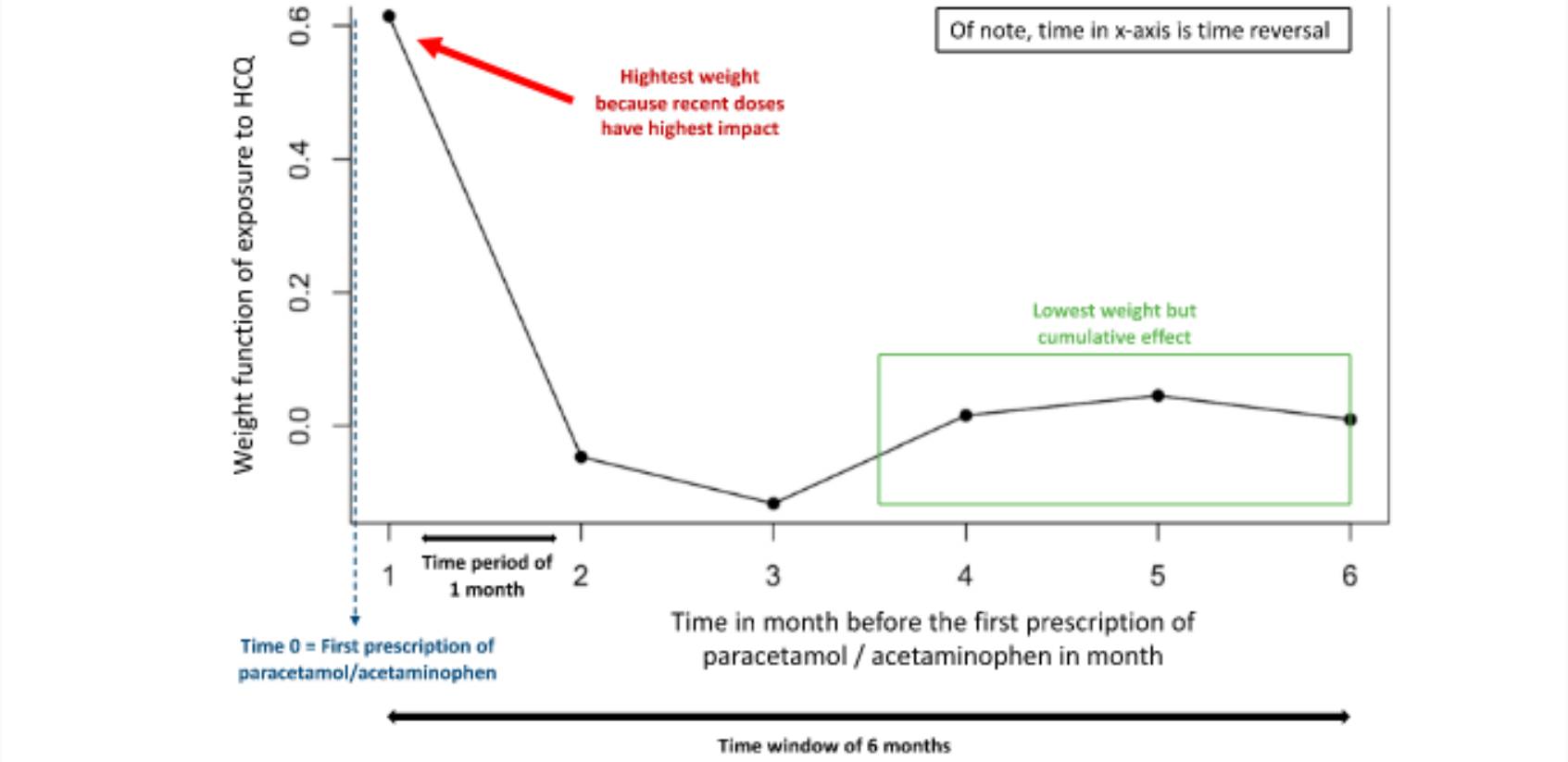
# The SNDS


PMSI
PMSI MCO
Médecine-Chirurgie-Obstétrique
PMSI RIM-P
Recueil d'information médicalisé de psychiatrie
PMSI SSR
Soins de suite et réadaptation
PMSI HAD
Hospitalisation à domicile
Référentiel des bénéficiaires
DCIR
Datamart des Consommations Inter-Régimes
CepiDC
Causes médicales de décès

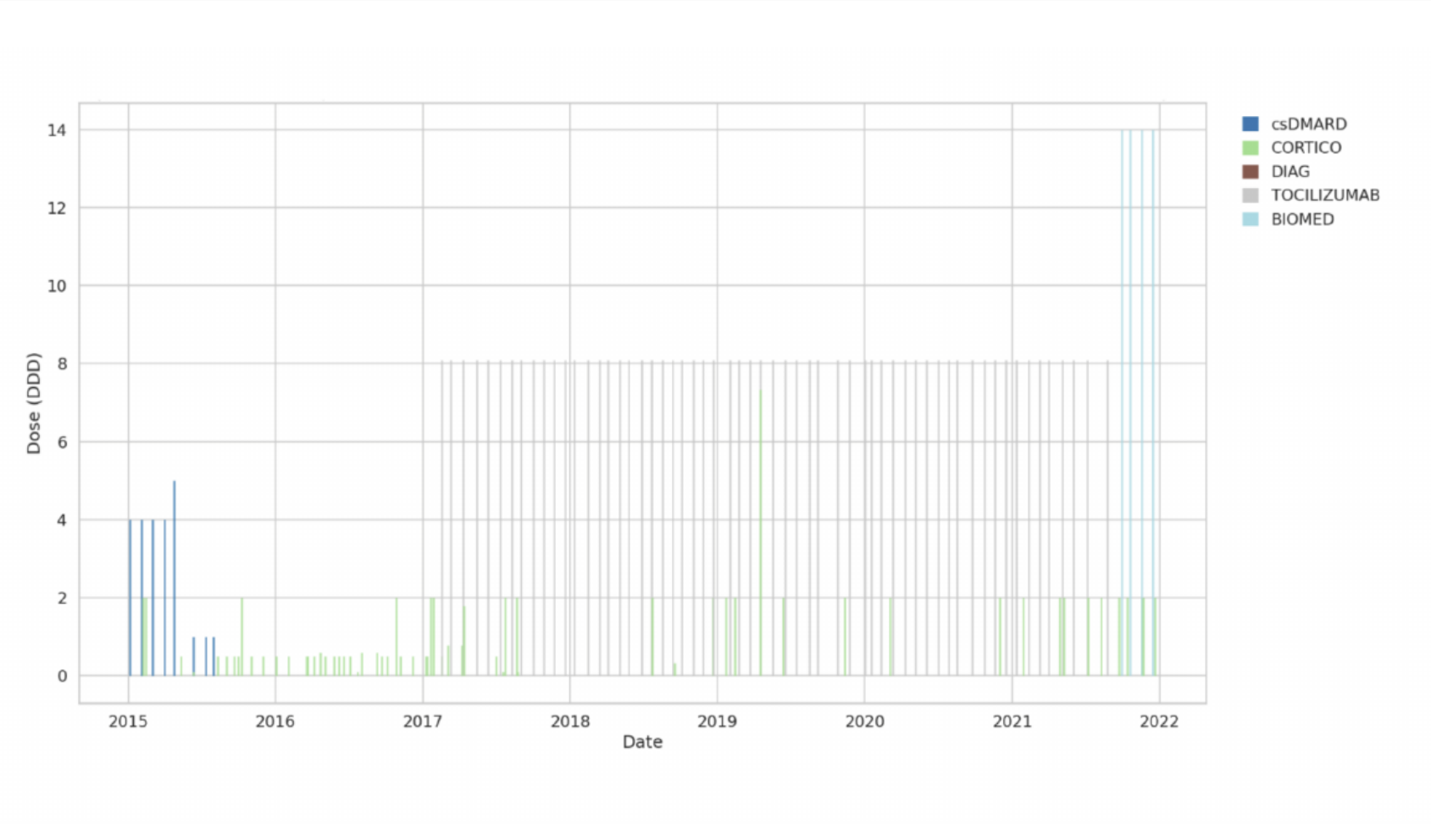
# Type of studies interested in exposure reconstitution

- Adherence study
- Prevalence study
- Exposure effect on individual patient

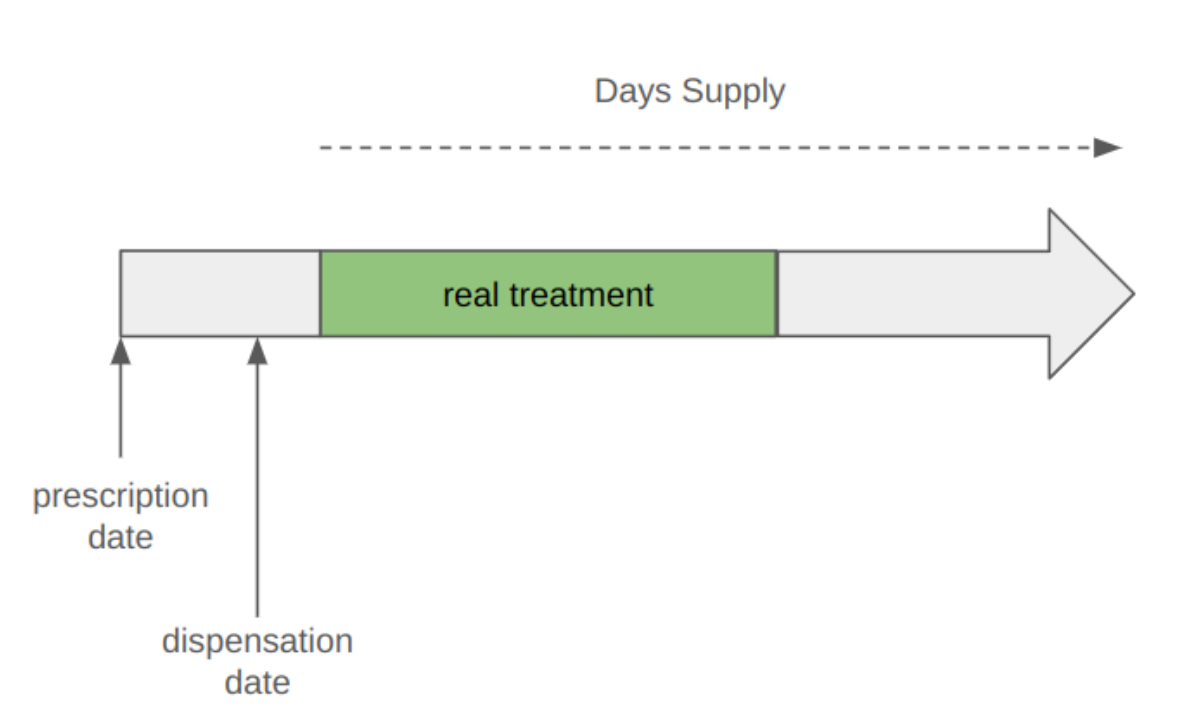
# The WCE model for analyzing



# Prescription in the SNDS



# A prescription event




**A prescription event**

# Information on the prescription event

**The real treatment of the patient**

- Consumed Daily Dose (CDD)
- Treatment duration

**If we assume that the patient take all his medication**

- Consumed Daily Dose (PDD)
- Days Supply

**The data we have**

- Number of boxes
- Unit per box
- Dose per unit

# Estimation objective

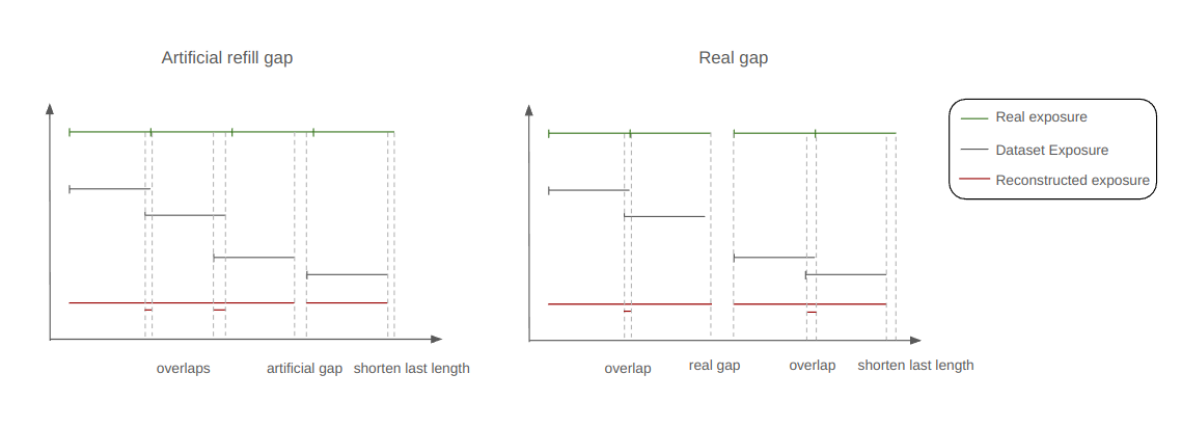
**We need to estimate the Days Supply or the PDD**

# Limits on the type of treatment

- Patient must use all the prescribed medication
- No gap in prescription event
- Limited to chronic treatment

# Reconstitution fo multiple prescription events

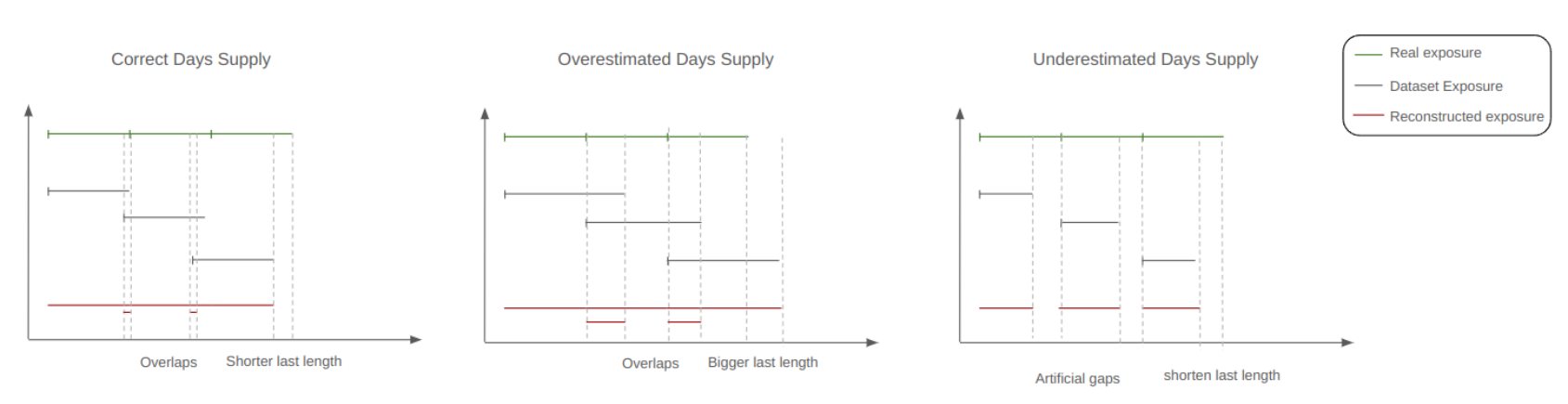
# Challenges of event reconstitution




**Early refill[1] and gap detection**

[1] Carnahan, "Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data".

# Challenges of event reconstitution




**Overestimation and underestimation of the Days Supply**

# The different methods of episode reconstitution


**Overlap Suppression**

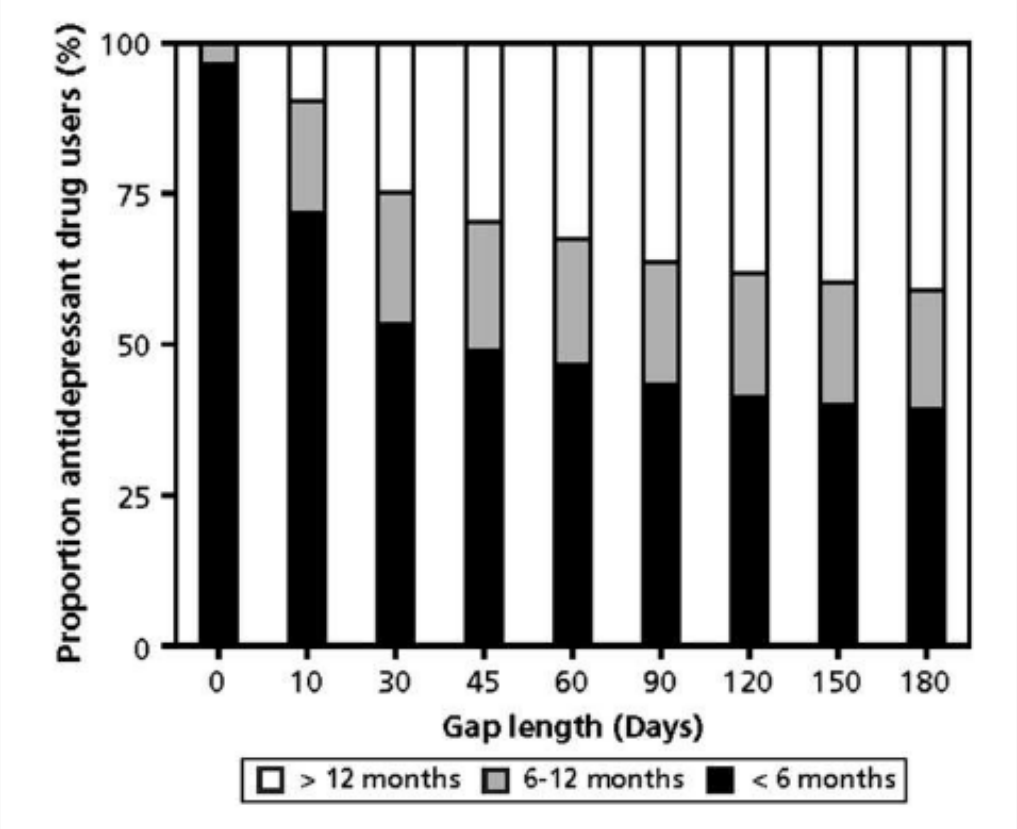
# The different methods of episode reconstitution


**Grace period**

# The different methods of episode reconstitution

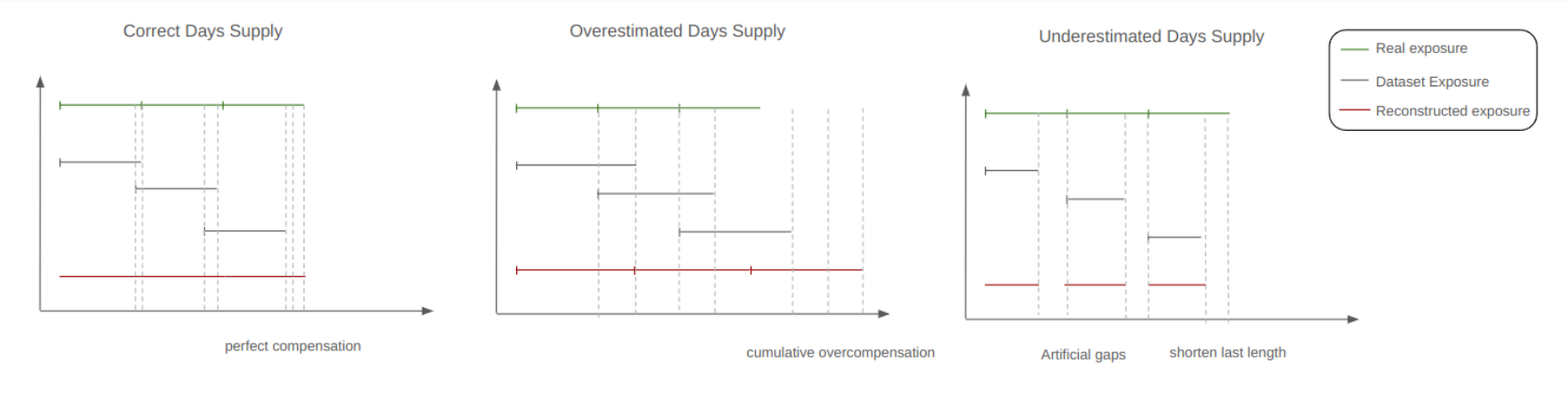

**Grace period and overlap suppression**

# The different methods of episode reconstitution




**Grace period duration influence**

# The different methods of episode reconstitution



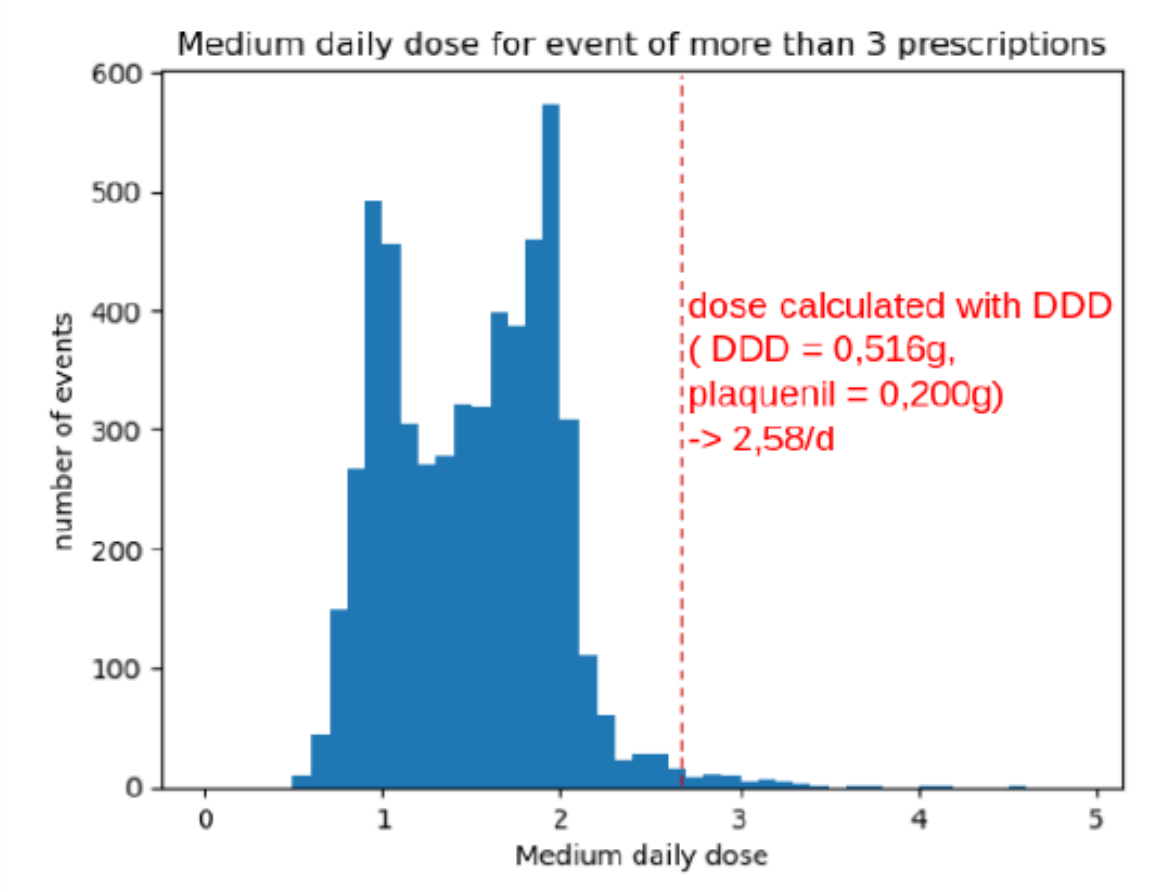

**Episode reconstitution**[1]

[1] Carnahan, "Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data".

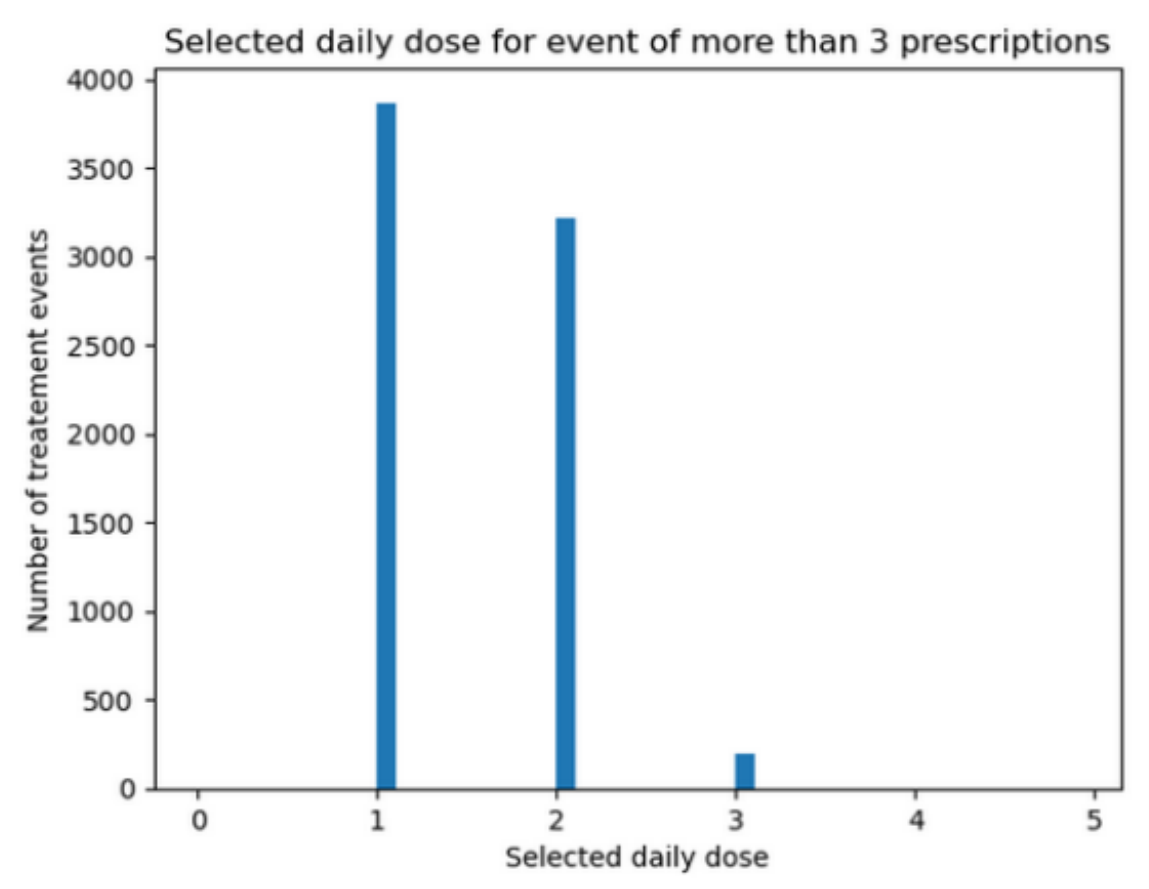
# The different methods of episode reconstitution

| | Underestimated Days Supply | Overestimated Days Supply | Early refill | Real Gaps detection | Computation complexity |
|---|---|---|---|---|---|
| **Naive data** | Artifical gaps<br>Shorten last event | Big overlaps<br>Bigger last event | Overlap<br>Artificial gaps<br>Shorten last event | Yes | |
| **Overlap suppression** | Artifical gaps<br>Shorten last event | Bigger last event | Artificial gaps<br>Shorten last event | Yes | Small |
| **Grace period** | Shorten last event | Big overlaps<br>Bigger last event | Overlap<br>Shorten last event | Suspectible to grace period definition | Small |
| **Overlap + Grace period** | Shorten last event | Bigger last event | Shorten last event | Suspectible to grace period definition | Small |
| **Event reconstitution** | Artificial gap<br>Shorten last event | Cumulative error for each event, can be badly overestimated | Realistic representation | Yes | High |
| **Event reconstitution + Grace Period** | Shorten last event | Cumulative error for each event, can be badly overestimated | Realistic representation | Suspectible to grace period definition | High |

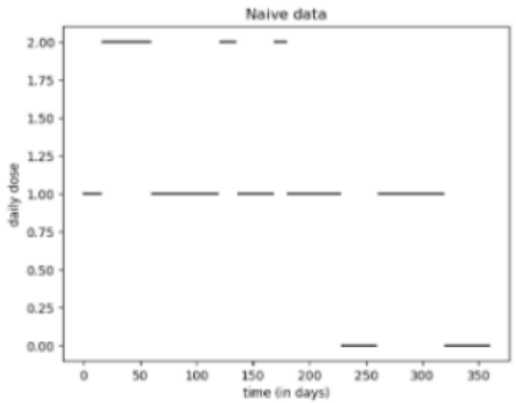
# Exemple of Hydroxychlorochin

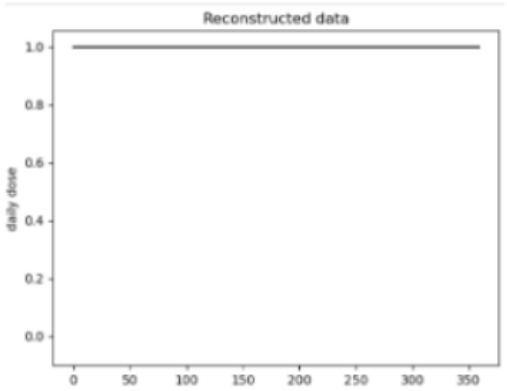


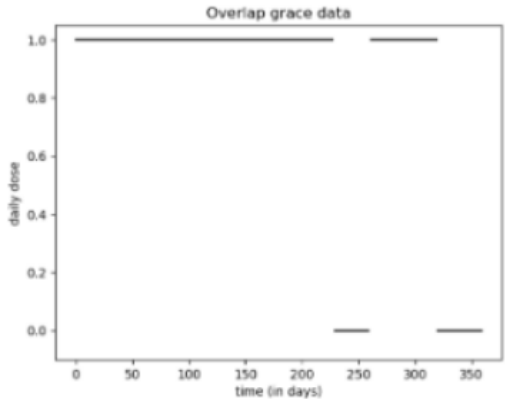
# Example of Hydroxychlorochin



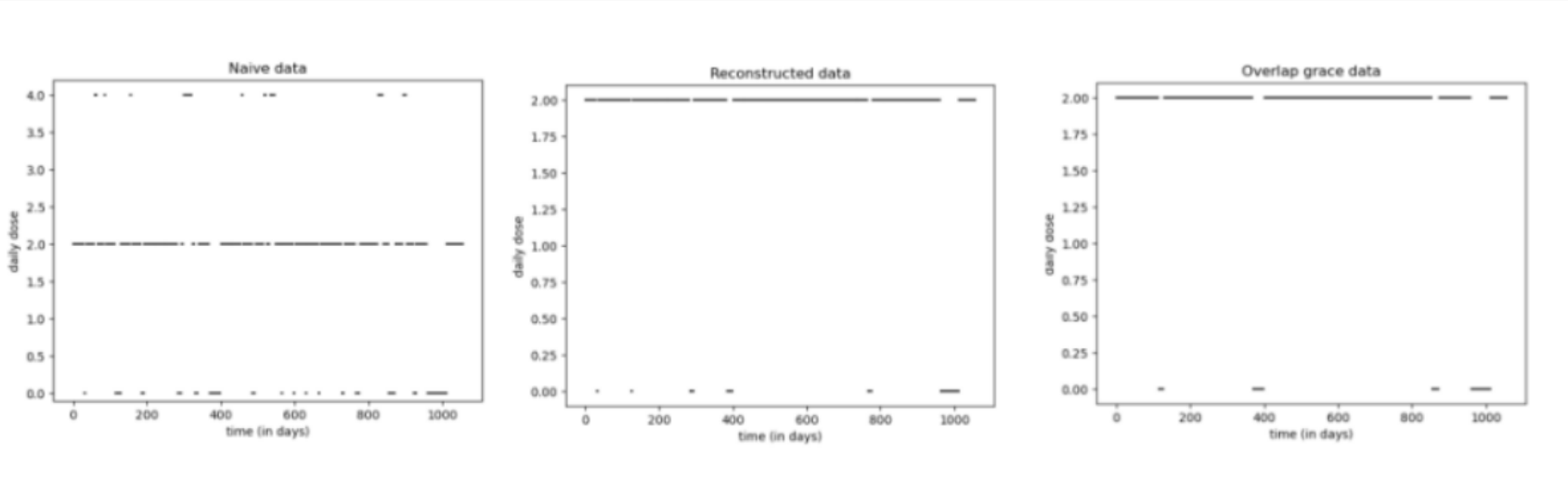
# Reconstitution of a exposure



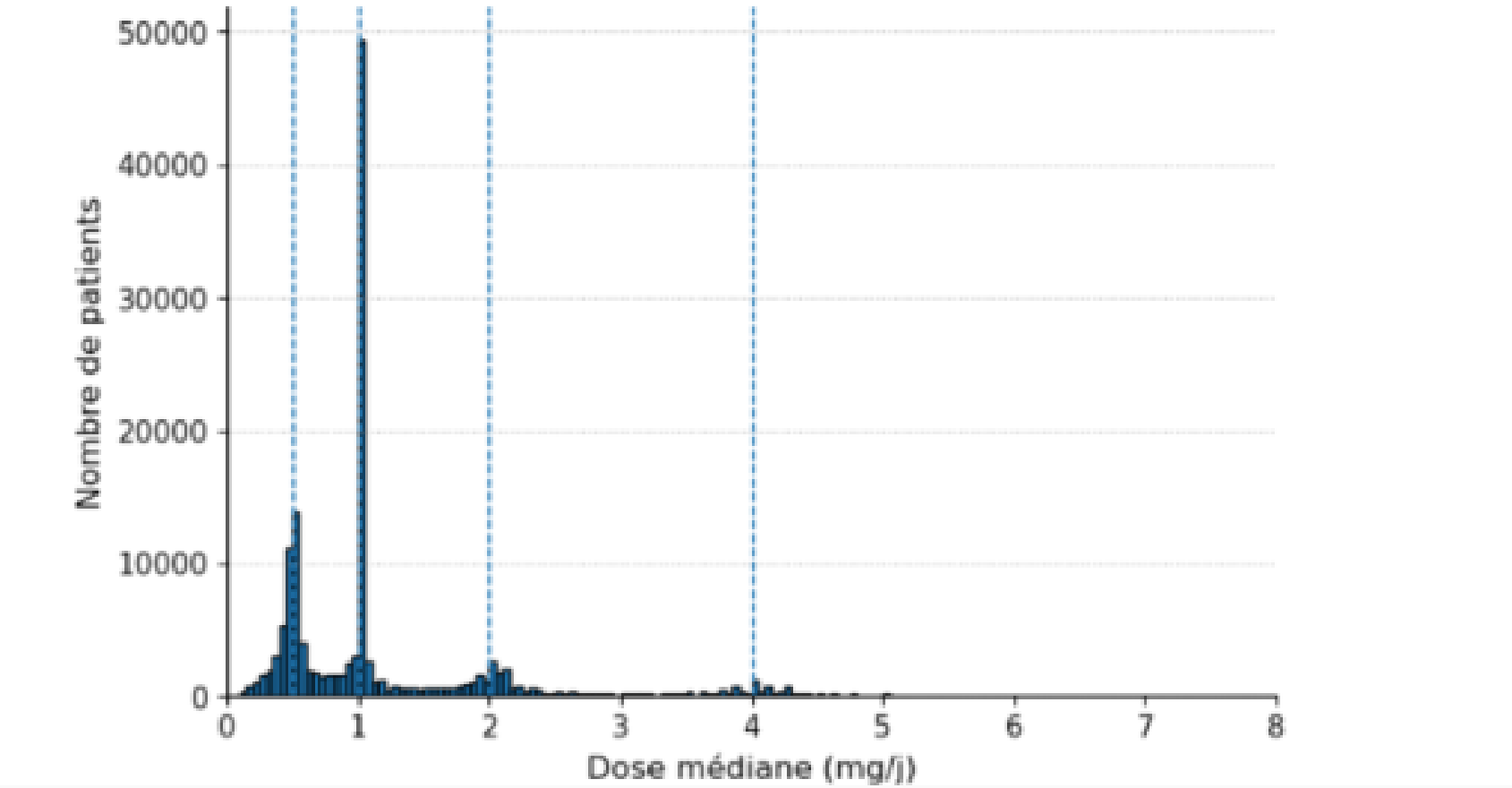



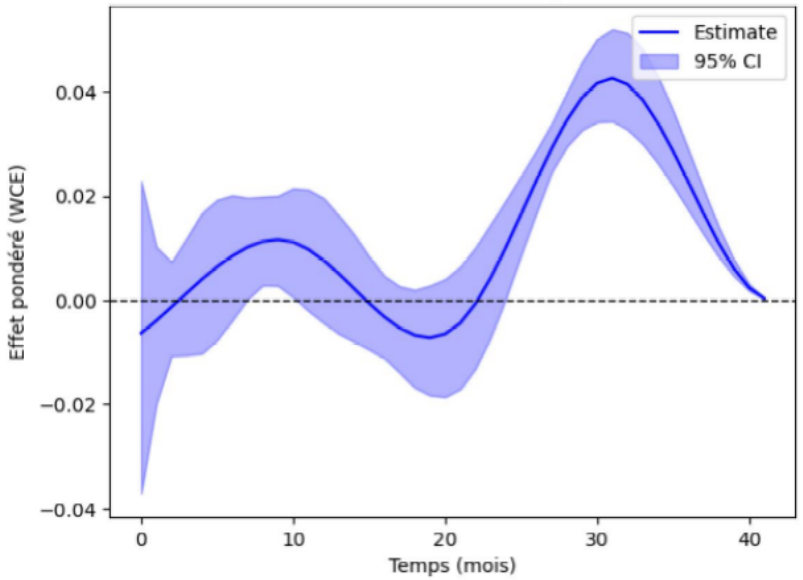
# Reconstitution of an exposure



# Atorvastatine adnd drug-induced hepathitis



# Atovastatine has an effect on drug-induced hepathitis




**Hyperparamètres**

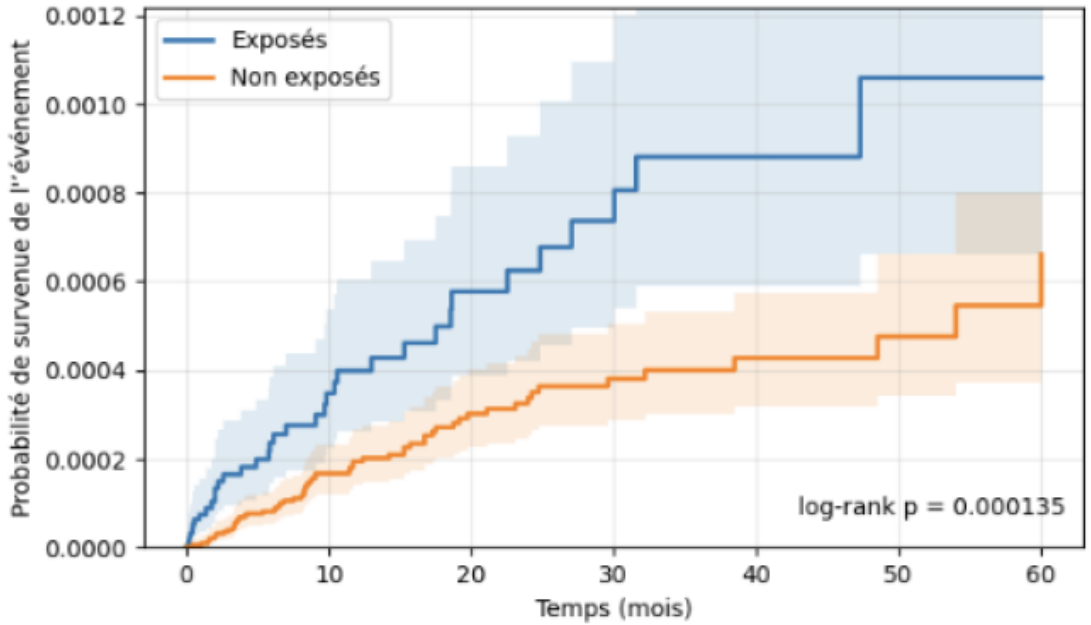
- Cutoff : 42
- Nknots : 3
- Constraint : Droite

**Bootstrap**

- Boostrap : 1000
- Batchsize : 10

**Hazard Ratio (IC95%) :** 1,62 [1,54 ; 1,69]

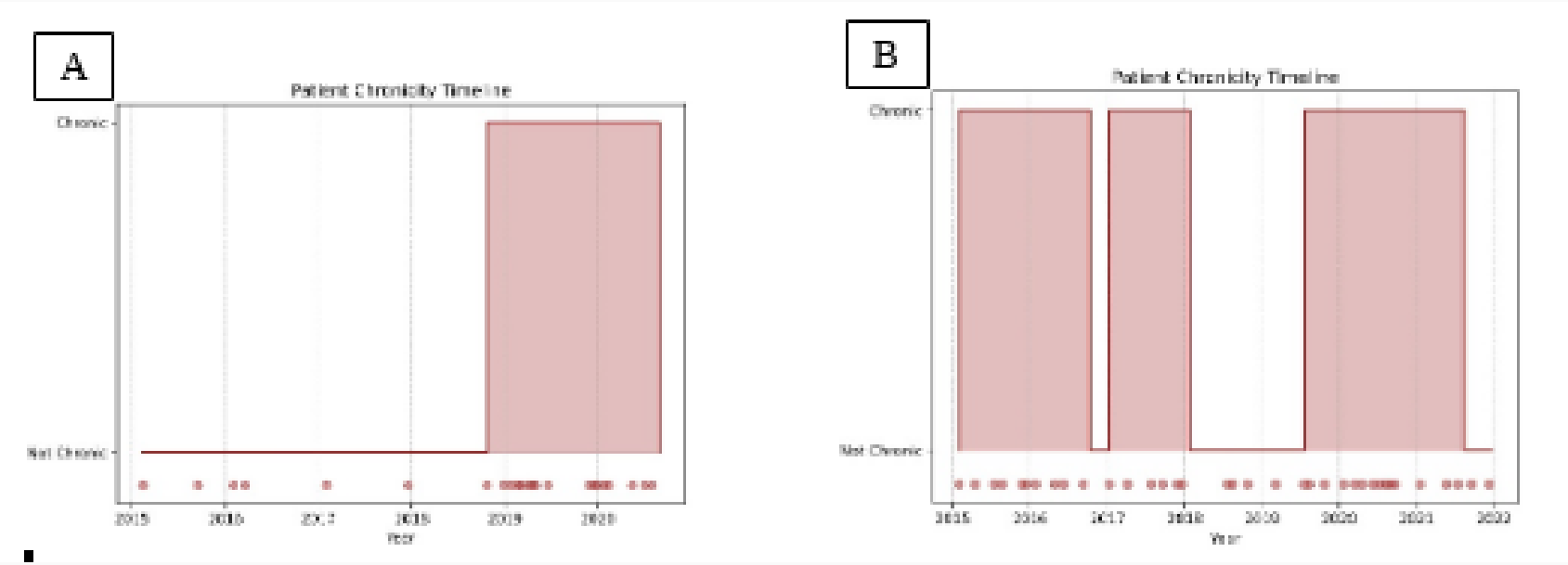
# Survival curve




| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Exposés | | | | | | |
| At risk | 95012 | 40407 | 23664 | 14489 | 8363 | 4694 | 0 |
| Censored | 0 | 54583 | 71319 | 80491 | 86615 | 90283 | 94977 |
| Events | 0 | 22 | 29 | 32 | 34 | 35 | 35 |
| | Non exposés | | | | | | |
| At risk | 379834 | 161599 | 94624 | 57925 | 33431 | 18761 | 0 |
| Censored | 0 | 218198 | 285157 | 321850 | 346342 | 361011 | 379770 |
| Events | 0 | 37 | 53 | 59 | 61 | 62 | 64 |

# A more complex simulation: corticoids

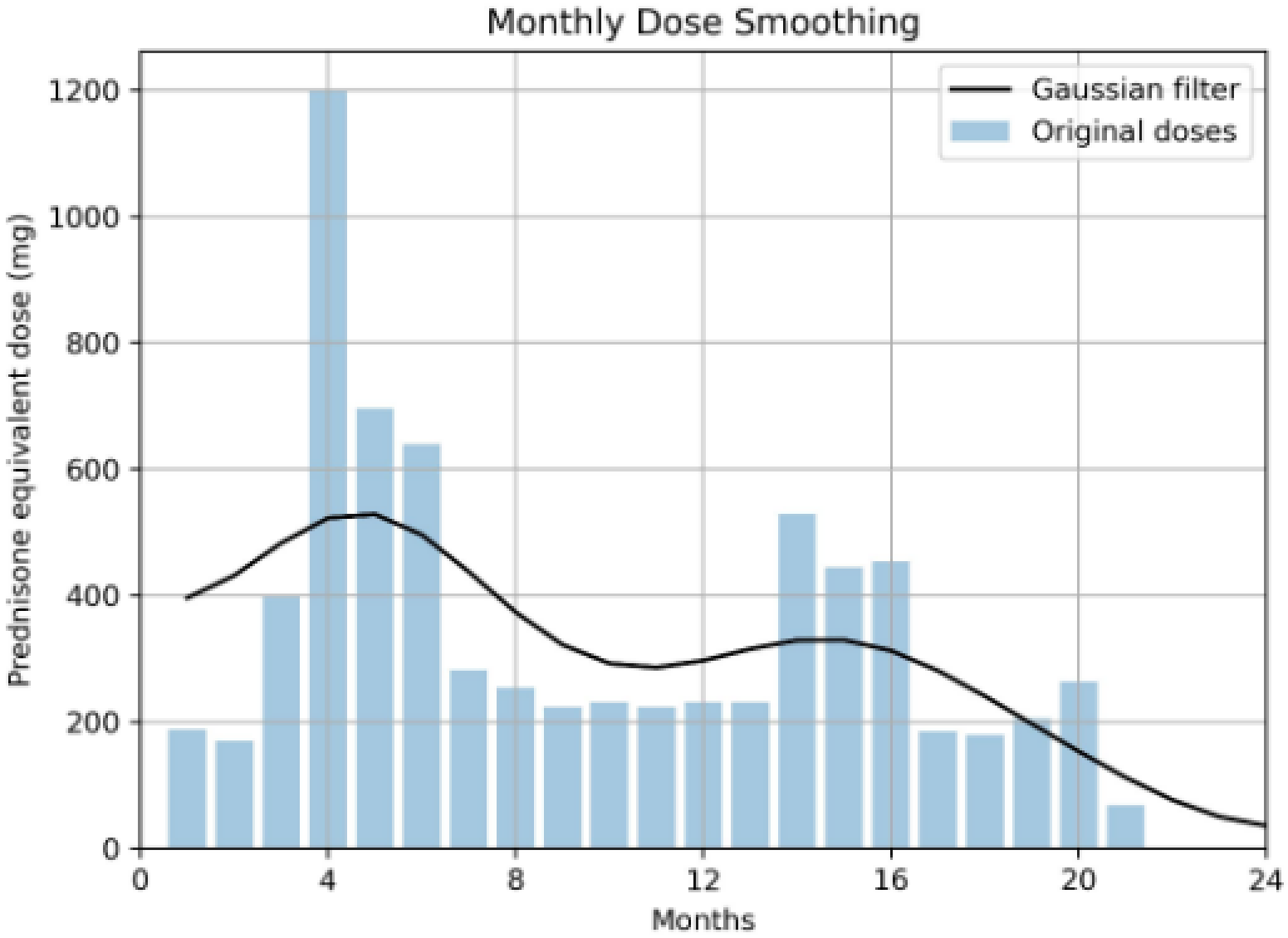
# Definition of a chronic exposition



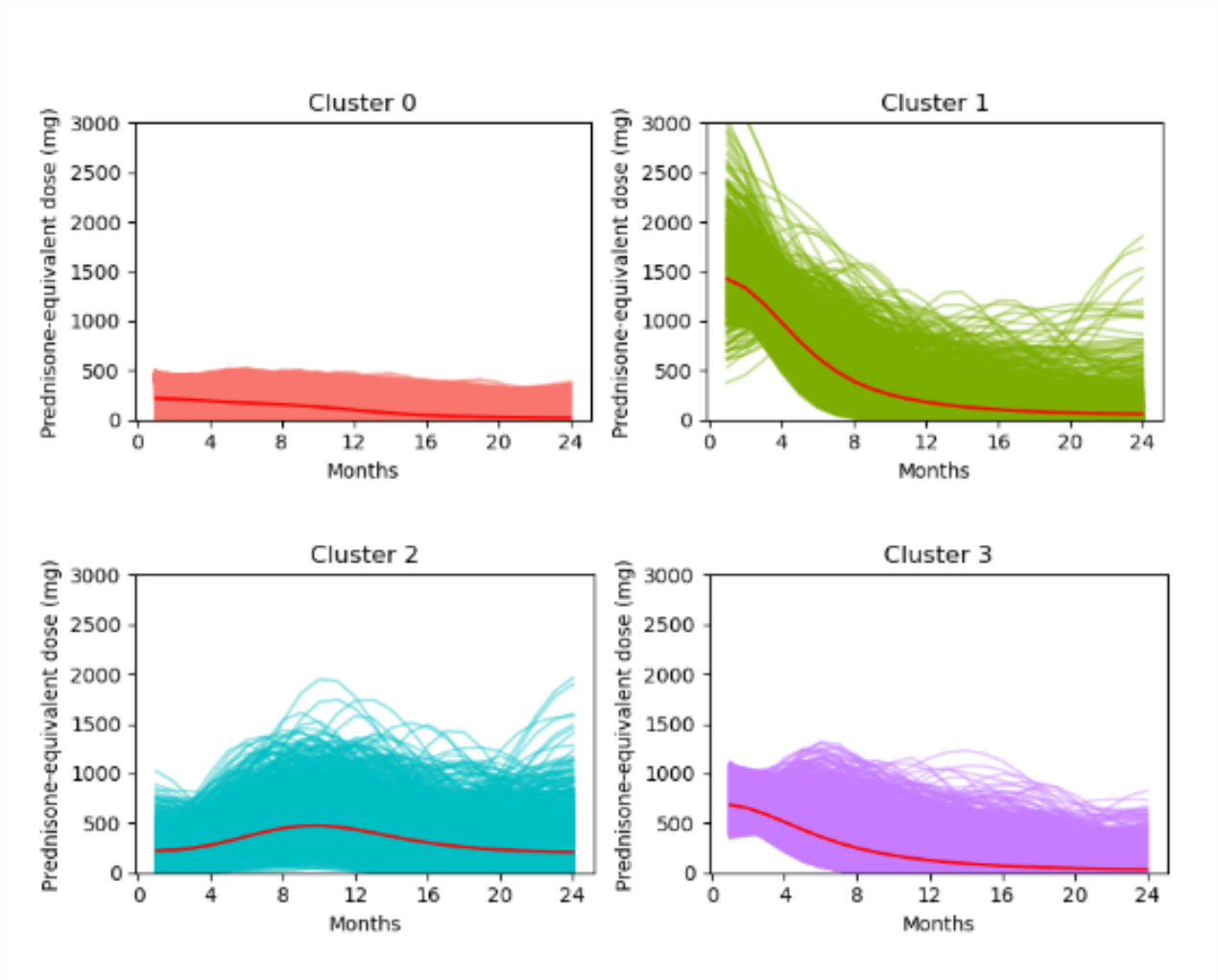
# Patient matching

| Variable | Chronic | Non-chronic |
|---|---|---|
| n | 53,400 | 1,014,453 |
| Female sex (%) | 32,144 (60.1) | 589,871 (58.1) |
| Age (mean (SD)) | 59.63 (17.55) | 46.76 (18.61) |
| Weighted Charlson score (mean (SD)) | 1.37 (2.36) | 0.38 (1.38) |

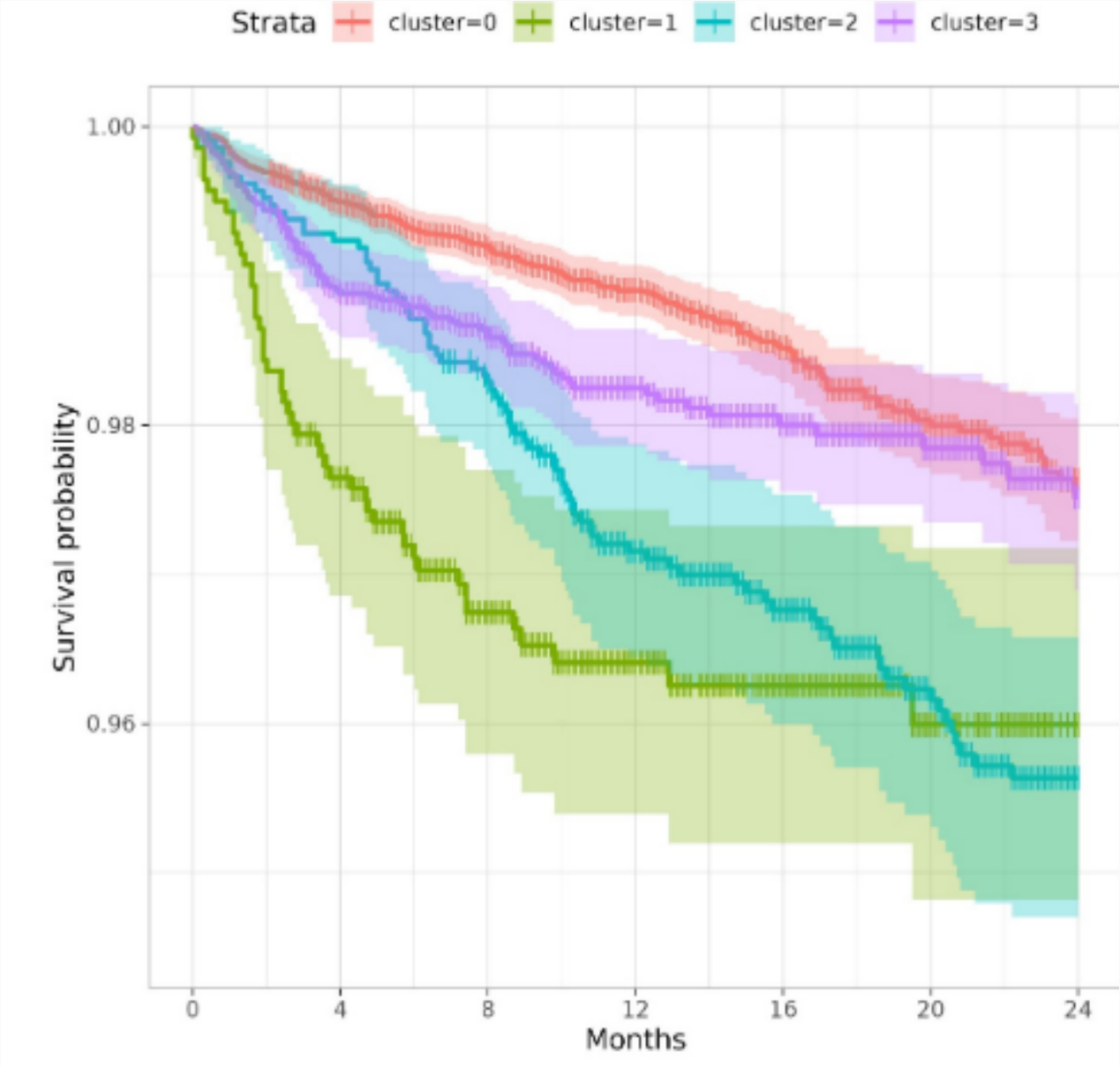
# Modelling doses with Gaussian Smoothing



# Modelling with Gaussian Smoothing



# Survival of the different clusters



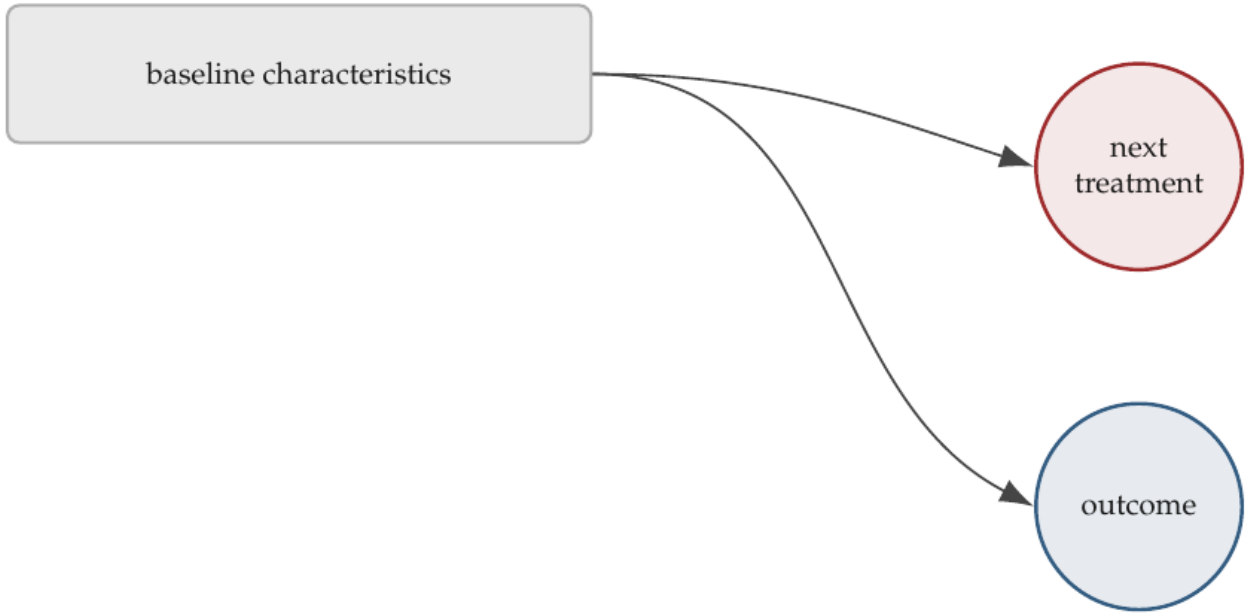
# Interpretation on the clusters

# Patient history for causal inference

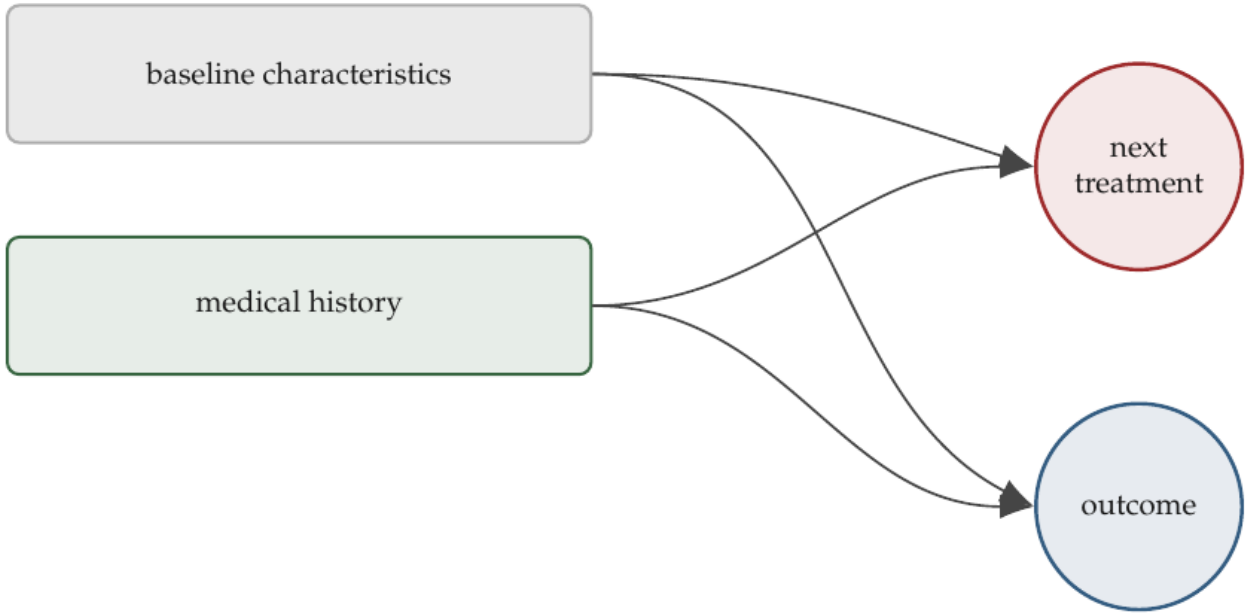
# Type of studies interested in exposure reconstitution

- Adherence study
- Prevalence study
- Exposure effect on individual patient
- Medical history for causal inference

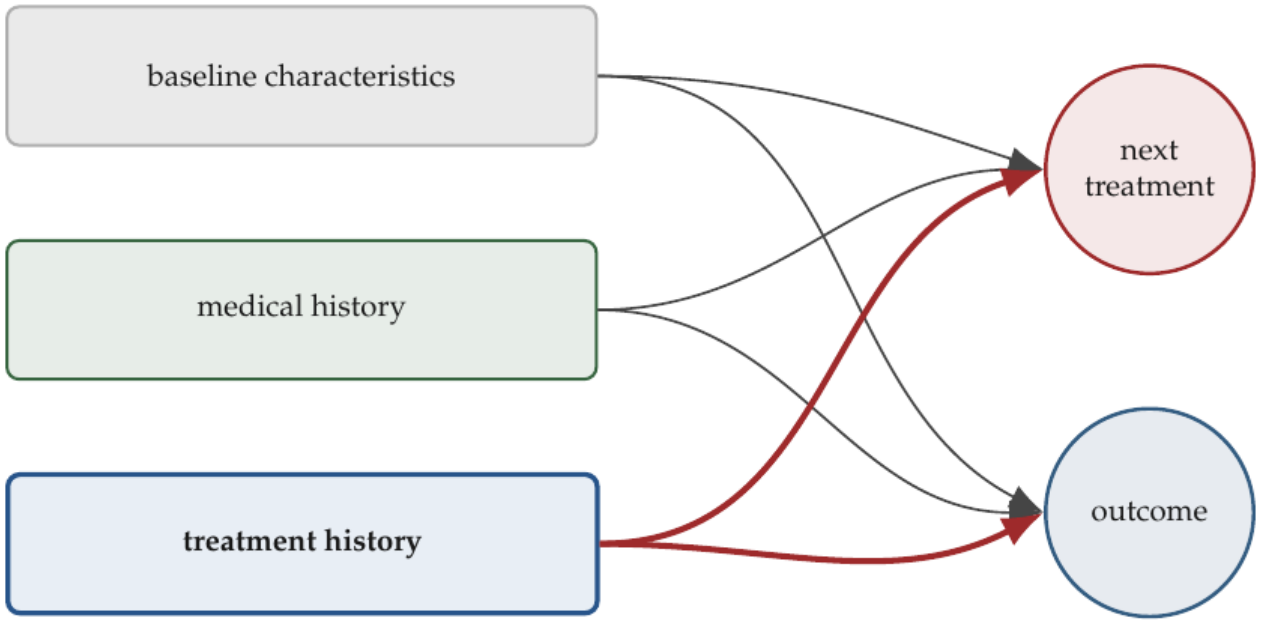
# Three layers of confounding biais in patients



# Three layers of counfounding biais in patients

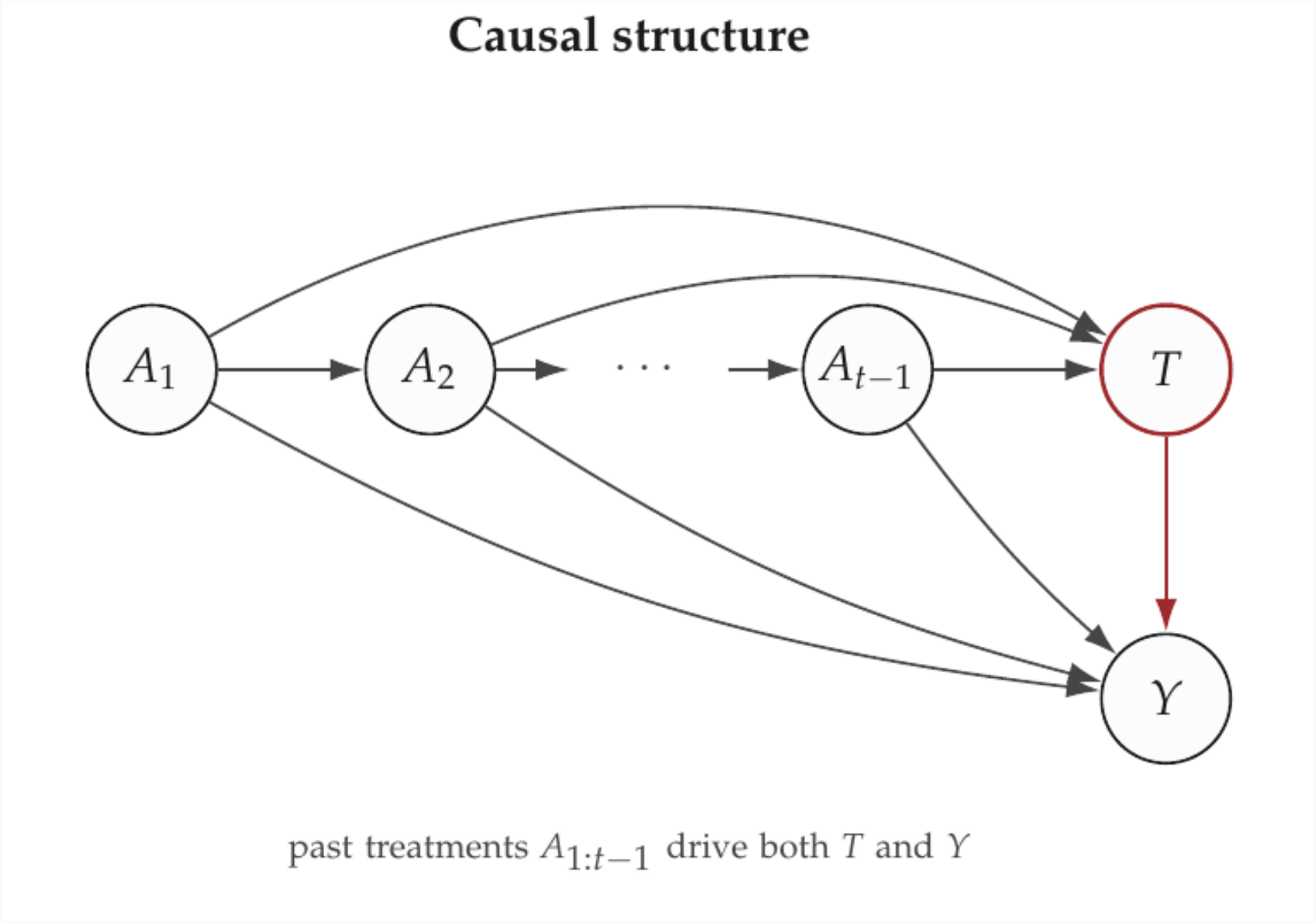


# Three layers of counfounding biais in patients

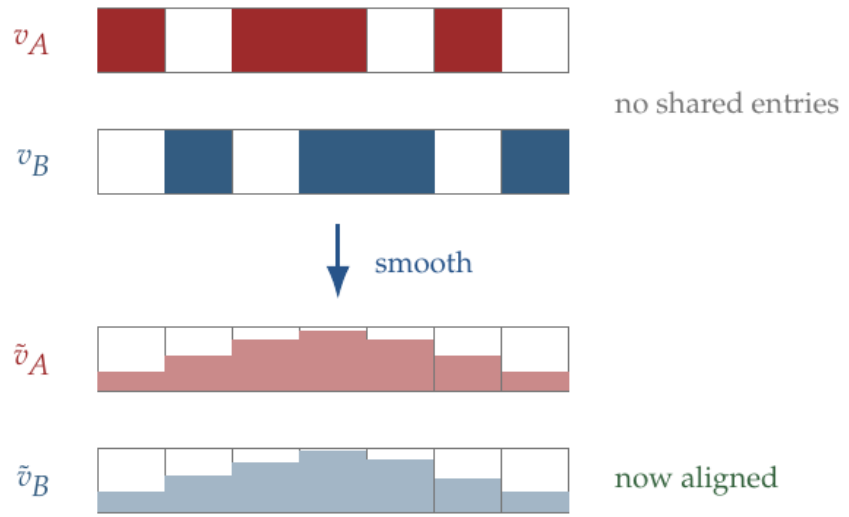



strongest driver of the next decision – and the hardest to summarise

# The causal structure

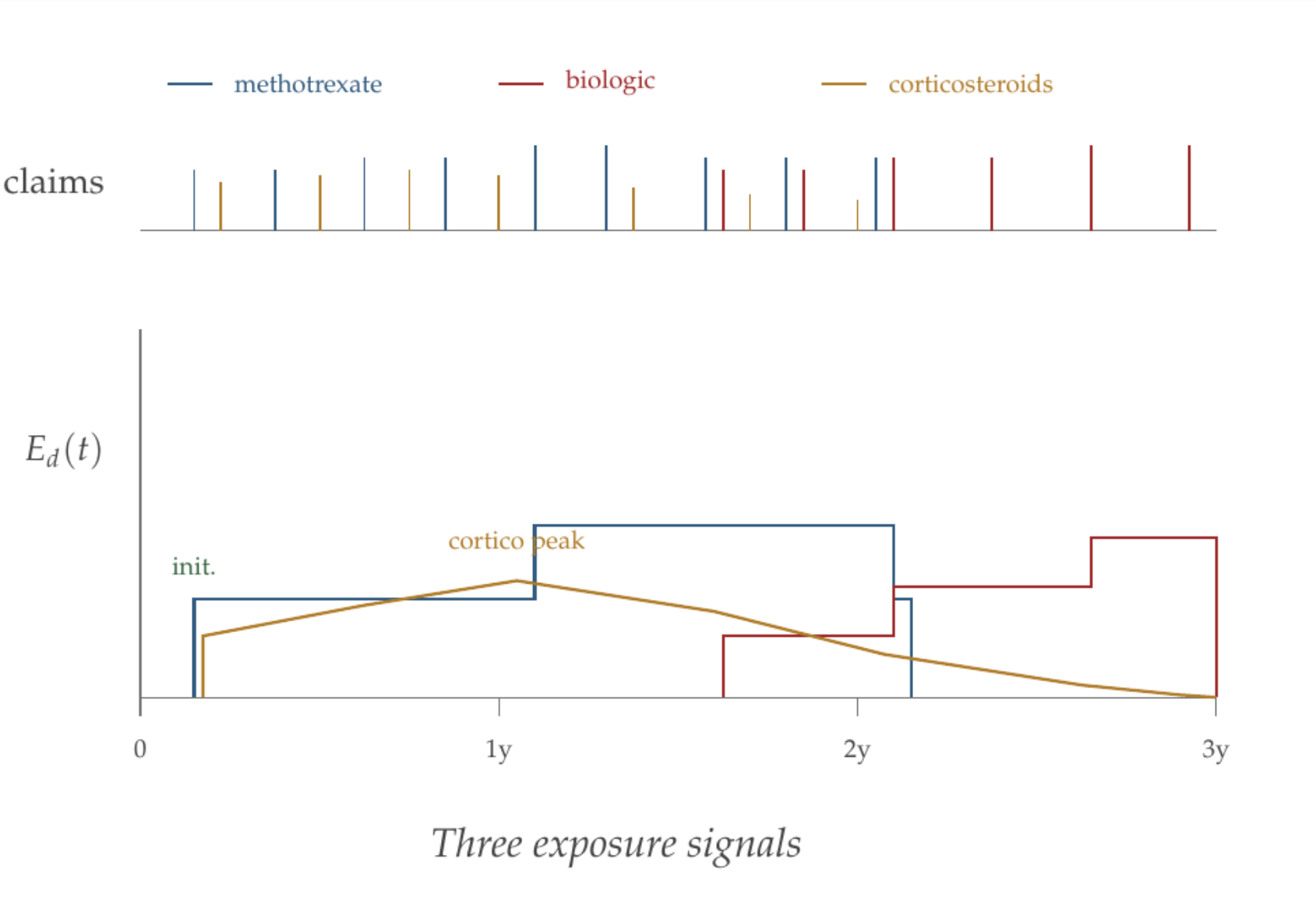


# Event reconstruction




*Different claim days ⇒ disjoint one-hot vectors; after smoothing, the exposures coincide.*

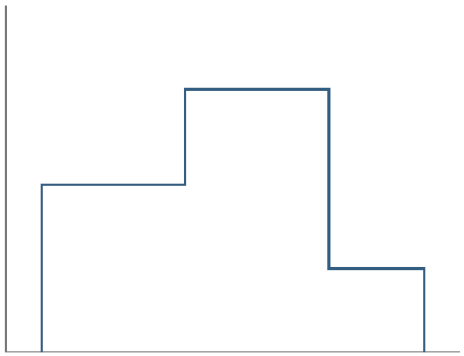
# Event reconstruction




*Three exposure signals*

# 3 different smoothing functions

**Step (coverage) function**

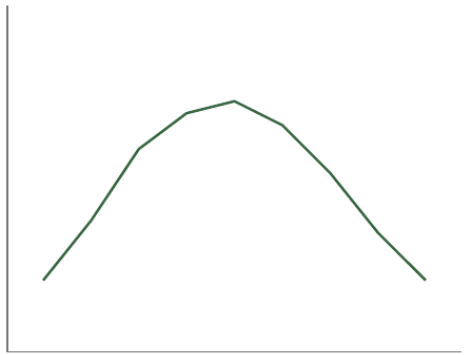


$$E_d(t) = \sum_k \frac{q_k}{\Delta_k} \mathbf{1}[t_k \leq t < t_k + \Delta_k]$$

**Decay kernel**

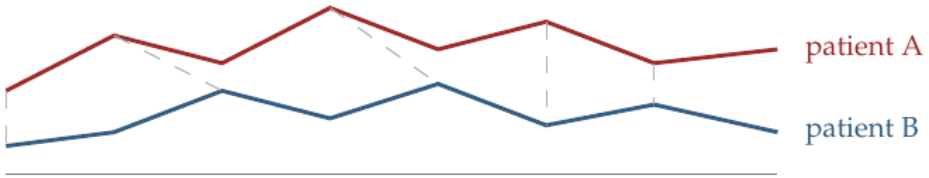
$$E_d(t) = \sum_k q_k e^{-(t-t_k)/\tau} \mathbf{1}[t \geq t_k]$$

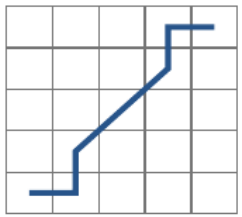
**Smoothing spline**



$$E_d = \arg\min_f \sum_k (f(t_k) - q_k)^2 + \rho \int (f'')^2$$

# Alignement matching

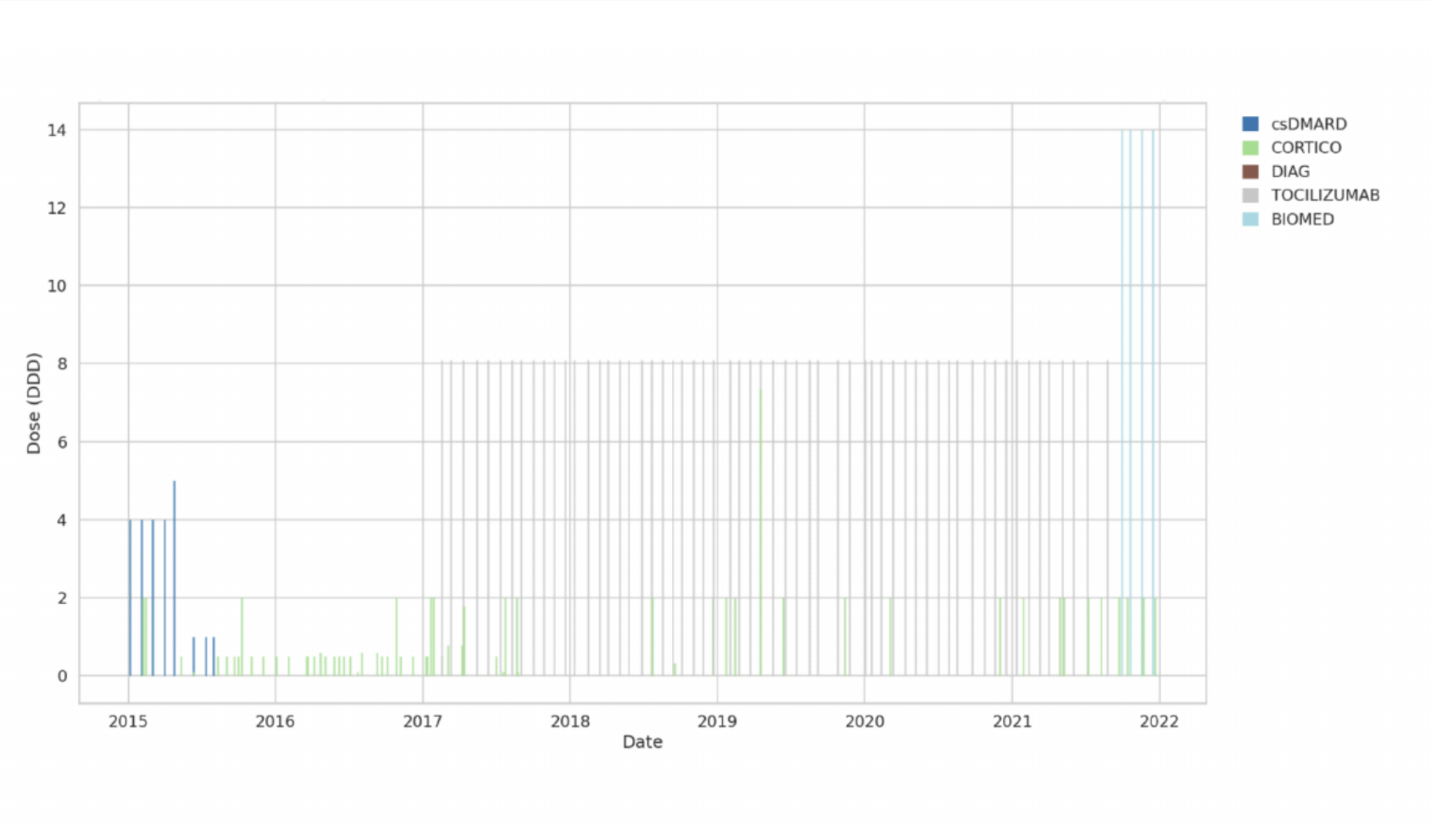


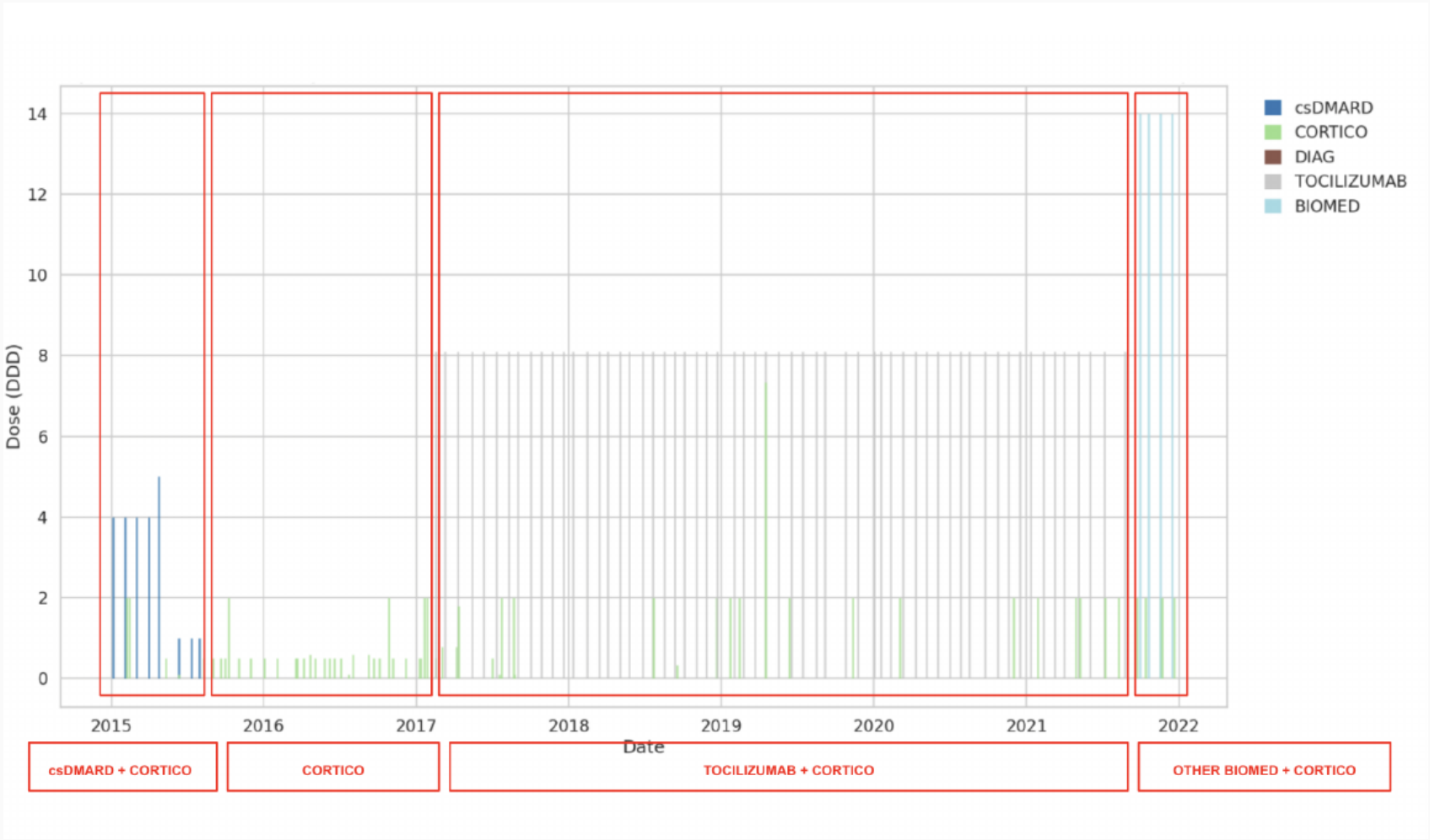
warping path

*Alignment matches comparable phases at different times – but it is slow.*

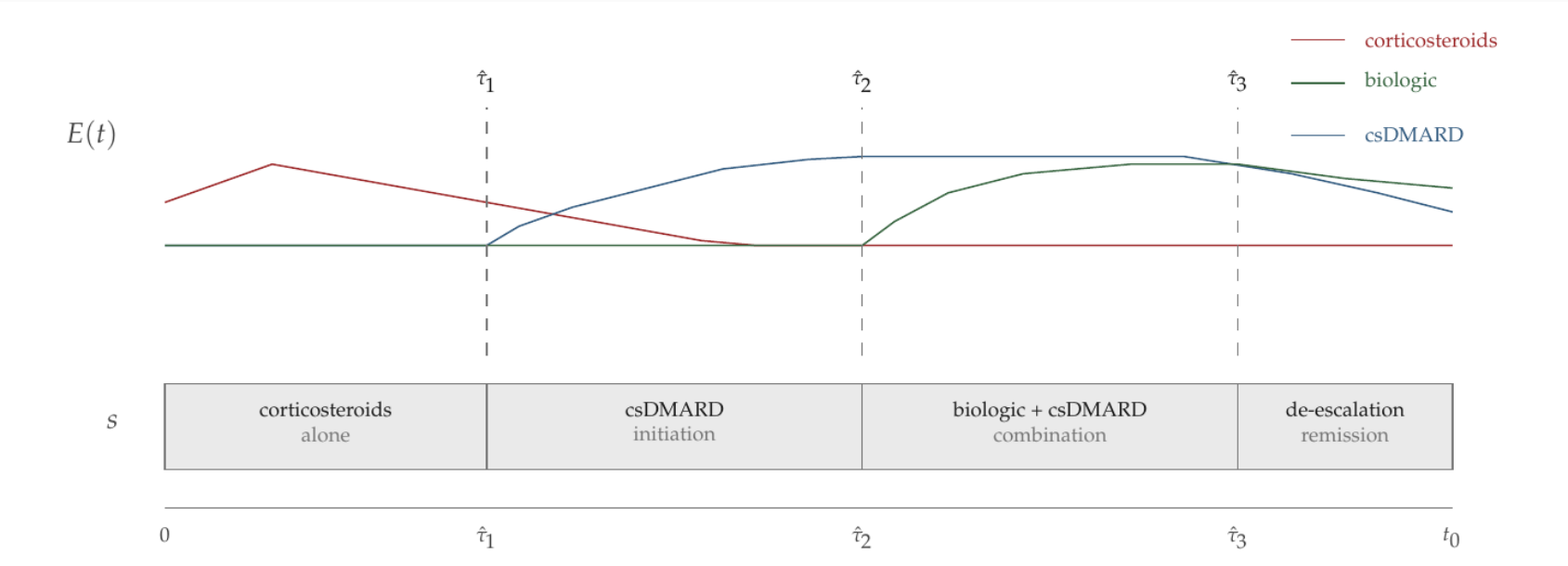
# Prescription in the SNDS



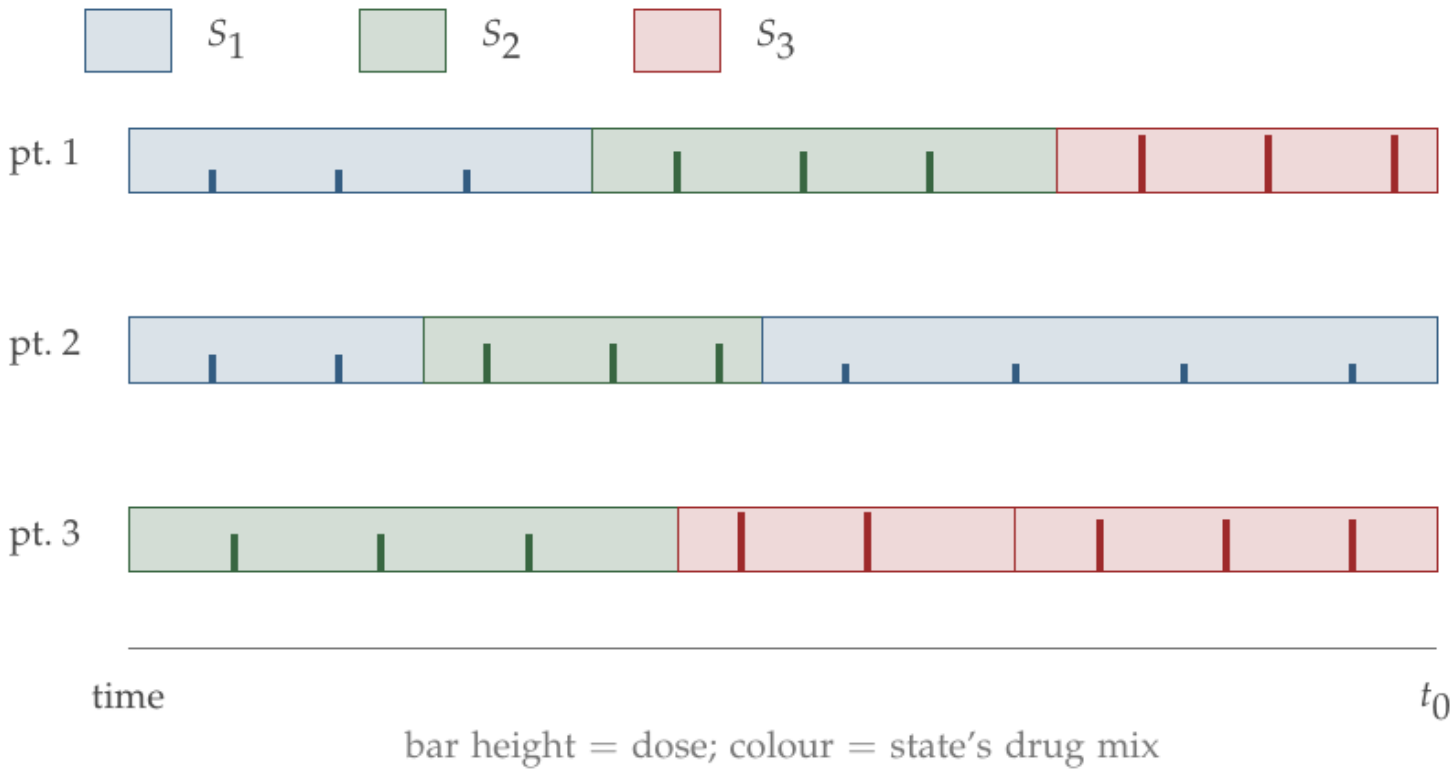
# Prescription in the SNDS



# Trajectory reconstruction

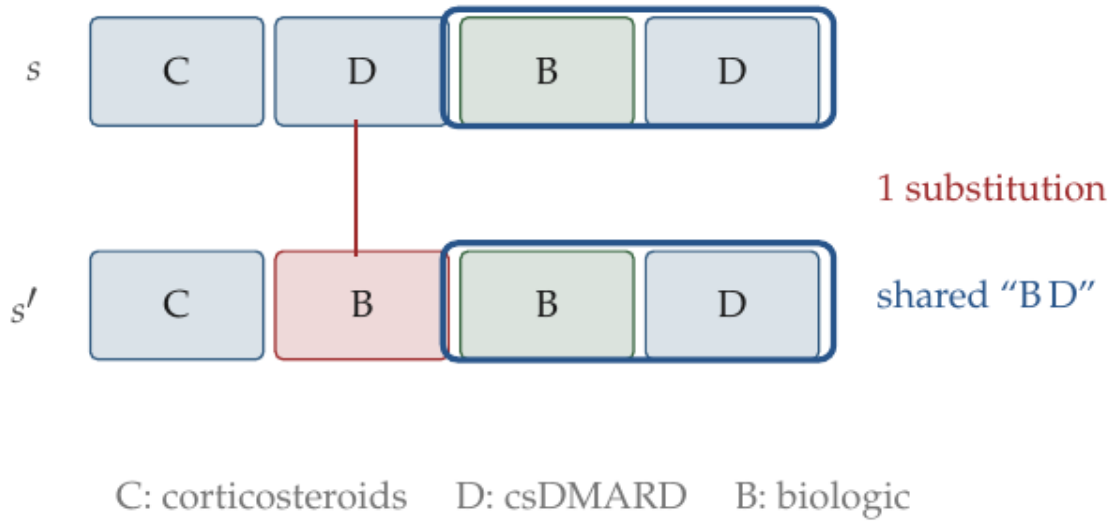


# Patient trajectory




*Unified states segment every patient and can generate synthetic histories.*

# Low dimension patient distance




*Distances on sequences live in far lower dimension than the raw signal.*