APPENDIX 1: List of the initial variables, their corresponding thresholds, and their status at the end of the macro-variable construction process

<u>51 initial variables:</u>

Variables	Threshold of
- _Hospitalization for neurological disease*_	
- _Hospitalization for bones disease*_	
- Laminectomy	≥1
- Kyphoplasty	≥1
- Radiofrequency ablation of liver metastases (RALM)	≥1
- Cementoplasty	≥1
- Radical prostatectomy	≥1
- Testicular pulpectomy	≥1
- Orchiectomy	≥1
- Brachytherapy**	≥1
- Cryotherapy**	≥1
- High intensity focused ultrasound (HIFU)**	≥1
- Pelvic or iliac lymphadenectomy	≥1
- Transurethral resection of the prostate (TURP)**	≥1
- Prostatic adenectomy**	≥1
- Prostate biopsy**	≥1
- Buserelin	≥1
- Goserelin	≥1
- Leuprorelin	≥1
- Triptorelin	≥1
- Degarelix**	≥1
- Bicalutamide	≥1
- Cyproterone	≥1
- Flutamide	≥1
- Nilutamide	≥1
- Diethylstilbestrol	≥1
- _Ketoconazole*_	
- Estramustine	≥1
- Abiraterone Acetate	≥1
- Enzalutamide	≥1
- Cabazitaxel	≥1
- Docetaxel (proxy)	≥1
- Denosumab	≥1
- Zoledronic acid	≥1
- Clodronic acid	≥1
- Radium 223	≥1
- Samarium 153	≥1
- Strontium 89	≥1
- Non-intensity modulated radiotherapy**	≥20
- Intensity modulated radiotherapy**	≥20
- Stereotactic radiotherapy*	
- Prostate magnetic resonance imaging**	≥1
- Laboratory test (except prostate-specific antigen test) *	
- Prostate-specific antigen test	≥1
- _Endocrinology lab test*_	
- _Visit to general practitioner*_	
- _Visit to oncologist*_	
- _Visit to urologist*_	
- _Visit to other medical specialist*_	
- Hospitalization with palliative care	≥1
- Hospitalization for metastasis management	≥1

* = removed variable
** = switched variable

Baulain R, Jové J, Sakr D, Gross-Goupil M, Rouyer M, Puel M, Blin P, Droz-Perroteau C, Lassalle R, Thurin NH. Clustering of prostate cancer healthcare pathways in the French National Healthcare database. Cancer Innovation. 2022. https://doi.org/10.1002/cai2.42

APPENDIX 2: List of all the potential patient statuses over healthcare pathways and their corresponding macro-variables combination

Surveillance	**Local treatment**	**Androgenic deprivation**	**Advanced treatment**	**Death**	**Label**	**Combination variable Code**
0	0	0	0	0	no treatment	**1**
0	0	0	1	0	advanced treatment	**2**
0	0	1	0	0	androgenic deprivation	**3**
0	0	1	1	0	androgenic deprivation + advanced treatment	**4**
0	1	0	0	0	local treatment	**5**
0	1	0	1	0	local treatment + advanced treatment	**6**
0	1	1	0	0	local treatment + androgenic deprivation	**7**
0	1	1	1	0	local treatment + androgenic deprivation + advanced treatment	**8**
1	0	0	0	0	surveillance	**9**
1	0	0	1	0	surveillance + advanced treatment	**10**
1	0	1	0	0	surveillance + androgenic deprivation	**11**
1	0	1	1	0	surveillance + androgenic deprivation + advanced treatment	**12**
1	1	0	0	0	surveillance + local treatment	**13**
1	1	0	1	0	surveillance + local treatment + advanced treatment	**14**
1	1	1	0	0	surveillance + local treatment + androgenic deprivation	**15**
1	1	1	1	0	surveillance + local treatment + androgenic deprivation + advanced treatment	**16**

mCRPC = metastatic castration-resistant prostate cancer

Baulain R, Jové J, Sakr D, Gross-Goupil M, Rouyer M, Puel M, Blin P, Droz-Perroteau C, Lassalle R, Thurin NH. Clustering of prostate cancer healthcare pathways in the French National Healthcare database. Cancer Innovation. 2022. https://doi.org/10.1002/cai2.42

APPENDIX 3: Description of Optimal Matching and related methods

Optimal matching is a method to compare sequences of states. To calculate the distance between two sequences, optimal matching computes the minimum number of operations to convert sequence 1 into sequence 2. There are three operations: substitution, insertion and deletion. The cost associated to *insertion* is the same to the one associated to *deletion* because deleting a state *k* in sequence 1 leads to the same result than to insert the state *k* in sequence 2: this cost is named the "indel cost"(for "*insertion/deletion cost*").

The cost of substituting a state by another can be constant (e.g., CONSTANT method) or derived from the observed transition rates (e.g., TRATE method). All the substitution cost are put in the substitution cost cell matrix. The cost of substituting the state *i* by the state *j* is the same than the cost of substituting state *j* by the state *i*, and can be found at the intersection of line *i* and column *j*. Substituting a state *i* by itself is equal to zero.

In this study, the indel cost is always set to 1. A substitution cost matrix was created for the TRATE method. The substitution cost matrix is based on the probability of transition between states *i* and *j*; the cost will decrease as the transition between state *i* and *j* will be usual, based on the assumption that if there is a lot of transitions between two states, then these two states are quite similar.

TRATE method substitution matrix:

$$\forall\, i,j = 1,2,\dots,n \quad \begin{cases} C(i,j) = C(j,i) = \ 2 - \mathrm{p(i|j)} - \mathrm{p(j|i)} & if\ i \ \neq j \\ C(i,j) = C(j,i) = 0 & if\ i = j \end{cases}$$

Where:

- $\mathrm{C(i,j)}$ *is the cost of transition between state i and state j*
- $\mathrm{p(i|j)}$ *is the rate of transition from state j to i*
- *n is the number of distinct states*

A cost matrix corresponding to the CONSTANT method, would have constant substitution cost set to X, where x is set manually (each swap of states has a cost of

X regardless of the beginning/ending states). Then, the substitute cost matrix would have every cell set to X, except the diagonal cells equal to 0.

CONSTANT method substitution matrix:

$$\forall\, i,j = 1,2,\dots,n \quad \begin{cases} C(i,j) = C(j,i) = X & if\ i\ \neq j \\ C(i,j) = C(j,i) = 0 & if\ i = j \end{cases}$$

Reference:

Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software. 2011;40:1-37. doi:10.18637/jss.v040.i04

APPENDIX 4: Description of the silhouette metric

The silhouette metric assesses for each subject *i* belonging to the cluster *k* if the subject is on average nearer of the cluster *k* (i.e., its own cluster) or nearer of the neighbor cluster *j ≠ i* (i.e. the nearest cluster of the point *i* among the clusters different from k).

Let

- *C_k be the set of subjects belonging to the cluster k*
- *I_k the number of subjects belonging to the cluster C_k*
- *a(i) the average distance of i with its cluster*

$$a(i) = \frac{1}{I_k - 1} \sum_{j \in C_k} d(i,j)$$

- *b(i) the average distance of i with its neighbor cluster noted k'*

$$b(i) = \frac{1}{I_{k\prime}} \sum_{j\prime \in C_{k\prime}} d(i,j')$$

Each subject *i* has his own silhouette value *$s_{sil}(i)$*

$$s_{sil}(\mathrm{i}) = \frac{\mathrm{b(i)} - a(i)}{\max\,(\mathrm{a(i)}, \mathrm{b(i)})}$$

s_{sil} (i) = 1 means all points belonging to the i-cluster coincide, i.e. the subject is well classified,
s_{sil} (i) = -1 means the subject i coincides with all the points of its neighbor cluster, the subject is wrong-classified.

The quality of the cluster partition is assessed by the average of all the individual silhouettes noted $S_{sil} \in [-1;1]$

Baulain R, Jové J, Sakr D, Gross-Goupil M, Rouyer M, Puel M, Blin P, Droz-Perroteau C, Lassalle R, Thurin NH. Clustering of prostate cancer healthcare pathways in the French National Healthcare database. Cancer Innovation. 2022. https://doi.org/10.1002/cai2.42

$$S_{sil} = \frac{1}{I} \sum_{i \in C} s_{sil}(\mathrm{i})$$

Where I is the number of subjects in the whole data and C is the set of subjects.

References:

Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7

Kaufman L, Rousseeuw PJ: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, 2009