HEALTH SCIENCE ECOSYSTEMS
GRADUATE SCHOOL AND RESEARCH

Risques
Epidémiologie
Territoire
INformations
Education et
Santé

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

MODÈLES DE PREDICTION : ANALYSE DE SURVIE

P Staccini

Plan du cours

- Introduction
- Définitions
  - Cohorte
  - Cohorte incipiente
  - Événement d'intérêt
  - Durée de survie
  - Date d'origine
  - Date de point
  - Date des dernières nouvelles
  - Perdu de vue
  - Censure
  - Temps de recul
  - Temps de participation
- Fonction de survie
  - Loi exponentielle
  - Fonction de survie
  - Courbe de survie
- Estimation de la survie
  - Recueil des données
  - Calcul des durées de suivi
  - Calcul de la survie
  - Analyse actuarielle
  - Méthode de Kaplan Meier
  - Choix d'une valeur résumée
- Comparaison de deux fonctions de survie

Introduction 1 /6

- Les méthodes d'analyse de survie sont les méthodes de références pour décrire les **données longitudinales** recueillies lors d'un **suivi** de sujets ou de groupes de sujets.
- Une étude de survie est une étude :
  - longitudinale
  - prospective
  - observation d'un groupe de sujets : une cohorte

Introduction 2/6

- On rencontre un très grand nombre de situations pratiques dans lesquelles le centre d'intérêt est la **survenue d'un événement.**
- Il peut s'agir :
  - d'un décès,
  - de la survenue d'une complication après un geste opératoire,
  - de la rechute d'une maladie après une période de rémission,
  - de la disparition de symptômes sous traitement, etc.

Introduction 3/6

- La méthodologie introduite dans ce cours s'appliquera sans modification à tout type d'événement à la survenue duquel on s'intéresse.
- Cependant, pour la commodité de l'expression, on parlera généralement dans la suite de survie, considérant ainsi que l'événement d'intérêt est le décès.

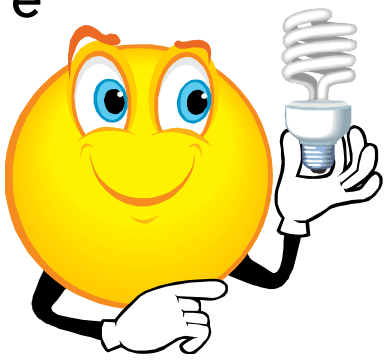
L'événement considéré doit être défini de la même manière pour tous les sujets.

Introduction 4/6

- S'intéresser à la survenue - dans le temps - d'un événement, c'est s'intéresser au **délai** de survenue de cet événement, délai compté à partir de **l'instant de référence (ou date d'origine)** :
  - en cas d'un décès pris comme évènement d'intérêt, dire d'un patient qu'il survit au moins un certain temps c'est dire que le délai de survenue du décès est supérieur à ce temps.

Introduction 5/6

- Les objectifs d'une analyse de survie sont **d'estimer** et **d'expliquer** la **durée de survie** en fonction de certains **facteurs** (on parle de **facteurs pronostiques**) et, souvent, de **comparer la survie** entre deux groupes de sujets ou plus.
  - *Un facteur pronostique est un facteur susceptible d'expliquer la survenue du décès (ou d'un autre événement) au cours du temps*



Introduction 6/6

- Au total on s'intéresse à :
  - la **probabilité de survivre au moins un certain temps t** à compter d'un instant de référence, ou encore à
  - la **probabilité pour que l'événement d'intérêt survienne après un délai t** à compter de l'instant de référence.

DEFINITIONS

Cohorte

Une cohorte est un ensemble de sujets qui vivent les mêmes événements au même moment.

- En matière de recherche médicale, c'est un ensemble de sujets inclus dans une étude au même moment, et suivis dans des conditions standardisées pendant une durée prédéfinie.

Cohorte « incipiente »

- Néologisme inspiré de l'expression anglaise « inception cohort » qui exprime le fait que la cohorte des patients qui rentrent dans l'étude doit inclure des **sujets observés au début de leur affection** à un point uniforme de l'évolution de leur maladie (« cas incidents »).

Evénement d'intérêt 1/2

- Contrairement à ce que le terme « survie » laisse penser, l'événement d'intérêt n'est pas forcément le décès, mais peut être aussi la survenue d'une maladie, la récurrence de symptômes après traitement, ou encore, en dehors d'un contexte médical, la durée de vie de composants électroniques, etc.
- En pratique, **les méthodes d'analyse de « survie » doivent donc être appliquées à chaque fois qu'il existe une notion de durée jusqu'à l'événement d'intérêt.** Dans ce cours, la terminologie « survie » sera utilisée quel que soit le type d'événement d'intérêt.

Evénement d'intérêt 2/2

- Lorsque l'événement d'intérêt est le décès :
  - on peut s'intéresser au décès toutes causes et, dans ce cas, chaque décès de patient compte comme un événement.
  - on peut également ne s'intéresser qu'au décès pour une cause spécifique (par exemple, décès par accident coronaire) et, dans ce cas, les décès d'autres causes (par exemple, décès par cancer) ne comptent pas comme un événement, mais comme une censure (cf. ci-après). Ceci n'est possible que lorsque les « autres causes de décès » sont indépendantes du phénomène étudié.

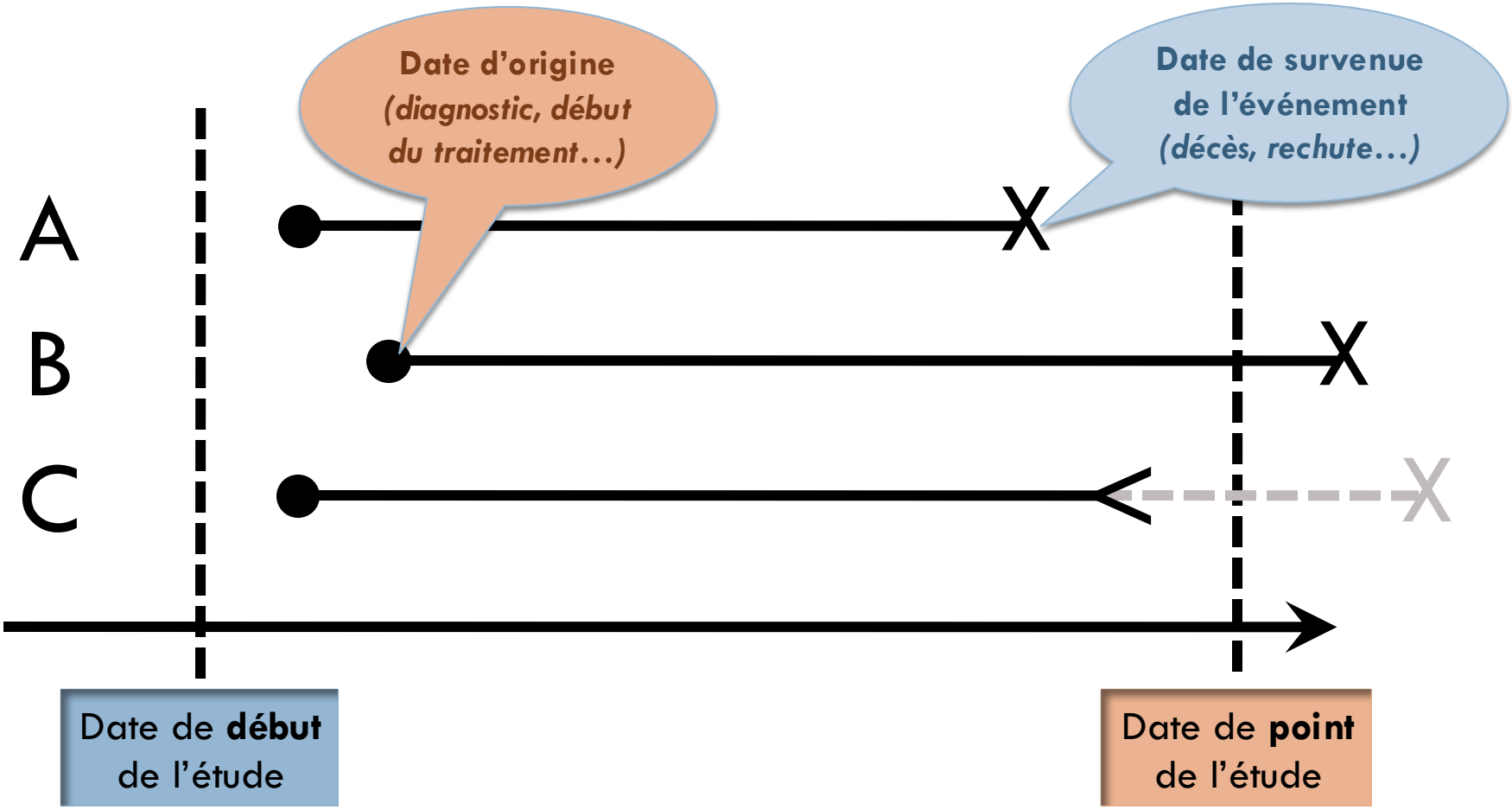
Durée de survie

- La **durée de survie** est donc un **délai entre deux dates,** mais lesquelles ?
- Prenons un exemple schématique : trois patients A, B et C sont inclus dans une étude de suivi longitudinal (étude de cohorte…)

Date d'origine

- La date d'origine est une date calendaire indiquant le **point de départ de la surveillance** : par exemple, la date de randomisation dans un essai thérapeutique.
- Cette date d'origine peut être identique ou différente pour chaque sujet en fonction des modalités d'inclusion des sujets.

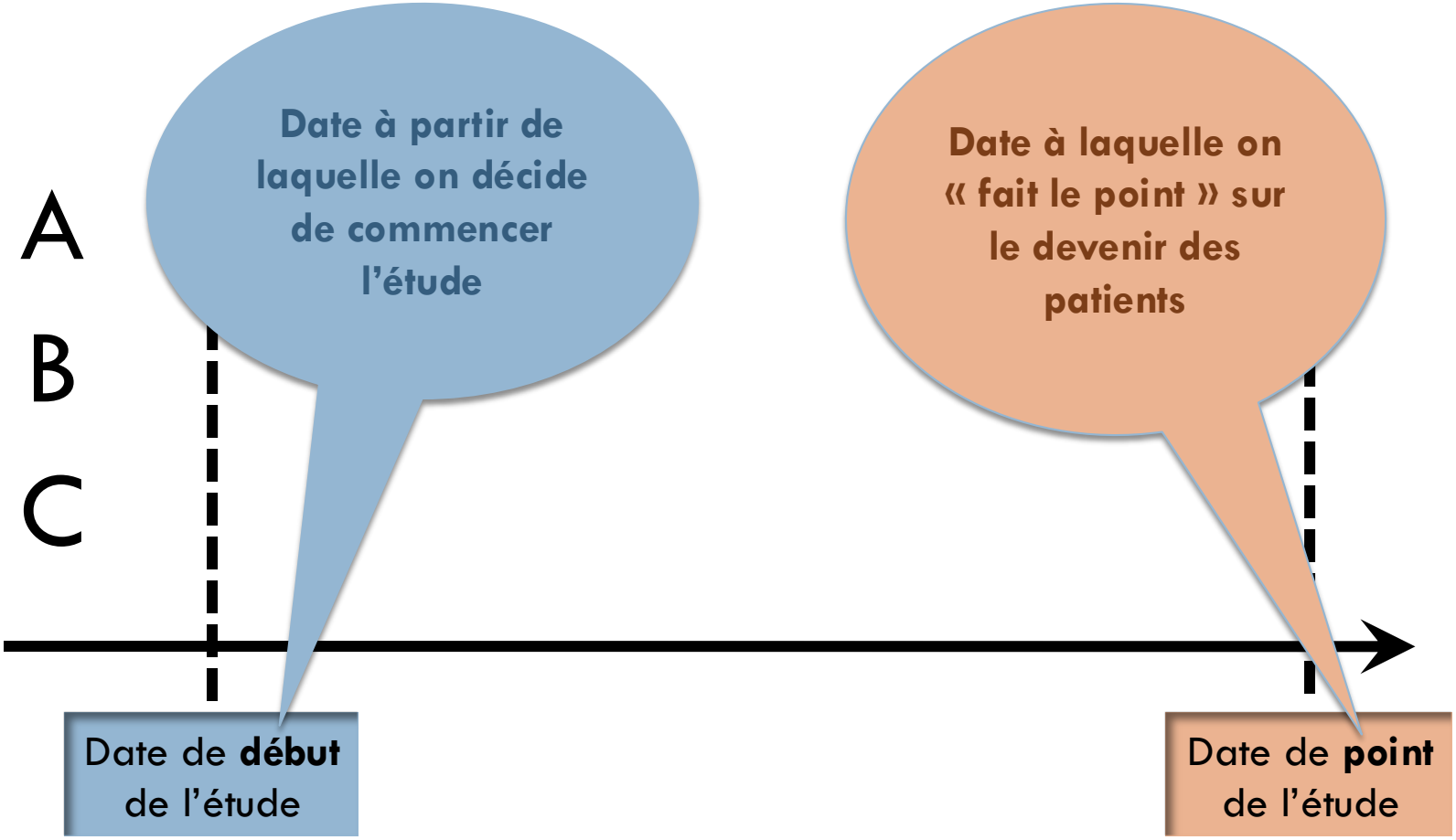
Date d'origine



Date de point (*end-point*) 1/2

- La date de point est une date fixe calendaire et correspond à la **date choisie pour faire le bilan,** au-delà de laquelle les informations recueillies ne sont plus considérées dans l'analyse

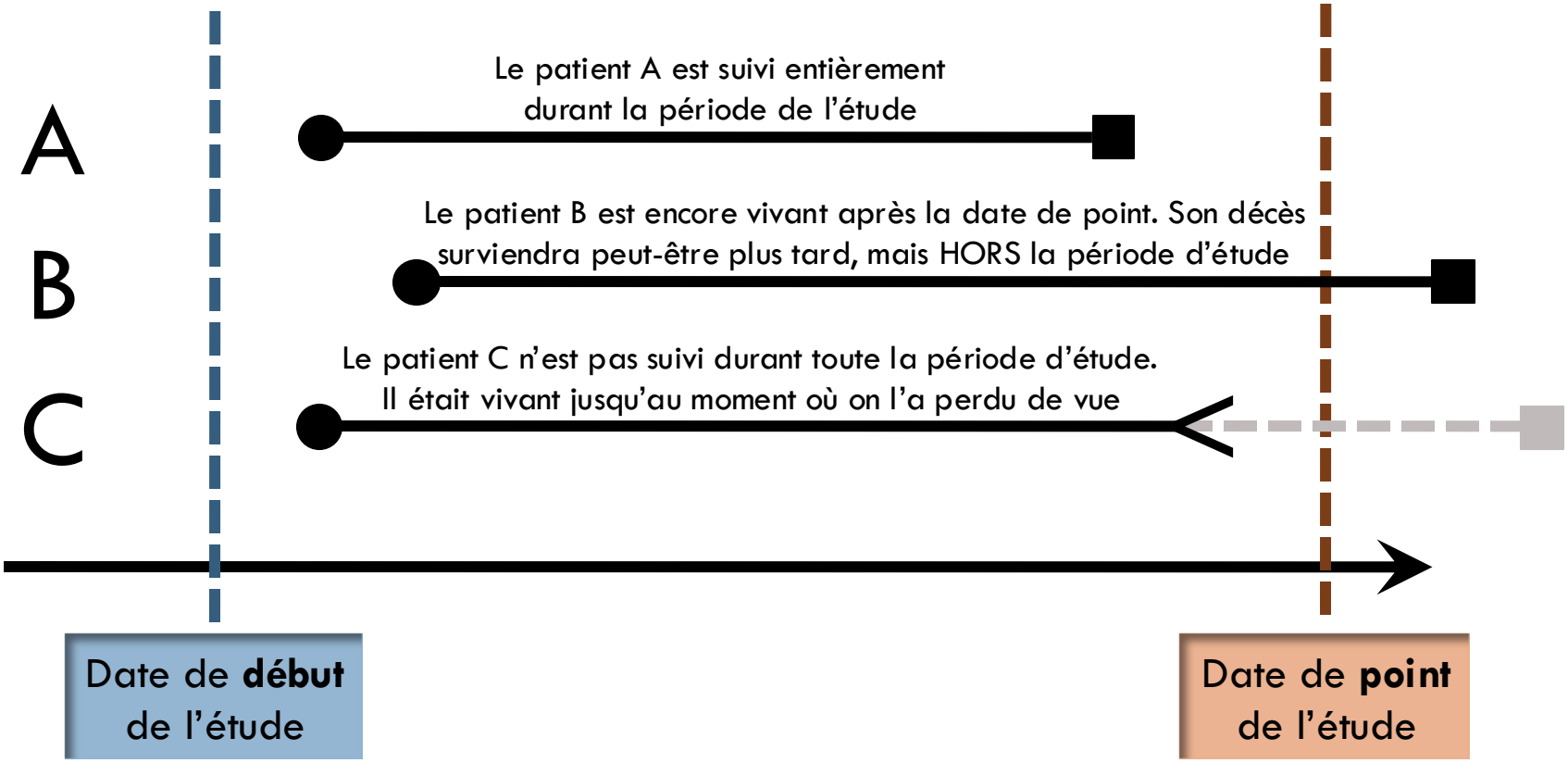
Date de point (*end-point*) 2/2



Date des dernières nouvelles 1/2

- La date de dernières nouvelles est la date la plus récente à laquelle on a recueilli des informations sur le patient, notamment sur la survenue ou non de l'événement étudié.

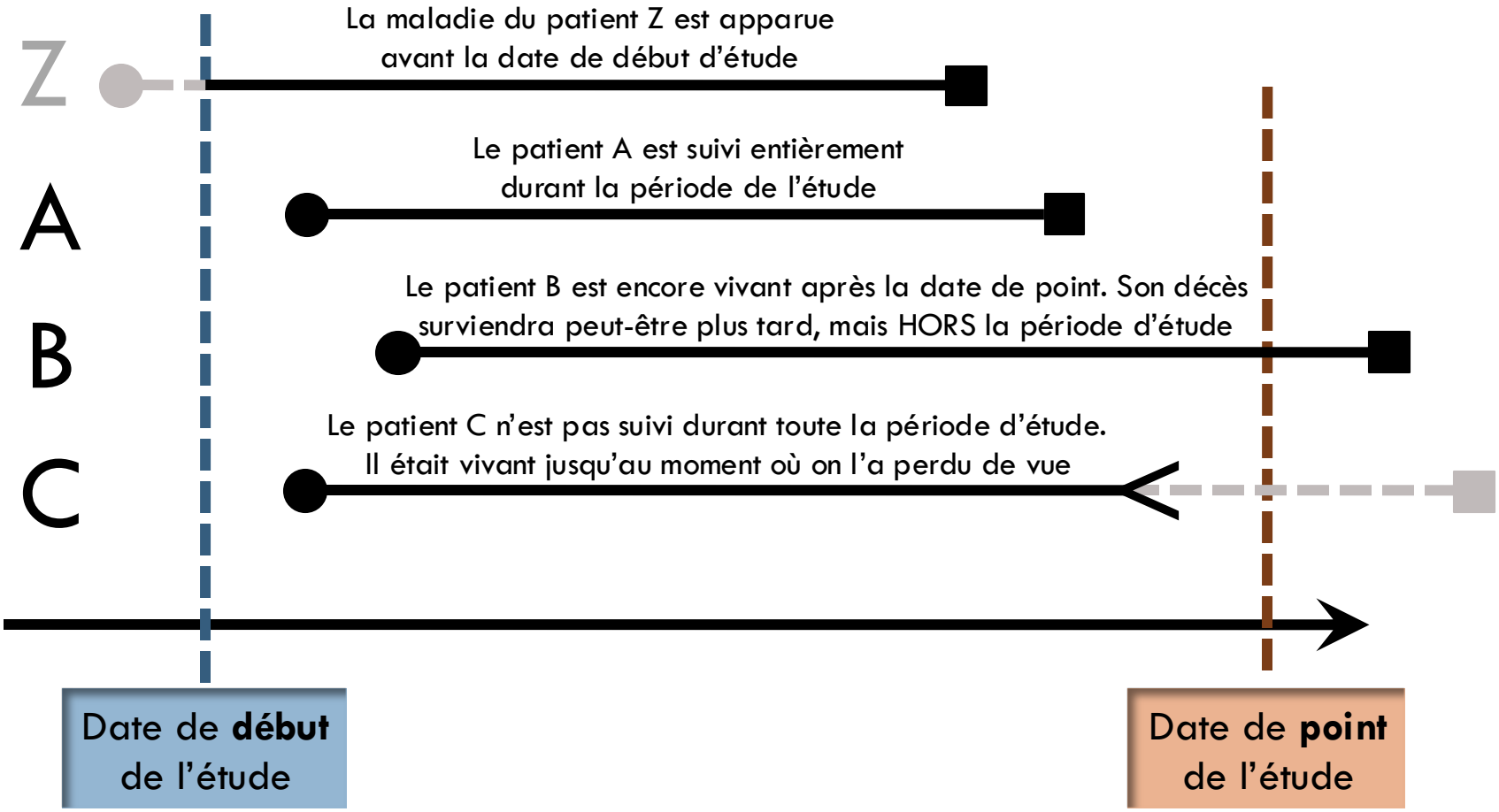
Date des dernières nouvelles 2/2



Cas particulier 1/2

- Dans certains cas, la date d'origine peut être antérieure à l'inclusion dans l'étude (on parle alors de cohorte « historique »).
- Par exemple, il peut s'agir de la date de découverte d'une hypertension artérielle dans une étude de cohorte portant sur les facteurs de risque de mortalité cardio-vasculaire.

Cas particulier 2/2



Perdu de vue (*lost of follow-up*)

- Un sujet est dit perdu de vue lorsque sa surveillance est interrompue avant la date de point et que l'événement ne s'est pas produit (cf. patient C)
- *Un cas particulier concerne les sujets inclus dans l'étude mais n'ayant fait l'objet d'aucun suivi. Ces sujets ne seront pas comptabilisés dans l'analyse. On parle alors de « perte de vue » d'emblée.*
- Dans tous les cas, il est d'usage de vérifier que le processus de perte de vue pour l'ensemble des sujets (perte de vue d'emblée, ou après une durée de suivi) n'est pas lié à l'événement d'intérêt, par exemple en comparant les caractéristiques de ces patients à celles des sujets ayant fait l'objet d'un suivi complet.

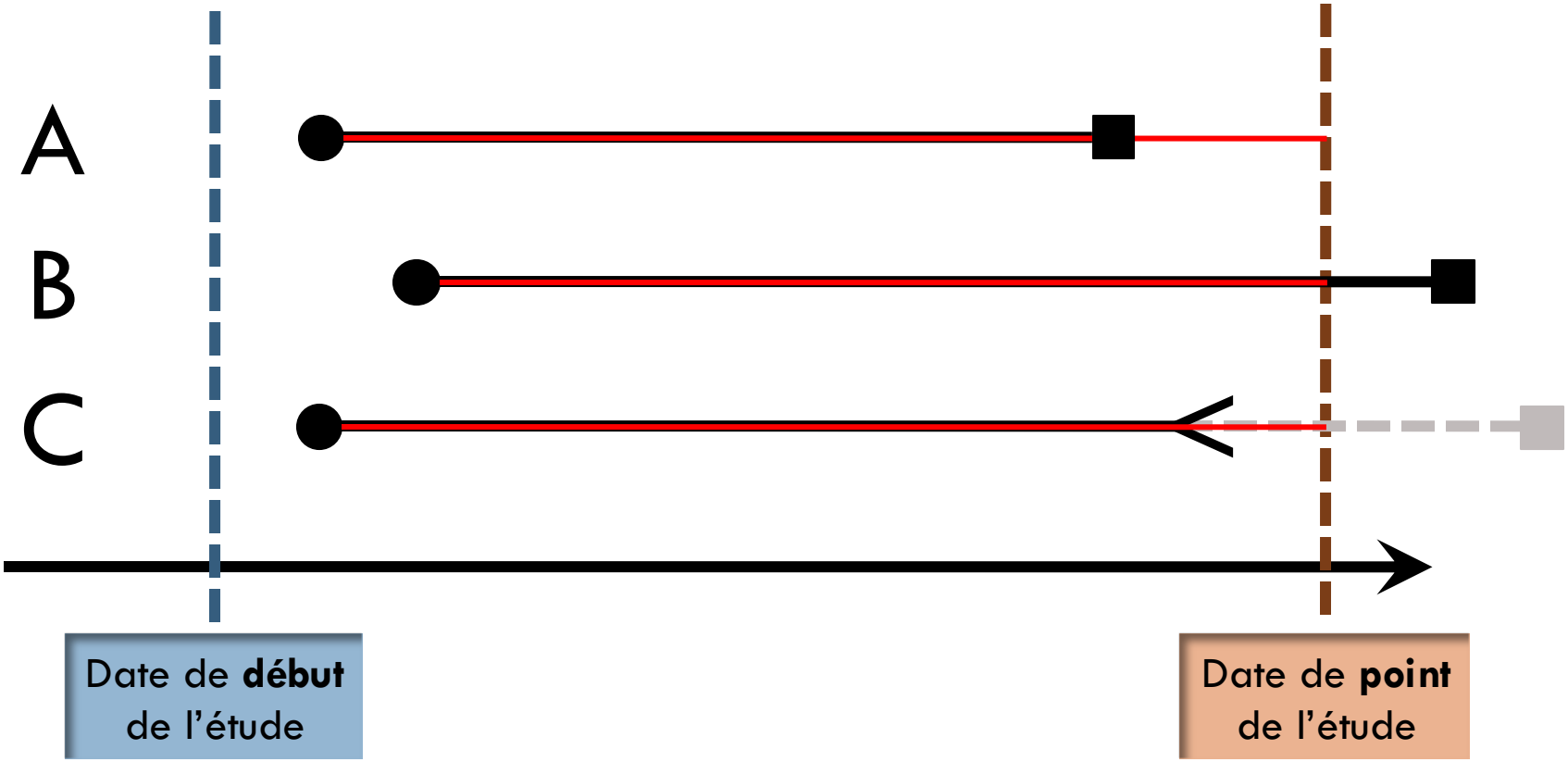
Censure (*censored data*)

- Une durée de survie d'un individu est dite censurée lorsque l'événement d'intérêt n'a pas été observé pour cet individu.
- Elle concerne donc :
  - les sujets perdus de vue (patient C)
  - les sujets vivant à la date de point (souvent appelés exclus-vivants) (patient B)
- Ces deux mécanismes de censure sont de nature différente. En effet, on ne peut assimiler les perdus de vue aux exclus-vivants, car la raison de leur « disparition » peut être liée à l'évolution de la maladie (décès méconnu de l'investigateur par exemple).

Temps de recul 1/2

- Le recul est le délai entre la date d'origine et la date de point, c'est-à-dire le délai maximum potentiel de suivi pour un sujet.
- Les reculs minimum et maximum d'une série de sujets définissent donc l'ancienneté de cette série.

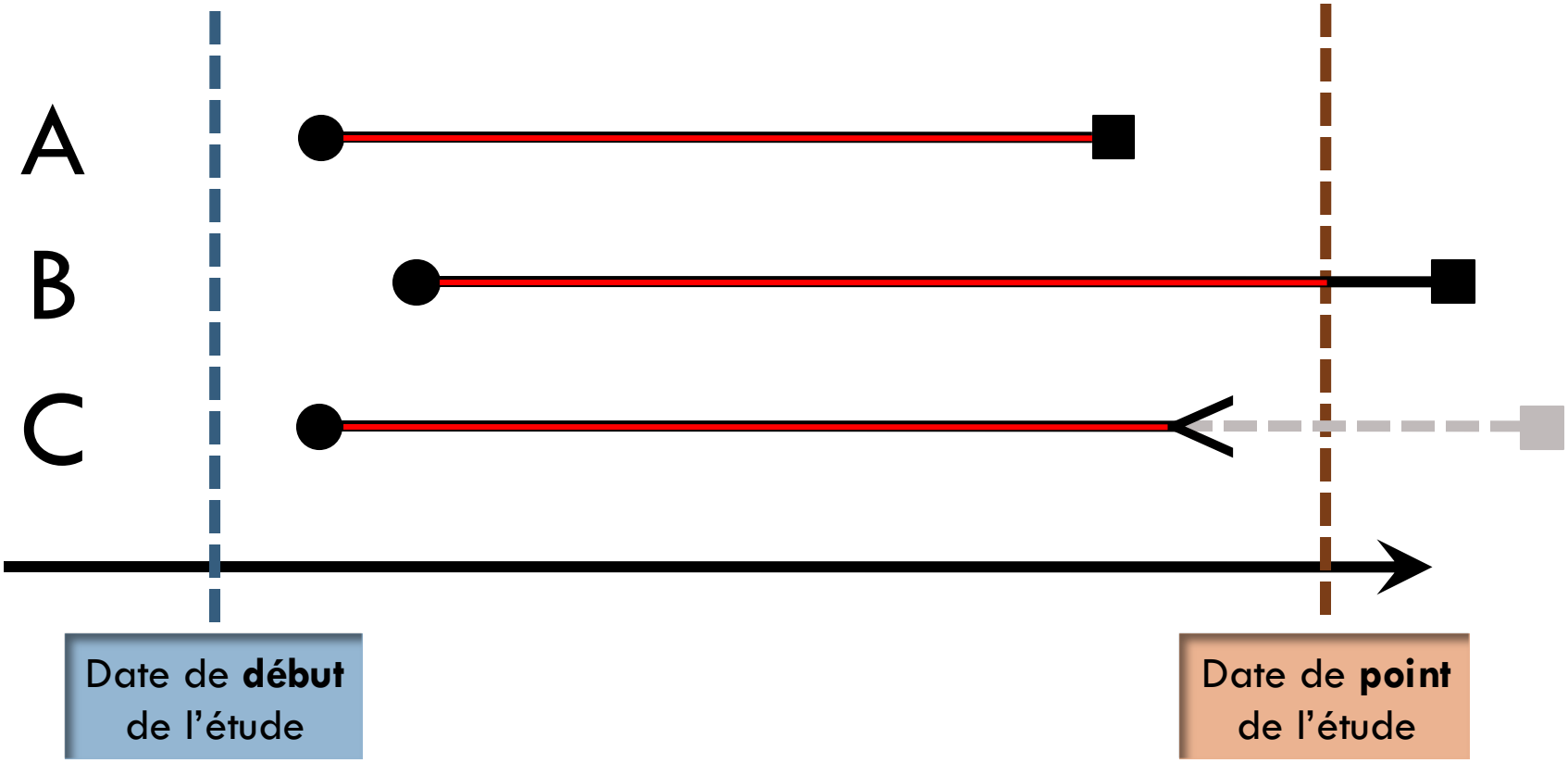
Temps de recul 2/2



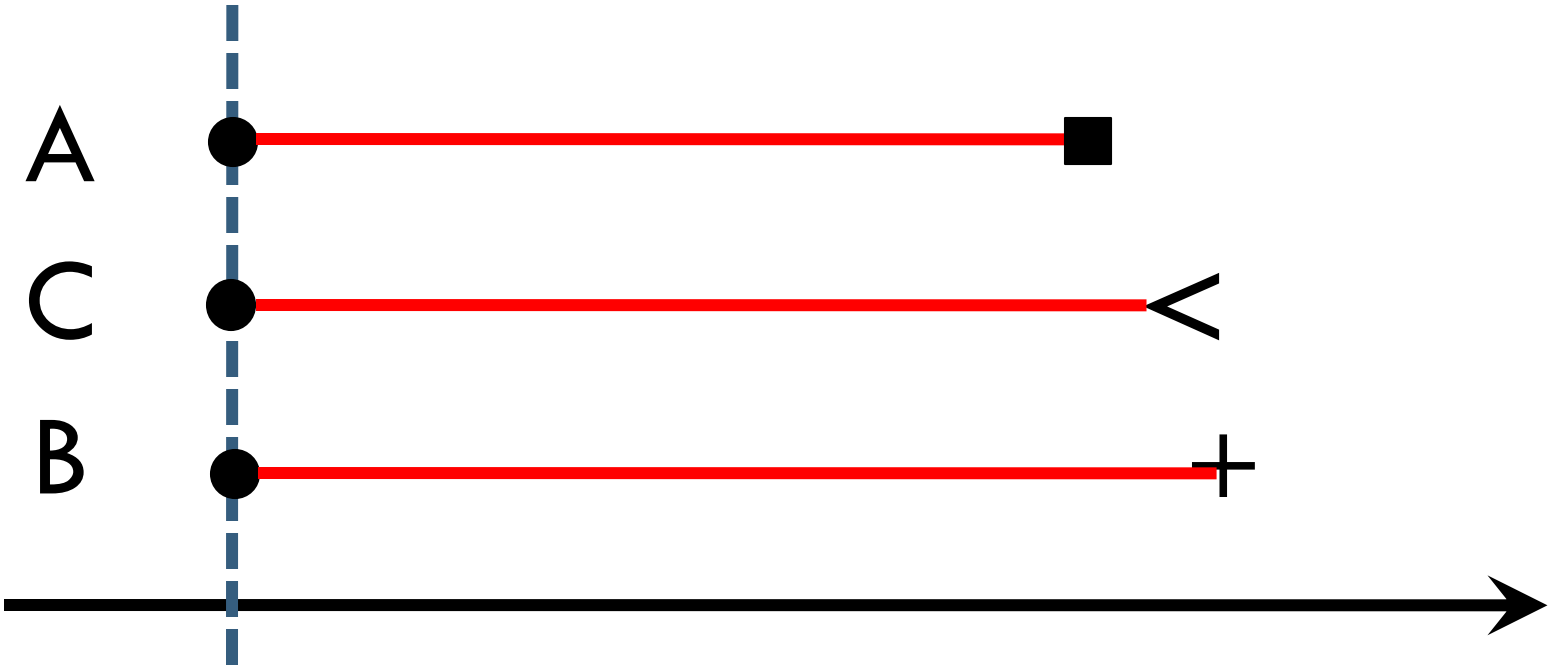
Temps de participation 1/3

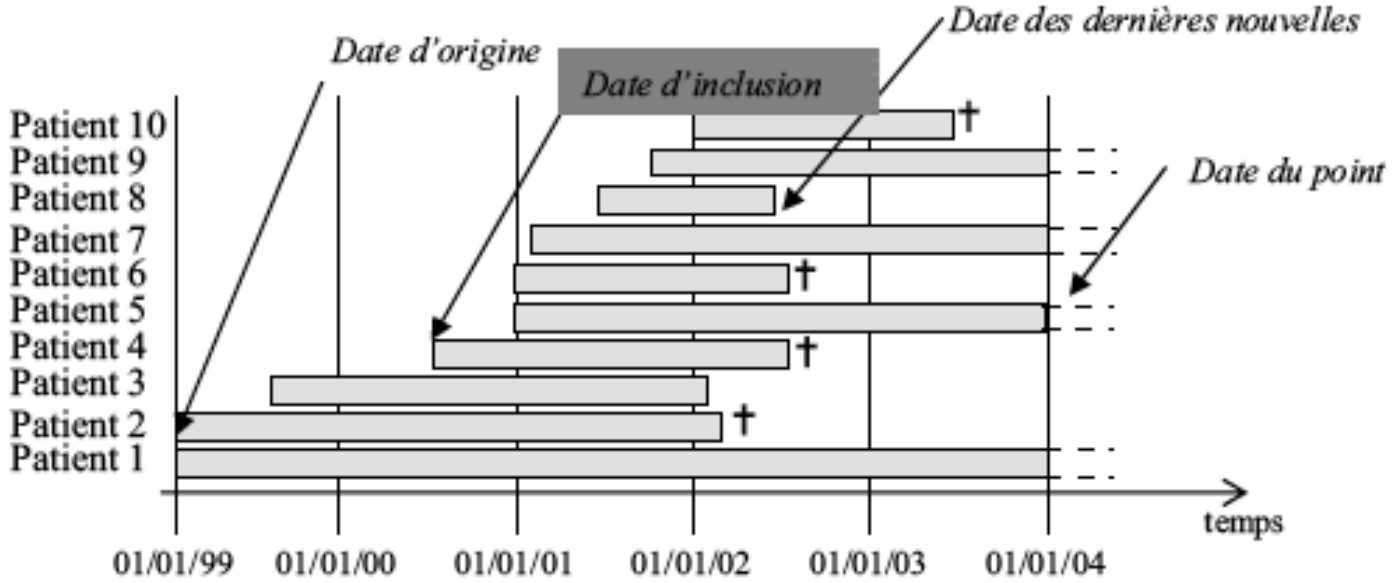
- Le temps de participation correspond à la durée de surveillance pour chaque sujet utilisée dans l'estimation de la survie.
- Trois situations peuvent se produire :
  - l'événement s'est produit au cours de la surveillance : le temps de participation est le délai entre la date d'origine et la survenue de l'événement (patient A)
  - le sujet est vivant à la date de point : son temps de participation est le délai entre la date d'origine et la date de point (patient B)
  - le sujet est perdu de vue : dans ce cas, son temps de participation est défini par le délai entre la date d'origine et la date de dernières nouvelles (patient C)

Temps de participation 2/3

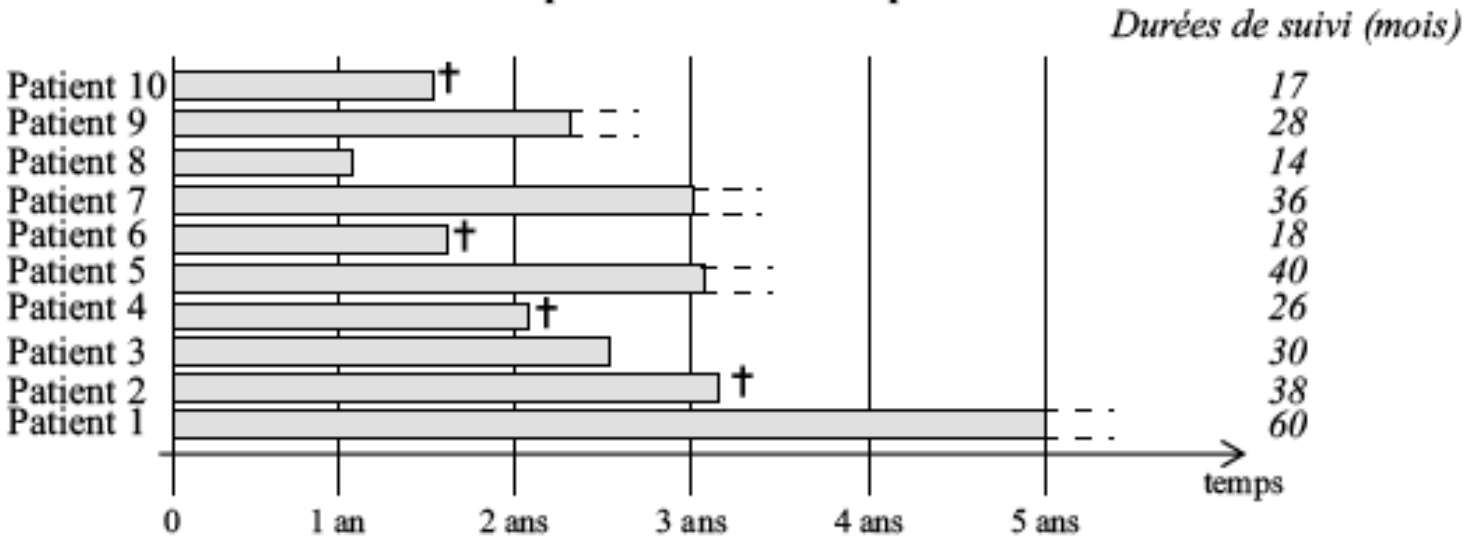


Temps de participation 3/3





Echelonnement dans le temps de l'inclusion des patients dans la cohorte



Description des durées de suivi

FONCTION DE SURVIE

Comprendre une fonction de survie

Loi exponentielle 1/3

- Rappel du cours sur les distributions de probabilité : loi de Poisson qui régit la survenue d'un événement par unité de mesure (temps, volume, surface…)
- On démontre que si un événement se réalise selon une loi de Poisson (de paramètre $\lambda=\mu=\sigma^2$), le temps entre deux réalisations consécutives de l'événement considéré est distribué selon une loi exponentielle d'espérance $1/\lambda$ (λ est appelé le taux de défaillance instantané)
- Fonction de densité de la loi exponentielle :
  - pour tout $x \geq 0$, $f(x) = \lambda e^{-\lambda x}$

Loi exponentielle 2/3

- La loi exponentielle est utilisée couramment pour représenter la durée de vie de composants ou d'équipements pour lesquels l'hypothèse d'un taux de défaillance constant au cours du temps peut être justifiée.
- Cela implique que les défaillances sont dues uniquement au hasard et qu'elles se produisent selon un processus de Poisson.

Loi exponentielle 3/3

La fonction de répartition de la loi exponentielle est donnée par l'équation :

$$F(t) = P(X \leq t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}$$

$F(t)$ représente la proportion d'équipements (de composants, etc.) qui tombent en panne avant le temps t (c'est la *fonction de "défaillance"*).
La quantité $1 - F(t)$ représente donc la quantité d'équipements qui fonctionnent pendant une durée au moins égale à t. Cette quantité est notée $S(t)$ et s'appelle la *fonction de survie* :

$$S(t) = 1 - F(t) = P(X > t) = e^{-\lambda t}$$

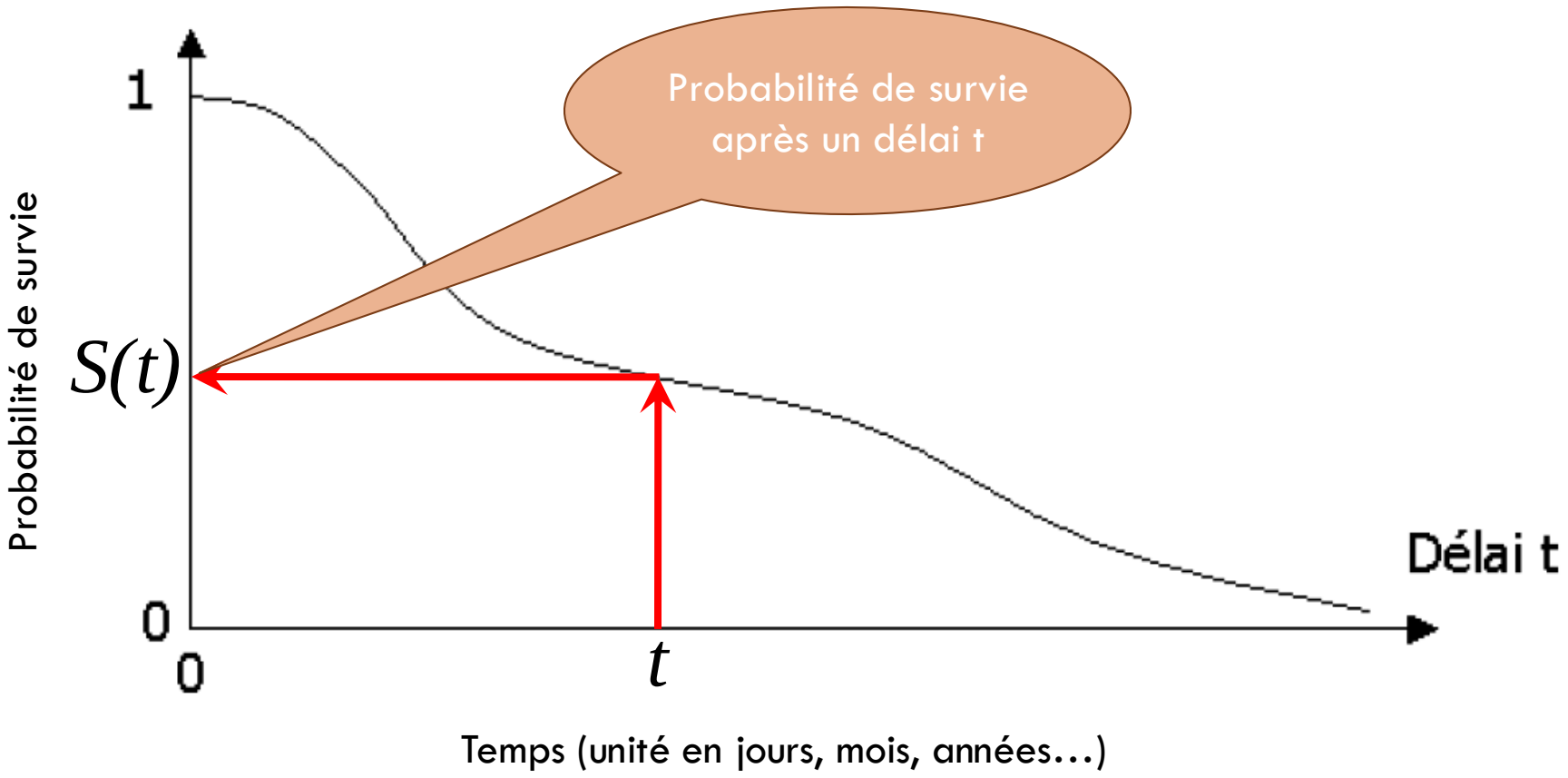
Fonction de survie 1/5

- En épidémiologie clinique, la durée résiduelle de vie d'un patient, à compter de l'instant de référence (date d'origine), est une caractéristique variable d'un patient à l'autre ; c'est donc une variable aléatoire, que nous noterons T.
- De façon analogue à la présentation de la loi exponentielle, la probabilité pour que le décès (« la défaillance ») intervienne après un délai supérieur à t est donc la probabilité pour que T soit supérieure à t :
  - $S(t) = Pr(T>t) = 1-F(t)$
  - où *F est la fonction de répartition de la durée de vie résiduelle*.
- En épidémiologie clinique, la fonction de survie est donc une fonction de répartition. On la note également S(t). Elle représente :
  - la probabilité pour qu'un patient soit encore vivant après un délai t
  - ou encore la proportion « vraie » des survivants après un délai t.

Fonction de survie 2/5

- La fonction de survie, S(t), est la probabilité que l'événement d'intérêt ne survienne pas avant la date t.
  - *S(t) = Pr (délai de survenue de l'événement d'intérêt à compter de l'instant de référence > t)*
- Si l'événement d'intérêt est le décès, c'est la probabilité de survivre au moins jusqu'à la date t.
- Si l'événement d'intérêt est la récurrence de symptômes après traitement, c'est la probabilité de survivre sans symptômes jusqu'à la date t (on parle alors de *disease free survival)*.
- La fonction de survie est représentée graphiquement par une **courbe de survie.**

Courbe de survie



Fonction de survie 3/5

La fonction de survie permet de calculer la probabilité pour que le décès survienne après un délai t_1 et avant le délai t_2 (t_2 plus grand que t_1).

Il s'agit de calculer $Pr(T \in]t_1; t_2])$.

Or : $Pr(T \in]t_1; t_2]) = F(t_2) - F(t_1) = S(t_1) - S(t_2)$

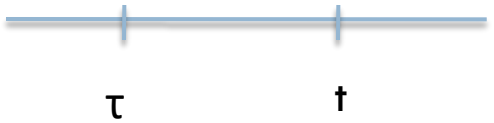
Fonction de survie 4/5

La fonction de survie donne aussi une information essentielle pour la suite : la probabilité de survivre encore après un délai t sachant que l'on est survivant après un délai τ ($\tau < t$), que l'on notera $S(t/_\tau)$. On a :

$$S(t) = Pr(X > t) \quad \text{et} \quad S(\tau) = Pr(X > \tau)$$

$$t = \tau + s \quad \text{avec} \quad s > 0$$

Or nous avons l'égalité d'événements suivante :



$$\{X > \tau + s\} \cap \{X > \tau\} = \{X > \tau + s\}$$

En appliquant la formule des probabilités composées il vient aisément que :

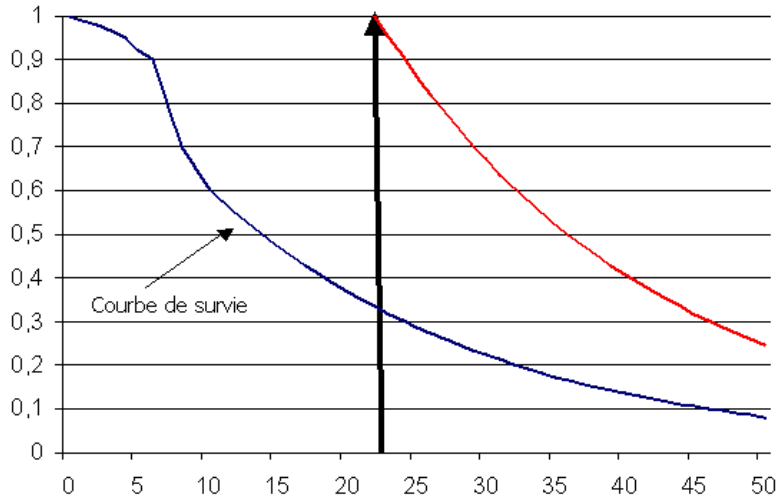
$$S(t/_\tau) = \frac{Pr((X > t) \cap (X > \tau))}{Pr(X > \tau)} = \frac{Pr((X > \tau + s) \cap (X > \tau))}{Pr(X > \tau)} = \frac{Pr(X > \tau + s)}{Pr(X > \tau)} = \frac{S(\tau + s)}{S(\tau)}$$

Au final :

$$S(t/_\tau) = \frac{S(t)}{S(\tau)}$$

Fonction de survie 5/5

Supposons que l'on veuille calculer la probabilité de survivre après (un délai de) $t = 33$ ans sachant que l'on est vivant à $t = 23$ ans.



A la lecture de la courbe de survie on remarque qu'il y a 33% de survivants à 23 ans. On lit également que à 33 ans, la proportion de survivants de la population initiale est de 20%. Mais, ne nous intéressant qu'aux survivants à 23 ans, ces 20% représentent 0,2/0,33 de la population d'intérêt, c'est-à-dire $\frac{S(33ans)}{S(23ans)}$

ESTIMATION DE LA SURVIE

Estimation à partir de données d'observation : on parle de méthodes non paramétriques

Méthode actuarielle

Méthode de Kaplan-Meier

Recueil des données 1/2

une date origine, c'est-à-dire la date à laquelle a débuté l'observation

- par exemple : la date de diagnostic du cancer broncho-pulmonaire. Cette date doit avoir un sens clinique, afin que la « survie » analysée puisse être interprétée facilement par les lecteurs ;

la date des dernières nouvelles, c'est-à-dire :

- la date de décès pour les patients décédés
- ou la date à laquelle on dispose des dernières données relatives à l'état du patient sachant qu'il n'est pas décédé ;

Recueil des données 2/2

la date de point, c'est-à-dire la date à laquelle on fait le point ou date de fin d'observation.

- Tout patient chez qui l'événement d'intérêt n'a pas été observé à la date de point est censuré à cette date. Un sujet perdu de vue à la date de point sera censuré à la date de dernières nouvelles.

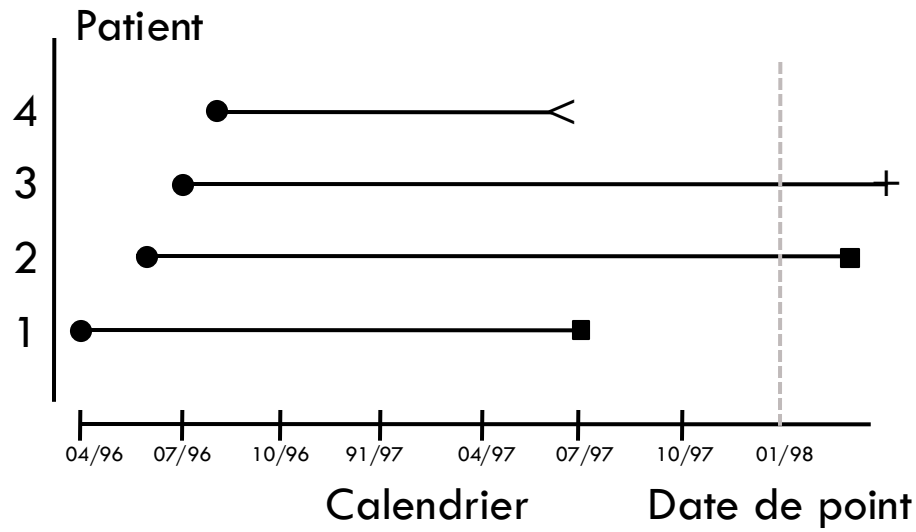
un événement « en tout ou rien » (binaire) correspondant à l'état du patient en deux éventualités (vivant ou décédé) à la date des dernières nouvelles

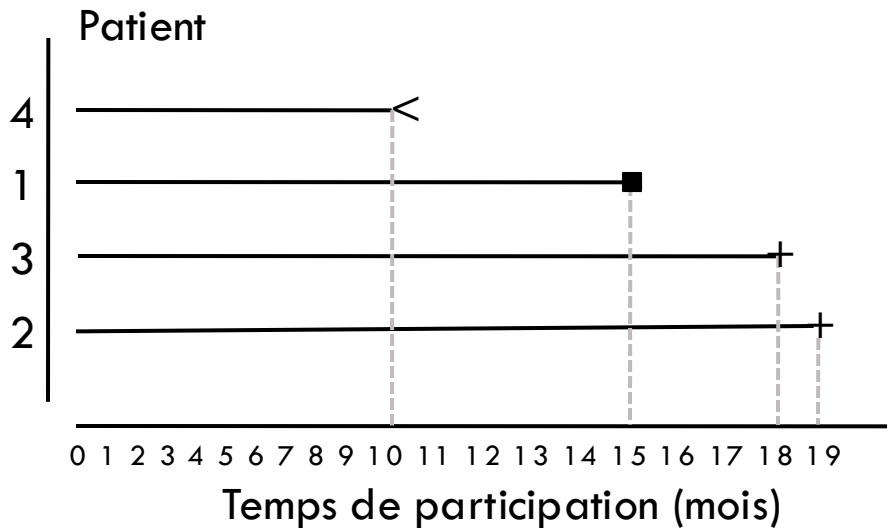
- Tout événement binaire autre que le décès associé à un délai de survenue peut être analysé en délai de survie. Par exemple on peut étudier la survenue de la rechute ou de la récidive tumorale après traitement ou la survenue de métastases.

Calcul des durées de suivi

- À partir de ces données, les durées de suivi (ou temps de participation) de chaque patient sont calculées par différence.
- Elles correspondent au délai entre la date d'origine et la date des dernières nouvelles qui sera :
  - la date de décès en cas de décès,
  - la date de point pour les patients vivants pour lesquels le suivi est assuré
  - ou la date de perte de vue pour les patients vivants n'étant plus suivi dans la cohorte à la date de point.

Exemple





● Date d'origine ■ Patient décédé + Patient vivant (censuré) < Patient perdu de vue (censuré)

Patient	Date d'origine	Date événement ou DDN	Etat aux dernières nouvelles	Etat à la date de point	Temps de participation	Temps de recul
1	04/96	07/97	Décédé	Décédé	15	21
2	06/96	03/98	Décédé	Vivant	19	19
3	07/96	04/98	Vivant	Vivant	18	18
4	08/96	06/97	Vivant	?	10	17

Calcul de la survie 1/3

- Si aucune variable n'est censurée, la fonction de survie se calcule par le pourcentage de survivants en fonction du temps.
- Cependant, cela ne se produit jamais, car un certain nombre de sujets seront perdus de vue, et un certain nombre seront encore vivants à la date de point.

Calcul de la survie 2/3

- Deux méthodes d'analyse de survie sont de préférence utilisées : l'analyse actuarielle et la méthode de Kaplan-Meïer, qui sont deux **méthodes non paramétriques** (*non-parametric* ou *distribution-free*), puisqu'elles ne nécessitent aucune hypothèse sur la distribution des temps de survie.
- **L'analyse actuarielle** est moins utilisée que la méthode de Kaplan-Meïer, et s'applique principalement lorsqu'il y a un *grand nombre de sujets* (plus de 200 par groupe) et de nombreux événements.
- **La méthode de Kaplan-Meïer** est donc la méthode de choix pour les *échantillons de taille plus réduite*.

Calcul de la survie 3/3

- Ces deux méthodes supposent une hypothèse forte : les probabilités de survie sont supposées indépendantes du calendrier. Ceci revient à supposer, par exemple, que la survie à 1 an d'un groupe de patients inclus en 1970 est identique à celle d'un groupe de patients inclus en 1990. Cette hypothèse n'est pas forcément vérifiée pour les études disposant d'un recul maximum très important, notamment en raison des progrès thérapeutiques vis-à-vis de la maladie étudiée.
- La fonction de survie estimée peut être résumée soit par le taux de survie à un délai fixé (1 an, 5 ans, etc.), soit par une valeur de durée : médiane de survie (*median survival time*) et quantiles (percentiles).

Analyse actuarielle 1/5

- La fonction de survie est calculée sur des **intervalles de temps fixés a priori** (mois, trimestre, semestre, année...).
- Schématiquement, le mode de calcul est le suivant.
- Pour chaque intervalle de temps (par exemple [0,1an], [1,2ans], etc.), on définit :
  - le nombre de sujets vivants au début de l'intervalle, V ;
  - le nombre de sujets décédés dans l'intervalle, D ;
  - le nombre de sujets vivants aux dernières nouvelles, dont le temps de participation s'arrête dans l'intervalle (censure), C.
    - l'hypothèse actuarielle (*actuarial assumption*) suppose que ces sujets sont exposés au risque d'événement sur la moitié de l'intervalle (6 mois dans notre exemple).

Analyse actuarielle 2/5

- Le nombre de sujets exposés au risque d'événement sur l'intervalle est : N = V - (C/2)
- La probabilité d'événements durant l'intervalle est simplement estimée par le rapport du nombre d'événements sur le nombre de sujets à risque : D / N
- La survie sur cet intervalle est : (N - D) / N
  - cette probabilité est appelée **survie instantanée.**
- La fonction de survie est obtenue en faisant le **produit des survies instantanées sur l'ensemble des intervalles** : par exemple, la survie à 3 ans = (survie instantanée entre 2 et 3 ans) x (survie instantanée entre 1 et 2 ans) x (survie instantanée entre 0 et 1 an) (cf. diapo 39)

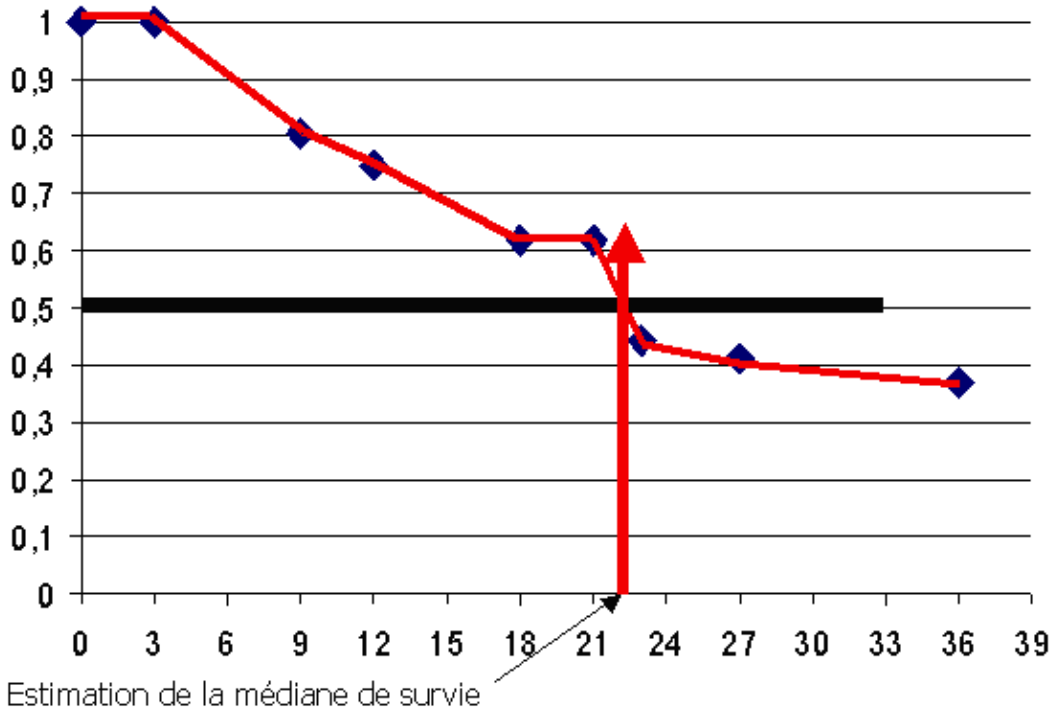
Analyse actuarielle 3/5

Instants	V	C	D	N = V – C/2	(N - D) / N	S(t)
0	-	-	-	-	-	1
3	210	0	0	210	**1**	1x1 = **1**
6	210	10	40	210 - 5 = 205	(205-40)/205 = **0,805**	0,805 x 1 = **0,805**
9	160	30	10	160 - 15 = 145	(145-10)/145 = **0,931**	0,931 x 0,805 = **0,749**
12	120	10	20	120 – 5 = 115	(115-20)/115 = **0,826**	0,826 x 0,749 = **0,619**
15	90	20	0	90 – 10 = 80	**1**	1 x 0,619 = **0,619**
18	70	0	20	70	(70-20)/70 = **0,714**	0,714 x 0,619 = **0,442**
21	50	18	3	50 – 9 = 41	(41-3)/41 = **0,927**	0,927 x 0,442 = **0,410**
24	29	8	4	29 – 4 = 25	(25-2)/25 = **0,920**	0,920 x 0,410 = **0,377**

V : nombre de sujets vivants au début de l'intervalle
C : nombre de sujets vivants censurés dans l'intervalle
D : nombre de sujets décédés dans l'intervalle
N : nombre de sujets exposés au risque de décès

Analyse actuarielle 4/5

Pour chaque intervalle de temps, on représente l'estimation de la survie S(t) par un point. Les coordonnées du premier point sont 0 (j0) en abscisse, et 1 (100 %) en ordonnée. Tous les points consécutifs sont reliés par un segment de droite.



Analyse actuarielle 5/5

- L'inconvénient majeur de cette méthode est qu'elle estime la survie à chaque borne supérieure des intervalles constitués a priori, et considère chaque censure, survenant dans un intervalle, de manière équivalente, c'est-à-dire qu'un sujet suivi pendant 21 jours apporte la même information qu'un sujet suivi pendant 29 jours pour la survie à 30 jours dans l'exemple présenté. C'est la raison pour laquelle cette méthode est **à réserver à de grands échantillons.**

Méthode de Kaplan-Meïer 1/3

- Contrairement à l'analyse actuarielle, les intervalles ne sont pas fixés a priori, mais sont définis par les instants auxquels les événements sont observés.
- Ces intervalles sont donc inégaux, débutent à l'instant d'un événement et s'arrêtent juste avant l'événement suivant.
- Pour chaque intervalle entre deux événements, on définit V, D et C comme précédemment (avec la particularité que D vaut souvent 1, sauf dans le cas où plusieurs événements surviennent au même temps de participation).
- Dans l'analyse de Kaplan-Meïer, N = V - C et la probabilité de survie instantanée calculée sur cet intervalle vaut : (N - D) / N
- L'estimation de Kaplan-Meïer de la fonction de survie s'obtient, comme dans l'analyse actuarielle, en faisant le produit des survies instantanées.

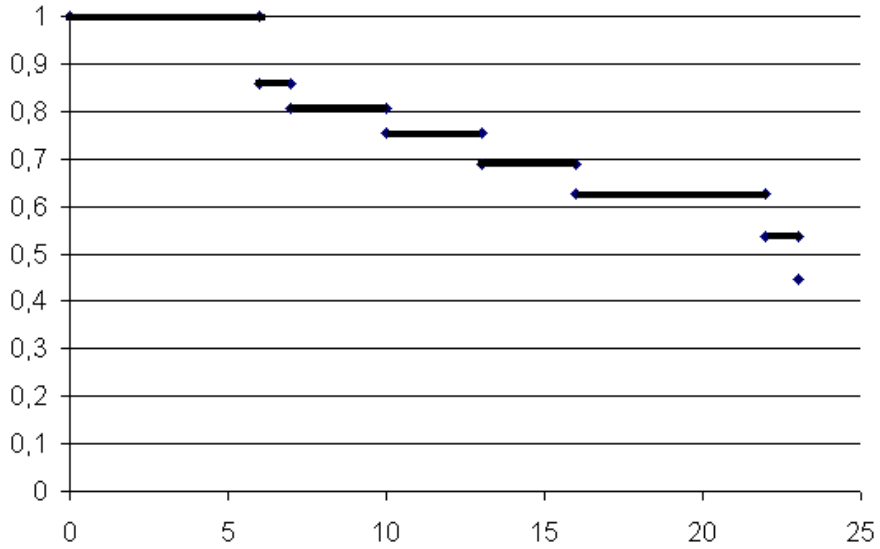
Méthode de Kaplan-Meïer 2/3

Instants	V	C	D	N = V - C	(N - D) / N	S(t)
0	21	-	-	-	-	1
6	21	0	3	21	**0,857**	**0,857**
7	18	1	1	17	**0,941**	**0,807**
10	16	1	1	15	**0,933**	**0,753**
13	14	2	1	12	**0,917**	**0,690**
16	11	0	1	11	**0,909**	**0,627**
22	10	3	1	7	**0,857**	**0,537**
23	6	0	1	6	**0,833**	**0,448**

V : nombre de sujets vivants au début de l'intervalle
C : nombre de sujets vivants censurés dans l'intervalle
D : nombre de sujets décédés dans l'intervalle
N : nombre de sujets exposés au risque de décès

Méthode de Kaplan-Meïer 3/3

La courbe de survie se compose de paliers successifs, où les probabilités de survie sont constantes entre deux temps d'événements consécutifs. Le premier palier vaut 1 depuis l'origine jusqu'au délai de survenue du premier événement. Il s'abaisse ensuite à la première valeur calculée pour constituer un second palier jusqu'au délai de survenue de l'événement suivant, etc. Il est possible de relier les paliers successifs par des segments verticaux, mais il n'est pas correct de les relier par des segments obliques. La courbe ainsi obtenue présente une allure en « marches d'escalier ».



Choix d'une valeur résumée

- **Médiane de survie**
  - La courbe de survie apporte des renseignements importants, mais il est utile de disposer d'indicateurs synthétiques ou résumés de cette courbe. La moyenne de survie n'est pas un bon indicateur, pour des raisons d'ordre statistique, notamment liées à l'existence de censures.
  - La médiane de survie lui est préférée. Elle représente la durée t pour laquelle la probabilité de survie S(t) est de 50 %. À cause de la distribution par paliers de la fonction de survie, il est souvent impossible de connaître la durée correspondant à une survie exacte de 50 %. En pratique, la médiane est estimée par la plus petite durée pour laquelle la survie est inférieure à 50 %.
  - Il arrive que la fonction de survie soit toujours supérieure à 50 %. Dans ce cas, la médiane ne peut être estimée. On estime alors les quantiles (NB : un quartile = 25 %) : pour le $p^{ième}$ quantile on estime la durée pour laquelle la probabilité de survie est de 100-p. Par exemple, le 25^{e} quantile (ou 1^{er} quartile) correspond à la plus petite durée pour laquelle la survie est inférieure à 75 %.
- **Survie à date fixée**
  - Un autre indicateur fréquemment utilisé pour résumer l'information d'une courbe de survie est l'estimation de la survie à un temps donné (survie à 5 ans par exemple…)

Comparaison de deux fonctions de survie

Méthode du log-rank

Contexte

- Il arrive fréquemment que l'on souhaite montrer qu'une action (intervention, traitement) ou une classification ont un lien avec la survie.
- Il s'agira de conduire une étude comparative et de mettre en oeuvre un test d'hypothèses.
- Le principe du test du log-rank (*ou test de Mantel-Cox ou de Peto-Mantel-Haenszel*) est de comparer, dans chaque groupe, le nombre observé et le nombre attendu d'événements si la survie était identique dans les deux groupes, sur l'ensemble de la période étudiée.

Attention 

- Une erreur importante, et souvent retrouvée, consiste à assimiler l'efficacité du traitement à la réponse des patients à ce traitement, et à comparer la survie, non plus entre les patients traités et les patients non traités, mais entre les sujets qui répondent au traitement et les sujets qui ne répondent pas (*comparison of survival by response*).
- Cette méthode est à proscrire et peut provoquer des biais et des conclusions fausses :
  - les sujets répondeurs sont en général en meilleure santé que les sujets non répondeurs et sont donc susceptibles - indépendamment de tout traitement - de vivre plus longtemps ;
  - la comparaison de la survie par la réponse au traitement peut être biaisée puisque les patients doivent vivre suffisamment longtemps pour avoir la possibilité de répondre au traitement (*guarantee-time bias*).

Principe du test du log-rank

- Pour chaque intervalle de temps (qu'il s'agisse de l'analyse actuarielle ou de Kaplan-Meier), le nombre attendu d'événements, sous l'hypothèse nulle d'égalité de la survie entre les deux groupes, s'obtient en appliquant, au nombre de sujets exposés au risque d'événements, la proportion d'événements observés sur l'ensemble des deux groupes.
- Le test du log-rank, évaluant l'écart entre le nombre observé et le nombre attendu d'événements sur les deux groupes, est un chi carré à 1 degré de liberté (ddl).
- Ce test est généralisable au cas de k groupes et permet de tester si globalement la survie est différente entre les groupes.

Exemple

- Imaginons par exemple que l'on souhaite faire la preuve qu'un traitement adjuvant à la chirurgie dans le carcinome hépatocellulaire améliore la survie des patients. Les grands traits de l'étude sont les suivants :
  - la survie sera comptée à partir de la date de la chirurgie.
  - des patients ont été inclus pendant une année dans une étude qui a duré 3 ans et répartis par tirage au sort dans un des deux groupes de traitement : chirurgie seule (groupe A) ou chirurgie + traitement adjuvant (groupe B).
  - la durée de suivi des patients (durée de participation à l'étude ou recul) varie d'un patient à l'autre
- A la fin de l'étude on dispose pour chaque patient :
  - du groupe auquel il a appartenu, A ou B
  - des temps de suivi pour chaque patient selon son groupe et selon le fait que le patient soit décédé ou bien que le patient soit censuré, qu'il soit encore vivant ou perdu de vue.

Exemple

- Supposons que l'on dispose des observations suivantes.
  - Dans le groupe A, les tA_i et tA_i* sont : 1; 1; 2; 2; 3; 4; 4; 5; 5; 8; 8; 8; 8; 11; 11; 12; 12; 15; 17; 22; 23
  - Dans le groupe B, les tB_i et tB_i* sont : 6; 6; 6; 6,1*; 7; 9*; 10; 10,1*; 11,2*; 13; 16; 17,3*; 19*; 20*; 22; 23; 25*; 32*; 32*; 34*; 35*
- Les ensembles des tA_i et tB_i (patients décédés) constituent l'ensemble des temps de décès observés, quelque soit le groupe ; on les notera t_i et on les considérera ordonnés par valeurs croissantes. Ici les t_i sont : 1; 2; 3; 4; 5; 6; 7; 8; 10; 11; 12; 13; 15; 16; 17; 22; 23

Hypothèses

- H_0 : les fonctions de survie sont les mêmes dans les deux populations d'où sont issus les groupes A et B. $S_A(t)=S_B(t)$
- H_1 : les deux fonctions de survie diffèrent

Estimation des décès 1/2

- Le principe est d'abord d'estimer, tous groupes confondus, **la probabilité de décéder à t_i sachant que l'on est vivant à t_{i-1}**, c'est-à-dire estimer $(1- S(t_i / t_{i-1}))$ et ceci pour chacun des temps de décès observés t_i.
- On utilise ici l'estimateur de Kaplan-Meïer de $S(t_i / t_{i-1})$.
- On obtient ainsi la dernière colonne du tableau ci-après.

Estimation des décès 2/2

t_i	V	C	N = V - C	D	$S(t_i / t_{i-1}) = (N - D) / N$	$1 - S(t_i / t_{i-1})$
1	42		42	2	0,952	0,048
2	40		40	2	0,950	0,050
3	38		38	1	0,974	0,026
4	37		37	2	0,946	0,054
5	35		35	2	0,943	0,057
6	33		33	3	0,909	0,091
7	30	1	29	1	0,966	0,034
8	28		28	4	0,857	0,143
10	24	1	23	1	0,957	0,043
11	22	1	21	2	0,905	0,095
12	19	1	18	2	0,889	0,111
13	16		16	1	0,938	0,062
15	15		15	1	0,933	0,067
16	14		14	1	0,929	0,071
17	13		13	1	0,923	0,077
22	12	3	9	2	0,778	0,222
23	7		7	2	0,714	0,286

Calcul des décès attendus 1/3

- On estime ensuite le nombre de décès que l'on attend dans chacun des groupes A et B, à chaque t_i, en supposant que la probabilité conditionnelle de décès estimée s'applique identiquement à chacun des deux groupes.
- Pour cela on évalue à chaque t_i l'effectif à risque à cette date.
- On obtient les deux dernières colonnes du tableau suivant.

Calcul des décès attendus 2/3

t_i	V	C	N = V - C	D	$S(t_i / t_{i-1}) = (N - D) / N$	$1 - S(t_i / t_{i-1})$	N_A	N_B	E_A	E_B
1	42		42	2	0,952	0,048	21	21	1,000	1,000
2	40		40	2	0,950	0,050	19	21	0,950	1,050
3	38		38	1	0,974	0,026	17	21	0,447	0,553
4	37		37	2	0,946	0,054	16	21	0,864	1,136
5	35		35	2	0,943	0,057	14	21	0,799	1,201
6	33		33	3	0,909	0,091	12	21	1,092	1,988
7	30	1	29	1	0,966	0,034	12	17	0,408	0,579
8	28		28	4	0,857	0,143	12	16	1,714	2,286
10	24	1	23	1	0,957	0,043	8	15	0,344	0,656
11	22	1	21	2	0,905	0,095	8	13	0,760	1,240
12	19	1	18	2	0,889	0,111	6	12	0,666	1,334
13	16		16	1	0,938	0,062	4	12	0,249	0,751

Calcul des décès attendus 3/3

- Ces nombres sont notés E_{Ai} et E_{Bi}. On remarque que l'on utilise ici, comme toujours, la justesse supposée de l'hypothèse nulle puisque les probabilités de décès, et donc la survie, sont supposées ne pas dépendre du groupe.
- Sous l'hypothèse nulle ces nombres doivent être voisins des nombres de décès réellement observés. En particulier le total de ces nombres de décès au cours du temps (noté E_A et E_B selon le groupe) doit être voisin du nombre total de décès observés (noté D_A et D_B selon le groupe), et ceci dans chacun des groupes.
- Dans l'exemple, on obtient : $E_A=10,74$; $E_B=19,26$; $D_A=21$; $D_B=9$.

Test du Chi^2

Le paramètre du test est construit à partir de ces quatre valeurs (aléatoires normalement à ce stade de la construction) :

$$Q_c = \frac{(D_A - E_A)^2}{E_A} + \frac{(D_B - E_B)^2}{E_B}$$

Sous H_0, Q suit une distribution de χ^2 à un degré de liberté
Condition de validité : E_A et $E_B \geq 5$
On construit l'intervalle de pari de niveau 0,95 : $IP_{0,95} = [0; 3,84]$
On met en place la règle de décision. Si la valeur calculée $Q_c \in [0; 3,84]$, on ne pourra conclure à une différence entre les fonctions de survie dans les deux population considérées. Si la valeur Q_c excède $3,84$ on conclura au risque de 5% que les fonctions de survie diffèrent.
Dans l'exemple traité, on obtient $Q_c = 15,26$. On rejette donc l'hypothèse d'égalité des fonctions de survie. La survie est meilleure dans le groupe dans le quel $D < E$, c'est le groupe B. La preuve est faite (au risque d'erreur de 5%) que le traitement adjuvant améliore la survie des patients à compter de la date de chirurgie.

Exercices

Exercice 1a

On a suivi le devenir d'un grand groupe de malades atteints d'une maladie M, à partir de la date de diagnostic. On considère alors que l'on dispose des probabilités suivantes : au bout d'un an, 20 % des malades sont morts ; au bout de 2 ans, 50 % des malades sont morts ; au bout de 3 ans, 70 % des malades sont morts ; au bout de 4 ans, 80 % des malades sont morts ; au bout de 5 ans, 80 % des malades sont morts. **La probabilité qu'un malade ayant déjà survécu 2 ans survive moins de 3 ans est :**

- A. 20 %
- B. 30 %
- C. 40 %
- D. 50 %
- E. 60 %

Exercice 1a

On a suivi le devenir d'un grand groupe de malades atteints d'une maladie M, à partir de la date de diagnostic. On considère alors que l'on dispose des probabilités suivantes : au bout d'un an, 20 % des malades sont morts ; au bout de 2 ans, 50 % des malades sont morts ; au bout de 3 ans, 70 % des malades sont morts ; au bout de 4 ans, 80 % des malades sont morts ; au bout de 5 ans, 80 % des malades sont morts. **La probabilité qu'un malade ayant déjà survécu 2 ans survive moins de 3 ans est :**

- A. 20 %
- B. 30 %
- **C. 40 % : (1-0,2) x (1-0,5) = 0,8 x 0,5 = 0,4**
- D. 50 %
- E. 60 %

Exercice 2a

On s'intéresse à une population de femmes atteintes d'un cancer du sein. Le taux de survie 5 ans après la découverte du cancer est de 65 %. Lors de la découverte du cancer, on peut définir la gravité du cancer par son stade (1 à 4). 45 % des femmes sont de stade 1, 30 % de stade 2, 15 % de stade 3, et 10 % de stade 4. La probabilité qu'une femme de cette population soit de stade 4 et survive au moins 5 ans est 0,03. **Quel est le % de femmes de stade 4 qui décèdent dans les 5 ans après la découverte de leur cancer ?**

- A. 30%
- B. 50%
- C. 70%
- D. 90%
- E. Les propositions A, B, C, et D sont fausses

Exercice 2b

On s'intéresse à une population de femmes atteintes d'un cancer du sein. Le taux de survie 5 ans après la découverte du cancer est de 65 %. Lors de la découverte du cancer, on peut définir la gravité du cancer par son stade (1 à 4). 45 % des femmes sont de stade 1, 30 % de stade 2, 15 % de stade 3, et 10 % de stade 4. La probabilité qu'une femme de cette population soit de stade 4 et survive au moins 5 ans est 0,03. **Quel est le % de femmes de stade 4 qui décèdent dans les 5 ans après la découverte de leur cancer ?**

- A. 30%
- B. 50%
- **C. 70%**
- D. 90%
- E. Les propositions A, B, C, et D sont fausses

Exercice 2c

Evénement A : être de **stade 4** : P(A) = 0,10

Evénement B : **survivre** à 5 ans

P(B/A) = P(A et B) / P(A) = 0,03 / 0,10 = 30%

Donc la probabilité qu'une femme de stade 4 **décède** dans les 5 ans = 1 – P(B) = 70%

Exercice 2d

On s'intéresse à une population de femmes atteintes d'un cancer du sein. Le taux de survie 5 ans après la découverte du cancer est de 65 %. Lors de la découverte du cancer, on peut définir la gravité du cancer par son stade (1 à 4). 45 % des femmes sont de stade 1, 30 % de stade 2, 15 % de stade 3, et 10 % de stade 4. La probabilité qu'une femme de cette population soit de stade 4 et survive au moins 5 ans est 0,03. **En cas de décès dans les 5 ans, quelle est la probabilité que la femme ait été de stade 4 ?**

- A. 20%
- B. 30%
- C. 50%
- D. 70%
- E. Les propositions A, B, C et D sont fausses

Exercice 2e

On s'intéresse à une population de femmes atteintes d'un cancer du sein. Le taux de survie 5 ans après la découverte du cancer est de 65 %. Lors de la découverte du cancer, on peut définir la gravité du cancer par son stade (1 à 4). 45 % des femmes sont de stade 1, 30 % de stade 2, 15 % de stade 3, et 10 % de stade 4. La probabilité qu'une femme de cette population soit de stade 4 et survive au moins 5 ans est 0,03. **En cas de décès dans les 5 ans, quelle est la probabilité que la femme ait été de stade 4 ?**

- **A. 20%**
- B. 30%
- C. 50%
- D. 70%
- E. Les propositions A, B, C et D sont fausses

Exercice 2f

Evénement A : être de stade 4

Evénement C : décéder dans les 5 ans

La probabilité qu'une femme de stade 4 décède dans les 5 ans = 70% (= P(C/A))

Or on cherche P(A/C)

P(A et C) = P (C/A) x P(A) = 0,70 x 0,10 = 0,07

P(A/C) = P(A et C) / P(C) = 0,07 / 0,35 = 0,20

Sources, Crédits

- Alberti C, Timsit JF, Chevret S. Analyse de survie : comment gérer les données censurées ? Méthode de Kaplan-Meier. *Rev Mal Respir* 2005 ; 22 : 333-7
- Deuffic S. Comparer la survie entre deux groupes. *Sang Thrombose Vaisseaux* 1998 ; 10(8) : 515-20
- Golmard JL, Mallet A, Morice V. Biostatistique PCEM1. 2009-2010. Université Paris VI.
- Paolaggi JB, Coste J. Le raisonnement médical, de la science à la pratique clinique. Editions ESTEM, 2001. 265 p.

Mentions légales

- L'ensemble de ce document relève des législations française et internationale sur le droit d'auteur et la propriété intellectuelle.
- Tous les droits de reproduction de tout ou partie sont réservés pour les textes ainsi que pour l'ensemble des documents iconographiques, photographiques, vidéos et sonores.
- Ce document est interdit à la vente ou à la location par un tiers autre que l'Université Côte d'Azur.
- La diffusion, la duplication, la mise à disposition du public (sous quelque forme ou support que ce soit), la mise en réseau, de tout ou partie de ce document, sont strictement réservées à l'Université Côte d'Azur.
- L'utilisation de ce document est strictement réservée à l'usage privé des étudiants inscrits aux cours et au tutorat organisés par l'UFR de Médecine de l'Université Côte d'Azur, et non destinée à toute autre utilisation privée ou collective, gratuite ou payante.