

L'ANALYSE DES DONNÉES DE SURVIE DANS LE CONTEXTE ÉPIDÉMIOLOGIQUE*

J. Estève

International Agency for Research on Cancer. Lyon.

Résumé

L'article révisé la méthodologie de l'analyse de survie dans le contexte épidémiologique, tout spécialement en relation avec les registres du cancer. Les principaux concepts utilisés pour l'analyse de survie sont présentés; les problèmes relatifs à la mortalité compétitive, et à l'ajustement pour d'autres variables, sont aussi discutés. Malgré que l'estimation et la comparaison de distributions de données de survie obtenues à partir

d'observations épidémiologiques est techniquement possible de nos jours grâce à diverses méthodes statistiques qui tiennent en compte de multiples facteurs pronostiques, il est proposé que soient développées des études pilotes autour des principaux problèmes regardant les biais de sélection potentiels, ainsi que la standardisation des facteurs pronostiques à évaluer.

Mots clés: Analyse de survie. Données épidémiologiques. Registres du cancer. Méthodes statistiques. Facteurs pronostiques.

SURVIVAL ANALYSIS IN THE EPIDEMIOLOGICAL CONTEXT

Summary

The present work reviews the methodology of survival analysis in the epidemiological field, especially within the framework of cancer registries. The main concepts used in survival analysis are presented, and the problems of competing mortality, as well as variable and stage adjustment are also discussed. Although the estimation and comparison of survival data distributions obtained from epidemiologic observations is technically possible nowadays thanks to a variety of statistical methods which take into account multiple prognostic factors, it is proposed that pilot studies should be undertaken to look at the main problems related with potential selection biases and the standardization of prognostic factors to be evaluated.

Key words: Survival analysis. Epidemiological data. Cancer registries. Statistical methods. Prognostic factors.

EL ANÁLISIS DE LOS DATOS DE SUPERVIVENCIA EN EL CONTEXTO EPIDEMIOLÓGICO

Resumen

El presente trabajo revisa la metodología de análisis de supervivencia en el campo de la epidemiología, especialmente en el contexto de los registros de cáncer. Se presentan los principales conceptos del análisis de supervivencia, así como los problemas relativos a la mortalidad competitiva y al ajuste según estadio. Aunque la estimación y comparación de las distribuciones de los datos de supervivencia obtenidos a partir de observaciones epidemiológicas es hoy en día técnicamente posible gracias a un arsenal de métodos estadísticos que tienen en cuenta múltiples factores pronósticos, se propone la realización de estudios piloto para solucionar las principales dificultades relacionadas con los sesgos de selección potenciales y con la normalización de los factores pronósticos evaluados.

Palabras clave: Análisis de supervivencia. Datos epidemiológicos. Registros de cáncer. Métodos estadísticos. Factores pronósticos.

Introduction

La méthodologie d'analyse des données de survie a considérablement progressé dans les deux précédentes décennies, grâce à une demande accrue des cliniciens qui souhaitaient des méthodes rigoureuses d'évaluation des traitements, et grâce à une réponse adéquate des biostatisticiens qui accumulaient des connaissances de plus en plus précises dans le domaine de la statistique des processus. Une méthodologie puissante s'est ainsi progressivement mise en place et est maintenant largement diffusée et utilisée dans le contexte des essais cliniques et des études pronostiques (voir Hill et al. pour une revue récente

Introducción

La metodología de análisis de los datos de supervivencia ha progresado considerablemente en las dos últimas décadas gracias a una demanda creciente de los clínicos que deseaban métodos rigurosos de evaluación de los tratamientos, y gracias a una adecuada respuesta de los bioestadísticos que acumulaban conocimientos cada vez más precisos en el campo de la estadística de procesos. Una potente metodología se ha puesto así progresivamente a punto, siendo muy difundida y utilizada en el contexto de los ensayos clínicos y de los estudios de pronóstico¹. A pesar de que esta metodología puede aplicarse sin

* Una revisión previa de este trabajo fue presentada en las Jornadas de la «Societat Catalano-Balear de Salut Pública» celebradas en Mallorca.

Correspondencia: J. Esteve. IARC. 150 Cours Albert Thomas. 69372 LYON. CEDEX 08. France.

Este artículo fue recibido el 18 de febrero de 1991 y fue aceptado, tras revisión, el 28 de octubre de 1991

en langue française¹). Bien que cette méthodologie puisse s'appliquer sans grand changement dans le contexte épidémiologique, et plus particulièrement dans le cadre des registres du cancer qui disposent de l'information nécessaire sur tous les cas d'une région ou d'un pays, l'interprétation des données de survie dans ce contexte et leur comparaison d'une région à l'autre nécessitent une réflexion spécifique.

Après avoir rappelé les principaux concepts utilisés pour l'analyse des données de survie, nous discuterons de façon plus approfondie le problème de la prise en compte de la mortalité compétitive et celui de l'ajustement pour les variables influençant la survie dans le contexte de l'étude de données de survie obtenues par les registres de cancer. Avant d'aborder cette discussion, on doit rappeler que, dans ce cas, l'objectif est d'évaluer la qualité de la prise en charge d'une population par les services de santé. Il est essentiel de rappeler ici qu'elle ne dépend pas seulement de la qualité des thérapeutiques, mais aussi de nombreux autres facteurs, tels que l'accessibilité géographique ou culturelle des systèmes sanitaires, par exemple. Une telle évaluation apparaît alors complémentaire de celle fournie par les services hospitaliers, qui est certainement basée sur des données plus homogènes, mais soumise à de nombreux effets de sélection qu'il est pratiquement impossible de contrôler.

Spécificité des données de survie

A priori, on pourrait penser que l'analyse des données de survie ne présente pas de difficultés particulières, puisqu'il s'agit essentiellement d'estimer la distribution de la durée de survie après un diagnostic donné et comparer entre elles deux ou plusieurs distributions de ce type correspondant à des populations de caractéristiques différentes. Ces problèmes statistiques sont *a priori* banals. En pratique cependant, trois éléments de complexité croissante rendent difficile l'analyse de ces données.

- a) La distribution de la durée de survie est très asymétrique. Des modèles paramétriques spécifiques doivent être construits pour la décrire, et le plus souvent des méthodes non-paramétriques ou semi-paramétriques doivent être utilisées.
- b) Les données de survie qu'on se propose d'analyser sont généralement incomplètes: l'observation relative à la plupart des individus inclus dans l'étude s'arrête avant que le décès ne se produise. L'information concernant ces patients est donc limitée: on sait seulement que leur survie est supérieure à la durée d'observation.
- c) La date qui marque le début de l'observation est définie de façon imprecise. Alors que dans les essais cliniques il est d'usage de définir la survie par le temps qui sépare le début du traitement du décès, en épidémiologie la survie est étudiée à partir du diagnostic. En d'autres termes, le début de l'observation dépend d'un grand nombre de facteurs qu'il faudra être en mesure de préciser et d'analyser. En particulier, les facteurs qui contribuent à avancer la date de diagnostic allongeront *de facto* la durée de survie sans nécessairement l'améliorer: bien qu'ayant commencé à observer la maladie dans une phase plus précoce,

grandes cambios en el contexto epidemiológico, y en concreto dentro del marco de los registros de cáncer que disponen de la información necesaria sobre todos los casos de una región o de un país, la interpretación de los datos de supervivencia en este contexto, y su comparación de una región a otra, exigen una reflexión específica.

Después de recordar los conceptos principales para el análisis de los datos de supervivencia, discutiremos de una manera más profunda el problema de tener en cuenta la mortalidad competitiva y el ajuste por las variables que influyen en la supervivencia en el contexto del estudio de datos de supervivencia obtenidos por los registros de cáncer. Antes de abordar esta discusión, conviene recordar que, en este caso, el objetivo es evaluar la calidad de la atención ofrecida por los servicios de salud. Es esencial recordar que no sólo depende de la calidad de los tratamientos, sino también de otros numerosos factores tales como, por ejemplo, la accesibilidad geográfica o cultural de los sistemas sanitarios. Una evaluación como esta es complementaria a la suministrada por los servicios hospitalarios, que sin duda está basada sobre datos más homogéneos, pero sometidos a numerosos efectos de selección que son prácticamente imposibles de controlar.

Especificidad de los datos de supervivencia

A priori, se podría pensar que el análisis de datos de supervivencia no presenta especiales dificultades ya que se trata esencialmente de estimar la distribución del tiempo de supervivencia después de un diagnóstico determinado, y de comparar entre ellas dos o más distribuciones de este tipo correspondientes a poblaciones con características distintas. Estos problemas estadísticos son *a priori* banales. Pero en la práctica, existen tres elementos de complejidad creciente que hacen difícil el análisis de estos datos:

- a) La distribución del tiempo de supervivencia es muy asimétrica. Hay que construir modelos paramétricos específicos para describirla, y la mayoría de las veces hay que utilizar métodos no paramétricos o semiparamétricos.
- b) Los datos de supervivencia que se van a analizar son generalmente incompletos: la observación de la mayor parte de los individuos incluidos en el estudio se define antes de que se produzca la defunción. La información que concierne a estos pacientes es pues limitada; sólo se sabe que su supervivencia es superior al tiempo de observación.
- c) La fecha que marca el principio de la observación está definida de manera imprecisa. Mientras que en los ensayos clínicos es habitual definir la supervivencia por el tiempo que separa el inicio del tratamiento de la defunción, en epidemiología la supervivencia está estudiada a partir del momento del diagnóstico. En otros términos, el comienzo de la observación depende de un gran número de factores que deben precisarse y analizarse. En particular, los factores que contribuyen a adelantar la fecha del diagnóstico alargarán de hecho el tiempo de supervivencia sin necesariamente mejorarlo. Mucho antes

le décès se produit néanmoins à la date à laquelle il se serait produit si le diagnostic avait été plus tardif. Dans une telle situation, le diagnostic précoce prolonge la durée de morbidité sans «améliorer» la survie.

de empezar a observar la enfermedad en una fase más precoz, la defunción se produce en la misma fecha que se hubiera producido si el diagnóstico se hubiera hecho más tarde. En tal situación, el diagnóstico precoz prolonga el tiempo de morbilidad sin «mejorar» la supervivencia.

L'approche statistique

La structure particulière des données de survie implique qu'il est presque toujours impossible d'estimer la survie globalement par une simple proportion de survivants à une date donnée. L'estimation de la survie d'un groupe de patients sera faite le plus souvent à partir d'une série d'estimations faites localement à des «dates» successives, en dénombrant les décès parmi les individus qui ont survécu la période correspondante et qui sont toujours observés à la date marquant pour eux* la fin de la durée considérée. Le concept de base permettant de faire l'analyse de cette façon est celui de force de mortalité, défini algébriquement par:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t < T < t + \Delta t | T > t) \quad (1)$$

Intuitivement, c'est le nombre de décès par unité de temps à une date t donnée dans la population qui a été observée jusqu'à cette date. Pour cette raison, la force de mortalité est aussi appelée «taux instantané de mortalité».

La probabilité que la durée de survie soit supérieure à t peut s'écrire¹:

$$S(t) = P(T > t) = \exp(-\Lambda(t)) \quad (2)$$

où $\Lambda(t) = \int_0^t \lambda(u) du$ est la force de mortalité cumulée jusqu'à la date t (appelée aussi taux cumulé); $R(t) = 1 - S(t)$ est alors le risque de décéder avant la date t (appelé aussi risque cumulé).

Ainsi, la distribution de la durée de survie est entièrement calculable à partir de la donnée de la force de mortalité sous-jacente. De nombreux modèles sont construits à partir de cette fonction. Le plus connu d'entre eux, la distribution de Weibull, est caractérisé par une force de mortalité proportionnelle à une fonction puissance (t^α). Une autre distribution pour laquelle les fonctions $S(t)$ et $\lambda(t)$ s'expriment simplement est celle pour laquelle le logarithme de la durée de survie a une distribution logistique; elle permet de prendre en compte avec assez de souplesse l'asymétrie observée. La distribution exponentielle enfin, la plus simple d'entre elles, est celle pour laquelle la force de mortalité est constante; il va sans dire que cette dernière ne convient pas pour les données épidémiologiques de cancer pour lesquelles on observe évidemment une forte décroissance de la force de mortalité avec le temps écoulé depuis le diagnostic.

El enfoque estadístico

La particular estructura de los datos de supervivencia determina que es casi siempre imposible estimar la supervivencia globalmente mediante una simple proporción de supervivientes en una determinada fecha. La estimación de la supervivencia de un grupo de pacientes se hará frecuentemente a partir de una serie de estimaciones hechas localmente en «fechas» sucesivas, contando las defunciones entre los individuos que han sobrevivido al período correspondiente y que aún se observan en la fecha que marca para ellos* el final del tiempo considerado. El concepto de base que permite hacer el análisis de esta manera es el de fuerza de mortalidad, definido algebraicamente por:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t < T < t + \Delta t | T > t) \quad (1)$$

Intuitivamente, es el número de defunciones por unidad de tiempo en una fecha t determinada en la población observada hasta esta fecha. Por esta razón, la fuerza de mortalidad también se llama «tasa instantánea de mortalidad».

La probabilidad de que el tiempo de supervivencia sea superior a t se puede escribir¹:

$$S(t) = P(T > t) = \exp(-\Lambda(t)) \quad (2)$$

donde $\Lambda(t) = \int_0^t \lambda(u) du$ es la fuerza de mortalidad acumulada hasta la fecha t (llamada también tasa acumulada); $R(t) = 1 - S(t)$ es el riesgo de morir antes de la fecha t (llamado también riesgo acumulado).

Así, la distribución del tiempo de supervivencia se puede calcular totalmente a partir del dato de la fuerza de mortalidad subyacente. Se han construido muchos modelos a partir de esta función. El más conocido de ellos, la distribución de Weibull, se caracteriza por una fuerza de mortalidad proporcional a una función potencial (t^α). Otra distribución para la cual las funciones $S(t)$ y $\lambda(t)$ se expresan simplemente es aquella en que el logaritmo del tiempo de supervivencia tiene una distribución logística; permite tener en cuenta con bastante flexibilidad la asimetría observada. Finalmente, la distribución exponencial, la más simple de todas, es aquella para la cual la fuerza de mortalidad es constante. Está claro que esta última no interesa para los datos epidemiológicos de cáncer, en los cuales se observa evidentemente un fuerte descenso de la fuerza de mortalidad con el tiempo transcurrido desde el diagnóstico.

* On notera que la date 0 est la date de diagnostic et est différente pour chaque individu qui a ainsi son propre «calendrier».

* Notar que la fecha 0 es la fecha de diagnóstico y es diferente para cada individuo, que de esta manera tiene su propio «calendario».

L'estimation de ces modèles par la méthode du maximum de vraisemblance est alors possible et simple lorsque les expressions algébriques de $S(t)$ et $\lambda(t)$ sont connues¹. En pratique, il est d'usage de ne pas faire d'hypothèse sur la distribution de la durée de survie, et on préfère estimer la distribution inconnue par une distribution empirique pour laquelle la force de mortalité est constante dans des intervalles définis *a priori*: le taux instantané est alors estimé par le nombre de décès observés dans l'intervalle, divisé par le nombre de personnes-années d'observation correspondant. Cette méthode, appelée méthode actuarielle*, est probablement la plus utilisée jusqu'à présent dans les études épidémiologiques². L'estimation de la survie aux extrémités de chaque intervalle s'accompagne d'une erreur-type qui permet d'évaluer la précision de l'estimation.

La comparaison de plusieurs distributions de survie a souvent été faite grossièrement en comparant les survies pour un nombre d'années donné (e.g., survie à cinq ans). On conçoit que cette comparaison est, soit peu puissante si la survie d'un groupe est systématiquement meilleure à toute date, soit trompeuse si l'avantage d'un groupe sur l'autre est éphémère. De plus, en pratique, on aimerait, au-delà de la comparaison de plusieurs groupes de patients, évaluer l'influence conjointe de plusieurs facteurs sur la durée de survie. Le modèle des taux proportionnels³ (méthode de Cox) permet de réaliser simplement ce type d'analyse; ce modèle stipule, dans le cas le plus simple d'un facteur à deux modalités, que la taux $\lambda_1(t)$, en présence du facteur de risque, est proportionnel au taux $\lambda_0(t)$, en l'absence du facteur de risque:

$$\lambda_1(t) = \alpha \lambda_0(t) \quad (3)$$

Le taux relatif α caractérise la différence entre les distributions de survie correspondantes. Lorsque plusieurs facteurs sont considérés, le modèle s'écrit ($Z = (Z_1, \dots, Z_p)$ étant un vecteur de variables pronostiques):

$$\lambda(t, Z) = \lambda(t, 0) \exp(\sum \beta_i Z_i) \quad (4)$$

ou encore:

$$\text{Log} \frac{\lambda(t, Z)}{\lambda(t, 0)} = \sum \beta_i Z_i \quad (5)$$

Le modèle de Cox est donc un modèle log-linéaire pour le taux relatif. Dans ce modèle, $\lambda(t, 0)$ est la force de mortalité pour le groupe des «patients de référence», pour lesquels tous les facteurs pronostiques sont égaux à 0. Ces facteurs peuvent correspondre à des variables catégorielles; k catégories génèrent alors $k-1$ variables indicatrices, la catégorie de référence étant celle pour laquelle toutes ces variables binaires sont nulles. Ces facteurs peuvent aussi être des variables quantitatives. Les coefficients β_i s'interprètent, soit comme le logarithme du taux relatif de la catégorie i par rapport à la catégorie de référence, soit comme la pente de la relation dose-effet entre le logarithme du taux relatif et

La estimación de estos modelos por el método de máxima verosimilitud es posible y simple si se conocen las expresiones algebraicas de $S(t)$ y $\lambda(t)$ ¹. En la práctica, no se acostumbra a hacer hipótesis sobre la distribución del tiempo de supervivencia y se prefiere estimar la distribución no conocida mediante una distribución empírica para la cual la fuerza de mortalidad es constante en intervalos definidos *a priori*: la tasa instantánea se estima entonces con el número de defunciones observadas en el intervalo, dividido por el número de personas-años de observación correspondientes. Este método, llamado actuarial*, es probablemente el más utilizado hasta la fecha en los estudios epidemiológicos². La estimación de la supervivencia en los extremos de cada intervalo se acompaña de un error estándar que permite evaluar la precisión de la estimación.

La comparación de varias distribuciones de supervivencia se ha realizado a menudo de forma cruda comparando las supervivencias para un número de años determinado (por ej., supervivencia a cinco años). Se concibe que esta comparación es poco potente si la supervivencia de un grupo es sistemáticamente mejor en toda fecha, o bien falsa si la ventaja de un grupo sobre otro es efímera. Además, en la práctica, sería deseable más allá de la comparación de varios grupos de pacientes, evaluar la influencia conjunta de varios factores sobre el tiempo de supervivencia. El modelo de riesgos proporcionales³ (método de Cox) permite realizar este tipo de análisis de forma simple; este modelo estipula, en el caso más simple de un factor con dos modalidades, que la tasa $\lambda_1(t)$, en presencia del factor de riesgo, es proporcional a la tasa $\lambda_0(t)$, en ausencia del factor de riesgo:

$$\lambda_1(t) = \alpha \lambda_0(t) \quad (3)$$

La tasa relativa α caracteriza la diferencia entre las distribuciones de supervivencia correspondientes. Si se consideran varios factores, el modelo se escribe ($Z = (Z_1, \dots, Z_p)$ siendo un vector de variables pronósticas):

$$\lambda(t, Z) = \lambda(t, 0) \exp(\sum \beta_i Z_i) \quad (4)$$

o también:

$$\text{Log} \frac{\lambda(t, Z)}{\lambda(t, 0)} = \sum \beta_i Z_i \quad (5)$$

El modelo de Cox es pues un modelo log-lineal para la tasa relativa. En este modelo, $\lambda(t, 0)$ es la fuerza de mortalidad para el grupo de «pacientes de referencia», para los cuales todos los factores pronósticos son iguales a 0. Estos factores pueden corresponder a variables categóricas; k categorías generan $k-1$ variables indicativas, siendo la categoría de referencia aquella en la cual todas estas variables binarias son nulas. Estos factores pueden también ser variables cuantitativas. Los coeficientes β_i se interpretan, bien como el logaritmo de la tasa relativa de la categoría i en relación a la categoría de referencia, bien como la pendiente de la relación dosis-efecto entre el logaritmo de la tasa relativa y la

* Cette méthode est *stricto sensu* une approximation de celle décrite ici.

* Este método es en sentido estricto una aproximación de lo aquí descrito.

la variable quantitative considérée. Le modèle de Cox a été construit pour tester l'influence des facteurs pronostiques, et on notera que la forme de la fonction $\lambda(t, o)$ n'est pas spécifiée. L'originalité de la méthode de Cox est de permettre l'estimation des paramètres β_i , en l'absence de cette information. Pour cette raison, on dit que le modèle est semi-paramétrique. Le test de l'influence d'un facteur revient alors à tester la nullité du ou des coefficients β_i , correspondants. La réalisation pratique de ces analyses est rendue possible par la disponibilité accrue de logiciels spécialisés, fonctionnant aussi sur micro-ordinateurs⁴.

Survie relative

La discussion précédente était limitée au cas où l'événement étudié est le décès, et ne distinguait pas les décès liés à la pathologie considérée des autres décès. Les patients atteints d'un cancer sont sujets aux autres risques de décès comme le reste de la population, et il semble souhaitable de corriger la survie calculée en prenant en compte la nature des décès observés. Cependant, l'attribution d'un décès à la pathologie considérée ou à une cause compétitive est en pratique difficile et sujette à de nombreux biais. Il arrive en outre que la cause de décès ne soit pas connue du registre qui souhaite analyser la survie des patients qu'il a enregistrés. Pour ces raisons, on a proposé de faire cette correction, en l'absence d'information sur la cause de décès, par référence à la mortalité de toute cause de la population dans laquelle les cas étudiés se sont produits. La force de mortalité observée est en effet la résultante de celle causée par le cancer considéré et de celle liée à l'ensemble des autres causes:

$$\lambda_o(t) = \lambda_c(t) + \lambda_a(t) \quad (6)$$

ce qui, en appliquant la formule (2) ci-dessus, conduit à:

$$S_o(t) = S_c(t) \cdot S_a(t) \quad (7)$$

i.e., la survie observée est le produit de la «survie nette» (S_o) et de la «survie attendue» (S_a).

On devine que, pour les études à court terme et pour beaucoup de sites de cancer, λ_o est pratiquement égal à λ_c , et il est à peine nécessaire de faire cette correction. Par contre, celle-ci devient nécessaire dans l'étude des survies à long terme; de plus, dans ces cas, les propriétés de la méthode choisie pour l'estimation de $S_c(t)$ doivent être connues, afin d'être en mesure d'interpréter correctement l'estimation obtenue.

La méthode la plus ancienne, proposée par Ederer⁵, est basée sur la table de mortalité de la population générale, qui permet de calculer la survie attendue $S_{ax}(t)$ pour un patient d'âge x . La survie nette est alors estimée par la «survie relative»:

$$S_r(t) = \frac{n S_o(t)}{\sum_x n_x S_{ax}(t)} \quad (8)$$

variable cuantitativa considerada. El modelo de Cox se ha construido para testar la influencia de los factores pronósticos; y se observará que la forma de la función $\lambda(t, o)$ no se ha especificado. La originalidad del método de Cox está en permitir la estimación de los parámetros β_i , en ausencia de esta información. Por esta razón, se dice que el modelo es semi-paramétrico. El test de la influencia de un factor consiste en testar la nulidad del o de los coeficientes β_i , correspondiente(s). La realización práctica de estos análisis es posible debido a la creciente disponibilidad de paquetes informáticos especializados, que también funcionan con microordenadores⁴.

Supervivencia relativa

La anterior discusión se limitaba al caso en el cual el suceso estudiado es la defunción, y no distinguía las defunciones ligadas a la patología considerada de las otras defunciones. Los pacientes que tienen un cáncer están sometidos a otros riesgos de defunción como el resto de la población, y parece deseable corregir la supervivencia calculada teniendo en cuenta la naturaleza de las defunciones observadas. Pero la atribución de una defunción a la patología considerada o a una causa competitiva es en la práctica difícil y está sometida a numerosos sesgos. También puede ocurrir que no se conozca la causa de la defunción en el registro que desea analizar la supervivencia de los pacientes que ha registrado. Por estos motivos, se ha propuesto hacer esta corrección, en ausencia de información sobre la causa de defunción, por referencia a la mortalidad por todas las causas de la población en la cual los casos estudiados se han producido. La fuerza de mortalidad que se observa es pues la resultante de la causada por el cáncer considerado y la ligada al conjunto de las restantes causas:

$$\lambda_o(t) = \lambda_c(t) + \lambda_a(t) \quad (6)$$

lo que, aplicando la anterior fórmula (2), conduce a:

$$S_o(t) = S_c(t) \cdot S_a(t) \quad (7)$$

es decir, la supervivencia observada es el producto de la «supervivencia neta» (S_o) y de la «supervivencia esperada» (S_a).

Se adivina que, para los estudios a corto plazo y para muchas localizaciones de cáncer, λ_o es prácticamente igual a λ_c , y no hace falta hacer esta corrección. En cambio, ésta es necesaria en el estudio de supervivencia a largo plazo; además, en este caso, las propiedades del método escogido para la estimación de $S_c(t)$ han de ser conocidas para permitir interpretar correctamente la estimación obtenida.

El método más antiguo, propuesto por Ederer⁵, se basa en la tabla de mortalidad de la población general, que permite calcular la supervivencia esperada $S_{ax}(t)$ para un paciente de edad x . La supervivencia neta es entonces estimada por la «supervivencia relativa»:

$$S_r(t) = \frac{n S_o(t)}{\sum_x n_x S_{ax}(t)} \quad (8)$$

où $S_o(t)$ est la survie observée estimée par la méthode actuarielle, n le nombre total, et n_x le nombre de patients d'âge x dans la population étudiée. On voit aisément que $S_r(t)$ aura tendance à estimer:

$$\bar{S}_c(t) = \frac{\sum_x n_x S_{ax}(t) S_{cx}(t)}{\sum_x n_x S_{ax}(t)} \quad (9)$$

moyenne des survies nettes aux différents âges, pondérée par le nombre de survivants attendus dans chaque groupe d'âge. Lorsque la survie nette dépend de l'âge, la survie relative à long terme tend plus ou moins à estimer la survie nette des sujets ayant la plus grande espérance de vie.

Plusieurs méthodes permettant de prendre en compte les facteurs pronostiques ont été proposées récemment⁶⁻¹⁰. Elles sont toutes basées sur un modèle de la force de mortalité observée dans lequel celle liée aux autres causes est supposée connue et accessible à partir des statistiques officielles. Par exemple, dans la dernière de ces références, on suppose que:

$$\lambda_o(t, z, x) = \lambda_c(t, z) + \lambda_a(x + t, z) \quad (10)$$

où la force de mortalité causée par le cancer étudié (λ_c) est un modèle de taux proportionnel dans lequel le taux de base est supposé constant par intervalle; x est l'âge et z est un sous-vecteur du vecteur des covariables pour lequel la force de mortalité générale (λ_a) est connue (sexe, groupe socioéconomique, etc.). Des programmes sont disponibles pour réaliser l'ajustement de tels modèles¹¹.

La qualité de ces méthodes dépend dans une certaine mesure de l'adéquation de la force de mortalité attendue. On sait qu'un groupe de personnes à risque élevé pour cancer donné n'est pas soumis *a priori* à la force de mortalité de la population générale, mais est au contraire le plus souvent soumis à des risques différents pour les autres causes de décès. La surmortalité observée peut évidemment être une conséquence directe du cancer considéré, mais peut parfois être dû à un effet de sélection non pris en compte. De plus, lorsque λ_c tend vers 0, il pourrait se faire que les patients survivants, qui sont de ce fait l'objet d'une forte sélection, soient soumis à une force de mortalité inférieure à celle de la population générale. Un facteur multiplicatif appliqué à λ_a pourrait être introduit dans le modèle, mais l'instabilité statistique de son estimation risque d'être importante. Ces considérations ne sont pas destinées à déconseiller ces méthodes, mais seulement à mettre en garde contre leur utilisation aveugle.

Ajustement pour le stade

Les études épidémiologiques n'étant pas randomisées, l'interprétation des facteurs pronostiques doit être faite avec prudence. Il est important également de se rappeler que l'ajustement pour les facteurs pronostiques doit aussi être faite avec la plus

donde $S_o(t)$ es la supervivencia observada por el método actuarial, n es el número total, y n_x es el número de pacientes de edad x en la población estudiada. Fácilmente se ve que $S_r(t)$ tendrá tendencia a estimar:

$$\bar{S}_c(t) = \frac{\sum_x n_x S_{ax}(t) S_{cx}(t)}{\sum_x n_x S_{ax}(t)} \quad (9)$$

la media de las supervivencias netas a edades diferentes, ponderada por el número de supervivientes esperados en cada grupo de edad. Cuando la supervivencia neta depende de la edad, la supervivencia relativa a largo plazo tiende más o menos a estimar la supervivencia neta de los individuos que tienen la esperanza de vida más larga.

Recientemente se han propuesto varios métodos que permiten tener en cuenta los factores pronósticos⁶⁻¹⁰. Todos se basan en un modelo de la fuerza de mortalidad observada en el cual la que está ligada a otras causas se supone conocida y accesible a partir de las estadísticas oficiales. Por ejemplo, en la última de estas referencias, se supone que:

$$\lambda_o(t, z, x) = \lambda_c(t, z) + \lambda_a(x + t, z) \quad (10)$$

donde la fuerza de mortalidad causada por el cáncer estudiado (λ_c) es un modelo de tasa proporcional en el cual la tasa de base se supone constante por intervalo; x es la edad y z es un subvector del vector de las covariables para el cual la fuerza de mortalidad general (λ_a) es conocida (sexo, clase socioeconómica, etc.). Existen programas disponibles para realizar el ajuste de estos modelos¹¹.

La calidad de estos métodos depende en cierta medida de la adecuación de la fuerza de la mortalidad esperada. Se sabe que un grupo de personas de riesgo elevado para un determinado cáncer no está sometido *a priori* a la fuerza de mortalidad de la población general, sino que, por el contrario, está frecuentemente sometido a diferentes riesgos para las otras causas de defunción. La sobremortalidad que se observa puede evidentemente ser una consecuencia directa del cáncer considerado pero puede, a veces, ser debida a un efecto de selección que no se ha tenido en cuenta. Además, cuando λ_c tiende hacia 0, podría suceder que los pacientes supervivientes, que son por ello objeto de una fuerte selección, estuviesen sometidos a una fuerza de mortalidad inferior a la de la población general. Un factor multiplicativo aplicado a λ_a podría introducirse en el modelo, pero la inestabilidad estadística de su estimación podría ser importante. Estas consideraciones no tienen por objeto desaconsejar estos métodos, sino únicamente poner en guardia contra su utilización a ciegas.

Ajuste por el estadio

Como los estudios epidemiológicos no son aleatorizados, la interpretación de los factores pronósticos se ha de hacer con prudencia. También es importante recordar que el ajuste por los factores pronósticos ha de realizarse también con la mayor de las

grande circonspection. On discutera brièvement ici le problème de l'ajustement pour le stade de diagnostic.

Les progrès des méthodes d'investigation contribuent à fournir une détermination du stade de plus en plus précise; une des conséquences de cette amélioration est le passage des plus mauvais pronostics d'un stade donné au stade suivant. En effet, des indices de gravité non détectables cliniquement conduisent en moyenne à une classification péjorative lorsque les méthodes modernes sont adoptées. Tout se passe donc comme si le passage d'un stade au suivant était avancé dans le temps. La survie de chaque stade est donc prolongée sans être améliorée¹². Ainsi, lorsque la survie est comparée entre deux groupes pour lesquels la détermination du stade n'a pas été normalisée, l'ajustement pour celui-ci risque de renforcer de façon artificielle une différence de survie non significative autrement. L'ajustement pour le stade fournira donc une fausse impression de sécurité. Cette difficulté peut avoir des conséquences importantes lorsqu'on souhaite comparer en particulier plusieurs filières de soin; une première tentation est en effet d'ajuster pour le stade pour prendre en compte le fait que le choix d'une filière particulière pourrait être associé à la gravité du pronostic. Si, comme il est probable, les filières comparées correspondent à des méthodes de détermination du stade différentes, cet ajustement est illusoire.

Conclusion

L'estimation et la comparaison de distributions de durées de survie, obtenue à partir d'observations épidémiologiques, est aujourd'hui techniquement simple grâce à un arsenal de méthodes statistiques avec lesquelles l'analyse simultanée de plusieurs facteurs pronostics est rendue possible. Il serait cependant imprudent de croire que ces méthodes sont en mesure de régler les vrais problèmes que pose l'analyse des données de survie dans le contexte épidémiologique¹³. L'analyse de routine des données de survies recueillies par les registres du cancer doit certainement être abordée car c'est la vocation de ces institutions de fournir aux organismes responsables de santé publique l'information sur le devenir des malades cancéreux. Des études pilotes destinées à régler les principales difficultés associées aux nombreux biais de sélection potentiels et à la normalisation des facteurs pronostics évalués est, à notre avis, la première étape d'une étude internationale dans ce domaine.

Bibliografía

1. Hill C, Com-Nougue C, Kramar A, Moreau T, O' Quigley J, Senoussi R, Chastang C. *Analyse statistique des données de survie*. INSERM Médecine-Sciences, Flammarion, Paris, 1990.
2. Raymond L, Levi F. Calcul de la survie dans le contexte épidémiologique: Principales options méthodologiques. *Médecine et Hygiène* 1986; 44: 392-4.
3. Cox D. Regression models and life tables (with discussion). *J Roy Stat Soc B* 1972; 34: 187-220.
4. BMDP Statistical Software: 1440 Sepulveda Blvd, Los Angeles, CA 90025; EGRET Software Division 909 NE 43 rd, Suite 310, Seattle, WA 98105, and others.

circunspecciones. Se discutirá brevemente aquí el problema del ajuste por el estadio del diagnóstico.

El progreso de los métodos de investigación contribuye a proporcionar una determinación del estadio cada vez más precisa. Una de las consecuencias de esta mejora es el paso de los pronósticos más malos de un estadio dado al estadio siguiente. En efecto, los índices de gravedad no detectables clínicamente, conducen en promedio a una clasificación peyorativa si se adoptan los métodos modernos. Ocurre como si el paso de un estadio al siguiente se hubiera avanzado en el tiempo. En consecuencia, la supervivencia de cada estadio se prolonga sin mejorarse¹². Así, cuando se compara la supervivencia entre dos grupos para los cuales la determinación del estadio no se ha normalizado, el ajuste por éste puede reforzar de manera artificial una diferencia de supervivencia de otro modo no significativa. El ajuste por el estadio proporcionará, pues, una falsa impresión de seguridad. Esta dificultad puede tener importantes consecuencias si se desea comparar en particular diversos tipos de tratamiento. Una primera tentación es en efecto ajustar por el estadio, para tener en cuenta el hecho de que la elección de un tipo de tratamiento particular podría asociarse a la gravedad del pronóstico. Si, como es probable, los distintos tipos comparados corresponden a diferentes métodos de determinación del estadio, este ajuste es ilusorio.

Conclusión

La estimación y la comparación de distribuciones de tiempo de supervivencia, obtenida a partir de observaciones epidemiológicas, es hoy técnicamente simple gracias a un arsenal de métodos estadísticos que hacen posible el análisis simultáneo de varios factores pronósticos. Pero sería imprudente creer que estos métodos están en condiciones de solucionar los verdaderos problemas que supone el análisis de datos de supervivencia en el contexto epidemiológico¹³. El análisis rutinario de los datos de supervivencia recogidos por los registros del cáncer debe sin duda abordarse, ya que es la vocación de estas instituciones el proporcionar a los organismos responsables de la salud pública la información sobre el futuro de los enfermos de cáncer. Estudios pilotes destinados a solucionar las dificultades potenciales asociados a los numerosos sesgos de selección potenciales y a la normalización de los factores pronósticos evaluados deben constituir, en nuestra opinión, la primera etapa de un estudio internacional en este campo.

Traducción: I. Casas

5. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: A statistical methodology. *Nat Cancer Inst Monogr* 1961; 6: 101-21.
6. Pocock SJ, Gore SM, Kerr GR. Long term survival analysis: The curability of breast cancer. *Stat Med* 1982; 1: 93-104.
7. Andersen PK, Borch-Johnsen K, Decker T, et al. A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics* 1985; 41: 921-32.
8. Hill C, Laplanche A, Rezvani A. Comparison of the mortality of a cohort with the mortality of a reference population in a prognostic study. *Stat Med* 1985; 4: 295-302.
9. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Appl Stat* 1987; 36: 309-17.

10. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Stat Med* 1990; 9: 529-38.

11. Hakulinen T, Gibberd R, Abeywickrama K, Söderman B. *A computer program package for cancer survival studies, version 1.0*. Cancer Society of Finland Publication 39, Finnish Cancer Registry and University of Newcastle, Australia, 1988.

12. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *New Engl J Med* 1985; 312: 1604-8.

13. Raymond L, Obradovic M, Fioretta G. Facteurs pronostiques de survie après cancer: Le modèle épidémiologique. *Méd Soc Prév* 1988; 33: 269-73.

