Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 239 a 254

¿Cuándo inicia la enfermedad de Alzheimer? Kaplan-Meier versus Turnbull: una aplicación a datos con censura arbitraria

¿When does Alzheimer's Disease Begin? Kaplan-Meier versus Turnbull: An Application to Arbitrary Censoring Data

Carlos Mario Lopera-Gómez[1,a], Mario César Jaramillo-Elorza[1,b], Natalia Acosta-Baena[2,c]

[1]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

[2]Grupo de Neurociencias de Antioquia y Grupo Académico de Epidemiología Clínica-GRAEPIC, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia

Resumen

La mayoría de los análisis de supervivencia se basan en tiempos de falla exactos y observaciones censuradas a la derecha, utilizándose métodos ampliamente difundidos como el método de Kaplan-Meier (KM). Para estimar la edad de inicio de la Enfermedad de Alzheimer (EA) familiar cuando las censuras son arbitrarias (censura a derecha, a izquierda o en intervalo), ¿cuál es el cambio en los resultados clínicos, si se utiliza el método de KM mediante imputación comparado con el método de Turnbull sugerido para este tipo de datos?

El método de Turnbull se comparó con el método de KM mediante un estudio de simulación y una aplicación con datos reales. Se realizó KM con imputación a través del punto medio del intervalo (PM) y en el extremo derecho (ED). Se analizaron diferentes tamaños de muestra y diferentes tiempos entre visitas.

En todos los escenarios de simulación, las funciones que fueron estimadas, usando imputación de datos, difieren significativamente de la verdadera función de supervivencia $S(t)$.

La edad de inicio de la EA determinada a través de un método de imputación tiene implicaciones clínicas relevantes que afectarían la toma de decisiones a la hora de iniciar una terapia preventiva. El método de Turnbull

[a]Profesor asistente. E-mail: cmlopera@unal.edu.co
[b]Profesor asociado. E-mail: mjarami@unal.edu.co
[c]Profesora. E-mail: natalia.acosta@neurociencias.udea.edu.co

presenta un menor sesgo cuando se necesita realizar un análisis de supervivencia con censuras arbitrarias.

***Palabras clave*:** análisis de supervivencia, censura de intervalo, edad de inicio, enfermedad de Alzheimer familiar.

Abstract

Most of the survival analysis are based on exact failure times and right censored observations, using methods widely known as the Kaplan-Meier (KM). To estimate the onset age of familial Alzheimer's Disease (AD) when the censor times are arbitrary (right, left or interval censor), what is the change in clinical outcomes, using the KM method with data imputation compared with procedure proposed by Turnbull for this kind of data?

Turnbull's method was compared with KM method in a simulation study and an application with real data. KM method was based on data imputation through the midpoint of the interval (MP) and the right side of the interval (RS), considering several sample sizes and different times between visits.

In all simulation scenarios estimated functions using data imputation differ significantly from the actual simulated survival function $S(t)$.

The estimated onset age of AD through data imputation methods has relevant clinical implications that would affect decision-making in initiating preventive therapy. Turnbull's method has fewer bias when was compared with KM with imputation to perform a survival analysis with arbitrary censure data.

***Key words*:** Age of onset, Familial Alzheimer's disease, Interval censoring, Survival analysis.

1. Introducción

El análisis de supervivencia es un conjunto de procedimientos estadísticos para el análisis de datos en los que la variable de resultado es el tiempo hasta que ocurre un evento de interés. La función de supervivencia es quizás la función más importante en los estudios de medicina y salud. Como es usual en el análisis de datos de supervivencia, es de interés estimar la función de supervivencia $S(t)$ y evaluar la importancia de factores potenciales de pronóstico o características individuales sobre este tiempo de supervivencia.

La gran cantidad de estudios epidemiológicos realizados en enfermedades como el cáncer, entre muchas otras, y la cantidad de estudios longitudinales con desenlaces que involucran el tiempo demuestran la importancia del análisis de supervivencia. Alternativamente al desenlace de supervivencia o tiempo hasta la muerte, el tiempo puede hacer referencia al momento en que una persona presenta cualquier otro evento. Si el evento se presenta en todos los individuos, se podrían aplicar muchos métodos. Sin embargo, lo habitual es que al final del seguimiento algunas de las personas no han desarrollado el evento de interés, por lo que el verdadero tiempo trascurrido hasta el evento es no observado. Además, los datos de supervivencia rara vez se distribuyen de forma "normal", y se componen generalmente de muchos eventos al inicio del seguimiento, y los eventos tardíos son

relativamente pocos. Estas características de los datos son las que hacen necesario un método especial como el análisis de supervivencia.

Las dificultades específicas relacionadas con el análisis de supervivencia surgen en gran medida por el hecho de que sólo algunas personas han experimentado el evento; por lo tanto, el tiempo de supervivencia se desconoce en un subconjunto de sujetos del estudio. Este fenómeno se llama censura y sus mecanismos pueden deberse a que el individuo no ha experimentado el desenlace en el momento de cierre del estudio; porque se pierde del seguimiento: o porque el sujeto presenta un evento diferente que hace imposible un seguimiento posterior (riesgo competitivo). En este último caso, las censuras deben estimarse de manera distinta y requiere un análisis especial de los datos. Pero al visualizar el proceso de supervivencia de un individuo como una línea de tiempo pueden verse tres tipos de censuras: si el evento (suponiendo que llegara a ocurrir) está más allá del final del período de seguimiento, esta situación se conoce como censura a derecha. Otro caso se presenta cuando se observa el evento de interés antes de la primera evaluación, pero no se sabe exactamente cuándo ocurrió. Este tipo de censura es la censura a izquierda. Y por último, el tiempo trascurrido hasta el evento también puede ser censurado en intervalo; cuando los individuos salen y entran del seguimiento (por ejemplo, cuando los individuos se presentan a controles médicos con cierta frecuencia), el individuo presenta el evento de interés al regreso del seguimiento pero la única información que se tiene en este caso, es que el evento se produce dentro de un intervalo de tiempo dado.

La mayoría de los datos de supervivencia incluyen solo observaciones censuradas a derecha y tiempos de falla exactos, utilizándose métodos ampliamente difundidos como el método de Kaplan-Meier (KM), pruebas de logrank y regresión de Cox (análisis de riesgos proporcionales). Sin embargo, los métodos que soportan datos censurados a izquierda o en intervalo no son tan conocidos. Pocos paquetes estadísticos permiten estos datos, y por esta razón, la práctica común entre los investigadores consiste en simplemente ignorar y descartar las censuras a izquierda de los datos, o realizar una imputación del desenlace para las censuras de intervalo. Es decir, asumir que el evento que ha ocurrido dentro del intervalo $(L_i, U_i]$ ocurrió ya sea en el límite inferior o superior del intervalo o en el punto medio del mismo. Autores como Rucker & Messerer (1988), Odell, Anderson & D'agostinho (1992), Dorey, Little & Schenker (1993) y Iceland (1997) manifiestan que asumir el tiempo de supervivencia de intervalo como si fuera exacto puede conducir a estimadores sesgados, así como a conclusiones y estimaciones parciales que no son completamente fidedignas. Estas afirmaciones motivan, de alguna manera, a propuestas distintas relacionadas con el tratamiento que se debe dar a estas censuras, con el fin de evitar sesgos y que se incorpore mayor información.

Los datos de la Cohorte Antioquia-E280A de 15 años de seguimiento, con sujetos en riesgo de enfermedad de Alzheimer familiar, incluyen los tres tipos de censuras mencionadas previamente. Conocer la edad de inicio de la enfermedad en estos sujetos que inevitablemente van a desarrollar la Enfermedad de Alzheimer (EA) exige métodos alternativos para dicha estimación. En este estudio se pretende difundir tales métodos e ilustrar qué tan erróneas serían las estimaciones en la edad de inicio de la EA, utilizando el método KM comparado con el método

de Turnbull para estimación bajo censura arbitraria (Peto 1973, Turnbull 1974, Turnbull 1976). También interesa determinar las implicaciones desde el punto de vista clínico cuando se incurre en un sesgo de medición y la importancia de los resultados para el diagnóstico del inicio de la EA familiar. Inicialmente se realiza un estudio de simulación y posteriormente la aplicación con los datos reales.

En la sección 2 se presenta el problema clínico y la base de datos que servirá para ilustrar los métodos que van a ser comparados. Los métodos estadísticos utilizados y el planteamiento de un estudio de simulación son presentados en la sección 3. La sección 4 recopila los resultados obtenidos a través del estudio de simulación y presenta la aplicación con los datos de enfermedad de Alzheimer. Finalmente, en la sección 5 se dan algunas conclusiones y recomendaciones con base en los hallazgos encontrados.

2. Problema y datos de enfermedad de Alzheimer

Conocer el tiempo hasta el inicio de la enfermedad de Alzheimer sólo es posible gracias a las formas genéticas de la enfermedad, con herencia autosómica dominante. En esta condición, todos los sujetos nacen portando una mutación que predispone a la enfermedad, expresándose en algún momento de la vida. Las manifestaciones consisten en quejas de memoria y deterioro cognitivo evidente en las evaluaciones neurosicológicas alrededor de los 50 años de edad. Conocer la edad más aproximada del inicio de la enfermedad es el primer paso para planear y desarrollar nuevos estudios en busca de terapias preventivas. El Grupo de Neurociencias de la Universidad de Antioquia ha seguido desde 1995 a este conglomerado poblacional, que es el más numeroso del mundo, con 5000 sujetos estimados, con riesgo de desarrollar EA genético mutación E280A en Presenilina 1 (PSEN1). Se identificaron, hasta enero del 2010, 1784 sujetos pertenecientes a 25 familias afectadas. Se detectaron 449 sujetos portadores de la mutación E280A-PSEN1. Los datos de estos últimos sujetos portadores fueron los utilizados para detectar el inicio de la enfermedad de manera retrospectiva (Acosta-Baena, Sepúlveda-Falla, Lopera-Gómez, Jaramillo-Elorza, Moreno, Aguirre-Acevedo, Saldarriaga & Lopera 2011).

3. Métodos

Con base en los datos descritos en la sección 2, se utilizaron los métodos de imputación, para comparar la función de supervivencia de Turnbull con KM.

Para medir la edad de inicio de la enfermedad, se realizó un análisis de supervivencia evaluando el tiempo transcurrido desde la fecha de nacimiento hasta la fecha de aparición del deterioro cognitivo leve o hasta la fecha de la última evaluación. Se utilizó el método de supervivencia desarrollado por Peto (1973), Turnbull (1974) y Turnbull (1976), que incluye los tres tipos de censuras, mediante el algoritmo implementado por Giolo (2004), para el software R versión 2.13.1 (R Development Core Team 2011). El código utilizado hace uso de la librería `SURVIVAL` del software R, y está disponible bajo pedido a los autores.

3.1. Estimador no paramétrico de Turnbull

En los estudios longitudinales, donde los individuos son monitoreados durante un lapso de tiempo prefijado, o visitados periódicamente un cierto número de veces, el tiempo $T_i,\ i = 1, \ldots, n$, hasta que ocurre el evento de interés para cada individuo, se desconoce. Sólo se sabe que está dentro de un intervalo entre dos visitas, es decir, entre la visita en el tiempo L_i y la visita en el tiempo U_i con $L_i < T_i \leq U_i$. Si el evento ocurre exactamente en el momento de una visita, lo cual es muy poco probable, pero puede ocurrir, se tiene un tiempo de supervivencia exacto. En este caso se asume que $L_i = T_i = U_i$.

Por otra parte, se sabe que para los individuos cuyos tiempos están censurados a derecha, el evento de interés no ha ocurrido hasta la última visita, pero puede ocurrir en cualquier instante desde ese momento en adelante. Por consiguiente, se supone en este caso que T_i puede ocurrir dentro del intervalo $(L_i, +\infty)$, con L_i igual al periodo desde el comienzo del estudio hasta la última visita y $U_i = +\infty$.

De modo semejante, para los individuos cuyos tiempos están censurados a izquierda, se sabe que el evento de interés ha ocurrido antes de la primera visita, y, por lo tanto, suponemos que T_i ha ocurrido en el intervalo $(0, U_i]$, con $L_i = 0$ representando el comienzo del estudio, y U_i es el tiempo hasta la primera visita. El método de Turnbull generaliza cualquier situación con combinaciones de tiempos de supervivencia (exacto o intervalo) y censuras a izquierda y derecha como datos de supervivencia de intervalo. Por lo tanto, los tiempos de supervivencia exacta, así como datos de censura a izquierda y derecha, son todos casos especiales de datos de supervivencia con censura de intervalo con $L_i = U_i$ para censuras exactas, $U_i = +\infty$ para las censuras a derecha y $L_i = 0$ para censuras a izquierda.

Como uno de los objetivos principales en análisis de supervivencia es estimar la función de supervivencia e investigar la importancia de factores potenciales de pronóstico bajo tiempos de supervivencia con censura a intervalo, el número de factores bajo estudio debería depender del propósito del estudio. Como lo sugiere Hougaard (1999), la estimación no paramétrica de la función de distribución acumulada $F(t)$, o en su defecto de la función de supervivencia $S(t)$, es preferible a su estimación paramétrica, por varias razones. Por ejemplo, una elección equivocada de la distribución paramétrica de T podría conducir a conclusiones erróneas de $S(t)$. Además, podría ser difícil encontrar una distribución paramétrica apropiada para ajustar los datos. Hougaard da el ejemplo de tiempos de vida de una población cuya función hazard muestra la llamada forma de bañera: la cual en un principio decrece pocos años, luego permanece constante durante muchos años y por último empieza a aumentar. En este caso, el mejor ajuste probablemente se obtendría de una mezcla de distribuciones.

En el caso de censura a derecha, se podría usar el estimador de Kaplan-Meier para obtener $S(t)$ (Kaplan & Meier 1958). Sin embargo, con datos censurados en intervalo, el método de Kaplan-Meier no puede ser aplicado, y han sido Peto (1973), Turnbull (1974) y Turnbull (1976) quienes han desarrollado el estimador no paramétrico de máxima verosimilitud (NPMLE, por su sigla en inglés) para estos datos.

El estimador de Turnbull, se basa en una muestra de intervalos observados $[L_i, R_i]\,,\ i = 1, \ldots, n$, los cuales contienen las variables aleatorias independientes $T_1, \ldots, T_n$. Como se mencionó antes, una observación exacta de T_i se da sólo si $L_i = R_i$.

Dado este ejemplo, la función de verosimilitud a ser maximizada es la siguiente:

$$L\left(F\right) = \prod_{i=1}^{n} \left[F\left(R_i+\right) - F\left(L_i-\right)\right] \tag{1}$$

Para resolver este problema de maximización, Peto (1973) define dos conjuntos $\gamma = \{L_i,\ i = 1, \ldots, n\}$ y $\kappa = \{R_i,\ i = 1, \ldots, n\}$ que contienen los extremos izquierdos y derechos de los intervalos, respectivamente. Si se denotan los incrementos de la función F dentro de los intervalos $[q_j, p_j]$ como $s_j, j = 1, \ldots, m$, entonces $L\left(F\right)$ debe ser maximizada como una función de $s_1, s_2, \ldots, s_m$, sujeto a las restricciones $s_j \geq 0$ y $s_m = 1 - \sum_{j=1}^{m-1} s_j$. Peto aborda este problema de maximización usando el algoritmo de Newton-Raphson.

Se puede probar que una función que maximice (1) es constante entre los intervalos $[q_j, p_j]$ e indefinida dentro de ellos. Note que esto implica que $\widehat{P}(T \in (p_{j-1}, q_j)) = 0$ para cualquier j. Como la función de distribución es no decreciente, la cual no es constante entre los intervalos, puede no maximizar a $L\left(F\right)$. Denote los incrementos de F dentro de los intervalos $[q_j, p_j]$ por $s_j\ j = 1, \ldots, m$, $L\left(F\right)$ debe ser maximizada como una función de $s_1, s_2, \ldots, s_m$ sujeto a $s_j \geq 0$ y $s_m = 1 - \sum\limits_{j=1}^{m-1} s_j$. Peto aborda este problema de maximización usando el algoritmo de Newton-Raphson. En contraste con Peto, Turnbull (1976) propone el uso del algoritmo de auto-consistencia para el mismo problema de maximización. La idea de este algoritmo fue presentada primero por Efron (1967), y su aplicación para la maximización en (1) es como sigue.

Sea $\alpha_{ij} = I_{\{[q_j,p_j]\in[L_i,R_i]\}},\ i = 1, \ldots, n,\ j = 1, \ldots, m$, las variables indicadoras que confirman si el intervalo $[q_j, p_j]$ se encuentra dentro o no del intervalo $[L_i, R_i]$; entonces, la probabilidad de que T_i se encuentre dentro del intervalo $[q_j, p_j]$ dado un vector $s = (s_1, s_2, \ldots, s_m)'$ está dada por:

$$\mu_{ij}\left(s\right) = \frac{\alpha_{ij} s_j}{\sum\limits_{k=1}^{m} \alpha_{ik} s_k} \tag{2}$$

puesto que $\widehat{F}$ es constante fuera de los intervalos $[q_j, p_j]$. La proporción de observaciones en el intervalo $[q_j, p_j]$ es igual a:

$$\pi_j\left(s\right) = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}\left(s\right) \tag{3}$$

y un vector $s = (s_1, s_2, \ldots, s_m)'$ es llamado auto-consistente, si

$$s_j = \pi_j\left(s\right),\ j = 1, \ldots, m$$

Siguiendo esta definición, el algoritmo de auto-consistencia de Turnbull para el cálculo del estimador no paramétrico de $F(t)$ se puede implementar siguiendo estos pasos:

1. Obtenga estimaciones iniciales de $\mathbf{s}$; por ejemplo, $s_j^{(0)} = \frac{1}{m},\ j = 1, \ldots, m$.
2. Para $i = 1, \ldots, n,\ j = 1, \ldots, m$, calcule $\mu_{ij}\left(\mathbf{s}^{(0)}\right)$ acorde a (2), y luego $\pi_j\left(\mathbf{s}^{(0)}\right)$ de acuerdo a (3).
3. Obtenga estimaciones mejoradas para $\mathbf{s}$ hallando $s_j^{(1)} = \pi_j\left(\mathbf{s}^{(0)}\right)$.
4. Retorne al paso 2, reemplazando $\mathbf{s}^{(0)}$ por $\mathbf{s}^{(1)}$ y continúe hasta que se logre la convergencia.

3.2. Estudio de simulación

Para establecer el efecto de la imputación de fallas exactas cuando en realidad se tiene una censura a intervalo, sobre la estimación de la función de supervivencia se utilizarán datos de falla lognormales con parámetros fijos para la simulación en valores $\mu = 3.78419$ y $\sigma = 0.133$, que se escogieron de tal forma que se emulan las condiciones de falla de los individuos presentes en el estudio de EA descrito en la sección 2 (tales valores son una estimación paramétrica de datos de fallas exactas generados de la función de supervivencia estimada mediante Turnbull, con el método de la transformación inversa de probabilidad integral; Kalbfleisch (1985)).

Se asume un punto de partida aleatorio para que el individuo comience sus visitas al estudio, en donde se registrará si éste tiene o no el evento. Así, se construyen intervalos de tiempo de una de las siguientes formas:

- $(0, U_i]$ un individuo llegó al estudio en el tiempo U_i pero ya tenía el evento de interés (esto constituye una censura a izquierda, la cual se puede ver como una censura a intervalo),
- $(L_i, U_i]$ un individuo llegó al estudio y asistió a visitas regulares, y en el tiempo L_i fue la última visita en la cual no tenía el evento pero al volver en la siguiente visita (al tiempo $U_i = L_i + \text{TEV}$, con TEV: el tiempo entre visitas) el individuo ya tiene el evento de interés (esto también constituye una censura a intervalo), y
- $(L_i, +\infty)$ un individuo llegó al estudio, asistió a varias visitas regulares, y en el tiempo L_i fue la última visita de la que se tiene registro del individuo en el estudio, sin que éste haya presentado el evento (esto constituye una censura a derecha).

Con este esquema de datos, no se tienen tiempos de falla exactos (aunque también las fallas exactas se pueden considerar como censuras a intervalo con $L_i = U_i$) y todos los datos deben entrarse al análisis como intervalos de tiempo.

Los factores de simulación que se van a variar son:

1. Método de imputación (MI): de acuerdo a la literatura se estudiarán los casos en que las censuras de intervalo son imputadas a través del punto medio del intervalo (PM) y utilizando el extremo derecho del mismo (ED). Lo cual lleva a tiempos de falla "exactos" y facilita los análisis, ya que la estimación de Kaplan-Meier (KM) para la curva de supervivencia puede ser estimada. Además, se considera el caso en que ninguna imputación es llevada a cabo (NI), es decir, usando los datos en forma de intervalos de tiempo, lo cual necesariamente lleva a utilizar el estimador de Turnbull (TB) para la función de supervivencia que tiene en cuenta censura arbitraria.

2. Tiempo entre visitas (TEV): indica con qué frecuencia los individuos asisten a los controles en el estudio. Interesan valores de TEV $= 1, 2, 4$ y 6 años.

3. Tamaño de la muestra (n): este factor tiene como objetivo establecer el efecto sobre el proceso de estimación del número de individuos en el estudio. Se tomarán valores de $n = 50, 100, 200, 500$.

Se utilizará como control para comparar el desempeño de las estimaciones el estimador KM, bajo los métodos de imputación "$\widehat{S}(t)_{\mathrm{PM}}$" y "$\widehat{S}(t)_{\mathrm{ED}}$", y el estimador de Turnbull "$\widehat{S}(t)_{\mathrm{TB}}$", a la función de supervivencia real, notada "$S(t)$". Esto permite, a través de las diferencias observadas entre cada una de las curvas "$\widehat{S}(t)_{\mathrm{TB}}$", "$\widehat{S}(t)_{\mathrm{PM}}$" y "$\widehat{S}(t)_{\mathrm{ED}}$", y la curva de supervivencia de referencia "$S(t)$", establecer el efecto de la imputación sobre la estimación.

Para comparar las curvas de supervivencia resultantes de la simulación, se generan $N = 1000$ muestras independientes para cada uno de los 16 escenarios de simulación (resultantes de las combinaciones de los niveles de los factores TEV y n). Luego, en cada escenario se realizan las estimaciones de la función de supervivencia, de acuerdo al factor de imputación: $\widehat{S}(t)_{\mathrm{PM}}$, $\widehat{S}(t)_{\mathrm{ED}}$ y $\widehat{S}(t)_{\mathrm{TB}}$, y se comparan con la función de supervivencia de control $S(t)$. Tal comparación se realiza usando el error cuadrático medio integrado (ECMI) como una medida global de error. Para calcular el ECMI con $N = 1000$ simulaciones en cada escenario, se utiliza la siguiente fórmula:

$$\mathrm{ECMI}_i = \frac{1}{N}\sum_{j=1}^{N}\int \left[\widehat{S}_j\left(t\right)_i - S\left(t\right)\right]^2 dt$$

donde $i = \mathrm{TB}, \mathrm{PM}, \mathrm{ED}$ representa el método de estimación de la función de supervivencia y $S(t)$ es la función de supervivencia real.

Adicionalmente, para establecer dónde se dan las diferencias entre las curvas de supervivencias estimadas con la real, se calculó el error cuadrático medio (ECM) en la estimación de los cuantiles $q_{0.05}$, $q_{0.1}$, $q_{0.25}$, $q_{0.5}$, $q_{0.75}$, $q_{0.9}$, $q_{0.95}$, de manera que se establece el correspondiente sesgo de estimación de los métodos estudiados (TB, ED y PM). El ECM se calculó para $i = \mathrm{TB}, \mathrm{PM}, \mathrm{ED}$ y $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ como:

$$\mathrm{ECM}_{i,h} = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{q}_{h,i,j} - q_h\right)^2$$

donde $\widehat{q}_{h,i,j}$ son $N = 1000$ estimaciones en cada uno de los métodos estudiados $i = \text{TB}, \text{PM}, \text{ED}$ de los cuantiles reales $q_h,\ h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ de la distribución lognormal con parámetros $\mu = 3.78419$ y $\sigma = 0.133$.

4. Resultados

4.1. Estudio de simulación

4.1.1. Diferencias en las funciones de supervivencia

La medida de error que se utiliza para comparar las estimaciones basadas en imputación $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{TB}}$, con la función de supervivencia verdadera $S(t)$, es el ECMI definido en la sección anterior. Un valor pequeño del ECMI indica que el método de estimación correspondiente produce una curva de supervivencia estimada que es muy cercana a la curva de supervivencia real a lo largo del tiempo; por el contrario, valores altos del ECMI indican que las curvas comparadas tienen diferencias a lo largo del tiempo.

La tabla 1 muestra los ECMI obtenidos en cada uno de los 16 escenarios de simulación considerados.

Tabla 1: ECMI estimado con los métodos TB, PM y ED.

n	TEV	ECMI$_{TB}$	ECMI$_{ED}$	ECMI$_{PM}$
50	1	3.04	42.11	32.51
50	2	2.24	30.25	32.04
50	4	2.50	29.69	31.68
50	6	2.95	34.16	32.06
100	1	1.58	41.45	31.85
100	2	1.18	28.82	31.30
100	4	1.35	29.28	31.53
100	6	1.64	33.21	31.24
200	1	0.87	41.17	30.88
200	2	0.71	28.82	30.63
200	4	0.76	28.64	30.86
200	6	0.96	33.15	30.79
500	1	0.45	40.73	30.78
500	2	0.38	28.31	30.52
500	4	0.42	28.36	30.53
500	6	0.50	32.74	30.62

En todos los escenarios de simulación el ECMI muestra que las funciones estimadas $\widehat{S}(t)_{\text{PM}}$ y $\widehat{S}(t)_{\text{ED}}$ difieren significativamente de $S(t)$, lo cual indica que las estimaciones basadas en estas curvas pueden estar muy alejadas de la realidad. Por otro lado, el ECMI asociado a la estimación de Turnbull ($\widehat{S}(t)_{\text{TB}}$) tiene los valores más pequeños en todos los escenarios, lo cual sucede sin importar el tamaño de muestra. Sin embargo, a medida que el tamaño de muestra aumenta, este error disminuye su valor. En el análisis del tiempo entre visitas (TEV) se puede observar que hay un patrón consistente en todos los valores del tamaño de muestra

considerados, que indica que TEV= 2 años provoca un ECMI menor que en los demás valores de TEV.

La figura 1 ilustra uno de los escenarios considerados en el estudio de simulación ($n = 500$, TEV $= 2$), donde claramente se observan diferencias entre las curvas de supervivencia estimadas usando los diferentes métodos de imputación y la supervivencia real, mientras que la supervivencia estimada mediante Turnbull se ajusta bien a esta última.

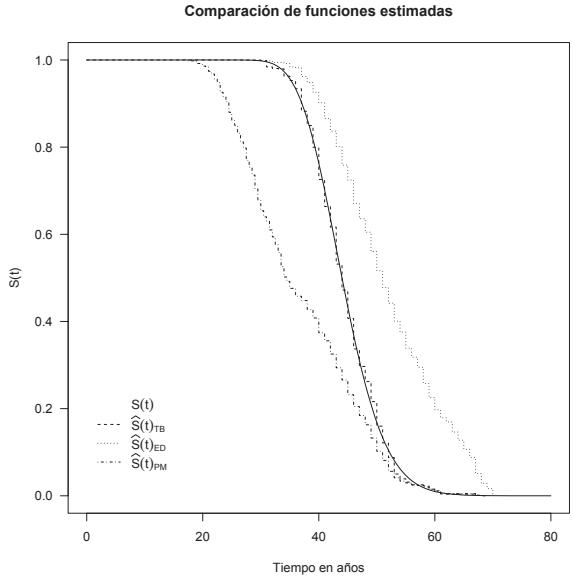

FIGURA 1: Diferencias entre la curva real y las curvas estimadas mediante Turnbull y KM(ED) y KM(PM). Una realización del caso simulado con $n = 500$ y TEV $= 2$.

4.1.2. Diferencias en las edades de inicio

Para el caso ilustrado en la figura 1, se estimaron la edad de inicio y sus respectivos límites de confianza en cada una de las curvas de supervivencia estimadas, mediante el método bootstrap percentil, lo cual se resume en la tabla 2. Detalles del proceso de estimación bootstrap se encuentran en Acosta-Baena et al. (2011), Meeker & Escobar (1998).

Observe que las edades estimadas de inicio de la EA obtenidas por imputación de datos (PM y ED) difieren significativamente del valor de referencia, mientras que el método de Turnbull estima bien. Esto se repite en todos los demás escenarios considerados.

Tabla 2: Estimaciones de la edad de inicio para datos simulados.

	Mediana	LI95 %	LS95 %
Referencia	44.00000	–	–
TB	44.00006	43.00002	44.99997
KM(ED)	51.00004	51.00002	52.00000
KM(PM)	34.00003	33.99999	38.99993

4.1.3. Sesgos de estimación de algunos cuantiles

A continuación se presentan los ECM calculados en los métodos estudiados.

Tabla 3: ECM para las estimaciones de los cuantiles q_h, $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método TB.

n	TEV	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
50	1	7.34	5.76	4.67	3.96	5.71	9.92	12.96
50	2	6.30	4.54	3.13	2.96	3.72	6.55	9.24
50	4	7.56	5.43	3.72	3.13	3.42	5.86	7.84
50	6	9.24	5.86	4.45	3.96	4.37	6.76	9.67
100	1	3.88	2.92	2.10	2.10	2.66	5.06	7.51
100	2	3.57	2.56	1.61	1.44	1.93	3.42	5.57
100	4	4.16	3.10	2.10	1.66	2.02	3.13	4.84
100	6	5.38	3.65	2.59	2.19	2.37	3.69	5.06
200	1	2.16	1.64	1.14	1.32	1.56	2.59	4.28
200	2	1.99	1.32	0.98	0.98	1.25	1.80	2.89
200	4	2.53	1.72	1.12	1.06	1.25	1.77	2.31
200	6	2.96	2.10	1.49	1.32	1.46	2.07	2.86
500	1	0.94	0.77	0.61	0.66	0.85	1.25	1.72
500	2	0.86	0.74	0.58	0.59	0.59	0.92	1.32
500	4	1.19	0.86	0.64	0.62	0.67	0.94	1.28
500	6	1.44	1.08	0.77	0.69	0.72	1.04	1.37

Note que los sesgos de estimación al utilizar el método TB (tabla 3) son menores que los obtenidos con los métodos de imputación PM y ED (tablas 4 y 5, respectivamente). En particular, los sesgos de estimación asociados al método de imputación PM (tabla 4) son mayores en los cuantiles más pequeños, mientras que para el método de imputación ED (tabla 5) los sesgos mayores se presentan en los cuantiles más grandes.

Ahora, en general (tablas 3, 4, y 5) observe que a medida que el tamaño de muestra aumenta, los sesgos medidos con el ECM disminuyen, y que los resultados señalan que el tiempo óptimo entre visitas sería de dos años, ya que en este caso los ECM resultaron menores que en los demás valores de este factor.

4.2. Aplicación con datos reales

Para los datos de EA, se aplicaron las diferentes técnicas de estimación de la función de supervivencia, y con base en ellas se calculó la mediana como estimador de la edad de inicio de la enfermedad.

Tabla 4: ECM para las estimaciones de los cuantiles q_h, $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método de imputación PM.

n	TEV	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
50	1	166.67	168.48	139.95	88.55	14.29	14.29	17.81
50	2	166.67	169.00	140.19	85.38	11.42	7.56	11.16
50	4	167.18	169.26	139.00	84.64	13.10	7.90	7.90
50	6	168.22	169.26	139.95	86.49	12.89	7.18	7.62
100	1	165.12	163.58	140.42	88.92	8.70	8.64	12.82
100	2	164.61	164.61	139.95	82.08	9.30	4.33	6.60
100	4	165.12	164.61	139.48	83.17	10.30	4.54	4.41
100	6	164.10	164.10	138.06	80.28	9.86	4.00	3.84
200	1	169.26	166.67	139.95	91.58	5.90	3.92	9.42
200	2	168.48	166.41	140.42	85.93	8.29	2.66	3.24
200	4	167.44	165.64	139.71	86.12	9.30	3.24	2.69
200	6	168.74	165.89	139.95	85.93	8.94	2.96	2.28
500	1	169.78	166.41	139.71	94.67	4.93	1.51	3.61
500	2	169.52	166.15	140.66	89.11	7.84	1.90	1.39
500	4	170.04	166.67	140.19	89.49	8.64	2.59	1.56
500	6	169.52	165.89	140.42	89.30	8.24	2.28	1.25

Tabla 5: ECM para las estimaciones de los cuantiles q_h, $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método de imputación ED.

n	TEV	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
50	1	23.33	27.56	45.83	102.01	180.63	199.66	176.62
50	2	16.32	16.65	27.98	58.06	135.02	177.69	159.52
50	4	22.28	23.33	32.95	54.17	114.06	157.25	155.75
50	6	27.88	29.92	40.20	64.32	121.66	160.78	156.50
100	1	18.23	22.75	42.51	97.81	183.60	205.92	178.49
100	2	15.29	16.16	26.01	56.40	133.40	181.44	166.41
100	4	20.98	22.75	31.81	53.44	111.94	163.58	158.26
100	6	26.94	30.03	39.94	62.25	117.07	164.61	158.00
200	1	16.08	21.44	40.83	94.67	181.98	207.65	181.98
200	2	13.76	15.60	24.70	54.61	133.86	184.14	170.82
200	4	19.36	21.90	30.80	53.29	111.72	165.64	160.78
200	6	25.00	29.48	39.44	63.36	116.86	167.96	162.05
500	1	15.37	20.70	39.69	94.09	183.87	209.38	183.33
500	2	13.10	15.21	24.30	53.58	132.94	184.42	171.87
500	4	18.75	21.53	30.47	52.85	111.30	167.44	163.58
500	6	24.11	28.73	39.44	62.73	117.29	169.52	163.84

La tabla 6 muestra cómo es la estimación de la edad de inicio de la enfermedad.

Tabla 6: Estimaciones de la edad de inicio para datos de EA.

	Mediana	LI95 %	LS95 %
TB	44.01006	43.01003	45.01003
KM(ED)	47.00998	46.00002	47.99997
KM(PM)	44.00499	42.00502	45.00498

Los resultados anteriores muestran que las estimaciones que usan TB y PM, estiman la edad de inicio a los 44 años, mientras que el método ED sobrestima tal valor. A nivel de intervalos de confianza, el método de Turnbull es más preciso que el método PM en la estimación de la edad de inicio de la enfermedad.

La figura 2 muestra las diferencias apreciables entre las curvas estimadas.

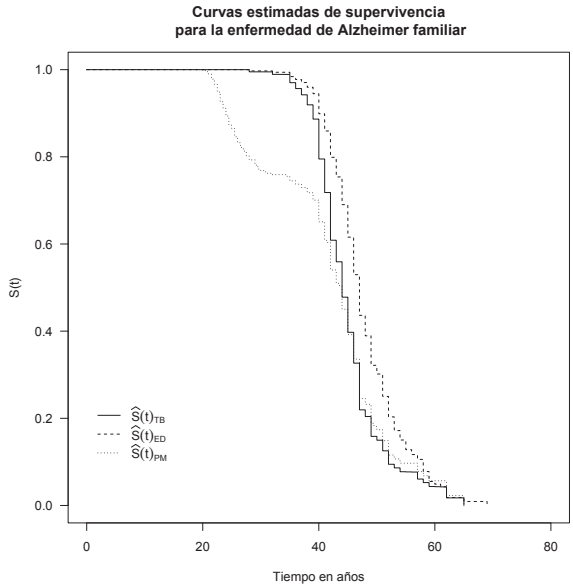

FIGURA 2: Funciones de supervivencia estimadas para los datos de EA.

Note que aunque las estimaciones de la edad de inicio que usan TB y PM son similares, estimaciones de otros cuantiles, particularmente cuantiles más pequeños, pueden llevar a errores apreciables. Esto puede deberse principalmente a que los datos de EA familiar incluyen un 21 % de datos con censura a izquierda (Acosta-Baena et al. 2011).

5. Conclusiones y recomendaciones

- En las últimas décadas existe gran interés en todo el mundo por definir adecuadamente el inicio de la EA, incluso etapas preclínicas y prodrómicas, con el objetivo de detectar la enfermedad de manera más temprana y ofrecer alternativas de tratamiento más oportuno (Petersen, Stevens, Ganguli, Tangalos, Cummings & DeKosky 2001, Reisberg, Ferris, Kluger, Franssen, Wegiel & de Leon 2008). Conocer adecuadamente la edad de inicio de esta cohorte de portadores de una mutación con irremediable inicio de la enfermedad de Alzheimer tiene utilidad para el diseño de ensayos clínicos dirigidos a tratamientos preventivos (Strobel 2011).

- En los análisis realizados, las edades de inicio de la EA obtenidas por imputación de datos (PM y ED) difieren significativamente de los datos reales en todos los tamaños de muestra y en los diferentes TEV, mientras que el método de Turnbull estima bien en todos los escenarios. También puede concluirse que un tiempo entre visitas igual a 2 años, independiente del tamaño de muestra, es óptimo para estimar la edad de inicio de la EA familiar, ya que en este caso se presentaron diferencias más pequeñas que las obtenidas en los escenarios restantes.
- El análisis de los resultados de la estimación de sesgos para los cuantiles q_h, $h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando los métodos TB, PM y ED (tablas 3, 4 y 5), muestra que en general el método TB presenta menores sesgos en la estimación que los métodos de imputación. También, como es de esperarse a medida que el tamaño de muestra aumenta, los sesgos medidos con el ECM disminuyen. Los resultados de la tabla 4 establecen que en general el método de imputación, usando el punto medio del intervalo, afecta la estimación de los cuantiles más pequeños, mientras que el método de imputación mediante el extremo derecho del intervalo afecta a los cuantiles más grandes (tabla 5).
- Aunque en la aplicación con datos reales se obtuvieron estimaciones de la mediana muy similares mediante Turnbull y usando la imputación PM, no se puede concluir que esto siempre va a ocurrir, de acuerdo a lo que se evidencia en el estudio de simulación. Sin embargo, el interés del investigador puede estar enfocado en otros cuantiles diferentes a la mediana, donde se podrían dar errores apreciables en la estimación, como se evidenció en la sección 4.1.3.
- De acuerdo a las edades de inicio encontradas con los métodos de imputación, el 50 % de los sujetos portadores de la mutación E280A para EA iniciará con deterioro cognitivo leve a los 47 años (según imputación por ED) o a la edad de 44 años (según imputación por TB y PM). La primera estimación, desde el punto de vista clínico, estaría retrasando un tratamiento preventivo.
- Tanto en los datos simulados como en los datos reales, los intervalos de confianza obtenidos usando TB son más estrechos que los calculados mediante KM, lo cual indica que el método de Turnbull es más preciso.
- La imputación de las censuras arbitrarias presentan grandes errores, con impacto clínicamente importante, como en el caso de esta cohorte de sujetos en riesgo de EA familiar, cuyos resultados sesgados implicarían un error en el diagnóstico, en el tratamiento y, por ende, en el pronóstico de la enfermedad.

Agradecimientos

Los autores agradecen de manera especial a los árbitros y a las editoras invitadas por sus valiosos comentarios que enriquecieron el texto.

Se agradece al CODI (Comité para el Desarrollo de la Investigación) de la Universidad de Antioquia y al programa de sostenibilidad 2010-2011.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

Referencias

Acosta-Baena, N., Sepúlveda-Falla, D., Lopera-Gómez, C. M., Jaramillo-Elorza, M. C., Moreno, S., Aguirre-Acevedo, D. C., Saldarriaga, A. & Lopera, F. (2011), 'Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: A retrospective cohort study', *The Lancet Neurology* **10**(3), 213–220.

Dorey, F. J., Little, R. & Schenker, N. (1993), 'Multiple imputation for threshold-crossing data with interval censoring', *Statistics in Medicine* **12**, 1589–1603.

Efron, B. (1967), 'The two sample problem with censored data', *University of California Press* pp. 831–853.

Giolo, S. R. (2004), 'Turnbull's nonparametric estimator for interval-censored data', *Department of Statistics, Federal University of Paraná* pp. 1–10. Consultado en septiembre 6, 2011.
*www.est.ufpr.br/rt/suely04a.pdf

Hougaard, P. (1999), 'Fundamentals of survival data', *Biometrics* **55**, 13–22.

Iceland, J. (1997), *The Dynamics of Poverty Spells and Issues of Left-Censoring*, PSC Research Report Series January 1997. Consultado en septiembre 6, 2011.
*http://www.psc.isr.umich.edu/pubs/pdf/rr97-378.pdf

Kalbfleisch, J. (1985), *Probability and Statistical Inference*, Vol. 1, 2nd edn, Springer-Verlag, New York.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**(282), 457–481.

Meeker, W. & Escobar, L. (1998), *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.

Odell, P., Anderson, K. & D'agostinho, R. (1992), 'Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model', *Biometrics* **48**, 951–959.

Petersen, R. C., Stevens, J. C., Ganguli, M., Tangalos, E. G., Cummings, J. L. & DeKosky, S. T. (2001), 'Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review)', *Neurology* **56**(9), 1133–1142.

Peto, R. (1973), 'Experimental survival curves for interval-censored data', *Journal of the Royal Statistical Society, Series C* **22**, 86–91.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Consultado en septiembre 6, 2011.
*http://www.R-project.org/

Reisberg, B., Ferris, S. H., Kluger, A., Franssen, E., Wegiel, J. & de Leon, M. J. (2008), 'Mild cognitive impairment (MCI): A historical perspective', *International Psychogeriatrics* **20**(1), 18–31.

Rucker, G. & Messerer, D. (1988), 'Remission duration: An example of interval-censored observation', *Statistics in Medicine* **7**, 1139–1145.

Strobel, G. (2011), Detecting Familial AD Ever Earlier: Subtle Memory Signs 15 Years Before, *in* 'Alzheimer Research Forum'. Consultado en septiembre 6, 2011.
*http://www.alzforum.org/new/detail.asp?id=2725

Turnbull, B. W. (1974), 'Nonparametric estimation of a survivorship function with doubly censored data', *Journal of the American Statistical Association* **69**(345), 169–173.

Turnbull, B. W. (1976), 'The empirical distribution function with arbitrarily grouped censored and truncated data', *Journal of the Royal Statistical Society, Series B* **38**(3), 290–295.