Universitat Autònoma de Barcelona

Propuesta metodológica para el análisis de supervivencia con eventos recurrentes en estudios epidemiológicos cuando el número de eventos previos es desconocido

Gilma Norela Hernández Herrera

Propuesta metodológica para el análisis de supervivencia con eventos recurrentes en estudios epidemiológicos cuando el número de eventos previos es desconocido

Tesis doctoral presentada por Gilma Norela Hernández Herrera para la obtención del título de Doctor en Metodología de la Investigación Biomédica y Salud Pública

Directores:
Albert Navarro i Giné
David Moriña Soler

Universidad Autónoma de Barcelona
Facultad de Medicina, Departamento de Pediatría, Obstetricia y Ginecología, Medicina Preventiva y Salud Pública
Cerdanyola del Vallés, España
2020

Propuesta de una alternativa de análisis de supervivencia para eventos recurrentes con dependencia de evento cuando se desconoce el número de eventos previos

Gilma Norela Hernández Herrera

Tesis presentada como requisito para optar al título de:
Doctora en Metodología de la Investigación Biomédica y Salud Pública

Directores:
Albert Navarro i Giné
David Moriña Soler

Línea de Investigación:
Aplicación de metodología estadística avanzada en Ciencias de la Salud

Universidad Autónoma de Barcelona
Facultad de Medicina, Departamento de Pediatría, Obstetricia y Ginecología, Medicina Preventiva y Salud Pública
Cerdanyola del Vallés, España
2020

A *Juan José*, la prolongación de mi existencia.

Agradecimientos

Esta tesis fue un trabajo en equipo con mis directores y por eso quiero agradecer principalmente a los dos *Albert Navarro y David Moriña*, por que sin su apoyo y trabajo continuo no hubiera sido posible esta tesis. Además me ofrecieron su amistad y afecto que hicieron que me sintiera en casa cuando estaba en un país lejano a mi lugar de origen.

Gracias a los dos por la paciencia y la gran capacidad para guiarme en el desarrollo de todo este trabajo. Gracias Albert por tus llamados de atención en los momentos en que parecía salirme del camino y me dejaba envolver por mis múltiples obligaciones laborales y demás compromisos. Gracias David, por tu soprendente generosidad para apoyarme con todo y para responder a mis inquietudes cada vez que las tuve. No tengo palabras para expresarles la inmensa gratitud que siento hacia ambos, ni mucho menos la forma de demostrarles que no se arrepentirán.

También quiero agradecerles a Miguel, a Mireia, a Sergio, a Claudia y a Fulvio por su hospitalidad en la universidad y su grata compañía en el despacho que me albergó cuando estuve por allí. Fueron grandes compañeros en este camino y espero volverlos a encontrar.

A Ceci, gracias por apoyarme y acompañarme durante todas mis estancias en la UAB, siempre estuviste ahí para escucharme y compartir mis alegrías y tristezas y para ayudarme con la gestión de todos los aspectos administrativos de mi doctorado. Nunca te olvidaré.

A mi familia: Juan Carlos y Juan José, por el apoyo incondicional, por soportar mis largas ausencias y por la confianza que siempre me han tenido y que me ha impulsado a luchar por muchos objetivos. A mi padre (QED), por su amor y por creer en mí siempre.

A mis amigos: Juan Carlos C, Marcela V, René C, Cristina O, Carlos T, Pilar P, Claudia V, Nora R, Vianey, Diego Ch y muchos otros, quienes han estado ahí siempre para reir y llorar conmigo, escucharme y apoyarme en todos los proyectos que emprendo.

A Carlos Palacio, decano de la facultad de medicina de la Universidad de Antioquia por su apoyo y respaldo para que pudiera llevar a cabo mis estudios de doctorado. Y por supuesto a mis compañeros docentes de la facultad que han estado ahí preguntando todos los días por

mis logros y tropiezos con este proyecto y a mis estudiantes quienes también han sido mi motivación para continuar estudiando y mejorando como docente.

Resumen

Introducción En muchos estudios clínicos, epidemiológicos y de salud pública, el evento de interés se presenta a menudo más de una vez y los tiempos entre ocurrencias pueden estar correlacionados. Para este tipo de datos, se requiere contar con métodos estadísticos específicos que permitan dar respuesta a preguntas que surgen en investigaciones epidemiológicas relacionadas con la probabilidad de ocurrencia de nuevos eventos y el impacto de las covariables añadidas. Cuando hay recurrencia de eventos, la experiencia ha mostrado que con mucha frecuencia se presenta dependencia de evento y heterogeneidad individual que conllevan a pensar en métodos de análisis que van más allá del modelo de Cox estándar. En este tipo de análisis, la censura a la izquierda juega un papel determinante, ya que la existencia de información previa puede ayudar a definir la dependencia de evento y con esta información determinar el método de análisis más apropiado y preciso. Sin embargo, la historia previa a menudo es desconocida y es importante establecer estrategias para su manejo.

Objetivo Proponer una alternativa de análisis de supervivencia para eventos recurrentes con dependencia de evento cuando se desconoce el número de episodios previos al inicio del seguimiento para parte (o todos) los sujetos de la muestra.

Metodología Para cumplir el objetivo principal de este estudio, se llevaron a cabo dos fases. La primera incluyó la revisión de métodos para el tratamiento de datos faltantes y la identificación de un método adecuado para imputar estos datos cuando la variable fuese de conteo, a través de un estudio de simulación que permitiera comparar diferentes métodos basados en la distribución Poisson y generalizaciones de esta distribución; teniendo en cuenta además diferentes escenarios de dispersión y diferentes porcentajes de datos faltantes. La segunda fase incluyó la revisión de modelos disponibles para el análisis de supervivencia cuando hay eventos recurrentes con dependencia de evento y heterogeneidad individual y se desarrolló un estudio de simulación que permitiera identificar el mejor modelo para ajustar a datos de eventos recurrentes con estas características, imputando la información previa (censura a izquierda) con el método encontrado en la primera simulación.

Resultados En la primera fase, se encontró que el modelo adecuado para imputar esta variable, cuando hay sobredispersión, equidispersión, infradispersión o exceso de ceros es la imputación múltiple usando la distribución COMPoisson.

En la segunda fase, se encontró que los modelos PWP-CP y PWP-GT incluyendo el compo-

nente de fragilidad, estratificando según si se había estado a riesgo previo, e imputando los episodios previos mostraron mejor rendimiento.

Adicional a los dos estudios de simulación, y teniendo en cuenta la necesidad de software disponible para este tipo de análisis, se construyó un paquete en R *miRecSurv* puesto a disposición de la comunidad científica para su uso.

Conclusiones Para la imputación de datos faltantes cuando se tiene una variable de conteo, la decisión del método se basa en el análisis de dispersión de la variable y en el porcentaje de datos perdidos. Según los resultados de simulación en este estudio, el modelo COMPoisson se comporta bien para la imputación de variable de conteo ya que es flexible en cuanto al manejo de esta variable con características de sobre y subdispersión, así como con equidispersión.

La propuesta de modelo realizada en esta tesis parece funcionar razonablemente bien para la mayoría de situaciones estudiadas, en general mucho mejor que la alternativa del uso de modelos con riesgo basal común.

Si el fenómeno de interés se genera a partir de funciones de riesgo constantes y cuando todos los sujetos de la muestra han estado a riesgo del evento antes del inicio del seguimiento, el modelo propuesto en su formulación *gap time* parece ser el más adecuado. Y si el fenómeno de interés se genera a partir de funciones de riesgo no constantes, el modelo más adecuado es el propuesto con formulación *counting process*, al menos hasta el 50 % de sujetos a riesgo previo al inicio del seguimiento.

Palabras clave: Eventos recurrentes, dependencia de evento, heterogeneidad individual, datos faltantes, imputación, COMPoisson, censura a izquierda

Abstract

Introduction In many clinical, epidemiological, and public health studies, the event of interest often occurs more than once, and the times between occurrences may be correlated. For this type of data, it is required to have specific statistical methods that allow answering questions that arise in epidemiological investigations related to the probability of occurrence of new events and the impact of the additional covariates. When there is a recurrence of events, experience has shown that event dependence and individual heterogeneity are very common, leading to the need of analysis methods that go beyond the standard Cox model. In this type of analysis, left censorship plays a determining role, since the existence of prior information can help define the event dependency and with this information determine the most appropriate and precise analysis method. However, the previous history is often unknown and it is important to establish strategies for its management.

Objective Propose an alternative survival analysis for recurrent events with event dependence when the number of episodes prior to the start of follow-up is unknown for part (or all) of the subjects in the sample.

Methodology To fulfill the main objective of this study, two phases were carried out. The first included the review of methods for the treatment of missing data and the identification of an adequate method to impute these data when the variable was discrete, through a simulation study that allowed the comparison of different methods based on the Poisson distribution and generalizations of this distribution, also taking into account different dispersion scenarios and different percentages of missing data. The second phase included the review of models available for survival analysis when there are recurrent events with event dependence and individual heterogeneity and a simulation study was developed to identify the best model to adjust data from recurrent events with these characteristics, imputing the previous information (left censorship) with the method found in the first phase.

Results In the first phase, it was found that the appropriate model to impute this variable, when there is over-dispersion, equidispersion, under-dispersion or excess of zeros is multiple imputation using the COMPoisson distribution.

In the second phase, it was found that the PWP-CP and PWP-GT models including the frailty component, stratifying according to whether they had been at previous risk, and imputing the previous episodes, showed better performance.

In addition to the two simulation studies, and taking into account the need for available software for this type of analysis, a package was built in R *miRecSurv* and made available to the scientific community for its use.

Conclusions For the imputation of missing data when dealing with a count variable, the decision of the optimal method is based on the dispersion analysis of the variable and on the percentage of missing data. According to the simulation results in this study, the COM-Poisson model behaves well for the imputation of the counting variable since it is flexible in terms of handling variables showing over and sub-dispersion, as well as with equidispersion.

The model proposal made in this thesis seems to work reasonably well for most of the situations studied, in general much better than the alternative of using models with common baseline risk.

If the phenomenon of interest is generated from constant risk functions, and when all the subjects in the sample have been at risk of the event before the start of the follow-up, the model proposed in its *gap time* formulation seems to be the more appropriate. And if the phenomenon of interest is generated from non-constant risk functions, the most appropriate model is the one proposed with a *counting process* formulation, at least up to 50 % of subjects at risk prior to the start of follow-up.

Keywords: recurrent events, dependence event, heterogeneity, missing data, imputation, COMPoisson distribution, left censoring

Resum

Introducció

En molts estudis clínics, epidemiològics i de salut pública, l'esdeveniment d'interès sovint es presenta més d'una vegada i els temps entre ocurrències poden estar correlacionats. Per a aquest tipus de situacions, es requereix comptar amb mètodes estadístics específics que permetin donar resposta a preguntes que sorgeixen en investigacions epidemiològiques relacionades amb la probabilitat d'ocurrència de nous esdeveniments i l'impacte de les covariables afegides. Quan hi ha recurrència d'esdeveniments, l'experiència ha mostrat que molt sovint es presenta dependència d'esdeveniment i heterogeneïtat individual que obliguen a pensar en mètodes d'anàlisi que van més enllà del model de Cox estàndard. En aquest tipus d'anàlisi, la censura a l'esquerra juga un paper determinant, ja que l'existència d'informació prèvia pot ajudar a definir la dependència d'esdeveniment i amb aquesta informació determinar el mètode d'anàlisi més apropiat i precís. No obstant això, la història prèvia sovint és desconeguda i és important establir estratègies per al seu maneig.

Objetiu Proposar una alternativa d'anàlisi de supervivència per a esdeveniments recurrents amb dependència d'esdeveniment quan es desconeix el nombre d'episodis previs a l'inici de l'seguiment per part (o tots) els subjectes de la mostra.

Metodologia Per assolir l'objectiu principal d'aquest estudi, es van dur a terme dues fases. La primera va incloure la revisió de mètodes per al tractament de dades mancants i la identificació d'un mètode adequat per imputar aquestes dades quan la variable és de recompte, a través d'un estudi de simulació que permetés comparar diferents mètodes basats en la distribució Poisson i generalitzacions d'aquesta distribució; tenint en compte a més diferents escenaris de dispersió i diferents percentatges de dades mancants. La segona fase va incloure la revisió de models disponibles per a l'anàlisi de supervivència quan hi ha esdeveniments recurrents amb dependència d'esdeveniment i heterogeneïtat individual, i es va desenvolupar un estudi de simulació que permetés identificar el millor model per ajustar a dades d'esdeveniments recurrents amb aquestes característiques, imputant la informació prèvia (censura a esquerra) amb el mètode trobat a la primera simulació.

Resultats En la primera fase, es va trobar que el model adequat per imputar aquesta variable, quan hi ha sobredispersió, equidispersió, infradispersió o excés de zeros és la imputació múltiple basada en la distribució COMPoisson.

En la segona fase, es va trobar que els models PWP-CP i PWP-GT incloent el component

de fragilitat, estratificant segons si s'havia estat a risc previ, i imputant els episodis prèvia mostrava el millor rendiment.

Addicionalment als dos estudis de simulació, i tenint en compte la necessitat de programari disponible per a aquest tipus d'anàlisi, es va construir un paquet en R *miRecSurv* posat a disposició de la comunitat científica per al seu ús.

Conclusions Per a la imputació de dades mancants quan es té una variable de recompte, la decisió del mètode es basa en l'anàlisi de dispersió de la variable i en el percentatge de dades perdudes. Segons els resultats de simulació en aquest estudi, el model COMPoisson es comporta bé per a la imputació de variable discreta ja que és flexible pel que fa a l'ús d'aquesta variable amb característiques de sobre i subdispersió, així com amb equidispersió.

La proposta de model realitzada en aquesta tesi sembla funcionar raonablement bé per a la majoria de situacions estudiades, en general molt millor que l'alternativa de l'ús de models amb risc basal comú.

Si el fenomen d'interès es genera a partir de funcions de risc constants, i quan tots els subjectes de la mostra han estat a risc de patir l'esdeveniment abans de l'inici del seguiment, el model proposat en la seva formulació *gap time* sembla ser el més adequat. I si el fenomen d'interès es genera a partir de funcions de risc no constants, el model més adequat és el proposat amb formulació *counting process*, almenys fins al 50 % de subjectes a risc previ a l'inici del seguiment.

Paraules clau: Esdeveniments recurrents, dependència d'esdeveniment, heterogeneïtat individual, dades mancants, imputació, COMPoisson, censura a l'esquerra

Contenido

Lista de Figuras

Lista de Tablas

1. Presentación

El presente trabajo se enfoca en el estudio y propuesta de modelos de supervivencia para eventos recurrentes cuando hay dependencia de evento, que son fenómenos muy comunes en investigación en salud. Es frecuente encontrar, en el área de la salud, situaciones donde se requiera analizar la recurrencia de un evento y los tiempos entre ocurrencias, sin embargo, tal vez con la misma frecuencia, se desconoce información de la historia previa de los individuos, necesaria para estos análisis. En este trabajo se propone una alternativa de análisis cuando ante un evento recurrente, con dependencia de evento, se desconocen los episodios previos para algunos (o todos) los sujetos de la muestra, combinando técnicas de imputación y el uso de términos de fragilidad.

El trabajo se estructura en doce capítulos partiendo de la definición de eventos recurrentes y ejemplos en salud, así como de los conceptos de Dependencia de evento y Heterogeneidad individual. En el segundo capítulo "Modelos de supervivencia para eventos recurrentes", se presenta una revisión de algunos modelos utilizados en el análisis de supervivencia para eventos recurrentes cuando se presentan los fenómenos de heterogeneidad individual y dependencia de evento. En el capítulo 4 se detalla el concepto de censura a la izquierda, haciéndo énfasis en la historia previa de eventos y lo que puede ocurrir cuando esta es desconocida. En los capítulos 5 y 6 se presentan la justificación del estudio, la pregunta de investigación y los objetivos planteados. Luego en el capítulo 7 se ilustra la propuesta de la tesis: métodos para el análisis de un evento recurrente cuando se desconocen los eventos previos. Dentro de la propuesta del capítulo anterior, se plantea el uso de métodos de imputación de estos eventos previos y por ello en el capítulo siguiente se presenta una revisión del tema de *missing data* y algunas estrategias de manejo de datos perdidos por imputación. Ya en el capítulo 9 se muestra el rendimiento del método propuesto de análisis de supervivencia para eventos recurrentes imputando la historia previa. En el capítulo 10 se describe la librería de R desarrollada para permitir el análisis de datos de eventos recurrentes imputando la información previa y que se pondrá a disposición de la comunidad académica e investigadora para su uso, luego en el capítulo 11 se muestra la discusión de los resultados del estudio, estructurando este capítulo en varios apartados. Finalmente en el capítulo 12 se encuentran las conclusiones de la investigación.

Si bien la tesis que se presenta lo es en formato manuscrito, se ha optado por incluir literalmente, y en el formato exigido por la revista en cuestión, los artículos que ya se encuestran publicados o sometidos, y que verán que ocupan apartados o capítulos enteros del texto, con una breve descripción inicial en cada caso. Así, el capítulo 8 está compuesto por dos artículos, uno publicado y otro actualmente en revisión en una revista especializada. Igualmente el contenido de los capítulos 9 y 10 está presentado en formato artículo, uno para cada capítulo. En los anexos de esta tesis se encuentra disponible el material suplementario de cada uno de estos trabajos.

2. Eventos recurrentes, definición y ejemplos en salud

2.1. Definición evento recurrente

En la investigación en salud, generalmente, se encuentran dos tipos de eventos, no reversibles y reversibles. Los eventos no reversibles son de naturaleza crónica y le ocurren a un individuo solo una vez, por ejemplo, hipertensión, SIDA, diabetes. Para este tipo de eventos, existen diversos métodos estadísticos que permiten su análisis cuando se pretende conocer el tiempo hasta presentar este evento, tal como un análisis de supervivencia para tiempo al evento. Los eventos reversibles que son de naturaleza aguda y pueden ocurrirle a un individuo más de una vez, pueden además presentarse como eventos múltiples o como eventos recurrentes. Los eventos múltiples son aquellos eventos repetidos que no son exactamente del mismo tipo pero están algo relacionados, como la hospitalización repetida por diferentes situaciones, mientras que los eventos recurrentes son aquellos eventos repetidos, que son del mismo tipo [Yadav *et al.*, 2018].

Los datos de eventos recurrentes tienen dos características principales que son: correlación intra sujeto y covariables dependientes del tiempo. La correlación intra-sujeto puede deberse a la dependencia de evento o la heterogeneidad individual. La correlación intra-sujeto debida a la dependencia de evento se refiere a una situación en la que un evento en sí mismo acelera o desacelera la tasa de eventos posteriores [Box-Steffensmeier y De Boef, 2006]. Por ejemplo, después que le ocurra un primer ataque cardíaco a un sujeto, las posibilidades de que ocurra un segundo ataque cardíaco aumentan porque durante el primer ataque cardíaco se daña una parte del corazón. Ahora, la correlación intra sujeto debida a la heterogeneidad individual se refiere a la situación en la que algunos sujetos son más propensos a experimentar un mayor número de eventos que otros sujetos por razones desconocidas, no consideradas o que no se pueden medir.

En salud, fenómenos como la epilepsia, las hospitalizaciones por problemas de falla cardíaca, recaídas por cáncer, crisis asmáticas, entre otros, son fenómenos que se pueden presentar más de una vez en un mismo individuo durante un periodo de seguimiento t. Para estos fenómenos el interés se centra en el tiempo hasta el primer evento y el tiempo entre los eventos sucesivos, así como el número de ocurrencias durante este intervalo de tiempo t; además, en la identificación de los factores que pueden aumentar o disminuir la probabilidad de sufrir más eventos en dicho intervalo para ayudar a los investigadores en la predicción de los tiempos interocurrencias y la probabilidad de nuevas ocurrencias.

En todas estas situaciones, es importante analizar tanto el tiempo al primer evento, como el tiempo entre ocurrencias y el número de ocurrencias durante un cierto tiempo de seguimiento. Los resultados de estos análisis se convierten en información útil para la evaluación de tratamientos o intervenciones, en la medida que las estimaciones de los tiempos entre ocurrencias y el número de recurrencias sean explicadas por factores propios de cada individuo durante el seguimiento e información de la exposición previa a cada ocurrencia.

Para el análisis de supervivencia el modelo clásico utilizado en investigación en salud es el modelo de Cox, indicado para modelar *tiempo al evento*, en el caso concreto, cuando el evento es posible una única vez en un mismo individuo, lo que representa una limitación para analizar eventos recurrentes, más aún cuando los tiempos entre ocurrencia de un mismo individuo están correlacionados, ya sea por heterogeneidad individual o por la dependencia de evento.

Los objetivos frecuentes en el análisis de los datos de sucesos recurrentes implican, los siguientes aspectos: comprender y describir los procesos de eventos individuales, identificar y caracterizar la variación a través de una muestra de procesos, comparar grupos de procesos y por último determinar la relación de covariables, que son tratamientos y/o factores variables en el tiempo que explican la ocurrencia del evento de interés.

Durante las últimas décadas, el desarrollo de métodos estadísticos para el análisis de datos de eventos recurrentes ha crecido de manera importante y se han propuesto varios enfoques. A pesar de varias técnicas poderosas disponibles para el análisis de los datos de eventos recurrentes, la mayoría de los investigadores todavía utilizan técnicas estadísticas tradicionales para analizar sus preguntas de investigación donde el resultado de interés es de naturaleza recurrente [Yadav *et al.*, 2018].

2.2. Algunos ejemplos en investigación en salud de fenómenos recurrentes

Algunos ejemplos de fenómenos con eventos recurrentes reportados en la literatura biomédica son, entre otros:

- Los pacientes con enfermedad renal crónica a menudo son hospitalizados repetidamente por diferentes eventos de enfermedades cardiovasculares (por ejemplo, insuficiencia cardíaca congestiva [ICC], infarto de miocardio o accidente cerebrovascular) [Yang *et al.*, 2017]. Se dan infecciones repetidas en pacientes en diálisis, infecciones del tracto urinario, peritonitis o episodios de rechazo de trasplante y las recurrencias de infecciones en este tipo de pacientes pueden presentar dependencia y heterogeneidad individual. La dependencia de evento, en este caso, se puede dar debido a que la probabilidad de que un paciente presente una infección en un tiempo t está relacionada con el número de infecciones anteriores que haya presentado el paciente. Además, los individuos pueden presentar diferentes tiempos de ocurrencia para una infección dado que unos tienen más vulnerabilidad que otros.

- En pacientes diabéticos los eventos de hipoglicemia se consideran como eventos recurrentes y estos pueden ser explicados por factores como hábitos alimenticios, uso de insulina, edad, entre otros. La presentacioń de estos eventos de manera recurrente puede estar explicada por la falta de adherencia a tratamientos o por comportamientos inadecuados para alimentación o incluso por comorbilidades que presenta el paciente y que hacen que cada individuo presente una heterogeneidad por factores no medidos cuando se pretende analizar los eventos recurrentes [Rojas *et al.*, 2011].

- Un problema de salud pública, que afecta principalmente a jóvenes es el consumo de drogas psicoactivas y para este existen múltiples tratamientos, que en general fallan, debido a las recaídas después de las terapias. En este caso, se requiere la estimación de los tiempos interocurrencias y los factores relacionados con tales recaídas. Las recaídas pueden estar relacionadas con el tipo de droga con el cual se inicia el consumo y con otras variables propias del individuo, lo que lleva a suponer que hay dependencia de evento y heterogenidad individual. En algunos estudios publicados se han utilizado análisis de supervivencia clásico con el modelo de Cox para modelar el tiempo a la primera recaída después del diagnóstico [Ødegård y Rossow, 2004] y análisis de regresión lineal para identificar los factores relacionados con la dosis consumida en la recaída.

- En población menor de 5 años y población adulta mayor las infecciones respiratorias son recurrentes, tales como el asma. En este caso el mayor interés está en las rehospitalizaciones y las múltiples consultas al médico atribuidas a episodios de asma [Cai y Schaubel, 2004]. En general los análisis de este fenómeno suponen independencia de eventos, sin embargo, en los pacientes con asma se puede presentar correlación intra sujeto por las diferencias existentes en las características fisiológicas o genéticas o incluso sociales que llevan a pensar que el supuesto de independencia de evento no se cumpla.

- En deportistas, las lesiones a menudo son recurrentes, y cada ocurrencia puede estar relacionada con lesiones anteriores e influenciar futuras lesiones y, por lo tanto, se debe tener en cuenta la correlación entre eventos para el análisis de dichos datos [Ullah *et al.*, 2014]. Muchos enfoques estadísticos ingenuos restringen la probabilidad inicial de lesión, y la influencia de las covariables es considerada igual para todas las lesiones subsiguientes, cuando, en realidad, tanto la probabilidad de una nueva lesión y la relación de las covariables puede variar según la persona y la clase de lesión. Esta variabilidad implica que algunas personas pueden tener probabilidades de nuevas lesiones más altas o más bajas que otras, lo que significa que en los datos de lesiones recurrentes habrá diferentes correlaciones dentro de la persona entre las personas y que los tiempos de lesiones dentro de la persona serán dependientes.

2.3. Dependencia de evento

Para entender la dependencia de evento, es necesario entender el concepto como la relación existente entre la ocurrencia de eventos consecutivos, es decir, la presentación de un evento en un tiempo t depende de los eventos previos, en otras palabras, el riesgo de una nueva ocurrencia puede modificarse por la historia del individuo. En otras palabras, podemos decir que existe dependencia temporal intrapaciente o modelo de contagio.

$$h_{0k}(t) \neq h_{0k-1}(t) \tag{2-1}$$

En presencia de la dependencia de evento, una intervención puede tener un impacto en la tasa de eventos en los no curados a través de dos vías: un efecto primario directamente en el evento de resultado y un efecto secundario mediado por la dependencia del evento. El efecto primario combinado con el efecto secundario es el efecto total [Xu *et al.*, 2014].

En otras palabras, la dependencia de evento ocurre cuando el riesgo de un evento particular depende de eventos previamente experimentados. La dependencia de evento puede ser positiva ($h_{0k}(t) > h_{0k-1}(t)$) o negativa ($h_{0k}(t) < h_{0k-1}(t)$). Por ejemplo en lesiones de deportistas puede haber dependencia negativa, dado que el deportista al modificar su comportamiento se le reduce la probabilidad de nuevas lesiones; mientras que en pacientes con infarto de miocardio, el número de eventos previos de esta naturaleza, aumenta la probabilidad de que se presenten nuevos episodios y en tal caso decimos que hay dependencia positiva, [Xu y Cheung, 2015], en otras palabras, la dependencia de ocurrencias se puede producir por un debilitamiento biológico o por un fortalecimiento, ambos producen correlación dentro de sujetos [Barceló, 2002]. Para tratar la dependencia de evento en los modelos de supervivencia, se han utilizado modelos con riesgo basal específico que depende de la historia previa del individuo, con limitaciones para su aplicación en el ámbito epidemiológico, debido a que en los estudios de fenómenos recurrentes esta historia generalmente es desconocida. Si tratamos la observación correlacionada como no correlacionada, se exagera la cantidad de información que proporciona cada observación, lo que llevaría a estimaciones incorrectas de errores estándar [Box-Steffensmeier y De Boef, 2006] y suele obtenerse estimadores sesgados, a veces de forma muy evidente, incluso con dependencia de evento no excesivamente intensa [Navarro *et al.*, 2017].

Por ejemplo, en el análisis de ocurrencia de caídas en personas adultas institucionalizadas, [Navarro y Ancizu, 2009] encuentran que el tiempo mediano de supervivencia disminuye dependiendo del número de caidas experimentadas previamente, mientras que el riesgo de caida aumentó en función de dichas caídas previas. En este ejemplo, la dependencia de evento es evidente y los autores mostraron las consecuencias de no tenerla en cuenta al ajustar modelos de supervivencia para el fenómeno de eventos recurrentes. Veáse la siguiente figura:

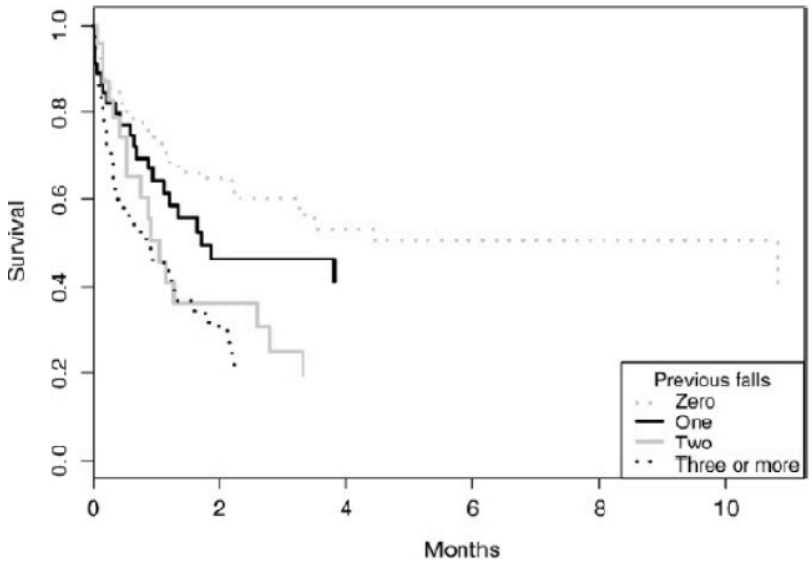


Figura 2-1.: Dependencia de evento (Tomada de [Navarro y Ancizu, 2009])

2.4. Heterogeneidad individual

La heterogeneidad individual se refiere a las diferencias que existen entre indivivuos en el riesgo de presentar una ocurrencia, debido a efectos desconocidos no medidos.En otras palabras, la heterogeneidad individual es lo que hace que dos individuos que comparten una o más covariables o factores de riesgo que se han medido, presenten diferente resultado en un desenlace de interés. Por ejemplo, algunas personas pueden ser más propensas a la enfermedad y esto conlleva a que los tiempos interocurrencias sean menores que en personas que son menos propensas, aún y compartiendo las mismas características. Esto introduce heterogeneidad entre los individuos y produce una correlación intra sujeto en la ocurrencia y el tiempo entre recurrencias dentro de un sujeto determinado. La heterogeneidad individual, conocida también con el nombre de fragilifad o propensión, en la práctica se analiza añadiendo una fragilidad al modelo, ν_i, es decir, una componente de efecto aleatorio individual para incluir esta *extra* variabilidad. Dado que ν_i es un efecto multiplicativo, este puede representar el efecto acumulado de una o más covariables omitidas [O'Quigley y Stare, 2002].

Los efectos de la heterogeneidad en un modelo pueden ser severos. Aalen [Aalen, 1988] muestra que la heterogeneidad puede hacer que un riesgo relativo poblacional pase de r a $1/r$ en el tiempo, aún cuando el verdadero riesgo relativo sigue siendo r y por esto es importante incluir el análisis de la heterogeneidad en los modelos.

En general, cuando se hacen estudios en salud, la población no puede considerarse homogénea ya que los individuos presentan diferencias en rasgos genéticos, estilos de vida y otras variables que pueden modificar la probabilidad de presentar ocurrencias y que a menudo no son medidos. Si se ignora el efecto de la heterogenidad en modelos de supervivenca, se pueden tener algunas consecuencias como: sesgos en la estimación de la tasa de riesgo, ya que en el modelo de Cox, esta estimación es general y no permite observar el riesgo a nivel individual o en subgrupos poblacionales más o menos vulnerables; subestimación del efecto de las covariables, o estimaciones de las covariables observadas y distribuciones marginales sesgadas, entre otras [Henderson y Oman, 1999].

La condición de dependencia y heterogeneidad de eventos se debe considerar en los estudios de fenómenos recurrentes por dos razones. Primero, parece probable que ambas fuentes de correlación puedan describir simultáneamente muchos de los procesos en investigación en salud. En segundo lugar, incluso si solo existe una fuente de correlación, normalmente no es posible identificar el tipo de fuente que impulsa la correlación a priori y, por esto, se dificulta la decisión del modelo a utilizar [Hougaard, 1995], [Lawless, 2011].

Teniendo en cuenta que la diferencia entre la dependencia real y la aparente, es que la primera proviene de una población homogénea y la segunda de una heterogénea, el análisis de los datos debe poder recoger la evolución de las ocurrencias, de lo contrario, será difícil distinguir entre los dos tipos de contagio. Una forma de valorar a priori la dependencia de ocurrencia es, ajustar un modelo de Cox estándar utilizando como factor la variable que podría estratificar el riesgo basal. Si los Hazard Ratio (HR) entre los valores de esta variable difieren, se puede pensar en dependencia de ocurrencia que podría ser, incluso, causada por la heterogeneidad individual. Para valorar la heterogeneidad individual se puede hacer a través de la evaluación de la varianza asociada a un modelo de fragilidad, si esta varianza no resulta significativa, es posible descartar la existencia de heterogeneidad. También puede ser de utilidad, después de ajustar los modelos, ver el análisis de las varianzas asociadas a las estimaciones de los coeficientes, donde una disminución en el error estándar estimado de forma robusta en los modelos marginales, respecto a su correspondiente error estándar *naive*, indica la existencia de mayor variación en los sujetos que entre ellos, lo cual podría mostrar una preponderancia de la heterogeneidad individual sobre la dependencia de ocurrencia. O si el error estándar robusto se incrementa, la variación intra-sujeto puede considerarse inferior que la variación entre-sujeto [Kelly y Lim, 2000], lo que destacaría la importancia de la dependencia de ocurrencia.

3. Modelos de supervivencia para el análisis de eventos recurrentes

Para la modelación de fenómenos donde se presentan eventos recurrentes, se pueden tratar mediante dos enfoques, uno acumulado y el otro instantáneo. El primero abarca modelos de conteo basados en la regresión Poisson o Binomial Negativa, el segundo a partir de análisis de supervivencia (tiempo al evento).

El enfoque basado en distribuciones de conteo como la distribución Poisson, que trata los eventos recurrentes como observaciones independientes, se puede usar en situaciones en las que la información sobre el tiempo no está disponible o el momento del evento no juega ningún papel en el tratamiento de la pregunta de investigación. En este caso se modela el número de eventos en un intervalo de tiempo o la tasa de eventos en función de algunas variables explicativas. Los parámetros del modelo se estiman por el método de máxima verosimilitud que proporciona una estimación razonablemente buena para un parámetro, siempre y cuando se asuma una tasa de eventos homogénea. La validez de estas estimaciones depende del supuesto de tasa de eventos homogénea entre los individuos, que en salud, no siempre se cumple, ya que, es muy frecuente encontrar individuos más propensos a desarrollar eventos recurrentes, lo que conlleva a que las estimaciones del modelo de Poisson ya no sean válidas.

Cuando el supuesto de tasa de eventos homogénea no se cumple, se ha utilizado el modelo de regresión binomial negativa [Jahn-Eimermacher, 2008], que supone que cada paciente tiene eventos recurrentes de acuerdo con la tasa de eventos de Poisson individual y las tasas de Poisson varían de acuerdo con la distribución *Gamma* entre pacientes, esto es, la regresión binomial negativa proporciona una mejor predicción que la regresión de Poisson cuando la suposición de riesgo uniforme en todo el sujeto no es válida [Glynn y Buring, 1996], especialmente cuando se presenta fenómeno de sobredispersión o varianza extra para la variable de conteo. El uso de la regresión Binomial Negativa puede estimar parte de la varianza que no consigue estimar la regresión Poisson, lo que hace que se mantengan algunos

aspectos interesantes en el modelamiento de variables que registren el número de veces que se produce un episodio, como la asimetría, pero es más flexible en forma de su distribución y por lo tanto más adaptable a diferentes tipos de datos [Navarro *et al.*, 2001].

Ahora si el interés está en estudiar las propiedades de la función de distribución del tiempo entre ocurrencias, el enfoque es el de análisis de supervivencia, en el cual, se tiene en cuenta la naturaleza recurrente de los eventos, es decir, que para realizar las estimaciones es necesario considerar que estas tendrán influencia del tiempo hasta que se observa el evento, así como del número de ocurrencias y, de la misma forma que en el caso de un evento único, de la censura producida generalmente por el fin del estudio. En este sentido, el enfoque estadístico del análisis se centra en el riesgo instantáneo de ocurrencia del evento.

El enfoque de análisis de supervivencia hace uso de todos los datos disponibles, en este caso los dos enfoques más populares son, los modelos de supervivencia basados en el modelo de riesgos proporcionales de Cox con corrección de varianza y los efectos de fragilidad / efectos aleatorios [Box-Steffensmeier y De Boef, 2006], [Guo *et al.*, 2008]. Los modelos con corrección de varianza tienen en cuenta la correlación mediante el uso de errores estándar robustos (sandwich) y la fragilidad puede considerarse como una covariable aleatoria en el modelo que corrige la dependencia entre los tiempos de eventos recurrentes.

En esta sección, se describirá de manera general el modelo de Cox, los supuestos y las ecuaciones como punto de partida para explicar las modificaciones de este modelo que serán utilizadas en el análisis de eventos recurrentes.

3.1. Modelo de Cox

El modelo de Cox (años 70) es un modelo semiparamétrico muy popular en la investigación biomédica, por que a pesar de que no se especifica la función de riesgo basal $h_0(t)$, se pueden obtener estimaciones razonablemente buenas de los coeficientes de regresión, las razones de riesgo de interés y las curvas de supervivencia ajustadas para una amplia variedad de situaciones [Cox, 1972].

En el modelo de Cox se especifica la función de riesgo para un individuo i como:

$$h_i(t) = h_0(t)e^{X_i\beta} \tag{3-1}$$

donde $X_i\beta$ representa el vector de covariables y los coeficientes de la regresión para el indi-

viduo i, y $h_0(t)$ representa la función de riesgo basal que no se especifica.

Los supuestos del modelo de Cox son:

- Proporcionalidad de riesgos, esto es, la razón de riesgos instantáneos entre dos individuos es constante a lo largo del tiempo.
- Independencia de los tiempos de supervivencia entre individuos distintos en la muestra,
- Relación multiplicativa entre las covariables y la función de riesgo.
- Todos los individuos están a riesgo hasta que sufren el evento o se censuran.
- Todos los individuos presentarán el evento, a lo sumo, una vez durante el tiempo de seguimiento.

En este modelo $h_0(t)$ es la componente no paramétrica porque no es necesario especificarla y $e^{X_i\beta}$ la componente paramétrica. La estimación de parámetros se hace por el método de Máxima Verosimilitud para el cual es necesario construir la función de verosimilitud y el método que se utiliza es el que aportó Cox que se basa en la verosimilitud parcial. En la función de verosimilitud parcial introducida por Cox, la única contribución de los datos a la verosimilitud es en los tiempos en que se observan eventos ya que el modelo no tiene ningún supuesto sobre $h_0(t)$ [Cox, 1975].

Representemos por $t_1, ..., t_n$ los tiempos de seguimiento de los n individuos. Mediante δ_i indicamos si los t_i están censurados por la derecha, $\delta_i = 0$, o si no, $\delta_i = 1$. En este contexto la función de verosimilitud se puede expresar como sigue:

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{exp(x_i\beta)}{\sum_{i \in R(t_i)} exp(x_i\beta)} \right\}^{\delta_i} \tag{3-2}$$

donde $R(t_i)$ representa el conjunto de individuos a riesgo en t_i.

Aunque la verosimilitud parcial no es, en general, una verosimilitud en el sentido de ser proporcional a la probabilidad en conjunto observado de datos, puede ser tratada de todas formas como una verosimilitud para propósitos de inferencia asintótica [Therneau y Grambsch, 2000].

Note que la expresión 3-2 es equivalente a 3-3- ya que $h_0(t)$ multiplica a numerador y denominador:

$$\prod_{i=1}^{n} \left\{ \frac{h_0(t)exp(x_i\beta)}{\sum_{i \in R(ti)} h_0(t)exp(x_i\beta)} \right\}^{\delta_i} = \prod_{i=1}^{n} \left\{ \frac{h_0(t)exp(x_i\beta)}{h_0(t)exp(x_1\beta) + h_0(t)exp(x_2\beta)...h_0(t)exp(x_n\beta)} \right\}^{\delta_i} \tag{3-3}$$

La expresión 3-2 es pues fácil de interpretar: sin necesidad de especificar $h_0(t)$, se puede decir que el numerador se aproxima a la función de riesgo del individuo i cuando se produce su evento, $h(t_i)$, mientras que el denominador es la suma de las funciones de riesgo de todos los individuos que están a riesgo en ese momento. Además, note que los individuos censurados tienen incidencia en el denominador mientras se conoce su estado, pero no intervienen en el numerador puesto que la expresión estará elevada a 0 cuando se produce su censura. Luego, $PL(\beta)$ no es más que el producto de probabilidades condicionadas de que, dado un evento en t, éste se haya producido en el individuo i.

Linealizando $PL(\beta)$ obtenemos:

$$ln(PL(\beta)) = \sum_{i=1}^{n} \delta_i \left\{ x_i\beta - ln \sum_{l \in R(ti)} exp(x_i\beta) \right\} \tag{3-4}$$

Diferenciando respecto a β, se obtiene el vector , $px1$, al que se le llama *score* o gradiente [Allison, 2010], $U(\beta)$:

$$U(\beta) = \frac{\partial lnPL(\beta)^*}{\partial \beta^*} \tag{3-5}$$

Y diferenciando de nuevo se obtiene la matriz de información o Hessian, $I(\beta)$, pxp, o segunda derivada de $lnPL(\beta)$ en negativo.

Las estimaciones máximo-verosímiles de los parámetros β_k se pueden obtener maximizando la función log-verosímil, mediante la solución de la ecuación $U(\beta^*) = 0$, que se obtiene utilizando el método numérico con el algoritmo de Newton-Raphson.

Estos estimadores tienen las siguientes propiedades asintóticas [Barceló, 2002]: consistencia, normalidad asintótica ($\beta_x^* \approx MVN(\beta_0 I^{-1}(\beta_k^*)$, donde $I(\beta_k^*)$ es la matriz de información observada de la verosimilitud parcial), y eficiencia semiparamétrica (entre todos los estimadores semiparamétricos de β_k este es el de mínima varianza asintótica).

Las varianzas de las estimaciones de los parámetros se pueden calcular invirtiendo la matriz de información observada:

$$V(\beta^*) = I^{-1} = -\left[\frac{\partial^2 lnPL(\beta^*)}{\partial \beta^* \partial \beta^*}\right]^{-1} \tag{3-6}$$

Teniendo en cuenta que las estimaciones se obtuvieron por máxima verosimilitud, los contrastes de hipótesis para cada coeficiente se pueden realizar usando su normalidad asintótica a través del estadístico de Wald ($\frac{\beta^*}{EE(\beta^*}$), siendo EE el error estándar del coeficiente estimado; comparando dos modelos, uno con el parámetro de interés y otro exactamente igual al primero sin el parámetro de interés, a través del estadístico de la razón de verosimilitudes ($LR = -2ln(L)$) o mediante el test equivalente al *log-rank*, el test del *score*

$$\left(\left|\partial L/\partial \beta^*/\sqrt{I(\beta^*)}\right|_{\beta=0}\right) \tag{3-7}$$

El cálculo del intervalo de confianza para la razón de riesgos, HR, puede efectuarse mediante la fórmula clásica $exp(\beta_k^* \pm 1,96EE(\beta_k^*))$.

La función de verosimilitud parcial presentada en (3-2) asume que el tiempo es una variable continua, con lo cual se da por supuesto que la ocurrencia de dos o más eventos a la vez es imposible. Usualmente, sin embargo, los tiempos de supervivencia se "discretizan" al redondear el momento exacto de una ocurrencia al día, mes o año en que ésta se produjo. Además puede suceder que en un mismo tiempo se produzca un evento en una observación y una censura en otra. En esa situación se asume que la censura ha ocurrido justo después del evento [Collett, 1994].

En todo caso, cuando existen "empates" $PL(\beta)$ debe modificarse y varias son las propuestas al respecto. Las aproximaciones de Breslow [Breslow, 1974], y la de Efron [Efron, 1977] son propuestas clásicas.

Tabla 3-1.: Aproximaciones de $PL(\beta)$ cuando hay observaciones empatadas

Aproximación	Función de verosimilitud parcial
Breslow	$\prod_{i=1}^{m} \left\{ \frac{exp(s_i\beta)}{\sum_{i \in R(ti)} exp(x_i\beta)} \right\}^{d_i}$
Efron	$\prod_{i=1}^{m} \left\{ \frac{exp(s_i\beta)}{\prod_{k=l}^{di}} \left[\sum_{i \in R(ti)} exp(x_i\beta) - \frac{k-1}{di} \sum_{i \in D(ti)} exp(x_i\beta) \right] \right\}$

donde s_i es el vector de sumas de los p parámetros de las covariables de los individuos con evento en t_i $i = 1, ..., m$ y d_i indica el número total de eventos en t_i para la ecuación de Breslow y s_i es el vector de sumas de los p parámetros de las covariables de los individuos con evento en t_i $i = 1, ..., m$ y D_{t_i} el conjunto de individuos que presentan el evento en t_i y d_i indicando el número total de eventos en t_i para la ecuación de Efron.

El modelo de riesgos proporcionales de Cox estándar para el tiempo al primer evento dejará por fuera los eventos subsiguientes y los tiempos entre ocurrencias.

3.1.1. Estratificación en el modelo de Cox

Una extensión del modelo de Cox consiste en trabajar con diversos estratos. Los estratos dividen a los individuos en grupos excluyentes con diferentes funciones de riesgo basal, aunque con iguales valores para los coeficientes del vector β. La estratificación puede ser útil ya que a veces el supuesto de proporcionalidad de riesgos se incumple en el global de la muestra estudiada, pero sin embargo se mantiene en cada uno de los grupos generados [Collett, 1994].

Hay que tener en cuenta, de todas formas, que ante un análisis estratificado, si bien se asumen riesgos basales diferentes entre los estratos, se mantiene el supuesto de la proporcionalidad de riesgos en las variables explicativas dentro de cada estrato.

Suponga que los individuos incluídos en el estrato k tienen una función de riesgo basal $h_{0k}(t)$, para $k = 1, ...K$, siendo K el número total de estratos. La función de riesgo para el individuo i del estrato k, donde $i = 1, ..., n_k$ siendo n_k el total de individuos en el estrato k, puede ser representada por:

$$h_{ik}(t) = h_{0k}(t)exp(x_{ik}\beta) \tag{3-8}$$

donde x_{ik} es el vector de valores de las p variables explicativas del individuo i del estrato k.

La forma de la función de verosimilitud parcial para el estrato k, es idéntica a la expresada en (3-2):

$$PL_k(\beta) = \prod_{i=1}^{n} \left\{ \frac{exp(x_{ik}\beta)}{\sum\limits_{l \in R(t_{ik})} exp(x_{kl}\beta)} \right\}^{\delta_i} \tag{3-9}$$

donde t_{ik} representa el i-ésimo valor de tiempo observado en el estrato k, δ_{ik} es el valor de la variable censura asociada con t_{ik}, $R(t_{ik})$ indica los individuos en el estrato k a riesgo en el tiempo t_{ik} y x_{ik} es el vector de valores para los p parámetros de las variables explicativas.

El cálculo de la función de verosimilitud parcial global, (3-9), se corresponde con el producto de verosimilitudes parciales de cada estrato, (3-8), y equivalentemente, la función log-verosímil global, (3-10), no es más que la suma de cada función log-verosímil de cada estrato:

$$PL(\beta) = \prod_{k=1}^{K} PL_k(\beta) \tag{3-10}$$

$$lnPL(\beta) = \sum_{k=1}^{K} lnPL_k(\beta) \tag{3-11}$$

En la práctica la gran ventaja de la estratificación es que proporciona una forma sencilla de ajustar el efecto de variables confusoras. Pero, ante la existencia de un gran número de estratos la precisión de los coeficientes estimados puede disminuir [Lee *et al.*, 1992] y tal vez el inconveniente principal de la estratificación radica en que no existe una estimación directa de la importancia del estrato y no se valora la asociación de la variable por la cual se estratifica con el riesgo de padecer el evento. En ese sentido parece claro que las variables candidatas a estratificar los análisis deben ser aquellas cuyo efecto no tiene otro interés para el investigador que la obtención de un buen ajuste de las otras variables explicativas. En definitiva, se trata de variables cuyo efecto sobre la ocurrencia del evento de interés normalmente es conocido (bien sea por el propio diseño del estudio o por análisis anteriores) y que se desea que no interfiera en la estimación de los otros efectos. Allison [Allison, 2010], por ejemplo, identifica cuáles variables que representan un nivel superior que agrupa conjuntos de observaciones serán óptimas para estratificar: el ejemplo evidente sería el hospital en un estudio multicéntrico.

3.1.2. Limitaciones del modelo de Cox estándar para el análisis de eventos recurrentes

Un inconveniente muy importante al analizar ocurrencias agregadas de eventos es el no tener en cuenta el tiempo y por tanto no poder seguir la historia del fenómeno durante el periodo de seguimiento. Por la misma razón, tampoco es posible conocer si las características de los individuos varían en el tiempo.

Es claro que el modelo de Cox tiene en cuenta el tiempo lo cual permite modelar, por ejemplo, los cambios en las variables explicativas. Además, para este modelo, en los datos se registra en qué momento se produce el evento y esto permite distinguir entre dos individuos que han padecido el suceso, uno al inicio del seguimiento y otro al final.

Sin embargo, la limitación principal del modelo estándar de Cox ante el análisis de la ocurrencia de un evento es que no toma en cuenta la posibilidad de que el fenómeno de interés sea recurrente, debido a que, en principio, el modelo de Cox está indicado solamente en el caso concreto en que el evento es posible una única vez en un mismo individuo, para lo cual, de hecho, fue desarrollado, (estudios donde el evento de interés era la muerte).

Habitualmente, para un fenómeno recurrente los tiempos de ocurrencia de un mismo individuo están correlacionados, fundamentalmente por dos cuestiones: la heterogeneidad entre individuos y/o la dependencia de ocurrencia.

Si la dependencia entre los tiempos de supervivencia no es capturada por la especificación del modelo, puede suceder que:

- Las estimaciones de las varianzas de los parámetros del modelo sean sesgadas, generando tests de significación "hinchados" [Wei *et al.*, 1989], [Therneau y Hamilton, 1997] ,[Lee *et al.*, 1992]

- El efecto del parámetro esté sesgado o se atenúe hacia cero [Lagakos y Schoenfeld, 1984],, [Keiding *et al.*, 1997]

- Se observe una dependencia de duración espúrea [Petersen, 1991], [Petersen, 1995].

Ante el estudio de un fenómeno recurrente, la aplicación del modelo de Cox podría intentarse de varias formas: primera, asumiendo que los eventos de un mismo individuo son independientes, que en la práctica significa que se supone que son de individuos distintos. Tal solución acarrea consecuencias evidentes: los individuos con más sucesos, así como las características concretas de éstos, están sobrerrepresentados.

Un segundo intento de aplicación del modelo de Cox consiste en analizar exclusivamente el primer evento para cada individuo: en este caso, cada individuo está representado por un único episodio y los eventos son efectivamente independientes. Pero para que este análisis

sea realmente representativo del conjunto del fenómeno, se debe asumir que no existe dependencia de ocurrencia (la probabilidad de que ocurra un evento es independiente a que el individuo ya haya padecido algún evento, o más concretamente, es independiente al número de eventos previos) y también que los efectos de las variables son iguales entre el primer evento y el resto. No es difícil pensar que dichas condiciones se cumplen en pocas ocasiones.

También existe la posibilidad de incorporar una variable explicativa que sea el número de eventos previos y valorar su asociación con las demás variables. En ese caso, es más que probable que el resumen de los resultados sea realmente complejo y no se obtenga una visión global del fenómeno. Además esta alternativa sigue sin solucionar las variables explicativas que cambian en el tiempo y, en consecuencia, el hecho de que habrá observaciones que serán del mismo individuo pero tratadas de forma absolutamente independiente.

Otra alternativa es el ajuste de un modelo para el primer evento y tantos modelos como recurrencias: con esta solución no se produce el problema de la dependencia entre observaciones, sin embargo sigue presentándose, y posiblemente incrementándose, la dificultad de interpretación. Además, los modelos para individuos con varios eventos se ajustarán sobre un tamaño muestral cada vez menor y en consecuencia las estimaciones para estos modelos serán más imprecisas y difícilmente comparables. En este contexto, en los años 90 a nivel teórico y principios del 2000 a nivel práctico con software disponible, varios autores desarrollan extensiones del modelo de Cox para eventos recurrentes que se presentan a continuación.

3.2. Componentes esenciales de los modelos de supervivencia para eventos recurrentes

Para un modelo de evento recurrente basado en Cox, se definen cuatro componentes : intervalo de riesgo, conjunto de riesgo; función de riesgo de base (común o específica) y manejo de la correlación intra sujeto [Kelly y Lim, 2000]

Intervalo a riesgo

Se define cuando un sujeto está a riesgo de padecer un evento a lo largo de una escala de tiempo determinada. Para este intervalo de riesgo existen tres formulaciones: gap time o tiempo de brecha, total time (tiempo total) y counting process (procesos de conteo).

En la formulación *gap time* (tiempo de brecha) se especifica cada uno de los intervalos respecto al inicio de los mismos, independientemente de los demás. El análisis de eventos recurrentes basado en *gap time* es usado frecuentemente cuando los eventos son relativamente infrecuentes, o cuando algún tipo de renovación individual ocurre después de un evento o cuando la predicción sobre el tiempo para el próximo evento es de interés para el investigador. Por ejemplo en salud, estudios en cáncer de vejiga donde los pacientes se someten a resección transuretral para extirpar todos los tumores recientemente eliminados y tratamiento profiláctico para retardar el desarrollo de nuevos tumores. Se incluyen además en esta formulación, estudios de las infecciones donde un paciente retorna a un estado similar después de que la infección ha sido "limpiada"y en episodios recurrentes de hospitalización [Cook y Lawless, 2007].

Los procesos de renovación son los típicos estudios que tienen la formulación *gap time* y son definidos como un proceso para el cual: $\lambda(t \mid H(t) = h(t - T_N(t-)$, donde $h(.)$ es la función de riesgo para el gap time entre eventos, que son independientes, idénticamente distribuidos.

El tiempo se considera siempre en relación al episodio anterior, por tanto el inicio de cada nuevo episodio para un mismo sujeto se fija en cero:

$$h_{ik}(t) = h_{0k}(t - t_{k-1})e^{X_i\beta} \tag{3-12}$$

La formulación *total time* (tiempo total) es el tiempo total desde un punto elegido como inicio de tratamiento.

La formulación *counting process* utiliza la misma escala de tiempo que el tiempo total, pero reconoce que un sujeto puede tener una entrada tardía o un período censurado antes de que el sujeto corra el riesgo de sufrir el evento; con eventos recurrentes, un sujeto no se considera en riesgo para el k-ésimo evento hasta después del (k-1)-ésimo evento.

Esta formulación, permite incluir en este análisis los eventos recurrentes, múltiples eventos, múltiples escalas de tiempo, intervalos a riesgo discontinuos o variables dependientes del tiempo [Therneau y Hamilton, 1997].

Un proceso contador, es un proceso estocástico, $N(t), t \geq 0$ con valores enteros no negativos, y no decreciente, esto es:

- $N(t) \geq 0$

- $N(t)$ es un entero
- Si $s \leq t$ entonces $N(s) \leq N(t)$
- Si $s < t$, entonces $N(t) - N(s)$ es el número de eventos que ocurren en el intervalo $(s, t]$.

Counting process es muy utilizado cuando los individuos experimentan el evento de interés frecuente e incidentalmente en el sentido que su ocurrencia no altera materialmente el proceso. Por ejemplo, los ataques de epilepsia o de asma moderado son eventos incidentales y los infartos de miocardio, ataques en estudios cardiovasculares o el desarrollo de nuevos sitios de enfermedad metastásico en pruebas de cáncer son ejemplos de eventos recurrentes pero que no son incidentales [Cook y Lawless, 2007].

Para ejemplificar estos tres conceptos, pensemos en un individuo seguido en el tiempo, el cual presenta dos episodios del evento de interés, el cuarto mes y el noveno mes desapareciendo en el mes 12. En la formulación *counting process* los intervalos a riesgo para el individuo 1 serían: (0,4] , (4,9] y (9,12]; intervalos que se identifican con su tiempo inicial y tiempo final. En la formulación *gap times*, estos intervalos serían (0,4], (0,5] y (0,3], intervalos que inician todos en cero con extremo superior la duración del intervalo. Y en la formulación *total time* los intervalos a riesgo son: (0,4], (0,5] (0,12]; intervalos que indican que el individuo está en riesgo para cualquier episodio desde el inicio del seguimiento e independiente del número de episodios previos.

En la figura siguiente se muestra el esquema de los tres tipos de intervalos de riesgo y representa un individuo con 5 recurrencias.

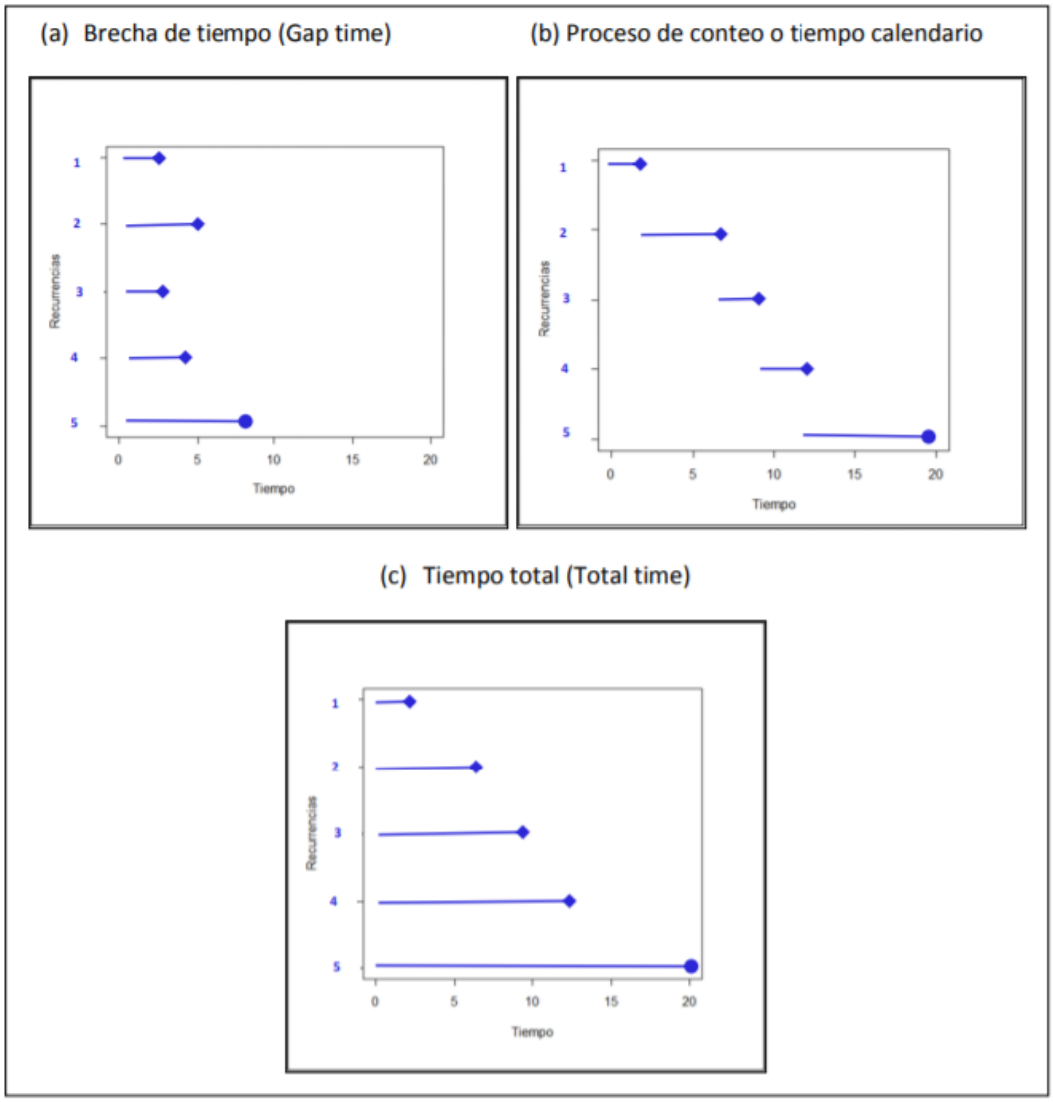


Figura 3-1.: Intervalos de riesgo

En la figura (a) se puede observar que siempre el inicio de cada nuevo episodio inicia en cero, en la figura (b) representa un proceso de conteo en el que se muestra que cada intervalo de riesgo se identifica con su tiempo inicial y su tiempo final respecto al inicio del seguimiento y en la figura (c) se muestra la estructura total time en la que el individuo está a riesgo para cualquier ocurrencia desde el inicio del seguimiento e independiente del número de eventos previos.

En la siguiente tabla se muestra la estructura de los datos para eventos recurrentes en cada uno de los intervalos de riesgo mostrados en las figuras.

Tabla 3-2.: Estructura de datos en *counting process, gap time, total time.*

Estructura de datos Counting process			
Intervalo de riesgo	**Estado**	**Estrato**	**Covariable**
(0,3]	1	1	1
(3,7]	1	2	1
(7,10]	1	3	1
(10,12]	1	4	1
(12,20]	0	5	1

Estructura de datos Gap time			
Intervalo de riesgo	**Estado**	**Estrato**	**Covariable**
(0,3]	1	1	1
(0,4]	1	2	1
(0,3]	1	3	1
(0,2]	1	4	1
(0,8]	0	5	1

Estructura de datos Total time			
Intervalo de riesgo	**Estado**	**Estrato**	**Covariable**
(0,3]	1	1	1
(0,7]	1	2	1
(0,10]	1	3	1
(0,12]	1	4	1
(0,20]	0	4	1

Hazard basal

En el análisis de supervivencia para eventos recurrentes, la función de riesgo basal (hazard basal) $h_0(t)$ puede ser común o específica. En el caso de función de riesgo basal común, se asume que todos los eventos tienen el mismo riesgo subyacente. Mientras que si la función de riesgo basal es específica, se asume que la función de riesgo basal se estratifica para cada k-ésimo evento, lo que significa ajustes separados para cada k-ésimo evento [Kelly y Lim, 2000].

Risk Set

El k-ésimo conjunto de riesgo contiene los individuos que están a riesgo para el k-ésimo evento. Hay tres posibles conjuntos de riesgo: sin restricciones; restringido o semi-restringido. La definición del conjunto de riesgo incorpora la elección de la función de riesgo basal. El conjunto de riesgos en un momento dado depende de las personas incluidas en el conjunto y cuando esas personas están en riesgo, es decir, el intervalo de riesgo [Kelly y Lim, 2000].

3.3. Extensiones del modelo de Cox

Los tiempos de eventos individuales generalmente se analizan utilizando el modelo de riesgos proporcionales de Cox, pero teniendo en cuenta la dependencia de evento y la heterogeneidad individual, este modelo no mantiene el supuesto de proporcionalidad cuando se presentan estos dos fenómenos. Las alternativas que existen para resolver este problema son extensiones del modelo de Cox que pueden ser clasificadas en modelos marginales y modelos condicionales. Los modelos marginales, también llamados modelos con varianza corregida, asumen que la función de riesgo queda totalmente especifcada por la función de riesgo basal y los valores de las covariables. Sin embargo, los modelos marginales, de entrada, estiman el modelo ignorando la posible dependencia (o la heterogeneidad) especificando su distribución de probabilidad e incorporándola al modelo [Barceló, 2002]. En otras palabras, estos modelos asumen que los eventos dentro de un mismo sujeto son independientes y por lo tanto no está condicionado a la historia de los eventos. Algunos de estos modelos intentan incorporar la dependencia de ocurrencia permitiendo que el riesgo basal varíe en función de los eventos previos de cada individuo. En estos modelos, aunque son marginales en relación con la estimación de los efectos, la función de riesgo puede ser especificada de forma condicional o marginal. Cuando se especifica de forma condicional, el riesgo se expresa condicionado a la historia previa $h(t; H(t))$ y cuando se especifica de forma marginal la función de riesgo se focaliza en el riesgo marginal $h(t)$.

Los modelos condicionales, denominados, modelos de fragilidad, tienen en cuenta la heterogeneidad individual, que se observa entre el conjunto de individuos que comparten las mismas características, permitiendo modelar el riesgo en poblaciones heterogéneas considerando efectos aleatorios tanto de asociación como de variabilidad y también permiten realizar ajustes de sobredispersión en los datos [Hougaard, 1995], [Wienke *et al.*, 2010].

A partir del intervalo de riesgo descrito en la formulación *counting process* y del conjunto a riesgo, Kelly y Lim [Kelly y Lim, 2000] clasifican los modelos de análisis de la supervivencia multivariante basados en el modelo de Cox: el de Lee, Wei y Amato (LWA) [Lee *et al.*, 1992], el de Wei, Lin y Weissfeld (WLW) [Wei *et al.*, 1989], el de Andersen y Gill (AG) [Andersen y Gill, 1982] y los de Prentice, Williams y Peterson [Prentice *et al.*, 1981], concretamente el formulado según *gap time* (PWP-GT) y según *counting process* (PWP-CP). Existen otras tres posibilidades que no han sido "formalizadas", que serían el modelo basado en *gap time* con riesgo basal común (GT-UR), el modelo con intervalo de riesgo *counting process* y conjunto a riesgo semi-restringido y el modelo *total time* con riesgo basal específico. La Tabla 3-2 muestra el conjunto de modelos según intervalo y conjunto a riesgo:

Tabla 3-3.: Modelos de supervivencia multivariante según el intervalo a riesgo y el conjunto a riesgo

Conjunto a riesgo/ Riesgo basal			
Intervalo de riesgo	No restringido/común	Semi-restringido-específico	Restringido/específico
Total time	LWA 3	WLW	Posible (TT-R)
Gap time	Posible (GT-UR)	Imposible	PWP-GT
Counting process	AG	Posible	PWP-CP

La generalización de la función de riesgo de todos estos modelos es similar a la expresada en (3-10) para el modelo estándar. De hecho existen dos posibilidades, en función de si se considera que el riesgo basal varía según el número de ocurrencias previas:

$$h_{ik}(t) = Y_i(t)h_0(t)exp(x_i\beta) \tag{3-13}$$

o

$$h_{ik}(t) = Y_{ik}(t)h_{0k}(t)exp(x_i\beta) \tag{3-14}$$

La diferencia entre las dos ecuaciones se aprecia en las funciones de riesgo basal. La primera especifica un riesgo basal común para todas las ocurrencias $h_0(t)$, mientras que la segunda considera un riesgo basal específico para cada ocurrencia $h_{0k}(t)$. En la primera ecuación estamos ante un modelo de Cox no estratificado y en la segunda el modelo estratificado. $Y_i(t)$ indica si un individuo i está a riesgo en el instante t, mientras que $Y_{ik}(t)$ señala lo mismo pero en concreto para la ocurrencia k-ésima.

En la práctica, considerar el riesgo basal específico por ocurrencia significa estratificar el modelo por la variable "estrato" utilizada en la formulación *counting process*, y que se corresponde con el número de ocurrencia (o recurrencia, según se desee) al que está expuesto el individuo.

De igual forma que para el modelo estándar, las estimaciones de los parámetros β_k se obtienen a través de la función de verosimilitud parcial. Para (3-13) y (3-14):

$$PL(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{Y_{ik} exp(x_{ik}\beta)}{\sum_{l=1}^{n} \sum_{h=1}^{K} Y_{lh}(t_{ik}) exp(x_{lh}\beta)} \right\}^{dN_i(t_{ik})} \tag{3-15}$$

$$PL(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{Y_{ik} exp(x_{ik}\beta)}{\sum_{l=1}^{n} Y_{lk}(t_{ik}) exp(x_{lk}\beta)} \right\}^{dN_i(t_{ik})} \tag{3-16}$$

y las correspondientes funciones score:

$$U(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \left[x_{ik} - \frac{\sum_{l=1}^{n} Y_{lk}(t_{ik}) exp(x_{lk}\beta) x_{lk}}{\sum_{l=1}^{n} Y_{ik}(t_{ik}) exp(x_{lk}\beta)} \right] \tag{3-17}$$

y

$$U(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \left[x_{ik} - \frac{\sum_{k=1}^{K} \sum_{l=1}^{n} Y_{lk}(t_{ik}) exp(x_{lk}\beta) x_{lk}}{\sum_{k=1}^{K} \sum_{l=1}^{n} Y_{lk}(t_{ik}) exp(x_{lk}\beta)} \right] \tag{3-18}$$

En (3-6) se presentó el cálculo de las varianzas de los coeficientes estimados de los parámetros del modelo de Cox estándar. Dichas estimaciones son inconsistentes si existen observaciones correlacionadas [Wei *et al.*, 1989], [Lin, 1994]. Al ignorar la correlación intra-sujeto se exagera la cantidad de información entre-sujetos [Carlin *et al.*, 1999]. Con el fin de corregir las

varianzas, en caso de dependencia entre observaciones, se puede obtener un estimador de la varianza "empírico" mediante un estimador "sandwich". En general, este estimador tendrá la forma siguiente:

$$V_R(\beta^*) = V(\beta^* BV(\beta^*)) = I^{-1}BI^{-1} \tag{3-19}$$

donde $I^{-1} = V(\beta^*)$ es el estimador (3-6) basado en la matriz de información y B es un factor de corrección basado en la correlación entre casos [White, 1980].

Para el modelo estándar de Cox, B puede derivarse de varias formas. Un estimador equivalente al mencionado por Reid y Crépeau [Reid y Crépeau, 1985] y al de Lin y Wei [Lin, 1994], extensión del propuesto por White, es el que describen Therneau y Grambsch [Therneau y Hamilton, 1997], estimador tipo *jackknife* que se aproxima mediante la matriz *delta-betas*, D.

$$D = LV(\beta^*) = LI^{-1} \tag{3-20}$$

donde $V(\beta^*) = I^{-1}$ es la matriz pxp de varianzas-covarianzas de los vectores de parámetros estimados en el modelo de Cox ajustado, y L es la matriz nxp de residuos *score*.

Entonces:

$$V_R(\beta^*) = I^{-1}(L'L)I^{-1} = D'D \tag{3-21}$$

siendo L la matriz $n \times p$ de vectores de residuos score de cada individuo, L_i, definidos en los residuos score. $D'D$ es un estimador asintóticamente insesgado de $V(\beta^*)$.

El estimador sandwich de la ecuación (3-20) asume que todas las observaciones, n, son independientes y que todas son usadas para el cálculo de la matriz D. Ya que cada individuo o cluster puede estar representado por más de una observación en el caso del análisis de la supervivencia multivariante, en este contexto, en realidad, no todas las observaciones son independientes, aunque sí hay q individuos o clusters que lo son:

$$n = \sum_{r=1}^{q} n_i \tag{3-22}$$

Asumiendo que las observaciones son independientes dentro de cada individuo/clúster, los *delta-betas* por individuo/clúster, $\tilde{D}$, se calculan colapsando los $\tilde{D}_{ik}$ en cada uno de los m individuos/clústers:

$\tilde{D}_{ik} = \sum_{i=1}^{n_1} D_{ik}, \tilde{D}_{2k} = \sum_{i=l+n_1}^{n_1+n_2} D_{ik}, \ldots$

para luego construir la varianza *jackknife* agrupada:

$$V_{R-A}(\beta^*) = \tilde{D}'\tilde{D} \tag{3-23}$$

$\tilde{D}'\tilde{D}$ resulta ser un estimador insesgado de la varianza en el contexto de múltiples eventos [Wei *et al.*, 1989] y Lee, Wei y Amato [Lee *et al.*, 1992] lo desarrollan para eventos correlacionados del mismo tipo.

La varianza robusta acostumbra a ser superior a la varianza *naive*, ya que normalmente la correlación intrasujeto no observada es positiva [Box-Steffensmeier y De Boef, 2006].

No todos los modelos expuestos en la Tabla 3-2 son útiles para el análisis de eventos recurrentes. En general, el uso de los modelos basados en intervalo de riesgo *total time* parecen ser, conceptualmente, inapropiados en este contexto ya que implica que cualquier ocurrencia del evento está a riesgo de producirse desde el inicio del intervalo, lo cual, no se ajusta a una de las características particulares de evento recurrente, la cual es, que las ocurrencias se producen secuencialmente. En realidad la única ocurrencia que está a riesgo en t=0, por ejemplo, es la primera, ninguna de las recurrencias del evento se pueden producir en ese momento. Este problema es más grave aún más cuando el primer evento ocurre de forma tardía: en ese caso, todas las recurrencias se consideran a riesgo durante un periodo de tiempo muy elevado cuando en realidad lo habrán estado durante un periodo más bien corto. Por otro lado, en el análisis de eventos recurrentes, no parece tener lógica que algún individuo esté a riesgo, en un momento t, de la k-ésima ocurrencia sin haber padecido aún la k-1. Esta consideración es aplicable a los modelos basados en *total time* y los modelos con conjunto a riesgo semi-restringido. Sin embargo, estos modelos parecen ajustarse perfectamente al análisis de múltiples eventos de distinto tipo. El modelo GT-UR, debido a su riesgo basal común y la forma de especificar los intervalos de riesgo también permite, en determinadas circunstancias, que un mismo individuo esté a riesgo en un momento de más de un evento, con lo cual tampoco parece de utilidad en el contexto de interés.

Siguiendo estas indicaciones, los modelos GT-UR, LWA, TT-R, el posible modelo *counting process-conjunto a riesgo semi-restringido* y el modelo WLW serían descartados de entrada como opciones a tener en cuenta para el análisis de eventos recurrentes.En este estudio, los modelos marginales que se tendrán en cuenta son AG, PWP-GT, PWP-CP.

A continuación se resumen las propiedades y características más importantes de los modelos que se utilizarán:

- **Modelo AG**:
  Conjunto a riesgo (evento s a tiempo t): Eventos independientes.

  Escala de tiempo: Desde el inicio del seguimiento.

  Función de verosimilitud parcial: $\prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\frac{exp(x_{ik}\beta)}{\sum_{l=1}^{n}\sum_{h=1}^{S}Y_{lh}(t_{ik})exp(x_{lk}\beta)}\right\}^{\delta_{ik}}$

  Función de riesgo: $h_{ik}(t) = Y_i(t)h_0(t)exp(x_i\beta)$ con $y_{ik}(t) = I(t_{ik-1} < t \leq t_{ik})$

- **Modelo PWP-CP**:
  Conjunto a riesgo (evento k a tiempo t): Todos los individuos que han experimentado el evento k-1, y no han experimentado el k, en tiempo t.

  Escala de tiempo: Desde el inicio del seguimiento.

  Función de verosimilitud parcial: $\prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\frac{exp(x_{ik}\beta)}{\sum_{l=1}^{n}Y_{lk}(t_{ik})exp(x_{ik}\beta)}\right\}^{\delta_{ik}}$

  Función de riesgo: $h_{ik}(t) = Y_{ik}(t)h_{0k}(t)exp(x_i\beta)$ con $y_{ik}(t) = I(t_{ik-1} < t \leq t_{ik})$

- **Modelo PWP-GT**:
  Conjunto a riesgo (evento k a tiempo t): Todos los individuos que han experimentado el evento k-1, y no han experimentado el k, en tiempo t.

  Escala de tiempo: Desde el final del evento previo.

  Función de verosimilitud parcial: $\prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\frac{exp(x_{ik}\beta)}{\sum_{l=1}^{n}Y_{lk}(t_{ik})exp(x_{ik}\beta)}\right\}^{\delta_{ik}}$

  Función de riesgo: $h_{ik}(t) = Y_{ik}(t)h_{0k}(t)exp(x_i\beta)$ con $y_{ik}(t) = I(t_{ik} - t_{ik-1} > t$; $t_{i0} = 0$

3.4. Modelo de Andersen Gill

Es una generalización del modelo de Cox, [Andersen y Gill, 1982] es un modelo no estratificado que se usó originalmente con intervalos de riesgo de tiempo calendario, esto es, se

divide el tiempo de seguimiento en segmentos definidos por la observación de eventos. Se utiliza cuando el orden de los eventos no se considera importante y, por lo tanto, los riesgos subyacentes son los mismos para todos los eventos.En esta aproximación AG, cada sujeto es tratado como un proceso contador con sucesos múltiples y con incrementos independientes, dada la historia de todas las variables observables hasta el tiempo de presentación de los eventos. Para este modelo cada individuo está representado por un conjunto de observaciones $s_{ij}, t_{ij}, d_{ij}, x_{ij}, k_{ij}, j = 1, ..., n_i$ siendo $(s_{ij}, t_{ij}]$ el intervalo de riesgo, abierto a la izquierda y cerrado a la derecha; $d_{ij} = 1$ si el individuo tuvo un evento en el instante t_{ij} y 0 en otro caso; x_{ij} es el vector de covariables en el intervalo, y k_{ij} es un estrato en el que puede estar el sujeto durante el intervalo. La ecuación del modelo AG está dada por:

$$h_{ik}(t) = Y_i(t)h_{0k}(t)e^{X_i\beta} \tag{3-24}$$

Con $Y_{ik}(t) = I(t_{ik-1} < t \leq t_{ik})$, donde $X_i\beta$ representa el vector de covariables y los coeficientes de la regresión, k es el k-ésimo evento del sujeto i y $h_{0k}(t)$ es la función de riesgo basal que depende de k y que es común para todas las ocurrencias. En este modelo el valor que toma $Y_i(t)$ seguirá siendo uno si ocurre el evento, dado que el individuo sigue en riesgo.

Es claro que la ecuación es igual a la del modelo de Cox estándar con una única diferencia: la definición del conjunto de individuos a riesgo. Para el modelo estándar en el momento en que un individuo presenta el evento deja de estar a riesgo de padecer otra ocurrencia, mientras que para el modelo AG el individuo puede reincorporarse al conjunto a riesgo.

La hipótesis clave de este modelo supone que las múltiples observaciones de un mismo individuo son independientes, condicionados a las variables explicativas y que un sujeto contribuye al conjunto de riesgo para un tiempo determinado todo el período en que este sujeto está bajo observación. Los datos de cada sujeto pueden describirse como de un sujeto distinto hasta cada tiempo de evento. Cada episodio se trata en forma independiente y no diferencia el primero de los siguientes, lo que se convierte en una limitación para su uso ya que puede llevar a obtener estimadores sesgados e ineficientes, con sobreestimación de la precisión de los mismos, debido a que, es más común encontrar que las observaciones de un mismo individuo estén correlacionadas.

Therneau y Hamilton [Therneau y Hamilton, 1997] afirman que el modelo AG es eficiente y da la estimación más fiable del efecto general del tratamiento, si se cumple el supuesto de independencia de evento (no hay cambio en el riesgo basal) y si el interés del análisis se centra en obtener el efecto global de la covariable.

En la siguiente figura se muestra el esquema del modelo. La flecha indica un estrato:

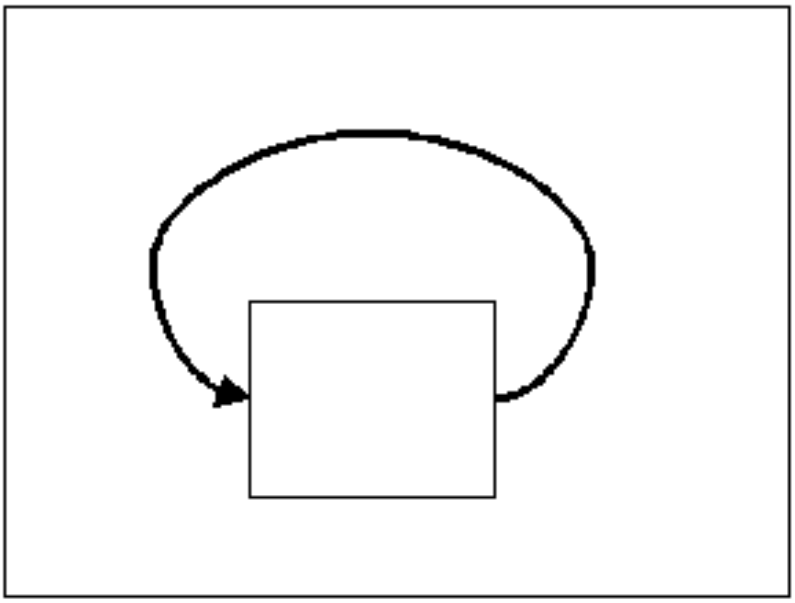


Figura 3-2.: Esquema del modelo AG

3.5. Modelos Prentice Williams Peterson - Counting Process PWP-CP y Gap Time PWP-GT

El modelo de Prentice, William y Peterson (1981), define claramente el orden de ocurrencia de los eventos. Un sujeto no se encuentra en riesgo para el k-ésimo evento si no ha experimentado el evento anterior $(k-1)$.

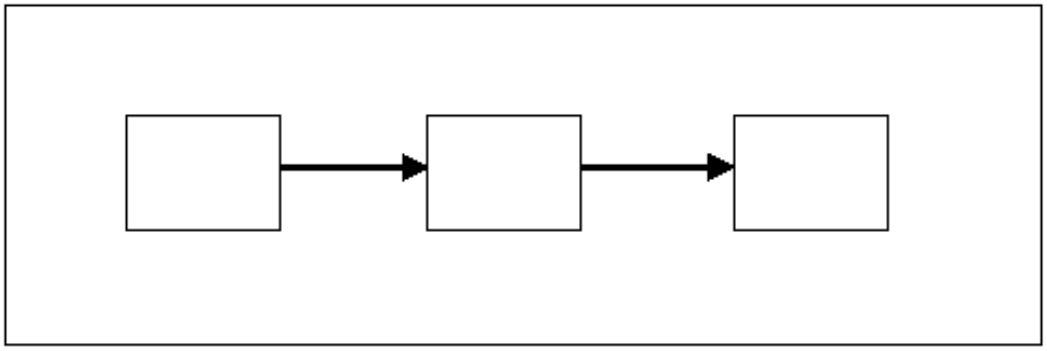


Figura 3-3.: Esquema del modelo PWP

Ellos consideran dos clases generales de modelos de regresión para eventos recurrentes los cuales relacionan el riesgo como una función de intensidad con covariables y la historia de falla. Los modelos consideran e incluyen el tiempo desde el origen del estudio hasta la ocurrencia de cada evento y el tiempo interocurrencia, respectivamente.

Este es un modelo marginal respecto a la estimación de los parámetros, pero condicional respecto a la construcción del conjunto de pacientes en riesgo.Es un modelo estratificado utilizado originalmente con intervalos de riesgo de tiempo calendario o de intervalo, el cual tiene en cuenta el orden de los eventos por lo que utiliza el número de eventos anteriores como una variable de estratificación [Prentice *et al.*, 1981]. Cada evento es asignado a un estrato diferente, esto es, el modelo PWP es un modelo de riesgos proporcionales con estratos dependientes del tiempo, donde la dependencia entre tiempos de ocurrencias se maneja estratificando por el número de ocurrencias previas, presentando cada estrato su propio riesgo basal.

En su documento de 1981, Prentice, Williams y Peterson (PWP) proponen dos modelos que pueden pueden considerarse como extensiones del modelo de riesgos proporcionales estratificado con estratos definidos por el tiempo o las periodicidades. Las formas de los dos modelos se definen de acuerdo a la especificación del intervalo de riesgo utilizada: PWP-CP, donde, como en el modelo AG, utiliza el proceso de conteo y la referencia se fija en el inicio del seguimiento (por *counting process*) o PWP-GT, donde se utiliza el tiempo desde el evento previo (gap time). Las ecuaciones que se utilizan para modelar el *función hazard* en función de las covariables y teniendo en cuenta el número de eventos previos son:

$$h_{ik}(t|N(t), X(t)) = Y_{ik}(t)h_{0k}(t)e^{x_i\beta_k} \tag{3-25}$$

$$h_{ik}(t|N(t), X(t)) = Y_{ik}h_{0k}(t - t_{n(t)})e^{X_i\beta_k} \tag{3-26}$$

donde $N(t)$ representa el proceso de conteo para el número de eventos previos, $X_i\beta_k$ representa el vector de covariables y los coeficientes de la regresión, k es el k-ésimo evento del sujeto i y $h_{0k}(t)$ es la función de riesgo basal que depende de k y $Y_{ik}(t) = I(t_{ik-1} < t \leq t_{ik})$ para counting process y $Y_{ik}(t) = I(t_{ik} - t_{ik-1} > t)$; $t_{i0} = 0$ para gap time.

El modelo puede incorporar el efecto general y el efecto específico del evento para cada

covariable. En la práctica, se puede requerir que los datos estén limitados a un número específico de eventos recurrentes si el riesgo ajustado se torna muy pequeño para estratos posteriores y la estimación específica de eventos se vuelve poco confiable.

Cuando se tiene un estrato con pocas observaciones, Therneau y Hamilton [Therneau y Hamilton, 1997] proponen tres opciones:

1. Tratar este estrato como los otros, sin olvidar que en ese estrato las estimaciones del riesgo pueden ser inestables y no permitan hacer extrapolaciones.

2. Hacer truncamiento de los datos a un número concreto de ocurrencias. Con esto los estratos con pocas observaciones posiblemente desaparezcan y con ellos también se perderá información.

3. Colapsar los estratos con menor número de observaciones. En este caso el riesgo basal se vuelve constante a partir de un número de ocurrencia determinado. Cuando se escoge esta opción, el modelo que resulta es una mezcla entre PWP y AG, condicional respecto a las primeras ocurrencias y con riesgo basal constante a partir de determinado número.

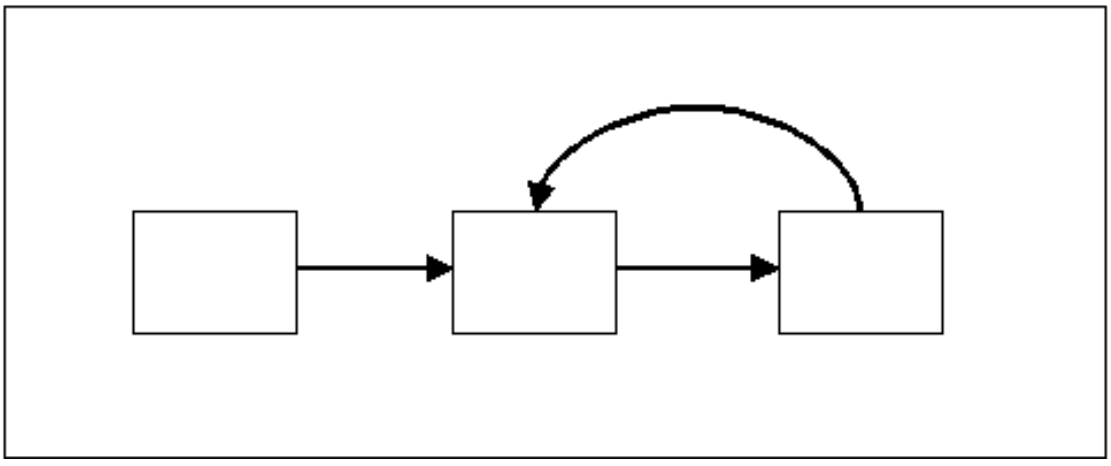


Figura 3-4.: Esquema del modelo híbrido PWP/AG

3.6. Modelos de fragilidad

Los modelos de fragilidad son modelos que introducen una o más variables aleatorias comunes a todos los individuos con el objetivo de permitir que las observaciones de eventos repetidos estén correlacionadas. Es un efecto multiplicativo en la función de riesgo o de intensidad que captura fuentes de variación diferentes pero relacionadas para modelar la dependencia entre los tiempos de recurrencia. Una de las fuentes de variación surge de las covariables consideradas, comunes a todos los individuos. La otra fuente proviene de covariables individuales no observadas, con un efecto equivalente a la dependencia serial [González y Peña, 2004].

Según Barceló [Barceló, 2002], el origen de los modelos de fragilidad puede encontrarse en dos contextos de investigación relacionados: el ámbito demográfico y el ámbito estadístico. En el ámbito demográfco, Vaupel, Manton y Stallard [Vaupel *et al.*, 1979] introdujeron el concepto de fragilidad y fueron los primeros en usar modelos de efectos aleatorios para capturar la heterogeneidad no observada. Mientras que en el ámbito estadístico, Clayton [Clayton y Cuzick, 1985] propone por primera vez un modelo que puede interpretarse como un modelo de riesgos proporcionales con un efecto aleatorio con distribución gamma.

Así pues, si en un modelo de riesgos proporcionales de Cox se omiten las covariables importantes del modelo, el supuesto de riesgos proporcionales para las demás covariables ya no es válido. No se puede esperar que la independencia entre una variable y las características no medidas se mantengan en el tiempo [Balan y Putter, 2020], lo que significa que en este caso el hazard ratio no sería una medida adecuada para el efecto causal para datos de supervivencia cuando hay heterogeneidad no observada.

De manera más general, si el modelo de riesgos proporcionales es válido para un vector de covariables $\mathbf{x} = (x_{incl}, x_{omit})$ con efectos de las covariables $\beta = (\beta_{incl}, \beta_{omit})$, entonces la ecuación del verdadero modelo es:

$$h(t|\mathbf{x}) = h_0(t)exp(x_{incl}\beta_{incl} + x_{omit}\beta_{omit}) \tag{3-27}$$

Si este modelo solo es ajustado incluyendo x_{incl}, solo estimamos β_{incl} entonces el resultado podría ser un efecto que es atenuado hacia cero y usualmente resultaría en una violación del supuesto de riesgos proporcionales y en consecuencia el modelo de riesgos proporcionales no tendría consistencia interna si la ecuación 3-27 es el modelo verdadero. En este caso las covariables x_{omit} son las que inducen heterogeneidad no observada y las diferencias entre

los individuos que son explicadas por x_{incl} son denominadas como heterogeneidad observada [Balan y Putter, 2020].

Si el efecto de x_{omit} no puede ser observado directamente, se puede definir una variable aleatoria $Z = exp(x_{omit}\beta_{omit})$ y escribir la ecuación del modelo como:

$$h(t|x) = Zh_0(t)exp(x_{inc}\beta_{inc}) \tag{3-28}$$

donde Z es definida como el término de la fragilidad que actúa multiplicativamente sobre el hazard.

Un modelo de fragilidad, también se define como, un modelo de riesgo multiplicativo que consta de tres factores: un término de fragilidad (efecto aleatorio), una función de riesgo base (paramétrica o no paramétrica) y un término que considera la influencia de algunas covariables observadas (efectos fijos). En este tipo de modelos se consideran dos aspectos importantes en el análisis de la historia de ocurrencia de los eventos: La situación de una heterogeneidad presente entre los individuos y la dependencia subyacente entre los tiempos de ocurrencia del mismo evento.

El modelamiento de efectos aleatorios en análisis de supervivencia puede utilizarse para tratar medidas repetidas en un individuo y estimar apropiadamente la dependencia dentro de sujetos. Por esta razón, algunos análisis que tratan de explicar la correlación entre tiempos de sobrevida fallan sobreestimando las varianzas de los parámetros [Rondeau *et al.*, 2012].

Los modelos de fragilidad son útiles para dos propósitos. Primero, los modelos de fragilidad univariados se pueden utilizar para explicar los efectos de la selección de sujetos más sanos a lo largo del tiempo y también para explicar la falta de ajuste, como las desviaciones del modelo de riesgos proporcionales. En este contexto, los modelos de fragilidad se pueden utilizar para ofrecer explicaciones alternativas del comportamiento del riesgo y de las razones de riesgo a lo largo del tiempo. En segundo lugar, las fragilidades también se pueden utilizar para modelar la dependencia de los tiempos de supervivencia en datos agrupados o eventos recurrentes. Aquí, el término de fragilidad se comparte entre individuos del mismo grupo o, en el caso de eventos recurrentes, entre eventos posteriores del mismo individuo [Balan y Putter, 2020].

Modelos de fragilidad Compartida

Las fragilidades son útiles para modelar correlaciones en datos multivariantes de supervivencia y, concretamente, para los eventos recurrentes, donde podemos pensar que las diversas ocurrencias en cada individuo comparten una fragilidad común y las fragilidades entre individuos son independientes. Este modelo se conoce como modelo de fragilidad compartida (shared frailty model).

Los modelos de fragilidad compartida son un caso especial de modelos de fragilidad más generales, como lo son los modelos de fragilidad correlacionadas [Petersen, 1998] o multivariantes [Vaida y Xu, 2000]. En los modelos de fragilidades correlacionadas, se utilizan dos variables aleatorias para caracterizar el efecto de la fragilidad en cada grupo, mientras que en los modelos multivariados, se asocian dos o más términos de fragilidad a cada observación en el grupo, con lo cual se elabora una estructura compleja en la asociación entre los tiempos de las observaciones [Nguti, 2003].

Note, pues, que cuando la medida de la dependencia de las ocurrencias de un evento recurrente en un individuo es de interés en un estudio, los modelos de fragilidad son especialmente adecuados [Clayton y Cuzick, 1985], [Hougaard, 1995]. Sin embargo, a los modelos de fragilidad se les realiza dos críticas:

1. La elección de la propia distribución. En la actualidad no existe suficiente teoría desarrollada que permita guiar la elección de una distribución concreta a imponer a las fragilidades [Box-Steffensmeier y De Boef, 2006].

2. La distribución de efectos aleatorios debe ser independiente de las covariables incluidas en el modelo. Este supuesto contradice la justificación que a menudo se ofrece ante la aplicación de modelos de efectos aleatorios. Según Vermunt [Vermunt, 1996]: “Si se asume que hay variables importantes que no se incluyen en el modelo, no es plausible asumir que éstas no mantienen relación alguna con los factores observados”.

Para el caso concreto del modelo de fragilidad compartida, objeto de estudio en este trabajo, el efecto aleatorio hace referencia al individuo y es constante en el tiempo. Cada valor de fragilidad se corresponde a un sólo individuo y es compartido por todos los tiempos de éste.

Así pues, un modelo de fragilidad compartida es un modelo de efectos aleatorios donde las fragilidades son comunes (o compartidas) entre grupos de individuos y se distribuyen

aleatoriamente entre grupos.[Hougaard, 1995]

Suponiendo que se tienen G grupos con n_i individuos cada uno, entonces el modelo de fragilidad compartida para el individuo j en el grupo i se expresa en términos de su función de riesgo dado el vector de covariables X_{ij} y la fragilidad del grupo al cual pertenece el individuo w_i :

$$h_{ij}(t|w_i, X_{ij}) = w_i h_0(t) e^{X_{ij}(t)\beta} \tag{3-29}$$

donde $h_0(t)$ es la función de riesgo de base y β es el vector de parámetros del modelo. Las fragilidades w_i son variables no observadas, no negativas, con función de densidad $f(w)$ y se asumen con media 1 y varianza σ^2.

Ya que el riesgo es forzosamente un valor positivo, la distribución del efecto aleatorio se escoge entre aquellas distribuciones que permitan cumplir este supuesto. En la investigación aplicada, éstas son, principalmente la distribución gamma y la Gaussiana, siendo la gamma la más utilizada por su flexibilidad [Box-Steffensmeier y De Boef, 2006],[Wei *et al.*, 1989], [Nguti, 2003] [Balakrishnan y Peng, 2006]. Cabe anotar que se han estudiado otras alternativas: positiva estable [Hougaard, 1995], Poisson compuesta [Aalen, 1988] [Aalen, 1992] [Aalen, 1994], Uniforme [Vaupel *et al.*, 1979], en dos puntos, etc. Para un resumen de las mismas se puede consultar a Hougaard [Hougaard, 1995].

La distribución gamma utilizada habitualmente, desde Clayton, [Clayton y Cuzick, 1985] tiene esperanza unitaria, $E(w) = 1$ y varianza desconocida $Var(w) = \gamma$. Así en este caso la función de densidad para w es :

$$f_w(w) = \frac{w^{1/\gamma - 1} exp(\frac{-w}{\gamma})}{\gamma^{1/\gamma}\Gamma(1/\gamma)} \tag{3-30}$$

En el contexto de eventos recurrentes, valores elevados de γ indican alto grado de heterogeneidad individual y si $\gamma = 0$ no existe heterogeneidad individual, es decir que no hay individuos con mayor vulnerabilidad que otros.

Los métodos de estimación de los modelos de fragilidad pueden ser paramétricos o semipa-

ramétricos. En el caso paramétrico se especifica la función de riesgo base y la estimación se realiza maximizando el logaritmo de la función de verosimilitud marginal, mientras que en la estimación semiparamétrica la función de riesgo base no se especifica y se utilizan otros métodos para estimar los parámetros.

Una posible limitación del modelo de fragilidad compartida, independientemente de la distribución asociada a la fragilidad, es que los tiempos de supervivencia que pertenecen a un mismo grupo solamente pueden estar correlacionados de manera positiva. [Hougaard, 1995], [Ripatti y Palmgren, 2000].

Modelos de fragilidad condicional

El modelo de fragilidad condicional combina un efecto aleatorio para incorporar heterogeneidad no observada con estratificación basada en eventos (riesgos iniciales variables) para incorporar la dependencia de eventos. El modelo está formulado en tiempo de brecha (*gap time*) para que las estimaciones de los parámetros se puedan interpretar como una estimación de riesgo para el evento k desde el evento anterior (solo para censuras a la derecha) [Box-Steffensmeier y De Boef, 2006].

El riesgo de que un evento particular k ocurra para un individuo específico i, (λ_{ik}), para el modelo de fragilidad condicional está dado por:

$$h_{ik}(t) = h_{0k}(t - t_{k-1})e^{X_i k\beta + w_i} \tag{3-31}$$

donde k denota el número de eventos, λ_{0k} es la función de riesgo basal y varía por número de eventos; $(t - t_{k-1})$ indica la estructura del intervalo de riesgo; X es un vector de variables independientes, que pueden variar en el tiempo; y β los coeficientes del modelo. La porción restante del riesgo incorpora el efecto aleatorio. Aquí, cada sujeto i tiene un efecto aleatorio que se comparte, es decir, constante, a lo largo del tiempo (a través de eventos) y w es un vector que contiene los efectos aleatorios desconocidos o las fragilidades.

El modelo de fragilidad condicional permite la posibilidad de que tanto la heterogeneidad como la dependencia de eventos hagan contribuciones importantes a la tasa de riesgo o al riesgo de que un individuo presente un evento en particular.

La elección de la alternativa de análisis adecuada para tratar datos de eventos recurrentes,

siempre depende de los tipos de datos que tenemos y de la pregunta de investigación de interés. Si la información sobre el tiempo del evento no se ha recopilado o no agrega nada para abordar la pregunta de investigación, se debe optar por un enfoque de no supervivencia. Entre la regresión de Poisson y la Binomial negativa, la última suele ser preferible porque permite considerar situaciones de sobredispersión. Si la información sobre el tiempo del evento juega un papel al abordar una pregunta de investigación, la opción obvia son los modelos de supervivencia. Bajo estas consideraciones, podemos elegir el modelo AG si estamos seguros de que no hay correlación entre eventos recurrentes dentro del sujeto. De lo contrario, se debería pensar en un modelo PWP-GT y el modelo de fragilidad condicional sobre AG, PWP-CP y el modelo de fragilidad estándar porque estos dos modelos abordan el proceso de eventos recurrentes de forma natural asumiendo que los eventos recurrentes no son independientes y el riesgo de un evento posterior no es el mismo que el del evento anterior.

4. Datos censurados a la izquierda: cuando se desconoce la historia previa

La censura a izquierda en el análisis de supervivencia surge cuando el evento de interés ya ha ocurrido para algunos individuos antes de que comience la observación. Para estas personas se conoce el hecho de que ya han presentado el evento, pero se desconoce su tiempo exacto de falla, además el evento lo pueden haber presentado, previamente, más de una vez.

Esto puede ser debido a dos situaciones: la primera, que el sujeto tenga el desenlace antes de iniciar el estudio, o la segunda forma sería que, aún estando bajo observación, no sea posible determinar el verdadero tiempo al evento [Cain *et al.*, 2011]. Lo anterior indica que, el tiempo real de supervivencia resulta ser menor o igual al tiempo de supervivencia observado.

Esta forma de censura no es tan común encontrarla descrita en los estudios de supervivencia, encontrando que muchas veces se prefiere omitir en los análisis [Gomez *et al.*, 1992]. De igual forma, existe poca información sobre la mejor manera de analizarla.

Este tipo de censura, al analizar eventos recurrentes, es una situación que puede ocurrir con cierta frecuencia en los estudios epidemiológicos, especialmente observacionales. Más concretamente, ocurrirá en estudios de cohorte que siga a individuos que ya estaban en riesgo de experimentar el resultado de interés antes del comienzo del seguimiento.

Esto puede llevar al desconocimiento de la historia previa de estos individuos, más específicamente del tiempo en riesgo, y si han experimentado el evento (si corresponde, y cuántas veces) al inicio de la cohorte. No conocer el tiempo que una persona ha estado en riesgo de sufrir el evento de interés será un problema, cuando el riesgo inicial de tener el evento depende del tiempo.

Con respecto al desconocimiento de si el evento ya ocurrió, podemos identificar dos situacio-

nes: primero, si el evento solo puede ocurrir una vez y ya se ha producido, el resultado para este individuo se determina independientemente del tiempo de seguimiento. Aquí estamos en presencia de censura a izquierda; en segundo lugar, si el resultado de interés se puede observar más de una vez en un mismo individuo, es decir, es un evento recurrente, se pueden observar uno o varios episodios nuevos del evento pero se desconoce el número de episodios previos. En este caso estaremos en una situación de censura por la izquierda donde la variable censurada es del tipo discreto, que también puede definir diferentes funciones de riesgo de base.

En el análisis de supervivencia para eventos recurrentes, la información previa puede ser determinante de las ocurrencias nuevas a partir del inicio del estudio, particularmente cuando estamos en presencia de dependencia de evento y heterogeneidad individual. Es necesario establecer estrategias para resolver el problema de censura a izquierda, porque contar con esta información mejorará las estimaciones en el análisis de eventos recurrentes.

En la mayoría de las ocasiones se ignoran los datos censurados por la izquierda y el experimento se analiza con los datos que el investigador recopila desde el comienzo del estudio. Si este es el caso, la omisión de estos datos podría dar lugar a graves sesgos en los estimadores [Gomez *et al.*, 1992]. En la investigación en salud, desconocer la historia previa de eventos es frecuente y lo que comúnmente se hace en los análisis es ignorar esta información, produciendo, probablemente estimadores sesgados de los parámetros de los modelos de supervivencia ajustados.

Existen varias formas de hacer frente a la censura izquierda, entre algunas alternativas se tienen las siguientes:

1. Analizar solo la información disponible.
   Una opción comúnmente empleada es simplemente omitir dichas observaciones y solo analizar la información completa. Esto puede tener consecuencias negativas en varios aspectos. Inicialmente, si se tienen muchos datos censurados, el disminuir el tamaño de muestra causaría disminución en el poder de las estimaciones en los análisis de supervivencia, lo que afectaría la precisión de los resultados. También, se tendría que tener en cuenta que las censuras sean aleatorias, ya que, de lo contrario, se tendrían resultados sesgados, con una sobre o subestimación de la probabilidad de supervivencia.

2. Imputar la información faltante. Puede que los distintos métodos de imputación sean adecuados cuando se tienen datos incompletos, pero no siempre es el caso si dichos datos incompletos de deben a censuras [Gomez *et al.*, 1992]. Existe en la literatura

la descripción de algunas formas de imputación claras cuando se trata de censuras a derecha o de intervalo, pero no ocurre lo mismo cuando se trata de censuras a la izquierda [Leung *et al.*, 1997].

3. Aproximación por verosimilitud. Con este método, se pretende estimar la máxima verosimilitud mediante supuestos del mecanismo de censura. Al igual que los anteriores, no se encuentra claramente descrito cuando se trata de censura a la izquierda y su uso se encuentra limitado a la información disponible y al tipo de supuestos realizados [Islam, 2016]

Adicional a estas formas de encargarse de la censura izquierda, también se han propuesto otras alternativas como el uso del estimador de Kaplan-Meir. Este método se puede utilizar para datos censurados a la izquierda de dos formas. La primera forma consiste en convertir los datos censurados por la izquierda a censurados por la derecha, calcular las probabilidades de supervivencia utilizando el método KM y luego voltearlo a la escala original [Tekindal *et al.*, 2017]

Otra estrategia de manejo con Kaplan Meier es tratar directamente con los datos censurados por la izquierda y se ha denominado estimador de Kaplan-Meier inverso (RKM) [Gillespie *et al.*, 2010] o, de manera equivalente, método de Turnbull, [Turnbull, 1974] que generaliza el estimador de Kaplan Meier para incluir censura tanto por la izquierda como por la derecha.

Es probable que muchos investigadores que utilizan análisis de supervivencia no tengan información sobre el mecanismo de censura o no tienen claro exactamente cuáles son las suposiciones sobre las censuras. En la medida que tales supuestos de censura se entiendan claramente, se podrá identificar aquellas situaciones en las que estos supuestos pueden ser ignorados o no ignorados.

5. Justificación

En los últimos años se ha evidenciado un creciente interés y una necesidad de realizar análisis de supervivencia en situaciones que presentan recurrencia de evento. En muchos estudios clínicos, epidemiológicos y de salud pública, el evento de interés se presenta a menudo más de una vez y los tiempos entre ocurrencias pueden estar correlacionados, por ejemplo en enfermedades infecciosas, respiratorias, neurológicas o psiquiátricas, entre otras. Para este tipo de datos, se requiere contar con métodos estadísticos específicos que permitan dar respuesta a preguntas que surgen en investigaciones epidemiológicas relacionadas con la probabilidad de nuevos eventos y las covariables relacionadas.

Según el análisis de supervivencia cuando hay recurrencia de eventos en la investigación clínica y epidemiológica, la experiencia ha mostrado que con mucha frecuencia se presenta dependencia de evento y heterogeneidad individual que conllevan a pensar en métodos de análisis que van más allá del modelo de Cox estándar [Jung *et al.*, 2018]. Ignorar estos dos aspectos conlleva a incurrir en sesgos en las estimaciones de la tasa de riesgo y el efecto de las covariables [Lancaster, 1990], [Navarro *et al.*, 2017].

Para atender esta necesidad, la recomendación más fuerte es el uso de modelos marginales y de fragilidad como primera alternativa para el análisis de eventos recurrentes, que pueden presentar gran flexibilidad en la formación de estratos y conjuntos a riesgo, manipulación de escalas de tiempo y además presentan una estimación de la varianza bien desarrollada [Kelly y Lim, 2000]. Estos modelos tienen como base el modelo de riesgos proporcionales de Cox y aunque en la literatura el uso de estos modelos se encuentra con mayor frecuencia en estudios de simulación o con datos muy controlados, se tiene la posibilidad de estudiar las propiedades y la disponibilidad de software para su aplicación con datos reales.

Otro aspecto a considerar en el análisis de eventos recurrentes es la censura a izquierda que generalmente es omitida para los análisis clásicos. Actualmente, no se encuentran recomendaciones claras en la literatura sobre metodologías para analizar los datos censurados por la izquierda. Particularmente la historia de eventos previos de un paciente puede impactar la

probabilidad de ocurrencia de un nuevo evento, y por lo tanto debe ser tenida en cuenta.

En el análisis estadístico de datos cuando se tiene información faltante, se recomienda revisar la forma "adecuada"para el tratamiento de estos datos. En análisis de supervivencia cuando hay censura a izquierda una de las recomendaciones para su tratamiento, evitando ignorarla, es el uso de métodos de imputación para conseguir un acercamiento a esta información e incluirla en los análisis de supervivencia.

Por lo expuesto en los párrafos anteriores, se considera necesario profundizar en métodos de imputación, puntualizando en esta estrategia para el manejo de censura a izquierda y en modelos de supervivencia para eventos recurrentes cuando hay dependencia de evento y heterogeneidad individual, que como se ha evidenciado es más frecuente de lo esperado en el ámbito de las Ciencias de la salud.

La propuesta de este estudio surge por la necesidad de generar alternativas de modelos más precisos que tengan en cuenta la información individual de los pacientes como los antecedentes de los eventos presentados en tiempos anteriores al inicio de un seguimiento y las características fisiológicas o clínicas que pueden producir vulnerabilidades distintas en cada paciente para la presentación de recurrencias, teniendo en cuenta estrategias adecuadas de imputación para la información de los eventos presentados en tiempos anteriores cuando esta es desconocida.

6. Pregunta de investigación y objetivos

Este trabajo se originó con el propósito de responder a las siguiente pregunta:

¿Cómo se debe modelar el riesgo de presentar un episodio de un evento recurrente cuando hay dependencia de evento, si se desconocen los episodios previos padecidos antes del inicio del seguimiento en algunos o todos los sujetos de la muestra?

6.1. Objetivo general

Proponer una alternativa de análisis de supervivencia para eventos recurrentes con dependencia de evento cuando se desconoce el número de episodios previos al inicio del seguimiento

6.2. Objetivos específicos

1. Identificar el método más eficaz que permita imputar los episodios del evento de interés previo en las observaciones en las cuales esta información es desconocida a través del uso de distribuciones generalizadas.

2. Proponer un método para el análisis de eventos recurrentes cuando hay dependencia de evento y el desconocimiento de los episodios previos en algunos o todos los sujetos del estudio y comparar su rendimiento con los modelos clásicos que no tienen en cuenta los episodios previos.

3. Diseñar un algoritmo en un software estándar que permita modelar la supervivencia en eventos recurrentes usando imputación para estimar las ocurrencias previas en las observaciones que no tengan esta información.

7. Propuesta de método para el análisis de un evento recurrente cuando se desconocen eventos previos

Nuestra propuesta parte del supuesto que si bien se desconoce la historia previa de todos o algunos sujetos incluidos en una cohorte cuando el interés se centra en un evento recurrente con dependencia de evento, sí se conoce cuáles de ellos estaban a riesgo previamente al inicio del seguimiento y desde cuándo. Así que para tratar la información previa faltante y luego ajustar un modelo de supervivencia, se tuvieron en cuenta las siguientes consideraciones para la solución al problema planteado:

1. Imputar el número de episodios previos, para aquellos sujetos a riesgo antes del inicio del seguimiento. Para establecer un método .ªdecuado"de imputación se evalúa el comportamiento en distribución de la variable de conteo *Número de eventos previos* y el comportamiento de la varianza de esta variable. Una vez identificado el mejor método de imputación, se aplicará en el caso concreto con el fin de imputar k en aquellos sujetos que previo al inicio del seguimiento ya estaban a riesgo de padecer el evento.Por tanto, la función de riesgo instantáneo será:

$$h_{ik}(t) = h_{0k}(t)exp(X_i\beta) \tag{7-1}$$

$$\text{donde } k = \begin{cases} k_{obs} & \text{si no hay datos perdidos} \\ k_{imp} & \text{si hay datos perdidos} \end{cases}$$

2. Tratar por separado la subpoblación de sujetos "Previamente a riesgo" de la " No previamente a riesgo".

$$h_{ikr}(t) = h_{0kr}(t)exp(X_i\beta) \tag{7-2}$$

donde k es como en la ecuación (7-1) y r es la subpoblación a la que pertenece, ya sea "Previamente a riesgo", o "No previamente a riesgo".

Las razones de este análisis por separado se explican porque 1) separamos las escalas temporales de los nuevos, donde la escala temporal usada coincide con la de "a riesgo" y de los que ya habían estado a riesgo, cuya escala temporal de seguimiento no coincide con la de *a riesgo* y 2) porque aseguramos que en los nuevos el riesgo basal según los episodios previos sea el que corresponde (el realmente observado), mientras que en los que ya estaban a riesgo, al ser imputado, podemos cometer ciertos errores que si mezcláramos con los nuevos impediría una estimación del riesgo basal *verdadero* (posiblemente sesgada).

3. Usar un término de fragilidad para captar el error que se cometerá al imputar el número de eventos previos, y para incorporar los efectos de otras variables omitidas en el análisis que tengan relevancia en la especificación de la función de riesgo basal, si es que las hubiera.

$$h_{ikr}(t) = \nu_i h_{0kr}(t)exp(X_i\beta) \tag{7-3}$$

donde k y r son como las descritas en las ecuaciones (7-1) y (7-2) y ν_i es el término de fragilidad o efecto aleatorio para cada sujeto i

Finalmente, dado que no se encuentra evidencia de que tipo de formulación funcionará mejor en cada caso, se valorará el modelo propuesto en los intervalos de riesgo *counting process* (CP), ecuación 7-3 y *gap time* (GT) de forma empírica, ecuación 7-4.

$$h_{ikr}(t) = \nu_i h_{0kr}(t - t_{k-1})exp(X_i\beta) \tag{7-4}$$

donde k será el número de episodios previos del sujeto i en caso que se conozca o su valor imputado en caso que no se conozca; r indica a qué subpoblación pertenece el sujeto: "Previamente a riesgo" o "No previamente a riesgo", ν_i es el término de fragilidad para cada sugeto, X_i el vector de covariables y β coeficientes del modelo.

8. Imputación de los episodios previos

8.1. Missing data e imputación de datos faltantes

En la investigación en salud es frecuente encontrar datos faltantes que representan falta de información en el contenido de una o varias variables en un conjunto de datos y que pueden deberse a factores como la no respuesta en una encuesta, la falta de alguna medición, la pérdida en el proceso de recolección, etc.

Una de las estrategias para encargarse de los datos faltantes es la Imputación. En la investigación en salud, es frecuente encontrar métodos de imputación, especialmente enfocada para variables continuas [Donders *et al.*, 2006], [Cañizares *et al.*, 2004]. En una revisión sistemática de literatura publicada en el año 2012, Karaholis et al realizaron una búsqueda de publicaciones de investigaciones en salud donde reportaran y trataran los datos perdidos en estudios de cohorte con mediciones repetidas [Karahalios *et al.*, 2012] y encontraron, que solo el 43 % de las publicaciones reportaron el número de datos perdidos, 83 % describieron como trataron los datos faltantes en sus análisis, entre los cuales, el 55 % refirieron el análisis solo con los datos completos y unos pocos reportaron uso de métodos de imputación múltiple u otros métodos. A partir del año 2016, se encuentra un crecimiento en el número de publicaciones sobre métodos de imputación de valores faltantes en investigación, encontrando que un 42.1 % corresponden a publicaciones en Medicina entre el año 2016 y 2020.

Los métodos de imputación se pueden clasificar en simples y múltiples según la forma de seleccionar el valor que reemplazará el valor ausente.

- **Imputación simple**:

  La idea que subyace en la imputación simple es la de sustituir un valor ausente por otro valor obtenido mediante una técnica de imputación. Estos valores imputados son

tratados posteriormente como si realmente hubiesen sido observados.

Para variables dicotómicas existen varios métodos:entre otros, generar una nueva categoría que agrupe los valores ausentes; asignar el valor del vecino más cercano; o el método HotDeck, que consiste en extraer al azar, del grupo de sujetos con las mismas características que el que presenta el valor ausente, uno de los valores observados (donador). Para variables continuas se ha utilizado la sustitución de datos faltantes por el promedio de los datos disponibles, a pesar de ser una práctica inapropiada, debido a que su aplicación afecta la distribución de probabilidad de la variable imputada, atenúa la correlación con el resto de las vaiables y subestima la varianza, entre otras consecuencias. Para resolver estos problemas, se ha utilizado una variante de este método, que consiste en hacer imputación por medias condicionadas para datos agrupados, esto implica, formar categorías a partir de covariables correlacionadas con la variable de interés e imputar los datos perdidos con observaciones de la submuestra que comparte características comunes [Acock, 2012].Como en el procedimiento anterior se supone que el mecanismo de pérdida es completamente aleatoria y existirán tantos promedios como grupos se formen, lo que puede contribuir a atenuar los sesgos pero no los elimina.

Existen otros métodos para imputación simple, que pueden ser útiles como la imputación mediante una distribución no condicionada *hot-deck*, el cual sustituye los registros vacíos (receptores) con información de registros completos (donantes), a partir de una selección aleatoria de valores observados, lo que no introduce sesgos en la varianza del estimador.

A pesar de que los métodos de imputación simple tienen muchas limitaciones, no es posible definir reglas para decidir cuándo es factible favorecer la aplicación de un método simple, por lo que se recomienda actuar con prudencia y asumir en cada caso las mejores decisiones.

- **Imputación múltiple**

La imputación múltiple es un método más complejo que el método simple. En este caso, cada valor ausente se reemplaza por diversos valores obtenidos de un modelo de imputación y posteriormente son combinados para obtener las estimaciones definitivas.

Su objetivo primario es mantener la variabilidad de la población preservando las rela-

ciones entre variables. Para aplicar este método se asumen los siguientes supuestos: (i) El patrón de pérdida de datos es aleatorio, (ii) Se requiere que el modelo estadístico utilizado para generar los datos imputados sea apropiado, es decir que exista correlación alta entre la variable a imputar y el vector de covariables que se utiliza para modelar los datos que se utilizarán como sustitutos, y (iii) Se requiere que el modelo de análisis guarde relación con el que se utilizó para hacer la imputación [Rubin, 1988].

A continuación se presenta la nota metodológica publicada en la revista Gaceta Sanitaria, que si bien ejemplifica el uso de métodos de imputación en variables dicotómicas, sirve aquí para introducir los conceptos principales de la imputación, como son el mecanismo que genera la pérdida y las diferencias entre la aproximación simple y múltiple de los métodos de imputación.

Gac Sanit. 2017;31(4):342–345

Nota metodológica

Imputación de valores ausentes en salud pública: conceptos generales y aplicación en variables dicotómicas

Gilma Hernández[a,b], David Moriña[c,d] y Albert Navarro[d,*]

[a] Instituto de Investigaciones Médicas, Universidad de Antioquia, Medellín, Colombia
[b] Programa de Doctorado en Metodología de la Investigación Biomédica y Salud Pública, Departament de Pediatria, d'Obstetricia i Ginecologia i de Medicina Preventiva, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès, Barcelona), España
[c] Unitat d'Infeccions i Càncer (UNIC), Programa d'Investigació en Epidemiologia del Càncer (PREC), Institut Català d'Oncologia (ICO)-IDIBELL, L'Hospitalet de Llobregat (Barcelona), España
[d] GRAAL-Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès, Barcelona), España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 23 de noviembre de 2016
Aceptado el 9 de enero de 2017
On-line el 15 de marzo de 2017

Palabras clave:
Valores ausentes
Imputación
Salud pública
Epidemiología

RESUMEN

Que haya valores ausentes en variables registradas en encuestas de salud es habitual, pero no lo es imputarlos posteriormente cuando se realiza el análisis. Trabajar con datos imputados puede tener ventajas en términos de precisión de los estimadores y de identificación sin sesgos de la asociación entre variables. Probablemente, el proceso de imputación sigue siendo desconocido para muchos profesionales no estadísticos, que le atribuyen una alta complejidad y quizás un objetivo que no es exactamente el que persigue. Para aclarar estas cuestiones, esta nota pretende ofrecer una visión amena, no exhaustiva, del proceso de imputación, que permita conocer sus bondades para el trabajo de un salubrista. Todo ello en el marco de variables dicotómicas, habituales en salud pública. Para ilustrar los conceptos se usa un ejemplo en el cual se trabaja con datos con valores ausentes, imputados de forma simple y múltiple.

Imputing missing data in public health: general concepts and application to dichotomous variables

Keywords:
Missing data
Imputation
Public health
Epidemiology

ABSTRACT

The presence of missing data in collected variables is common in health surveys, but the subsequent imputation thereof at the time of analysis is not. Working with imputed data may have certain benefits regarding the precision of the estimators and the unbiased identification of associations between variables. The imputation process is probably still little understood by many non-statisticians, who view this process as highly complex and with an uncertain goal. To clarify these questions, this note aims to provide a straightforward, non-exhaustive overview of the imputation process to enable public health researchers ascertain its strengths. All this in the context of dichotomous variables which are commonplace in public health. To illustrate these concepts, an example in which missing data is handled by means of simple and multiple imputation is introduced.

Introducción

Que haya valores ausentes es frecuente en salud pública. Ignorarlos conlleva la pérdida de potencia del estudio y la obtención de estimadores ineficientes y posiblemente sesgados. Los valores ausentes representan falta de información en el contenido de una o varias variables en un conjunto de datos, y pueden deberse a factores como la no respuesta en una encuesta, la falta de alguna medición, la pérdida en el proceso de recolección, etc. Algunos ejemplos en el ámbito de la salud pública son la imputación del instante de seroconversión al virus de la inmunodeficiencia humana[1] o el estado físico y mental en las personas mayores[2].

El abordaje más frecuente consiste en ignorar los valores ausentes y usar la variable sin mayor consideración. Al hacerlo conjuntamente con otra variable sin valores ausentes, el análisis tiene en cuenta solo aquellos casos completos (*listwise deletion* [LW]), descartando información disponible. Con esta estrategia, si el análisis es multivariado, incluso en situaciones en que el porcentaje de valores ausentes sea bajo en cada variable, puede suponer que el número de casos analizados sea sensiblemente inferior al tamaño muestral con el que se creía trabajar[3]. Ello implica estimaciones ineficientes y, a veces, sesgadas[4–7].

La alternativa consiste en imputar los valores ausentes, consiguiendo que no se descarten casos. Si bien se dispone de programas estándar, como SAS, R, Stata o SPSS, que cuentan con algoritmos de imputación, diríamos que su uso no es habitual.

Existe literatura sobre imputación en el ámbito de la salud, pero la mayoría se ocupa de la imputación de variables continuas[7,8] y

* Autor para correspondencia.
Correo electrónico: albert.navarro@uab.cat (A. Navarro).

http://dx.doi.org/10.1016/j.gaceta.2017.01.001

no dicotómicas, muy habituales en salud pública. El propósito de esta nota es ofrecer a profesionales no estadísticos una descripción general de la imputación de valores ausentes, enfatizando en variables de naturaleza dicotómica.

Mecanismos de pérdida

Existen tres mecanismos:

- *Missing Completely At Random (MCAR):* la probabilidad de observar un valor ausente en una variable no depende de las otras variables ni de ella misma. Los sujetos con y sin valores ausentes tienen las mismas características.
- *Missing At Random (MAR):* la probabilidad de observar un valor ausente depende de otras variables, no de los valores de la propia variable.
- *Missing Not At Random (MNAR):* la probabilidad de observar un valor ausente depende de los valores de la propia variable, una vez controladas el resto de las variables. En esta situación no pueden imputarse los valores ausentes.

Es importante identificar el patrón en que aparecen los datos ausentes, ya que esto puede determinar la viabilidad de imputar y, en caso afirmativo, el método más eficiente[3,5,7].

Imputación simple

Consiste en asignar un valor al valor ausente, que posteriormente es analizado exactamente igual que los realmente observados. Para variables dicotómicas existen varios métodos: entre otros, generar una nueva categoría que agrupe los valores ausentes; asignar el valor del vecino más cercano; o el método Hot-Deck, que consiste en extraer al azar, del grupo de sujetos con las mismas características que el que presenta el valor ausente, uno de los valores observados (donador). El lector interesado puede profundizar en imputación simple consultando varios trabajos[4,5].

Imputación múltiple

Su objetivo primario es mantener la variabilidad de la población preservando las relaciones entre variables. Tiene tres fases (fig. 1):

1. *Imputation step:* se crean *m >1* conjuntos de datos completos donde en cada uno se mantienen fijos los valores observados ($x1_i$), imputando los valores ausentes $x1_{i\_imp\_k}$. El valor imputado para una misma observación en cada conjunto no tiene por qué ser el mismo, lo cual incorpora variabilidad a estos valores (de los cuales nunca conoceremos el valor real). La obtención de valores plausibles se consigue mediante un modelo de imputación, que debería contener las variables que se analizarán posteriormente, incluida la respuesta, más aquellas que ayuden a explicar los valores ausentes.
2. *Completed-data analysis step:* cada conjunto de datos es analizado individualmente mediante procedimientos estándar, obteniendo estimadores particulares en cada conjunto $\left(\hat{\beta}_{X1\_k}\right)$ y $\left(\hat{\sigma}_{X1\_k}\right)$. Los estimadores diferirán en cada conjunto a causa de la variación introducida en la imputación de los valores ausentes.
3. *Pooling step:* combinando las estimaciones de los diversos conjuntos de datos mediante reglas simples[6] se obtienen los estimadores definitivos ($\hat{\beta}_{X1\_imp}$), así como los errores ($\hat{\sigma}_{X1\_imp}$) que incorporan la incertidumbre de los valores ausentes.

Para profundizar en la imputación múltiple pueden consultarse Rubin[6] y Van der Palm et al.[2].

Ejemplo

Tenemos una población con tres variables dicotómicas: la dependiente, Y∼Bin(N, π=0,207); la variable con valores ausentes, X_1∼Bin(N, π=0,399); y una sin valores ausentes, X_2∼Bin(N,

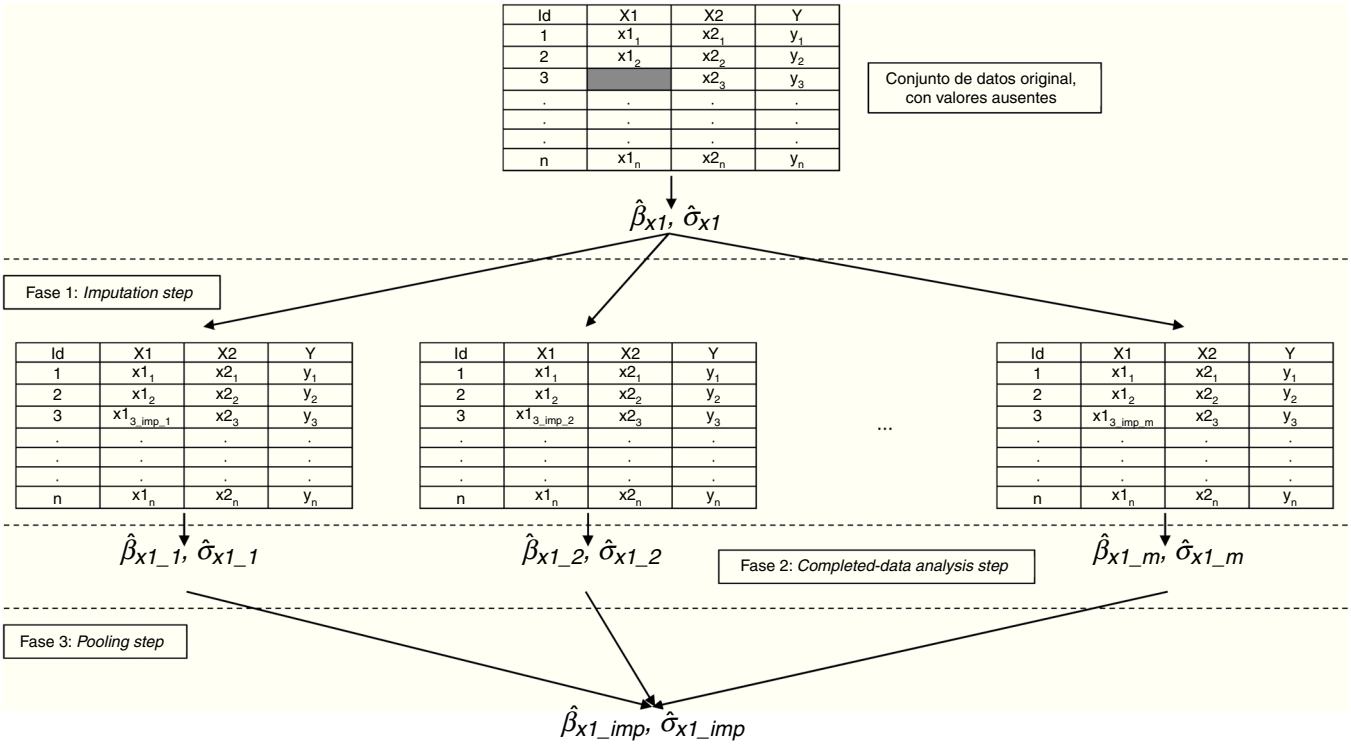


Figura 1. Esquema del proceso de imputación múltiple para una variable X1, con dos covariables sin valores ausentes (X2 e Y).

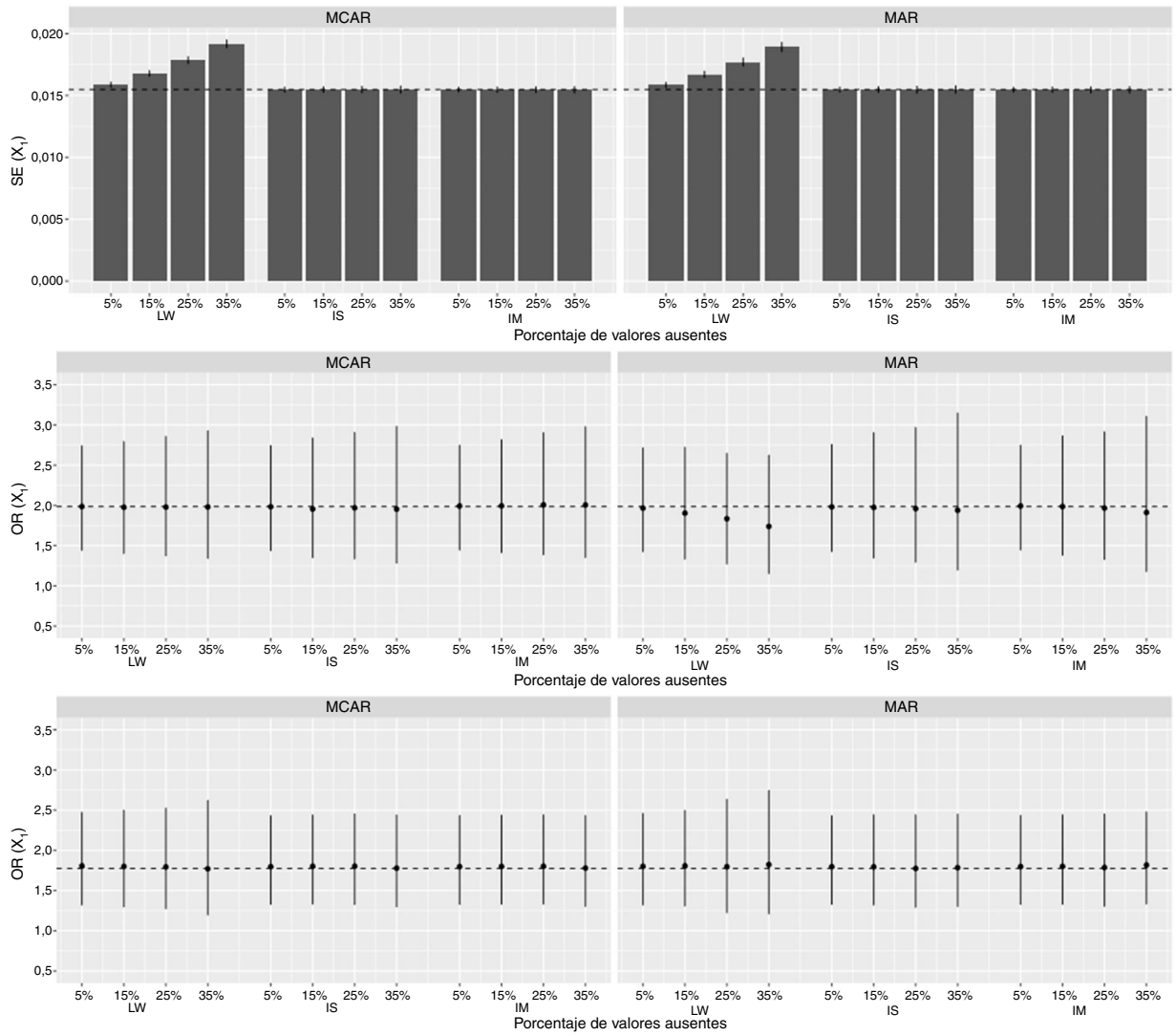


Figura 2. Resultados de las simulaciones: error estándar de X_1 (SE(X_1)), OR(X_1) y OR(X_2). La línea discontinua indica el valor poblacional.

π=0,442). Seleccionamos muestras de tamaño n = 1000 con diferentes porcentajes de pérdidas según MCAR y MAR (véase el Apéndice disponible *online* como Material suplementario). Se estiman los coeficientes de una regresión logística según LW, imputación simple (método Hot-Deck, librería R HotDeckImputation[9]) e imputación múltiple, mediante ecuaciones encadenadas[2,10] (librería R mice[10]). Se comparan los resultados en términos de precisión para la estimación de X_1 y de la asociación entre X_1 y X_2 con Y.

En la figura 2 se presentan los resultados de las simulaciones. En términos de precisión de X_1 puede observarse que, con LW, a mayor porcentaje de pérdida, peor precisión, mientras que al trabajar de forma imputada esta se mantiene. En términos de asociación de X_1 con Y se observa que, cuando el patrón de pérdidas es MCAR, todos los métodos realizan estimaciones cercanas al valor real. Sin embargo, cuando el patrón es MAR, LW obtiene estimadores con mayor sesgo al aumentar el porcentaje de valores ausentes. La imputación simple y la imputación múltiple arrojan estimadores cercanos al valor real en todos los casos, ligeramente con menor variabilidad con la imputación múltiple.

Discusión y conclusiones

En nuestra opinión, hay tres razones fundamentales por las que el uso de la imputación múltiple sigue siendo poco frecuente: 1) porque se cree que su objetivo consiste simplemente en sustituir un valor ausente por uno imputado; 2) por la percepción de que es una técnica compleja; y 3) por la creencia de que ante la incertidumbre que provoca un valor ausente lo más prudente es dejarlo como tal. La primera es falsa; sobre la segunda, creemos que puede afirmarse que hay técnicas más complejas cuyo uso está generalizado; y para la última opinamos que, a menudo, imputar puede ser más prudente que no hacerlo (con la información disponible e imputando podemos lograr estimadores más eficientes y menos sesgados, si no insesgados).

Trabajar con LW aumenta la imprecisión, y si el mecanismo de pérdida es MAR, generará estimadores sesgados[5,7]. Hay que distinguir entre imputación simple e imputación múltiple: la primera solo sustituye el valor ausente por otro que es tratado exactamente igual que uno observado; la segunda consiste en un proceso más elaborado que permite capturar la incertidumbre de los valores ausentes. A diferencia de cuando se trabaja con una variable continua, donde la imputación simple suele subestimar el error[5–7], según nuestros resultados para variables dicotómicas parecería que las diferencias entre imputación simple e imputación múltiple no son tan sensibles, siempre que el mecanismo de imputación reproduzca el patrón de pérdida. Y es que la validez de los resultados depende de que, en el caso de la imputación múltiple, el modelo de imputación se realice adecuadamente[3].

Nótese que la magnitud y la dirección del sesgo no siempre coincidirán con lo mostrado en nuestro ejemplo; dependerá de la relación entre las variables estudiadas. Siguiendo a Sterne et al.,[3] en la actualidad los procedimientos de imputación son ampliamente accesibles, por lo que no existe excusa para que los análisis potencialmente engañosos e ineficientes basados en LW sean considerados adecuados sin mayor atención.

Editora responsable del artículo

María Victoria Zunzunegui.

Contribuciones de autoría

Todas las personas firmantes contribuyeron a la concepción y el diseño del trabajo, el diseño de las simulaciones, el análisis y la interpretación de los datos, la escritura del documento y su revisión crítica con contribuciones intelectuales importantes, y aprobaron la versión final para su publicación.

Financiación

Si bien este trabajo no ha tenido financiación directa, el segundo autor ha sido parcialmente apoyado por becas del Instituto de Salud Carlos III (Gobierno de España), cofinanciado por fondos FEDER (Fondos para el Desarrollo Regional Europeo) - Una forma de hacer Europa (referencias: RD12/0036/0056, PI11/02090) y por la Agència de Gestió d'Ajuts Universitaris i de Recerca (2014SGR 756) y RecerCaixa 2015 (MD088652).

Conflicto de intereses

Ninguno.

Agradecimientos

Queremos agradecer a la Dra. Valeria Stuardo MA la lectura crítica y los posteriores comentarios a una de las versiones de este manuscrito.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en doi:10.1016/j.gaceta.2017.01.001

Bibliografía

1. Pérez-Hoyos S, Ferreros I, del Amo J, et al. Imputación del instante de seroconversión al VIH en cohortes de hemofílicos. Gac Sanit. 2003;17:474–82.
2. Van der Palm DW, van der Ark LA, Vermunt JK. A comparison of incomplete-data methods for categorical data. Stat Methods Med Res. 2016;25:754–74.
3. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
4. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 2002.
5. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77.
6. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley-Interscience; 2004.
7. Donders ART, van der Heijden GJMG, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087–91.
8. Cañizares M, Barroso I, Alfonso K. Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. Gac Sanit. 2004;18:58–63.
9. Joenssen DW. HotDeckImputation. Hot Deck Imputation Methods for Missing Data. 2015.
10. Van Buuren S, Groothuis-Oudshoorn K. MICE. Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45:1–67.

8.2. Imputación de datos faltantes para una variable discreta

En muchos estudios en salud, aparecen variables de conteo, por ejemplo, es común estudiar episodios de una enfermedad en particular y estos forman la base de estimaciones de incidencia, cuando también se tiene en cuenta el período de tiempo en el que se observan. Aunque en presencia de datos faltantes de este tipo, parece lógico que el modelo de imputación se base en distribuciones discretas, como Poisson o Binomial negativo. En una revisión de la literatura biomédica y epidemiológica, se encontró que muy pocos estudios emplearon estos métodos para imputar variables discretas. En la práctica, muchos de los procedimientos utilizados para manejar estos datos faltantes tratan la variable como si fuera una variable continua, o tratándola como categórica u ordinal, utilizando técnicas de regresión politómica y, en otros casos, utilizando la estrategia de aplicar alguna transformación de normalización a los datos, y posteriormente utilizando métodos de imputación para datos normales [Landerman *et al.*, 1997].

Mucho menos habitual es el uso de otras técnicas algo más sofisticadas que abordan situaciones específicas, como el problema de ceros inflados[Pahel *et al.*, 2011], una situación que puede ser explicada por la falta de software disponible para llevar a cabo este tipo de imputaciones. Sin embargo, en la actualidad se han desarrollado paquetes, por ejemplo en R, que permiten imputar esta clase de datos [Kleinke *et al.*, 2011] y en algunos estudios se enfatiza la utilidad de los modelos de ceros inflados de Poisson y Binomial Negativo para estudios epidemiológicos con diseños transversales y estudios longitudinales. [Lewsey y Thomson, 2004]. Pocos usan las distribuciones generalizadas que son distribuciones más flexibles que tienen en cuenta los problemas de sobre o subdispersión, también frecuentes en datos de salud. Después de revisar algunos métodos para imputación de variable discreta se construyó un segundo artículo para publicación actualmente sometido y que se muestra en el siguiente apartado [Hernández-Herrera *et al.*, 2020]. La principal conclusión de este trabajo es que la imputación mediante la distribución COMPoisson parece ser la que, en general, ofrece resultados más adecuados.

Regression-based imputation of explanatory discrete missing data

G. Hernández-Herrera[1,2], A. Navarro[1], and D. Moriña[*3,4]

[1]Research Group on Psychosocial Risks, Organization of Work and Health (POWAH), Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona
[2]Instituto de Investigaciones Médicas, Facultad de Medicina, Universidad de Antioquia
[3]Barcelona Graduate School of Mathematics (BGSMath), Departament de Matemàtiques, Universitat Autònoma de Barcelona
[4]Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona (UB)

Abstract

Imputation of missing values is a strategy for handling non-responses in surveys or data loss in measurement processes, which may be more effective than ignoring the losses. When the variable represents a count, the literature dealing with this issue is scarce. If the variable has an excess of zeros it is necessary to consider models including parameters for handling zero-inflation. Likewise, if problems of over- or under-dispersion are observed, generalisations of the Poisson, such as the Hermite or Conway-Maxwell Poisson distributions are recommended for carrying out imputation. In order to assess the performance of various regression models in the imputation of a discrete variable based on Poisson generalisations, compared to classical

*Corresponding Author: David Moriña (dmorina@mat.uab.cat)

counting models, this work presents a comprehensive simulation study considering a variety of scenarios and real data from a lung cancer study. To do so we compared the results of estimations using only complete data, and using imputations based on the Poisson, negative binomial, Hermite, and COMPoisson distributions, and the ZIP and ZINB models for excesses of zeros. The results of this work reveal that the COMPoisson distribution provides in general better results in any dispersion scenario, especially when the amount of missing information is large. When the variable presenting missing values is a count, the most widely used method is to assume that there is equidispersion and that a classical Poisson model is the best alternative to impute the missing counts; however, in real-life research this assumption is not always correct, and it is common to find count variables exhibiting overdispersion or underdispersion, for which the Poisson model is no longer the best to use in imputation. In several of the scenarios considered the performance of the methods analysed differs, something which indicates that it is important to analyse dispersion and the possibile presence of excess zeros before deciding on the imputation method to use. The COMPoisson model performs well as it is flexible regarding the handling of counts with characteristics of over- and under-dispersion, as well as with equidispersion.

1 Introduction

Missing data are practically unavoidable in research in any field, however their consequences for the validity of research findings are not considered in the majority of cases. Nowadays, many scientific journals emphasise the importance of including information about missing data and the strategy used to handle them ([5]), and yet it is still not common to find this, and very few publications explicitly describe missing data and the methods used to deal with them. Data can be missing due to non-responses or be caused by problems in study design, or even created deliberately by researchers as part of privacy policies. In order to identify the behaviour of missing data in a sample, in terms of quantity and mechanisms of data loss, the first step is to carry out a statistical description of the observed variables, identify the data lost in each one, and then identify the patterns of data loss which will help to make a decision about a method for handling the missing data. In many cases, researchers resort to using only data which are completely available,

ignoring those which are missing when performing analyses; but this decision can generate problems of bias and lack of precision in estimates and affect the power of the study due to the reduction in sample size.

In order to adequately deal with missing data it is necessary to identify the mechanism of data loss (process of non-response) defined as the origin, causes, moment, relationships or characteristics which give rise to the lack of information. Moreover, it is important to establish whether the observations have been lost randomly, or whether their loss is associated with definable causes, and to determine the percentage of missing data in the sample, since depending on these factors, the levels of uncertainty in working with imputed data can vary significantly ([12]).

The mechanism of data loss is classified depending on the probability of response: if this is independent of the observed and unobserved data, we say the non-response process is MCAR, missing completely at random. If however it is dependent on the observed data, we say the non-response process is MAR, missing at random. When the process is neither MAR nor MCAR it is termed *informative*, NMAR, not missing at random ([17]).

The advantage, statistically speaking, of the mechanism of loss being MCAR is that conclusions obtained from the analysis of these data can still be valid. The power of the study can be affected by the design, but the parameters estimated are not biased by the absence of data. However, MAR is the most common mechanism and thus it is important to take into account that the probability of data loss is conditionally independent; but if the mechanism is NMAR, this represents a difficulty for imputation, since the estimation of parameters requires knowledge of the model of data loss in order to achieve unbiased estimates.

[15], [21] and [8] have constructed a taxonomy of the most popular methods for handling missing data, which shows that when the data loss mechanism is completely random (MCAR) the methods stochastic regression, multiple imputation, maximum likelihood with Expectation-Maximization (EM) algorithm, Bayesian imputation, and weighted methods produce consistent estimates; when the mechanism of data loss is random (MAR) multiple imputation produces consistent estimates but only under certain conditions, just as with weighted imputation methods, whereas for the non-random mechanism (NMAR) none of the methods produce consistent estimates although it is possible to achieve imputations using some of these methods provided that the correct probability model of data loss can be identified ([16]).

In many studies, counting variables appear, for example in health research

it is common to study episodes of a particular disease and these form the basis of estimates of incidence, when the time-period in which they are observed is also taken into account. Although in the presence of missing data of this type, it would seem logical that the imputation model be based on discrete distributions, such as Poisson or Negative Binomial, in a bio-medical and epidemiological literature review, it was found that very few studies employed these specific methods to impute discrete variables. In practice, the procedures used to handle such missing data were as though the variable had been continuous, or treating it as categorical or ordinal, using polytomic regression techniques, and in other cases using the strategy of applying some normalising transformation to the data, and subsequently using imputation methods for normal data ([13]).

Much less common is the use of other more sophisticated techniques which tackle specific situations, such as the problem of inflated zeros ([20]), a situation which may be explained by the lack of software available to carry out imputations of this type. However, some packages have now been developed, for example in R, which allow imputation of these kinds of data ([12]) and some studies have stressed the utility of Poisson and Negative Binomial zero-inflated models for epidemiological studies with both cross-sectional and longitudinal designs ([14]). Few studies use generalised distributions, which are more flexible and take into account problems of over- and under-dispersion, also common in health data: for example no studies are found employing the Hermite distribution, a generalisation of the Poisson distribution, more flexible when there is over-dispersion for the imputation variables of count data. Nor are studies found employing the Conway Maxwell Poisson distribution which permits modeling count data in the presence of over- and under-dispersion.

It is also not common in scientific literature to find an exhaustive examination of missing information in a discrete variable acting as covariate in an analysis. However, this point is crucial in certain contexts such as survival analysis in the presence of recurrent events, where the usual analysis techniques introduced in [1] and [22] require knowledge of the number of previous events suffered by individuals, something which of often unknown (for example events which occurred prior to initiation of a cohort study). The aim of the present study is precisely to assess the performance of methods of imputation of missing data in a discrete covariate, based on generalisations of the Poisson distribution, in comparison to the classical Poisson and Negative Binomial counting methods, in different scenarios of dispersion and nature of

response variable and within a framework of multiple imputation, following the recent recommendations in many scenarios like confirmatory clinical trials ([3]). Although applied researchers were reluctant to using multiple imputation methods until recently, the implementation in most used data analysis software has increased their popularity in the latest years. [11] present an alternative based on regression models, but accounting only for continuous covariates. The considered regression models are described in the next section and their performance on real lung cancer clinical trial data and on a comprehensive simulation study are analysed in Section *Results*.

2 Methods

In this section we present some discrete variable regression models, on the basis of which to carry out imputation of missing data. The classical counting models including Poisson, negative binomial (parameterised in terms of its mean μ and dispersion index d) and their zero-inflated versions are described in the supplementary material, and only the less known distributions (Hermite and Conway-Maxwell Poisson) are presented here. These distributions are very flexible and may be adapted and used in any scenario of dispersion. A comprehensive description of most common count data modeling strategies with special focus on dispersion issues can be found in [7]. In order to carry out imputation of missing data using the regression models described in this section, two phases are required; firstly, a generalised linear model (GLM) is fitted using the covariate X as response and the response Y as covariate, based on the corresponding distribution (Poisson, NB, ZIP, ZINB, Hermite or COMPoisson). Imputed values are randomly sampled from the corresponding distribution with the parameters obtained in the previous step, including random noise generated from a normal distribution in all cases. In order to produce proper estimation of uncertainty, the described methodology can easily be extended to a multiple imputation framework. The results reported in Section *Results* correspond to the combination of $m = 5$ imputed data sets, according to the well known Rubin's rules ([27]) and based on the following steps in a Bayesian context:

1. Fit the corresponding count data model and find the posterior mean and variance $\hat{\beta}$ and $V(\hat{\beta})$ of model parameters β.

2. Draw new parameters β^* from $N(\hat{\beta}, V(\hat{\beta}))$.

3. Compute predicted scores p using the parameters obtained in the previous step (the actual expression depends on the count data model).

4. Draw imputations from the corresponding count data distribution and scores obtained in the previous step.

The performance of these models in different scenarios will be compared in the next section. The simulation strategy is described in detail in Section *Simulation study*.

2.1 Hermite distribution

The Hermite distribution results from the summation of two independent Poisson variables X_1 and X_2: $Y = X_1 + mX_2$. It is very useful for modeling count data with a multimodal distribution and with overdispersion ([10]).

The probability generating function (PGF) for Hermite distribution is:

$$P(s) = \exp(a_1(s-1) + a_m(s^m - 1)). \tag{1}$$

Setting the positive integer $m \geq 2$ (the *order* or *degree* of the distribution), the domain of the parameters is $a_1 > 0$ and $a_m > 0$.

As a_m tends to zero, this distribution tends to a Poisson.

The PGF $P(s)$ corresponds to the PGF of $X_1 + mX_2$, where X_i are independent random variables following a Poisson distribution with population mean a_1 and a_m respectively ([23]) The functions for calculating probabilities and fitting Hermite regression models have recently been implemented in R ([18]).

The Hermite distribution is said to be zero-inflated with respect to the Poisson distribution, because the probability of the variable taking value zero under Hermite is greater than under Poisson, when the two distributions have the same mean. This characteristic of the Hermite distribution allows proposing the use of a Hermite regression model for count variable with an excess of zeros, instead of using a classical Poisson regression model.

2.2 Conway-Maxwell Poisson distribution

A generalisation of the Poisson distribution, taking account of dispersion in the data. The regression model based on the Conway-Maxwell Poisson

distribution (COMPoisson) allows modeling count data in three dispersion scenarios: equidispersion, overdispersion and underdispersion, and is therefore an interesting alternative from the point of view of the present study.

The probability distribution function is:

$$P(x, \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu Z(\lambda, \nu)} \tag{2}$$

where $\lambda = E(x^\nu)$ with ν the dispersion parameter and $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^\nu}{(j!)^\nu}$ is a normalising constant. If $\nu = 1$ the distribution is equidisperse, and it is overdisperse (underdisperse) if $\nu < 1$ ($\nu > 1$).

The COMPoisson distribution is an extension of Poisson with two parameters which generalises certain discrete distributions, such as the Poisson distribution when $\nu = 1$, the Bernoulli distribution with probability $\frac{\lambda}{1+\lambda}$ when $\nu \to \infty$ and the geometric when $\nu = 0$ and $\lambda < 1$; it was first suggested in 1962 by Conway and Maxwell in [4] (see[29] for a recent review of its applications). This distribution may also be seen as a weighted Poisson distribution with weighting function $w_y = (y!)^{1-\nu}$. In this sense, [26] compared the COMPoisson with a weighted Poisson where the weights take the following form:

$$W_y = \begin{cases} e^{-\beta_1(\lambda - y)} & \text{if } y \leq \lambda \\ e^{-\beta_2(y - \lambda)} & \text{if } y > \lambda \end{cases} \tag{3}$$

There is underdispersion when $\beta_1, \beta_2 > 0$, overdispersion when $\beta_1, \beta_2 < 0$ and equidispersion when $\beta_1 = \beta_2 = 0$.

This is a flexible distribution which can take account of the excessive or insufficient dispersion often found in count data ([2]). The distribution is appropriate for use in imputation of count data when there are dispersion problems, replacing the Poisson distribution. For example, the COMPoisson distribution has been used in linguistics to model word length, to model count data in marketing, and eCommerce, and to model grocery shop sales, among other uses. The capacity to handle different types and levels of dispersion makes the distribution more useful in applications where the level of dispersion may vary ([28]).

2.3 Simulation study

In order to assess the efficiency of the methods considered in this study, we designed a simulation procedure based on the following algorithm.

1. Generate a population of size 1,000,000 with two variables. One variable Y, following a binomial, normal or Poisson distribution will be used as the response, depending on the scenario considered (binary, continuous or discrete response variable), and a second will be used as the explanatory variable X, consisting of a count of the number of events and based on a Poisson, negative binomial or zero-inflated distribution, depending on the scenario of dispersion and zero-inflation. The covariate X is generated first randomly sampling from the corresponding distribution and then the response Y is generated on the basis of a GLM with different intensities of association with the explanatory variable ($\beta = 0.5$, $\beta = -0.3$ and $\beta = -0.5$) although we only report results corresponding to $\beta = 0.5$ because no relevant differences were observed compared to the other values. These values of β were considered to keep the association between the response variable and the coviarate within the usual ranges, with a moderate and reasonable strength.

2. From the generated population, randomly select 1000 samples of size 2000 and generate missings in the explanatory variable using a MAR or MCAR mechanism, and for percentages of 5% to 30%. Regarding missings generated via MAR, the criteria was that 80% of missings corresponded to values of 0 for the binary variable or values below the mean for discrete or continuous variables and 20% of the missings corresponded to values of 1 for the binary variable or values above the mean for discrete or continuous variables.

3. For each sample, fit logistic, linear or Poisson regression models depending on the response variable, as follows:

    - **Listwise deletion (lw)**. Fit the regression model eliminating missing values, in other words, using only the information available.
    - **Poisson (pois)**. Fit a Poisson regression model using the count variable as response. Based on the estimated coefficients, impute missing values and subsequently fit the regression model, incorporating the set of imputed values.
    - **Negative Binomial (nb)**. Similar to the above procedure, but impute missing values using a Negative Binomial regression model.

- **Hermite**. In this case, use a Hermite regression model to impute the missing values in the discrete variable.
- **COMPoisson (cmp)**. Use Conway-Maxwell Poisson regression model to perform imputation of the missing data.
- **Zero-inflated Poisson (zpois)**. Use a Poisson regression model with zero-inflation to impute the missing values.
- **Zero-inflated negative binomial (znb)**. Use a Negative Binomial regression model to impute the missing values.

The efficiency of the proposed imputation methods was assessed by comparing relative bias with respect to the population parameter, the average length of confidence intervals for the parameter of interest (AIL) and the percentage of coverage. The same procedure was also carried out using samples of size $n = 200$ but the results showed no differences with those reported, and have been omitted. The following dispersion scenarios were considered:

- Equidispersion: the explanatory variable was generated following a Poission distribution with parameter $\lambda = 2$. The procedure described was performed using other values for λ but the results did not differ from those reported.

- Overdispersion: the explanatory variable was generated following a negative binomial distribution with mean $\mu = 2$ and dispersion index $d = 2$. The procedure described was performed using other values for μ and d but the results did not differ from those reported.

- Underdispersion: the explanatory variable was generated following a Poisson distribution with parameter $\lambda = 2$, and the underdispersion was generated through an iterative procedure, substituting values for the mean at random until a dispersion of $d = 0.5$ was obtained. The procedure described was performed with other values for λ and d but the results did not differ from those reported. In this case, the methods based on the Hermite distribution or on zero-inflated distributions were not considered due to convergence issues.

- Excess zeros: the explanatory variable was generated following a Poisson distribution with parameter $\lambda = 2$ and subsequently a random 10%

of values were replaced by zeros. The procedure described was performed using other values for λ and other proportions of zero-inflation but the results did not differ from those reported.

The tables including the results for all kinds of response variables and all dispersion scenarios are available as supplementary material.

3 Results

The performance of the considered imputation methods is compared in this section in a real data example from a randomised clinical trial (RCT) of two treatment regimens for lung cancer, first introduced in [9]. The considered variables are survival time (continuous response) and the number of months from diagnosis to randomisation (discrete explanatory variable), in which a quantity of random missing values were introduced 10%, 20%, 30% and 40%. Although the main interest in practice in a RCT would be to estimate the treatment effect, we focus here on the effect of the discrete covariate over the continuous response, as it was an observational study like the simulation study presented in Section *Simulation study*.

3.1 Lung cancer data

The explanatory variable number of months from diagnosis to randomisation is clearly overdisperse (the variance is 112.62 while the mean is 8.77). Table 1 shows the performance measures for each of the considered imputation methods (all methods included in a multiple imputation framework with $m = 5$ imputed data sets), and it can be seen that COMPoisson performs better than the rest of alternative methods regarding any of the considered measures and in any of the missing data scenarios (10% to 40%). Using the full data, the estimate is $\hat{\beta} = -0.69$ ($SD = 1.28$).
The 95% confidence intervals include the reference value of $\hat{\beta}$ in all cases.

3.2 Simulation study results

Below we present the results for estimations of the β coefficient and standard error of the regression models fitted, handling missing values in the cases of continuous response variables by the methods: listwise deletion (lw), imputation with Poisson regression (pois), imputation with negative binomial

Table 1: Average interval length (AIL) for each imputation method. p-values from χ^2 goodness of fit test comparing the full data and imputed distributions.

% of missing data	Model	AIL	p-value
10%	Listwise deletion	5.54	0.014
	Poisson	5.51	0.155
	Negative binomial	5.25	0.055
	Hermite	5.38	0.154
	COMPoisson	5.25	0.053
20%	Listwise deletion	5.78	0.002
	Poisson	5.79	0.158
	Negative binomial	5.43	0.121
	Hermite	5.44	0.152
	COMPoisson	5.21	0.001
30%	Listwise deletion	8.93	0.014
	Poisson	9.62	0.647
	Negative binomial	8.84	0.373
	Hermite	8.34	0.186
	COMPoisson	8.00	0.123
40%	Listwise deletion	9.79	0.157
	Poisson	10.37	0.626
	Negative binomial	8.40	0.822
	Hermite	8.59	0.156
	COMPoisson	8.37	0.306

regression (nb), imputation with Hermite regression (herm), imputation using the COMPoisson model (comp), imputation with zero-inflated Poisson model (zpois) and imputation with zero-inflated negative binomial model (znb). For the other types of response variable, results are provided as supplementary material. We also present biases in the estimations and true coverage indices of the confidence intervals for the same case in all scenarios.

According to these results, we observe a considerable increase in bias when over 20% of values are missing, in all scenarios of dispersion and response variable (see Tables S1, S2, S3 and S4) in the Supplementary material. Figure 1 shows the behaviour of bias for each imputation method under the scenario of equidispersion, and it may be seen that in the estimate using only the information available without imputation (listwise deletion), the bias is greater in comparison to the other methods. The same trends can be seen for the average length of the confidence intervals and their coverage (see Figure 2 and Figure 3).

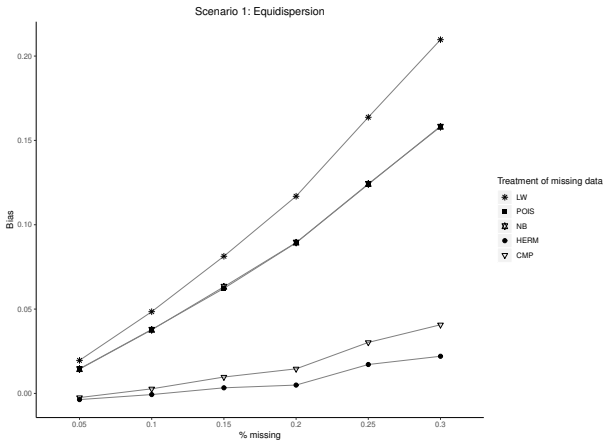


Figure 1: Bias in the coefficient estimate for the scenario of equidispersion with continuous response variable.

In the scenario of overdispersion with a continuous response variable, the behaviour of bias is similar when listwise deletion is used, or when using zero-inflated models to carry out the imputation; lower bias may be seen when imputation is done with the negative binomial or COMPoisson models, as Figure 4 shows, however coverage of the confidence intervals is low compared to the other methods (see Figure 6).

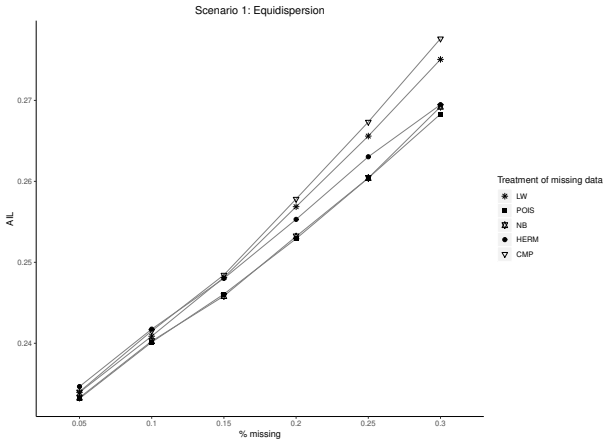


Figure 2: Average length of confidence intervals of the coefficient in the scenario of equidispersion and continuous response variable.

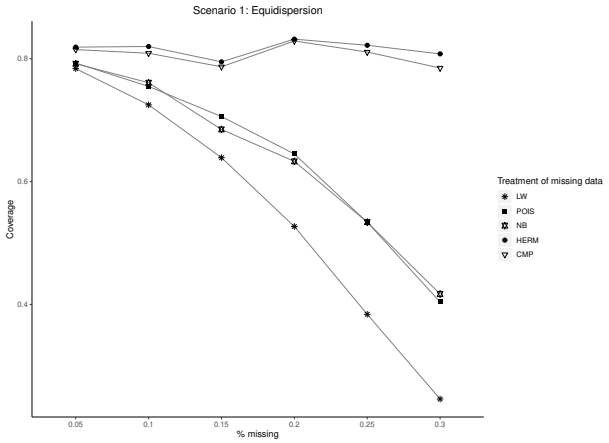


Figure 3: Coverage of confidence intervals for the coefficient in the scenario of equidispersion and continuous response variable.

In the case of underdispersion the results presented in Figure 7 show that in this scenario the biases are lower when imputing using the Poisson and Negative Binomial regression models (in this scenario it is not possible to obtain the maximum likelihood estimators corresponding to the Hermite distribution) and that the coverage of confidence intervals is greater for these two models (see Figure 9).

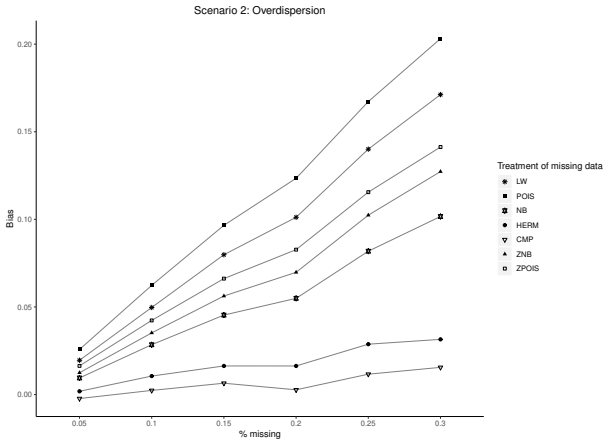


Figure 4: Bias in coefficient estimation when there is overdispersion and the response variable is continuous.

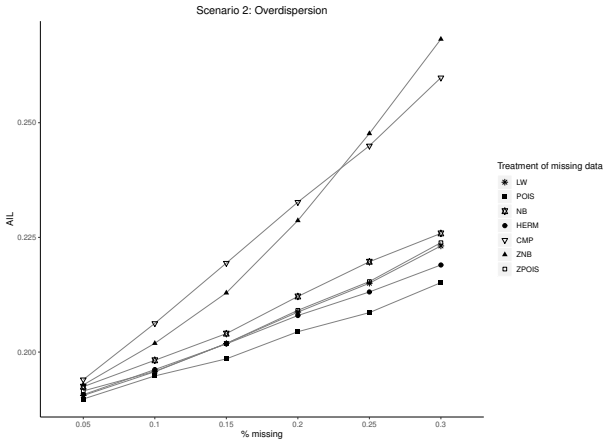


Figure 5: Average length of confidence intervals of the coefficient when there is overdispersion and the response variable is continuous.

When the data present, apart from missing values, an excess of zeros, the results show an improved estimation behaviour using the zero-inflated Poisson and Negative binomial models, as is to be expected, although the performance of the imputation methods considered is in general worse than in the other dispersion scenarios, as may be seen in Figure 10.

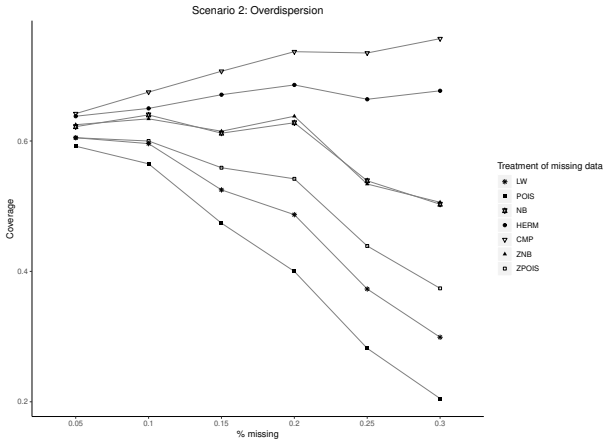


Figure 6: Coverage of confidence intervals of the coefficient in the scenario of overdispersion and continuous response variable.

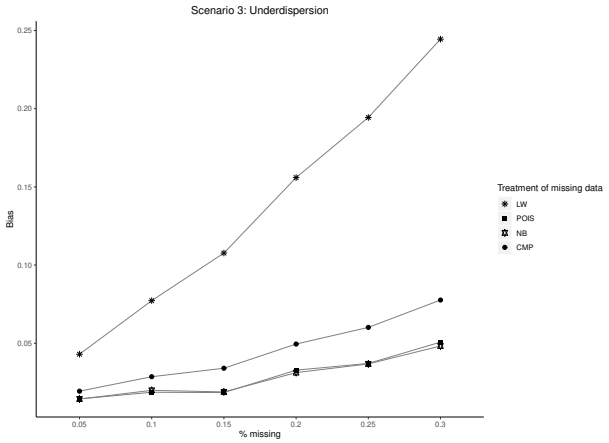


Figure 7: Bias in estimation of the coefficient when there is underdispersion and the response variable is continuous.

4 Discussion

In medical research it is common for data to be partially missing and it is necessary to cope with this problem in order to analyse the information in a coherent and consistent manner. Many methods are available which allow this, and proper handling of information which is lacking will depend on

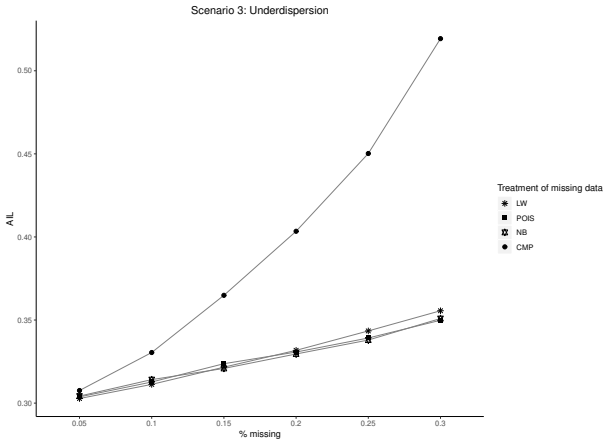


Figure 8: Average length of confidence intervals in the scenario of underdispersion and continuous response variable.

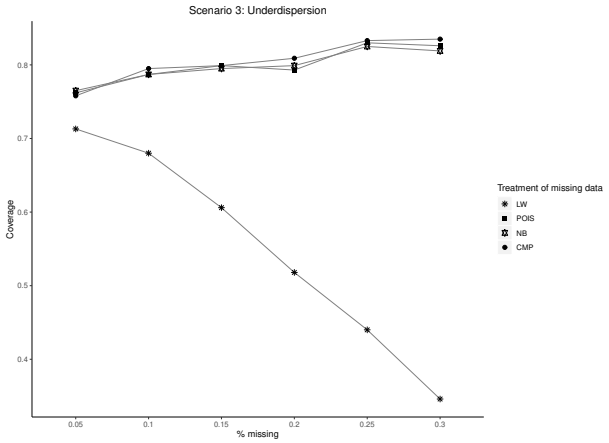


Figure 9: Coverage of confidence intervals in the scenario of underdispersion and continuous response variable.

choosing the most appropriate one.

When the variable presenting missing values is a count, there are various alternative ways to impute such missing values, which depend on the distribution characteristics of the count variable, particularly the behaviour of the mean and variance. The most widely used method is to assume that there is equidispersion and that a classical Poisson model is the best alternative

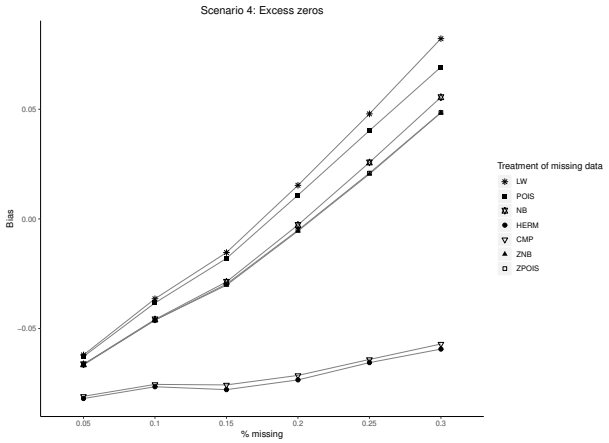


Figure 10: Bias in estimation of the coefficient when there is an excess of zeros and the response variable is continuous.

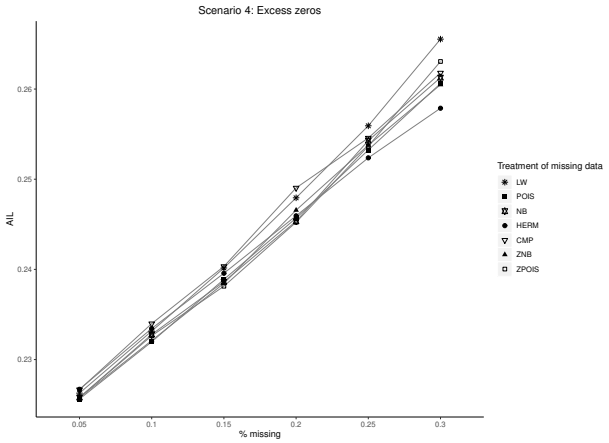


Figure 11: Average length of confidence intervals in the scenario of excess zeros and continuous response variable.

to impute the missing counts; however, in real-life research this assumption is not always correct, and it is common to find count variables exhibiting overdispersion or underdispersion, for which the Poisson model is no longer the best to use in imputation. If there is overdispersion the Poisson model underestimates the amount of dispersion. In recent years much work has been done on implementation of other counting models, which may be gen-

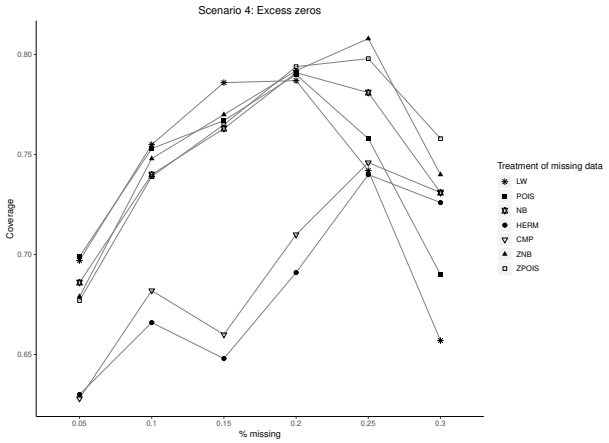


Figure 12: Coverage of confidence intervals when there is an excess of zeros and the response variable is continuous.

eralisations of the Poisson model and which take over- and under-dispersion into account, as well as the problem of excess zeros ([24]).

The lung cancer example shows that the COMPoisson distribution provides with a powerful alternative to missing data imputation in a realistic situation of overdisperse discrete explanatory variables, even with a large proportion of missing information, within the framework of multiple imputation, which is now implemented in many standard software and is therefore available to researchers in a straightforward way.

Moreover, in the simulation study we generated different scenarios with a variety of percentages of missing data and three distributions of the response variable: continuous, binary and discrete, and subsequently fitted the respective models with the imputed independent variable using one of the models mentioned. The results of this simulation allow us to affirm that in order to impute a count variable, it is not sufficient to assume the distribution is Poisson; it is necessary to identify the relationship between the variance and the mean of the data, as well as whether the presence of excess zeros might make it appropriate to use specific models for handling missing data in this scenario. Additionally, it can be seen that especially for continuous and discrete response variable when the proportion of missing information is very high (over a 20%), imputing the missing values can lead to inaccurate results regarding relative bias and extremely low coverage rates.

One of the most unexpected findings is that fitting models using only the available data in some cases produces estimator with less bias than performing imputation of the missing values ([6]), in the case of a dichotomic response variable. Although it is recognised that fitting a model with only the available data may affect the power of the study and produce imprecise estimations, it is evident that the effects on power and precision may be significant if the percentage of missings is low and the mechanism of data loss is completely random.

5 Conclusions

In several of the scenarios considered the performance of the methods analysed differs, something which indicates that it is important to analyse dispersion and the possibile presence of excess zeros before deciding on the imputation method to use. Specifically, in the scenario of equidispersion and binary response variable, when a logistic regression model is fitted imputing missing values with the MAR mechanism using the different models mentioned we observe, as expected, that the Poisson and Negative Binomial models produce estimations with low bias and acceptable coverage. Moreover, the COMPoisson model performs well as it is flexible regarding the handling of counts with characteristics of over- and under-dispersion, as well as with equidispersion ([28]). If, however, the count variable is overdisperse, as is often the case in health research, there are various alternatives for performing imputation, the Negative Binomial model being the most recommended. In our results, just as in the case of equidispersion, when the response variable is binary, estimating using available data without performing imputation produces good estimators and with low bias, however it is important to observe that in this case the imputation methods which perform best are those employing zero-inflated models, and the COMPoisson model works very well. For the case where the variable presents underdispersion we observe that imputation based on Poisson and Negative Binomial regression models perform similarly, although the Negative Binomial can present certain difficulties with convergence in this scenario. Furthermore, regard in size of confidence intervals, it is curious that in fitting these models, the higher the percentage of missings the smaller the confidence intervals (apparently an artifact) whereas the COMPoisson is able to maintain their size, and it is worth emphasising that with listwise deletion size increases, as in the other scenarios. When in ad-

dition to having missings the data has excess zeros the estimates of the β parameter in the case of binary response are more precise, i.e. they have less bias and greater coverage but, as in the previous scenarios, this result is similar to when no imputation is performed and only the available data is used. For the case of continuous response variable (Table S4) estimates of the β coefficient are always smaller than the value of the parameter and coverage of the confidence intervals is acceptable, coverage being even greater for listwise, although at the expense of some confidence intervals being considerably wider.

According to the results of the simulation obtained in this study, the choice of the best method of imputation for count variables depends on various factors such as the amount of missing data, the behaviour of the expected value in relation to the variance, i.e. whether there is equi-, under-, or overdispersion and the distribution of the response variable. In particular, if data present an excess of zeros, this represents an additional factor to be taken into account when choosing the missing data imputation method. Although in practice the exact distributional form of the incomplete covariates is unknown because, precisely, of the missing information, the behaviour of many phenomena in the context of public health are well established. For instance, it is well known that the risk of suffering a new sickness leave increases with the number of previous events (see [25] for instance), which would lead to overdispersed data, and the same behavior can be observed with the number of falls suffered by long-term centers residents (as in [19]).

Acknowledgements

David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and from Fundación Santander Universidades.

References

[1] P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10:1100–1120, 1982.

[2] S. Chakraborty and S. H. Ong. A COM-Poisson-type generalization of the negative binomial distribution. *Communications in Statistics - Theory and Methods*, 45(14):4117–4135, jul 2016.

[3] Committee for Medicinal Products for Human Use (CHMP). Guideline on Missing Data in Confirmatory Clinical Trials. Technical report, European Medicines Agency, London, 2010.

[4] R. W. Conway and W. L. Maxwell. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136, 1962.

[5] Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, 15(1):30, dec 2015.

[6] Gilma Hernández, David Moriña, and Albert Navarro. Imputing missing data in public health general concepts and application to dichotomous variables. *Gaceta Sanitaria*, 31(4):342–345, 2017.

[7] J. Hilbe. *Negative Binomial regression*. New York: Cambridge University Press, 2 edition, 2011.

[8] Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[9] D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 1980.

[10] CD Kemp and Adrienne W Kemp. Some properties of the hermite distribution. *Biometrika*, 52(3-4):381–394, 1965.

[11] Soeun Kim, Catherine A Sugar, and Thomas R Belin. Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in medicine*, 34(11):1876–88, may 2015.

[12] Kristian Kleinke, Roel de Jong, Martin Spiess, and Jost Reinecke. Multiple imputation of incomplete ordinary and overdispersed count data.

Bielefeld University, Faculty of Sociology and Centre for Statistics, 1, 2011.

[13] Lawrence R. Landerman, Kenneth C. Land, and Carl F. Pieper. An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. *Sociological Methods & Research*, 26(1):3–33, aug 1997.

[14] J. D. Lewsey and W. M. Thomson. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology*, 32(3):183–189, jun 2004.

[15] Roderick JA Little. Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[16] T. Martin Lukusa, Shen-Ming Lee, and Chin-Shang Li. Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, 79(4):457–483, may 2016.

[17] Geert Molenberghs and Els Goetghebeur. Simple fitting algorithms for incomplete categorical data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):401–414, 1997.

[18] David Moriña, Manuel Higueras, Pedro Puig, and María Oliveira. Generalized Hermite distribution modelling with the R package hermite. *R Journal*, 7(2):263 – 274, 2015.

[19] Albert Navarro and Iciar Ancizu. Analyzing the occurrence of falls and its risk factors: Some considerations. *Preventive Medicine*, 48(3):298–302, mar 2009.

[20] Bhavna T. Pahel, John S. Preisser, Sally C. Stearns, and R. Gary Rozier. Multiple imputation of dental caries data using a zero-inflated Poisson regression model. *Journal of Public Health Dentistry*, 71(1):71–78, jan 2011.

[21] Therese D Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383, 2001.

[22] Peterson A. V. Prentice R. L., Williams B. J. On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–379, 1981.

[23] Pedro Puig. Characterizing Additively Closed Discrete Models by a Property of Their Maximum Likelihood Estimators, With an Application to Generalized Hermite Distributions. *Journal of the American Statistical Association*, 98(463):687–692, 2003.

[24] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

[25] Ricardo J. Reis, Mireia Utzet, Poliana F. La Rocca, Fúlvio B. Nedel, Miguel Martín, and Albert Navarro. Previous sick leaves as predictor of subsequent ones. *International Archives of Occupational and Environmental Health*, 84(5):491–499, jun 2011.

[26] Martin S Ridout and Panagiotis Besbeas. An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89, 2004.

[27] D.B. Rubin. *Multiple Imputation for nonresponse in Surveys*. John Wiley & Sons, Inc., 1987.

[28] Kimberly F Sellers, Sharad Borle, and Galit Shmueli. The com-poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28(2):104–116, 2012.

[29] Galit Shmueli, Thomas P Minka, Joseph B Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005.

9. Evaluación del rendimiento del método propuesto

Una vez visto qué método de imputación para variable discreta parece más útil, se procede a probar la propuesta realizada en el capítulo 7 de esta tesis, comparando su rendimiento con la de otras posibles alternativas en varios escenarios, mediante un exhaustivo estudio de simulación. Así, se lleva a cabo una simulación de seis poblaciones, correspondientes a cohortes de trabajadores previamente descritas. Las tres primeras cohortes simulan la historia de los trabajadores en función de la ocurrencia de la baja por enfermedad de larga duración (SL) con varias intensidades de dependencia de evento. Estas simulaciones se realizaron asumiendo distribuciones Weibull con parámetro ancilar igual a 1 (es decir, distribuciones exponenciales), en cuyo caso el riesgo se puede suponer constante dentro de cada episodio, pero diferente entre episodios (es decir, con dependencia de eventos). Para las demás poblaciones, las cohortes de trabajadores se simulan por igual y el desenlace es *baja por enfermedad* según el diagnóstico (sistema respiratorio, sistema musculoesquelético y trastornos mentales y del comportamiento). En estos casos la función de riesgo no es constante entre episodios. Se simularon cuatro situaciones diferentes para cada población, que son combinaciones entre dos posibles tiempos de seguimiento (2 y 5 años) y dos tiempos máximos de riesgo previos al inicio de la cohorte (2 y 10 años) como se describe en el artículo que se presenta a continuación.

Dicho artículo que tiene como título *Left-censored recurrent event survival analysis in epidemiological studies: a proposal when the number of previous episodes are unknown*, está actualmente en consideración en una revista científica y concluye que en general la propuesta en que se basa esta tesis parece ser una alternativa más que razonable en la mayoría de situaciones posibles.

Left-censored recurrent event survival analysis in epidemiological studies: a proposal when the number of previous episodes is unknown

Gilma Hernández-Herra[1,2], David Moriña[3,4], Albert Navarro[5,6]

1. Instituto de Investigaciones Médicas, Facultad de Medicina, Universidad de Antioquia. Medellín, Colombia.
2. PhD program in Methodology of Biomedical Research and Public Health. Autnomous University of Barcelona, Cerdanyola del Vallès, Spain.
3. Barcelona Graduate School of Mathematics (BGSMath), Department of Mathematics, Autnomous University of Barcelona, Cerdanyola del Vallès, Spain.
4. Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, University of Barcelona.
5. Research group on Psychosocial Risks, Organization of Work and Health (POWAH), Autonomous University of Barcelona (UAB), Cerdanyola del Vallès, Spain
6. Biostatistics Unit, Faculty of Medicine, Autonomous University of Barcelona (UAB), Cerdanyola del Vallès, Spain

Corresponding author:

Albert Navarro. Biostatistics Unit, Faculty of Medicine, Autonomous University of Barcelona (UAB). Avda. Can Domènech S/N, 08193 Cerdanyola del Vallès, Spain. Mail: albert.navarro@uab.cat

Word Count: 2984

ABSTRACT

Background: Left censoring can occur with relative frequency when analysing recurrent events in epidemiological studies, especially observational ones. Concretely, the inclusion of individuals that were already at risk before the effective initiation in a cohort study, may cause the unawareness of prior episodes that have already been experienced, and this will easily lead to biased and inefficient estimates. The objective of this paper is to propose a statistical method that performs successfully in these circumstances.

Methods: Our proposal is based on the use of models with specific baseline hazard, imputing the number of prior episodes when unknown, with a stratified model depending on whether the individual had or had not previously been at risk, and the use of a frailty term. The performance of the method is examined in different scenarios through a comprehensive simulation study.

Results: The proposed methodology achieves notable performance even when the percentage of individuals at risk before the beginning of the follow-up is very elevated, with biases that are often under 10% and coverages of around 95%, sometimes somewhat conservative. If the baseline hazard is constant, it seems to be that the “Gap Time” approach is better; if it is not constant, the “Counting Process” seems to be a better choice.

Conclusions: Because of the lack of knowledge of the prior episodes that have been experienced by a part (or all) of the individuals in a sample, the use of common baseline methods is not advised. Our proposal seems to perform acceptably in the majority of the scenarios proposed, becoming an interesting alternative in this context.

KEY MESSAGES:

- Left censoring, when analysing recurrent events, is a situation that can occur with certain frequency in epidemiological studies, especially observational ones. More concretely, it will occur in every cohort study that follows individuals that were already at risk of experiencing the outcome of interest before the beginning of the follow-up.
- If the recurrent event presents event dependence, not knowing the number of episodes that each individual has had prior to the effective initiation of the follow-up, and analysing this through “classical” methods, is an important problem, as it will lead to biased and inefficient estimates.
- In this paper we propose a method based on imputation to approach those situations, which includes stratification depending on whether the individual has or has not previously been at risk, and the use of a term of frailty.
- The performance of the proposal is examined through a comprehensive simulation study based on real populations and is compared against a model that ignores the prior episodes.
- Our proposal achieves significant improvement over usual approaches, even when the percentage of individuals at risk before the beginning of the follow-up is considerable with small biases, and coverages of around 95%, sometimes slightly conservative.

Often, in cohort epidemiological studies, all or part of the individuals included have been at risk of the event of interest before the beginning of the follow-up of the study, especially in the case of observational designs. This can lead to the unawareness of the prior history of these individuals, more specifically of the time at risk, and if they have experienced the event (if corresponds, and how many times) at the beginning of the cohort. Not knowing the amount of time an individual has been at risk of the event of interest will be a problem, when the baseline hazard of having the event depends on time. With regards to the lack of knowledge of whether the event has already happened, we can identify two situations: first, if the event can only occur once, and it has already been produced, the result for this individual is determined regardless of how long we follow him or her. We are before what is known as left censoring with a series of specific statistical techniques for its analysis; second, if the outcome of interest can be observed more than once on a same individual, in other words, it is a recurrent event, one or several new episodes of the event can be observed but the number of prior episodes will be unknown. In this case we will be in a situation of left censoring where the censored variable is of the discrete type, that can also define different baseline hazards.

This paper is situated in the context in which the prior history is unknown for all, or some, of the individuals included in a cohort, when the outcome of interest is a recurrent event with event dependence. Specifically, we suppose that we know the moment from where all individuals are at risk, we do now know the number of prior episodes they have experienced. This is a realistic situation in the practice of several cases; for example, it is very probable that in a work force cohort we know when a worker started to work (thus, to be at risk of having a sick leave) and, however, and especially for people with ample trajectory, that we don´t know whether or not in effect they have already had sick leaves (and in this case, how many). Another example, of this situation is a study of cohorts with an outcome of incidence of infection form Human Papilloma Virus on adult women. It would be relatively simple to know how long they have been at risk (beginning of active sexual life), however, we will not be able to know the number of infections since most of the time, when they occur, they are asymptomatic.

Event dependence

When we analyse a recurrent event, we frequently find a phenomenon called event dependency which implies that the baseline hazard of having an episode depends on the number of episodes that have already been experienced. The phenomenon of the event dependence has been estimated for several outcomes such as falls,[1] sickness absence,[2,3] hospitalizations in heart failure[4] or cardiovascular readmissions post percutaneous coronary intervention,[5] showing in all the aforementioned cases that the baseline hazard increases most significantly according to the prior episodes experienced.

Methods to analyse recurrent events in epidemiological studies are based fundamentally on extensions of Cox's[6–10] classical model of proportional hazards. Specifically, the methods that consider the existence of event dependence are specific baseline hazard models, also called conditional models or Prentice, Williams and Peterson (PWP).[11] Through stratification according to the number of prior episodes, these models assume that the baseline hazard of having an episode of the event is different as a function of the episodes that the individual has already had, allowing then to calculate a general effect or specific effects for each episode. Therefore, all individuals are at risk for the first strata, but only those with an episode on the previous strata would be at risk for the following strata.

PWP models can be formulated two different ways according to the specification of the risk interval used,[8] that is to say, according to how time is considered. In the first one, called "Counting Process", time is considered in the standard way of survival analysis, being referenced always at the beginning of the follow-up, so that the beginning of the k_{th} episode is always posterior at the end of the $k\text{-}1_{st}$. Its hazard function is shown below:

$$\lambda_{ik}(t) = \lambda_{0k}(t)e^{X_i\beta} \quad (1)$$

where $X_i\beta$ represents the vector of covariables and the coefficients of the regression, k is the k_{th} episode of the event for individual i and $\lambda_{0k}(t)$ is the function of the baseline hazard that depends on k.

In the second format called "Gap Time", time is considered always in relation to the previous episode, thus being that the beginning of each new episode for a same individual is set at zero:

$$\lambda_{ik}(t) = \lambda_{0k}(t-t_{k-1})e^{X_i\beta} \quad (2)$$

If the phenomenon being studied should not present event dependence, the previous models could be "simplified" giving place to models as a function of common baseline hazard, also called Andersen-Gill Models,[12] that assign the same baseline hazard independent of the episodes that have already been experienced:

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta} \quad (3)$$

where $\lambda_0(t)$ is the common baseline hazard.

Individual heterogeneity

Other aspects to be considered are the unmeasured effects produced by between-subject variability, presumably due to unobserved exposures. This phenomenon is called individual heterogeneity and in practice is analysed adding a frailty to the model, v_i, in other words, an individual random effect to account for this "extra" variability. Since v_i is a multiplicative effect, we can imagine that it represents the cumulative effect of one or more omitted covariates.[13] The most commonly-adopted frailty term has $E[v_i]=1$ and $V[v_i]=\theta$.[14–16] In this context, the models specified in equations (1), (2) and (3) receive the names of "Conditional Frailty Model - Counting Process" (4), "Conditional Frailty Model - Gap Time" (5) and "Shared Frailty Model" (6):

$$\lambda_{ik}(t) = \upsilon_i\lambda_{0k}(t)e^{X_i\beta} \quad (4)$$

$$\lambda_{ik}(t) = \upsilon_i\lambda_{0k}(t-t_{k-1})e^{X_i\beta} \quad (5)$$

$$\lambda_i(t) = \upsilon_i\lambda_0(t)e^{X_i\beta} \quad (6)$$

The problem of being unaware of the previous history of some of the individuals

Specific baseline hazard methods, either with or without frailties, can be applied when all the required information is known, in particular, the number of prior episodes that each individual has had. In practice, however, this information is not always available, which prevents from including basic information to consider the event dependence.

See Figure 1. This figure represents two subjects; on the top part, according to the counting process formulation, and on the bottom part, according to gap time. Notice that finally the difference between both formulations is that gap time "restarts" time to risk at *t=0* every time an episode is produced. At the moment of starting our study, the first subject (id=1) had a considerable amount of time at risk of the event of interest and had had two episodes (this is, then, a left-censored observation). However, this information is not known, we only know that as of the moment he/she is effectively followed, he/she has two other episodes. The second one starts its exposure effectively on *t=0*, presents an episode on *t=5* and stops being followed on *t=7.* The data tables to the right show the data that would be analysed. Notice that both in the counting process as in the gap time, the previous history of individual 1 "disappears" and it would seem to be that he or she barely has two episodes and starts to be at risk exactly at the same instant as individual 2.

(Figure 1)

Thus, we find ourselves before two problems: first, if there is event dependence, and we don´t stratify by number of episode, we mix individuals that at a same instant have different baseline hazards (for example, when id=2 starts to be at risk at *t=0* for the first episode, id=1 is already at risk of the third); second, we also mix temporal scales. On one side there will be subjects, such as id=2, whose scale is follow-up time that at the same time corresponds with time at risk of the event, but on the other side there will be subjects, for example id=1, whose follow-up time does not correspond at all to his/her time at risk. As a consequence, what occurs is two subpopulations are mixed, whose baseline hazard is not the same, going against the assumptions of the Cox model and its extensions for recurrent phenomenon.

Because of the lack of knowledge of the previous history of some of the subjects, and because of the impossibility of using models of specific baseline hazard, the alternative usually is ignoring that history and adjusting models based on common baseline hazard (equations 3 and 6). However, these models assign the same baseline hazard to all episodes and, thus, do not consider the possible effect of the number of episodes that have already occurred, and it doesn't consider that it is comparing two

individuals at the same instant, with times to risk that can be radically different. In fact, the use of models with a common baseline hazard for the analysis of recurrent phenomena with event dependence has been shown to be highly inefficient, generating high levels of bias in the estimate of parameters and coverages of confidence intervals, very much under what was expected, even if the event dependence is of small intensity.[17] All of this shows the need to explore analytical alternatives that may work acceptably in this context.

Objectives

This study has two objectives: first, describe an analysis proposal for recurrent phenomena in the presence of event dependence when there are subjects whose previous history is unknown; and second, compare the performance of our proposal with that of the model that ignores event dependence.

Proposal

Our proposal starts from the assumption that, even though the previous history of all or some of the individuals is unknown, we do know which of these were at risk prior to the beginning of the follow-up and starting when, and that is based fundamentally on three considerations: 1) impute *k*, the number of previous episodes for those subjects at risk before the beginning of the follow-up; 2) treat the subpopulation of subjects "Previously at risk" separately from those "Not previously at risk", and 3) use a frailty term basically to capture the error that will be made when imputing *k.* Concretely, in the two formulations, "Counting process" (7) and "Gap time" (8), the ones we call "Specific Hazard Frailty Model Imputed", in its versions - Counting Process (SHFMI.CP) and Gap Time (SHFMI.GT) :

$$\lambda_{ikr}(t) = \upsilon_i \lambda_{0kr}(t) e^{X_i \beta} \quad (7)$$

$$\lambda_{ikr}(t) = \upsilon_i \lambda_{0kr}(t - t_{k-1}) e^{X_i \beta} \quad (8)$$

where *k* will be the number of previous episodes of individual *i* in case they are known or their imputed value in case they are not known; *r* indicates the subpopulation the individual belongs to: "Previously at risk" or "Not previously at risk". In both cases, information corresponding to time to risk prior to *t=0*

for each individual is included in $X_i\beta$, that will be zero for all those that start to be at risk as of the beginning of the follow-up and a value different than zero for those that were previously at risk. In practice, this proposal means that we stratify by the interaction between having been at risk or not before the beginning of the follow-up, and the number of previous episodes.

Therefore, the use of the term individual random error υ_i intends to capture the error that will be made when imputing, as well as the effect of any variable that, having a non-nil effect, would not have been considered in the analysis. Stratifying by the number of prior episodes intends to safeguard the event dependence, and doing it as an interaction with the fact that it is an individual previously at risk, or not, separates the two subpopulations so as to not mix times, that are not comparable, on the same scale.

The imputation of the number of previous events in individuals at risk before the beginning of the follow-up is done through the COMPoisson generalized distribution, that allows adjusting a regression model using the Conway-Maxwell Poisson distribution considering the dispersion of the data (sub, equi or overdispersion).[18] This imputation is carried out through multiple imputation calculating its parameters directly from the data observed.[19]

Simulation study

Six populations were simulated through the **R** package survsim,[20] all corresponding to previously described cohorts of workers.[17,21] The first three simulate the history of the workers depending on the occurrence of the long term sick leave (SL) with several intensities of event dependence. Since the simulations were carried out through Weibull distributions with ancillary parameter equal to 1 (that is, exponential distributions) the hazard is supposedly constant within each episode, but different between episodes (in other words, there is event dependence). For the fourth, fifth, and sixth populations, the worker cohorts are equally simulated and the outcome is SL according to the diagnosis (respiratory system, musculoskeletal system, and mental and behavioral disorders). In this case, the hazard functions are not constant within the episode. Table 1 shows the characteristics of each episode in each one of these populations. The maximum number of episodes that a subject may suffer was not fixed, although

the baseline hazard was considered constant when $k \geq 3$. X_1, X_2, and X_3 are covariates that represent the exposure, with $X_i \sim Bernoulli(0.5)$, and $\beta_1=0.25$, $\beta_2=0.5$, and $\beta_3=0.75$ being their parameters that represent effects of different magnitudes, set independently of the episode k to which the worker is exposed.

(Table 1)

Four different situations were simulated for each population, that are combinations between two possible follow-up times (2 and 5 years) and two maximum times at risk prior to the beginning of the cohort (2 and 10 years). In each case, samples of sizes 250, 500 and 1000 were simulated with different proportions of subjects at risk prior to the beginning of the cohort (0.1, 0.3, 0.5 and 1).

Our proposal was compared to a model with frailty and common baseline hazard in terms of the number of previous events, but different as per subpopulation (to previous risk or not). We could call this model "Common Hazard Frailty Model with stratification by subpopulation" (CHFM.strata), in other words, a model that does not take event dependence into account, as the one expressed in equation (6), but that separates individuals according to whether they have been previously at risk, or not:

$$\lambda_{ir}(t) = \upsilon_i \lambda_{0r}(t) e^{X_i \beta} \quad (9)$$

Results

The results presented are the ones that refer to n=1000 and follow-up of 5 years, since the observed differences for n=250 and n=500, as well as for a follow-up of 2 years are not considered very relevant (see supplementary material).

With regards to the bias, Figure 2 highlights that the CHFM.strata model only obtains values under 10% in population 1 and in some case when the percentage of individuals at previous risk is 100%. Models SHFMI.CP and SHFMI.GT generally obtain biases under 10% in most situations, except for SHFMI.GT in populations 4 and 5, which is slightly above, and when the percentage of individuals at previous risk

at t=0 is 100%, where SHFMI.CP seems to be more sensitive, especially in population 3. So, for the first three populations, SHFMI.GT shows equal or less bias than SHFMI.CP, whereas for 4, 5 and 6 it is more the opposite as long as there is at least 50% of the individuals that start their risk during the cohort.

(Figure 2)

The average length of 95%CI of the CHFM.strata model, except in population 1, is the largest of the three models for percentages of individuals at previous risk of up to 30%, always overcoming from there the preciseness of the SHFMI.CP model, as well as in the first three populations to SHFMI.GT, that is the most precise one in the last three populations, figure 3.

(Figure 3)

The coverage of 95%CI for the CHFM.strata model is clearly under 95% in most cases, and is lower still, the higher the event dependence; and, for the first four populations, the percentage of individuals at previous risk prior to t0. The coverages for models SHFMI.CP and SHFMI.GT are generally acceptable, even in some case, excessively conservative (over 95%). The SHFMI.CP model fails in excess for population 3 when there is 100% of the individuals at previous risk because of the high bias found, whereas SHFMI.GT obtains coverages even under 80% when in population 4 the percentage of individuals at previous risk is lower than 100%, figure 4.

(Figure 4)

Final remarks

From the results obtained we can make some considerations and suggestions:

1. Not having availability of the number of previous episodes that an individual has had should not be a justification for the use of models with a common baseline hazard, that on the large majority of

occasions show a higher bias and less coverage than specific baseline hazard models.

2. The first step, indispensable for a correct choice of the method to be used, should be the description of the baseline hazard form of the phenomenon being studied in each one of the episodes.
3. If the phenomenon of interest is generated from a function of constant risk, the best model to choose should be SHFMI.GT, if there are percentages of up to 30-50% of lack of information in terms of the number of previous episodes experienced, model SHFMI.CP would also be more than acceptable.
4. For phenomenon ruled by non-constant hazard functions, model SHFMI.CP should be selected for percentages up to 50% of lack of previous information.
5. If the number of previous episodes experienced is unknown for all the individuals, the most adequate choice seems to be model SHFMI.GT.
6. The implementation of the proposal presented in this article is already working on standard software and ready to be used by any user that could be interested. Specifically, in **R** package `miRecSurv`.[22]

Acknowledgements

D. Moriña acknowledges financial support from the Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and Fundación Santander Universidades.

References

1. Navarro A, Ancizu I. Analyzing the occurrence of falls and its risk factors: Some considerations. *Prev Med (Baltim)* 2009;**48**:298–302.
2. Navarro A, Reis RJ, Martin M. Some alternatives in the statistical analysis of sickness absence. *Am J Ind Med* 2009;**52**.
3. Reis RJ, Utzet M, Rocca PF La, Nedel FB, Martín M, Navarro A. Previous sick leaves as predictor of subsequent ones. *Int Arch Occup Environ Health* 2011;**84**:491–499.
4. Braga JR, Tu J V., Austin PC, Sutradhar R, Ross HJ, Lee DS. Recurrent events analysis for examination of hospitalizations in heart failure: Insights from the Enhanced Feedback for

Effective Cardiac Treatment (EFFECT) trial. *Eur Hear J - Qual Care Clin Outcomes* 2018;**4**:18–26.

5. Vasudevan A, Choi JW, Feghali GA, et al. Event dependence in the analysis of cardiovascular readmissions postpercutaneous coronary intervention. *J Investig Med* 2019;**67**:943–949.
6. Cox DR. Regression models and life table. *J R Stat Soc* 1972;**B34**:187–220.
7. Cox DR. Partial likelihood. *Biometrika* 1975;**62**:269–276.
8. Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Stat Med* 2000;**19**:13–33.
9. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000.
10. Amorim LDAFAF, Cai J. Modelling recurrent events: A tutorial for analysis in epidemiology. *Int J Epidemiol* 2015;**44**:324–333.
11. Prentice R, Williams B, Peterson A. On the regression analysis of multivariate failure time data. *Biometrika* 1981;**68**:373–379.
12. Andersen PK, Gill RD. Cox's regression model counting process: a large sample study. *Ann Stat* 1982;**10**:1100–1120.
13. O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. *Stat Med* 2002;**21**:3219–33.
14. Rondeau V, Commenges D, Joly P. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Anal* 2003;**9**:139–53.
15. Balakrishnan N, Peng Y. Generalized gamma frailty model. *Stat Med* 2006;**25**:2797–816.
16. Govindarajulu US, Lin H, Lunetta KL, D'Agostino RB. Frailty models: Applications to biomedical and genetic studies. *Stat Med* 2011;**30**:2754–64.
17. Navarro A, Casanovas G, Alvarado S, Moriña D. Analyzing recurrent events when the history of previous episodes is unknown or not taken into account: proceed with caution. *Gac Sanit* 2017;**31**:227–234.
18. Shmueli G, Minka TP, Kadane JB, Boatwright P. A Useful Distribution for Fitting Discrete Data: Revival of the COM-Poisson. *Appl Stat* 2005;**54**:127–142.
19. Hernández-Herrera G, Navarro A, Moriña D. Regression-based imputation of explanatory

discrete missing data. 2020; arXiv:2007.15031. Available at: https://arxiv.org/abs/2007.15031.

20. Moriña D, Navarro A. The R package survsim for the simulation of simple and complex survival data. *J Stat Softw* 2014;**59**:1–20.

21. Navarro A, Moriña D, Reis R, Nedel FB, Martin M, Alvarado S. Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses. *Scand J Work Environ Health* 2012;**38**:447–455.

22. Moriña D, Hernández-Herrera G, Navarro A. miRecSurv: Left-Censored Recurrent Events Survival Models; 2020. Available at: https://cran.r-project.org/

Tablas y figuras de la publicación

Tabla 9-1.: Characteristics of the simulated populations

Episode	**Distribution**	β_0	**Ancillary**	**HR**
1	Weibull	8.109	1	1
2	Weibull	7.927	1	1.20
≥ 3	Weibull	7.745	1	1.44
1	Weibull	8.109	1	1
2	Weibull	7.703	1	1.50
≥ 3	Weibull	7.298	1	2.25
1	Weibull	8.109	1	1
2	Weibull	7.193	1	2.50
≥ 3	Weibull	6.276	1	6.25
1	Log-normal	7.195	1.498	1
2	Log-logistic	6.583	0.924	1.77
≥ 3	Weibull	6.678	0.923	2.53
1	Log-logistic	7.974	0.836	1
2	Weibull	7.109	0.758	3.81
≥ 3	Log-normal	5.853	1.989	7.19
1	Log-normal	8.924	1.545	1
2	Log-normal	6.650	2.399	10.13
≥ 3	Log-normal	6.696	2.246	11.19

id	start	stop	event	episode
1	0	3	1	1
1	3	5	1	2
1	5	7	0	3
2	0	5	1	1
2	5	7	0	2

id	start	stop	event	episode
1	0	3	1	1
1	0	2	1	2
1	0	2	0	3
2	0	5	1	1
2	0	2	0	2

Figura 9-1.: Dependencia de evento

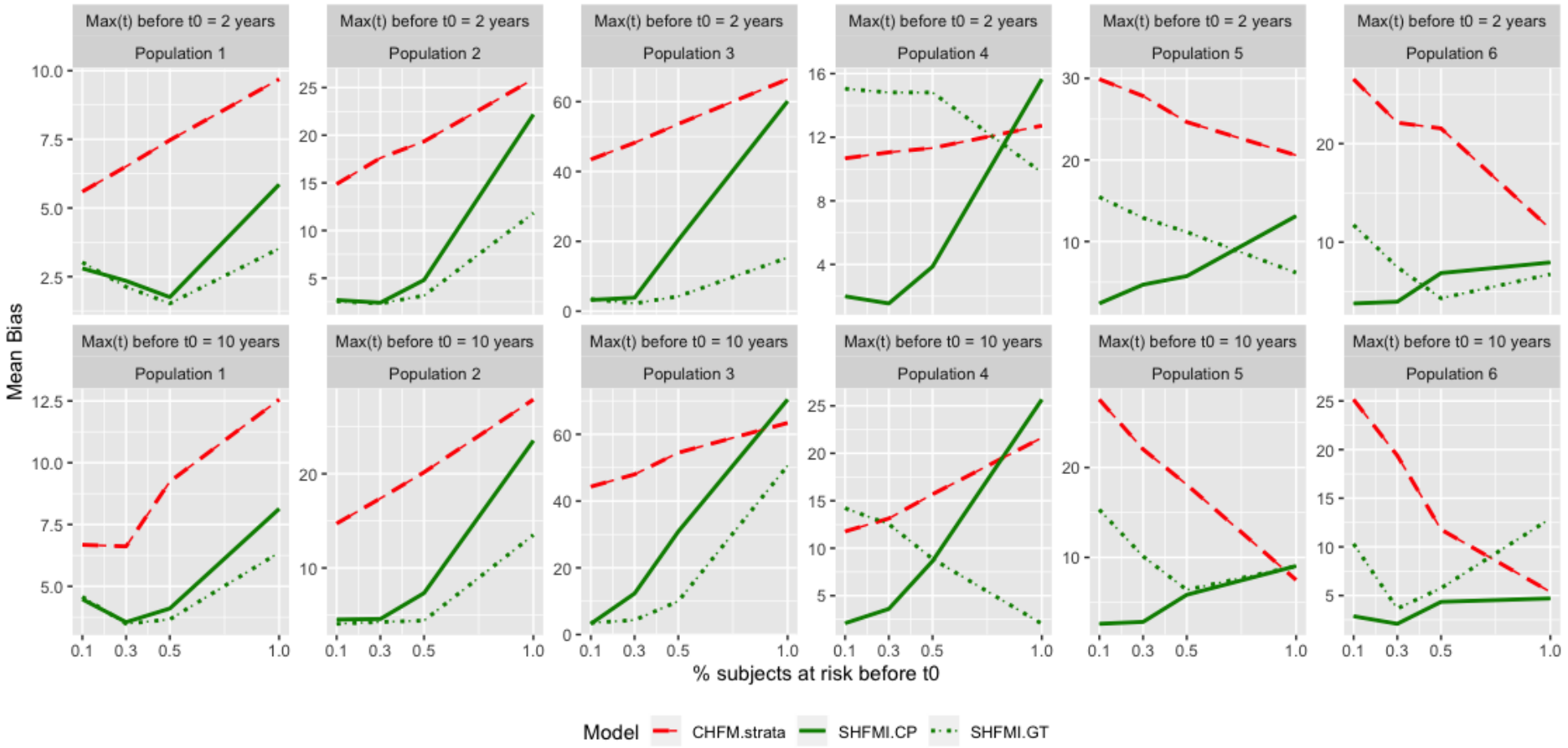


Figura 9-2.: Bias

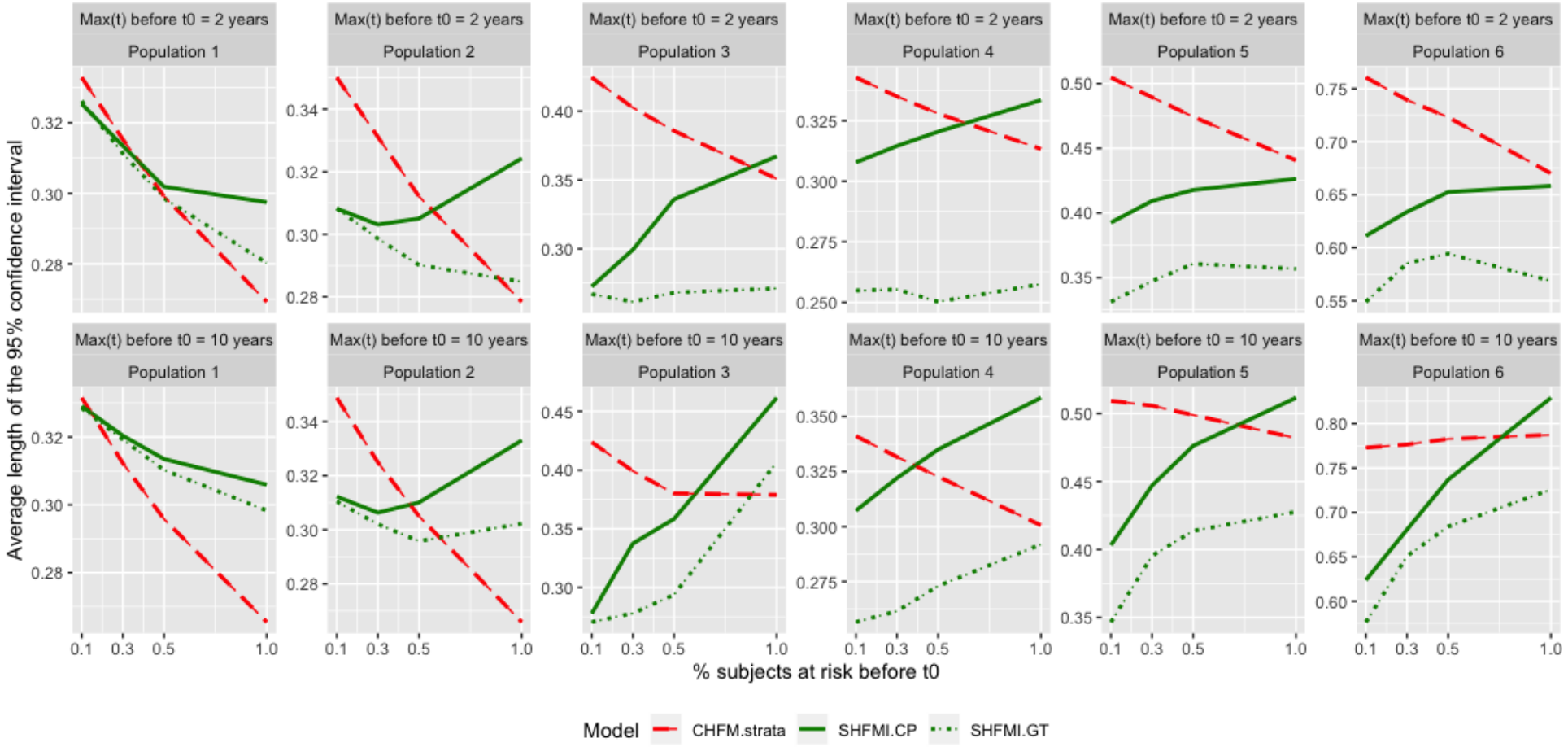


Figura 9-3.: LPI

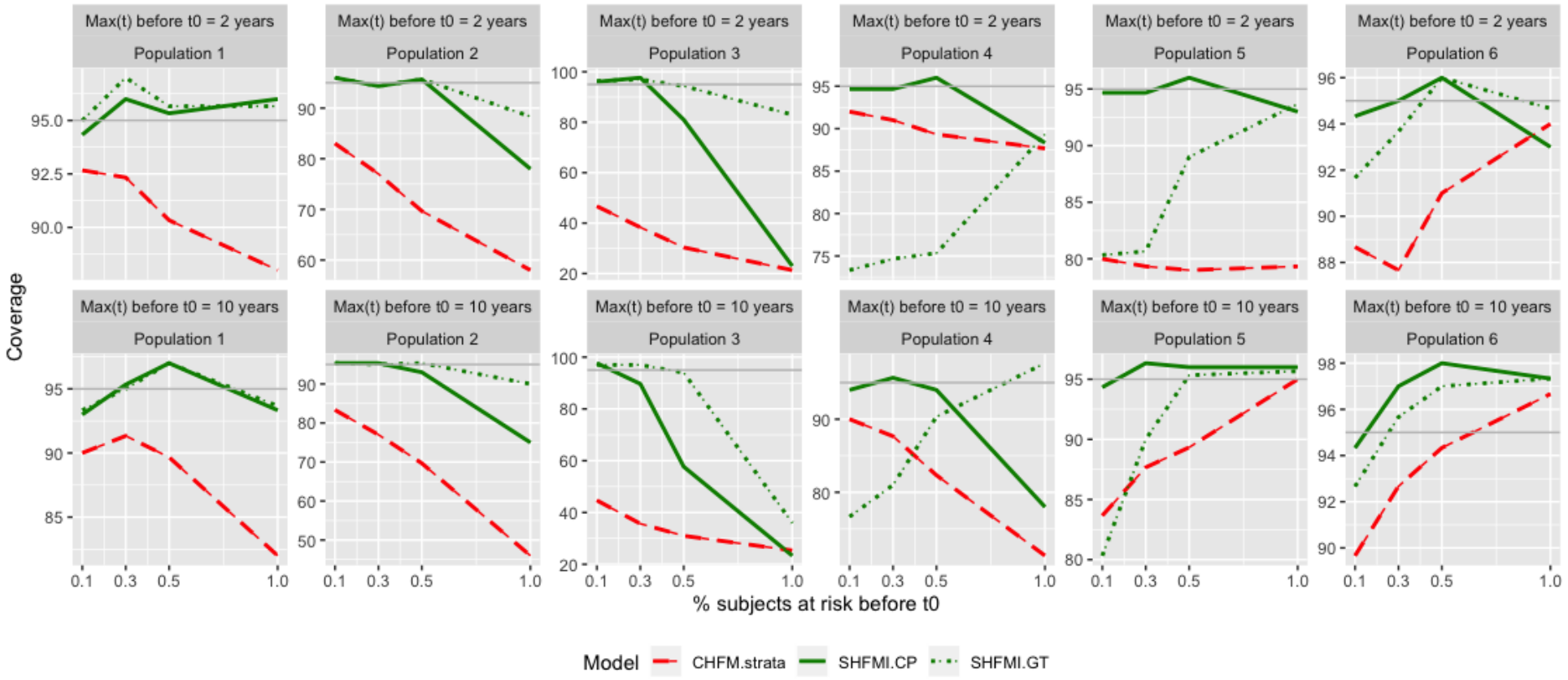


Figura 9-4.: Coverage

10. Paquete miRecSurv: Left-Censored Recurrent Events Survival Models

En la revisión de literatura sobre el uso de modelos de supervivencia para eventos recurrentes, se encontró que la baja frecuencia de uso de estos modelos existentes, puede deberse, entre otras razones, a la falta de herramientas de software para llevar a cabo este análisis. Particularmente para epidemiólogos o profesionales de la salud no estadísticos, es necesario generar este tipo de herramientas de fácil acceso y comprensión que les permitan profundizar en sus análisis de eventos recurrentes teniendo en cuenta la censura a izquierda, que generalmente es omitida.

Un factor determinante para que los investigadores incluyan en los análisis la censura a izquierda, es el manejo de estos datos cuando se desconoce el número de eventos previos, particularmente cuando se requiere el uso de algún método de imputación para una variable de conteo, para la cual no se encuentran reglas establecidas en la literatura que orienten a los investigadores en el método más adecuado para ello, ni tampoco se encuentra software disponible para llevar a cabo las imputaciones y posteriormente los ajustes de los modelos.

El paquete *miRecSurv* desarrollado se basa en el uso de modelos con función de riesgo basal específico, con imputación múltiple del número de episodios anteriores, cuando se desconoce, mediante la distribución COMPoisson, una distribución de conteo muy flexible que puede manejar sobre, sub y equidispersión, con un modelo estratificado en función de si el individuo había estado o no en riesgo anteriormente, y el uso de un término de fragilidad.

La construcción de este paquete se basó en los resultados de las dos simulaciones centrales de este estudio. La primera, para identificar un método adecuado para la imputación de variable discreta en diferentes escenarios de dispersión y con diferentes porcentajes de datos perdidos y la segunda, para identificar un modelo de supervivencia para eventos recurrentes que incluya la depedendencia de evento y la heterogeneidad individual con datos previos faltantes que fueron imputados con el resultado de la primera simulación.

En el repositorio CRAN está disponible la información general del paquete con las funciones que lo conforman y el modo de uso. A continuación se presenta el documento construido para la publicación del paquete *miRecSurv* que fue enviado a revista de software estadístico.

miRecSurv Package: Prentice-Williams-Peterson Models with Multiple Imputation of Unknown Number of Previous Episodes

by David Moriña, Gilma Hernández-Herrera and Albert Navarro

Abstract

Left censoring can occur with relative frequency when analysing recurrent events in epidemiological studies, especially observational ones. Concretely, the inclusion of individuals that were already at risk before the effective initiation in a cohort study, may cause the unawareness of prior episodes that have already been experienced, and this will easily lead to biased and inefficient estimates. The *miRecSurv* package is based on the use of models with specific baseline hazard, with multiple imputation of the number of prior episodes when unknown by means of the COMPoisson distribution, a very flexible count distribution that can handle over-, sub- and equidispersion, with a stratified model depending on whether the individual had or had not previously been at risk, and the use of a frailty term.

1 Introduction

It is not unusual in cohort epidemiological studies that part (or all) of the participants have experienced the event under study at least once before the beginning of the follow-up. This situation is particularly common in the case of observational designs. Under these circumstances, the prior history and prior time at risk of these individuals can be unknown or estimated on the basis of self recall questionnaires, which could lead to modelling issues when the baseline hazard of suffering the event is time-dependent. If the event of interest can only occur once and it occurred for an individual before the start of the follow-up, the result for this individual is fixed regardless of the duration of the follow-up, and therefore we are in the well known and well studied situation of left censoring of a binary variable, for which specific modelling techniques are available. On the other hand, if the event of interest can be suffered several times by the same individual (it is recurrent) and the number of events suffered by the individuals in the cohort before the beginning of the follow-up is unknown we face a left censoring situation with a discrete censored variable, that can define different baseline hazards depending on the episode an individual is at risk of.

This paper introduces the *miRecSurv*, useful when the prior history is unknown for all, or some, of the individuals included in a cohort and the outcome of interest is a recurrent event with event dependence. Specifically, we suppose that we know the moment from where all individuals are at risk, but the number of episodes experienced by the individuals in this time is unknown. This is a realistic situation in practice; for example, it is very probable that in a work force cohort we know when a worker started to work (thus, to be at risk of having a sick leave) and, however, and especially for people with ample trajectory, that we don't know whether or not in effect they have already had sick leaves (and in this case, how many). Another situation that might deal with this issue is a study of cohorts with an outcome of incidence of infection from human papillomavirus on adult women. It would be relatively simple to know how long they have been at risk (beginning of active sexual life) or to make a reasonable assumption, however, we will not be able to know the number of infections since most of the time, when they occur, they are asymptomatic [2].

2 Methods

2.1 Theoretical approach

Our proposal starts from the assumption that, even though the previous history of all or some of the individuals is unknown, we do know which of these were at risk prior to the beginning of the follow-up and starting when, and that is based fundamentally on three considerations: 1) impute k, the number of previous episodes for those subjects at risk before the beginning of the follow-up; 2) treat the subpopulation of subjects "Previously at risk" separately from those "Not previously at risk", and 3) use a frailty term basically to capture the error that will be made when imputing k. Concretely, in the two formulations, "Counting process" (Eq. 1) and "Gap time" (Eq. 2), the ones we call "Specific Hazard Frailty Model Imputed", in its versions - Counting Process (SHFMI.CP) and Gap Time (SHFMI.GT):

$$\lambda_{ikr}(t) = \nu_i \lambda_{0kr}(t) e^{X_i \beta} \tag{1}$$

$$\lambda_{ikr}(t) = \nu_i \lambda_{0kr}(t - t_{k-1}) e^{X_i \beta} \tag{2}$$

where k will be the number of previous episodes of individual i in case they are known or their imputed value in case they are not known; r indicates the subpopulation the individual belongs to: "Previously at risk" or "Not previously at risk". In both cases, information corresponding to time to risk prior to $t = 0$ for each individual is included in $X_i\beta$, that will be zero for all those that start to be at risk as of the beginning of the follow-up and a value different than zero for those that were previously at risk. In practice, this proposal means that we stratify by the interaction between having been at risk or not before the beginning of the follow-up, and the number of previous episodes.

Therefore, the use of the term individual random error ν_i intends to capture the error that will be made when imputing, as well as the effect of any variable that, having a potential effect, would not have been considered in the analysis. Stratifying by the number of prior episodes intends to safeguard the event dependence, and doing it as an interaction with the fact that it is an individual previously at risk, or not, separates the two subpopulations so as to not mix times, that are not comparable, on the same scale.

The imputation of the number of previous events in individuals at risk before the beginning of the follow-up is done through the COMPoisson generalized distribution, that allows adjusting a regression model using the Conway-Maxwell Poisson (COMPoisson) distribution [8] considering the dispersion of the data (sub, equi or overdispersion). This imputation is carried out through multiple imputation calculating its parameters directly from the observed data in two phases: Firstly, a generalised linear model (GLM) is fitted using the number of episodes observed during the follow-up as response variable, based on the COMPoisson distribution. Imputed values are randomly sampled from this distribution with the parameters obtained in the previous step, including random noise generated from a normal distribution. In order to produce proper estimation of uncertainty, the described methodology is included in a multiple imputation framework, according to the well known Rubin's rules [7] and based on the following steps in a Bayesian context:

1. Fit the COMPoisson count data model and find the posterior mean and variance $\hat{\beta}$ and $V(\hat{\beta})$ of model parameters β.
2. Draw new parameters β^* from $N(\hat{\beta}, V(\hat{\beta}))$.
3. Compute predicted scores p using the parameters obtained in the previous step.
4. Draw imputations from the COMPoisson distribution and scores obtained in the previous step.

The COMPoisson random number generation is based on the `rcom` function included in the archived package *compoisson* [1]. The overperformance of the COMPoisson distribution in this context compared to alternative discrete distributions (Poisson, negative binomial, Hermite and zero-inflated versions) is discussed in [3] by means of a comprehensive simulation study.

2.2 The miRecSurv package

The main function of the *miRecSurv* is `recEvFit`, which allows the user to fit recurrent events survival models. A call to this function might be

```
recEvFit(formula, data, id, prevEp, riskBef, oldInd,
         frailty=FALSE, m=5, seed=NA)
```

The description of these arguments can be summarized as follows:

- `formula`: a formula object, with the response on the left of a $\sim$ operator, and the terms on the right. The response must be a survival object as returned by the `Surv` function.

- `data`: a data.frame in which to interpret the variables named in the formula.
- `id`: subject identifier.
- `prevEp`: known previous episodes.
- `riskBef`: indicator for new individual in the cohort (`riskBef=FALSE`) or subject who was at risk before the start of follow-up (`riskBef=TRUE`).
- `oldInd`: time an individual has been at risk prior to the follow-up. This time can be positive or negative (time origin as the start of follow-up).
- `frailty`: should the model include a frailty term. Defaults to `FALSE`.
- `m`: number of multiple imputations. The default is `m=5`.
- `seed`: an integer that is used as argument by the `set.seed` function for offsetting the random number generator. Default is to leave the random number generator alone.

The output of this function is a list with two elements. The first element is the summary table of the fitted model and the second element of the list is the original `data.frame` with the columns corresponding to the multiple imputed values for the previous episodes. In order to facilitate the interpretation, the summary table is formatted in a very similar way to the `summary` tables of very well-known functions as `coxph`, as can be seen in the next section.

3 Example

To illustrate our proposal we use a simulated data generated with the parametres estimated in a worker cohort, where the outcome is the occurrence of sick leave due to any cause. Table 1 shows the characteristics of each episode in this population, estimated in a cohort study described in [6]. The maximum number of episodes that a subject may suffer was not fixed, although the baseline hazard was considered constant when $k \geq 4$. X_1, X_2 and X_3 are covariates that represent the exposure, with $X_i \sim Bernoulli(0.5)$, $i = 1, 2, 3$, and $\beta_1 = 0.25$, $\beta_2 = 0.5$, and $\beta_3 = 0.75$ being their parameters that represent effects of different magnitudes, set independently of the episode k to which the worker is exposed. All the simulations were conducted using R package *survsim* [4], and all the code to reproduce these analyses is available as supplementary material.

To illustrate the usage of the *miRecSurv* package, the results corresponding to a sample from the first scenario can be obtained by means of

```
library(survsim)
library(miRecSurv)
d.ev    <- c('llogistic','weibull','weibull','weibull')
b0.ev   <- c(5.843, 5.944, 5.782, 5.469)
a.ev    <- c(0.700, 0.797, 0.822, 0.858)
d.cens  <- c('weibull','weibull','weibull','weibull')
b0.cens <- c(7.398, 7.061, 6.947, 6.657)
```

Episode	Distribution	β_0	Ancillary
1	Log-logistic	7.974	0.836
2	Weibull	7.109	0.758
3	Log-normal	5.853	1.989
4	Log-normal	5.495	2.204

Table 1: Characteristics of the simulated population.

```
a.cens  <- c(1.178, 1.246, 1.207, 1.422)
set.seed(1234)
sample1 <- rec.ev.sim(n=1500, foltime=1095,
                      dist.ev=d.ev, anc.ev=a.ev, beta0.ev=b0.ev,
                      dist.cens=d.cens, anc.cens=a.cens, beta0.cens=b0.cens,
                      beta=list(c(-.25,-.25,-.25,-.25), c(-.5,-.5,-.5,-.5), c(-.75,-.75,-.75,-.75
                      x=list(c("bern", .5), c("bern", .5), c("bern", .5)),
                      priskb=.1, max.old=5475)
sample1$old2 <- -sample1$old
sample1$old2[is.na(sample1$old)] <- 0

### Shared frailty
ag_s1 <- coxph(Surv(start2,stop2,status)~as.factor(x)+as.factor(x.1)+as.factor(x.2)+old2+
               strata(as.factor(risk.bef))+frailty(nid), data=sample1)

### Counting process
shfmi.cp_s1 <- recEvFit(Surv(start2, stop2, status)~x+x.1+x.2, data=sample1,
                        id="nid", prevEp = "obs.episode",
                        riskBef = "risk.bef", oldInd = "old", frailty=TRUE, m=5, seed=1234)


### Gap time
shfmi.gt_s1 <- recEvFit(Surv(stop2-start2, status)~x+x.1+x.2, data=sample1,
                        id="nid", prevEp = "obs.episode",
                        riskBef = "risk.bef", oldInd = "old", frailty=TRUE, m=5, seed=1234)
```

The generated cohort including the estimated number of previous events (multiple imputed) can be obtained as

```
head(shfmi.cp_s1[[1]])
  nid real.episode obs.episode       time status    start     stop      time2   start2
1   1            1           1 119.673502      1   0.0000 119.6735 119.673502 124.5052
2   1            2           2 389.637244      1 119.6735 509.3107 389.637244 244.1787
3   1            3           3 244.130315      1 509.3107 753.4411 244.130315 633.8160
4   1            4           4 144.476145      0 753.4411 897.9172 144.476145 877.9463
5   2            1           1 107.943850      1   0.0000 107.9438 107.943850 681.4178
6   2            2           2   6.472291      1 107.9438 114.4161   6.472291 789.3617
      stop2 old risk.bef long z x x.1 x.2 EprevCOMPoissDef1 EprevCOMPoissDef2
1  244.1787   0    FALSE   NA 1 1   0   0                 0                 0
2  633.8160   0    FALSE   NA 1 1   0   0                 1                 1
```

```
3  877.9463   0    FALSE   NA 1 1   0   0                2                2
4 1022.4224   0    FALSE   NA 1 1   0   0                3                3
5  789.3617   0    FALSE   NA 1 1   1   0                0                0
6  795.8340   0    FALSE   NA 1 1   1   0                1                1
  EprevCOMPoissDef3 EprevCOMPoissDef4 EprevCOMPoissDef5
1                 0                 0                 0
2                 1                 1                 1
3                 2                 2                 2
4                 3                 3                 3
5                 0                 0                 0
6                 1                 1                 1
```

The coefficients table can be obtained as

```
shfmi.cp_s1[[2]]
                 coef exp(coef)     se(coef)     Pr(>|z|)
x        2.765129e-01  1.318524 3.245577e-02 1.499301e-17
x.1      4.337681e-01  1.543061 3.426452e-02 2.970463e-39
x.2      7.495489e-01  2.116045 3.727413e-02 8.875830e-95
old     -2.603888e-05  0.999974 5.883904e-05 7.830972e-01
frailty            NA        NA           NA 5.569193e-02
```

Simulated data include four scenarios of $n = 1500$ subjects, with a maximum follow-up time of 3 years and a maximum time at risk prior to the beginning of the cohort of 15 years. First scenario has a 10% of subjects at risk prior to the beginning of the cohort (i.e. we don't know the number of previous episodes in 10% of the subjects), whilst the second, third and fourth sample have 25%, 50% and 100%, respectively. Samples based on these settings were generated 100 times and the estimates were averaged across simulations for each sample.

To compare the results obtained we also estimate the shared frailty model (Eq 3). This model has a common hazard baseline (i.e. doesn't consider the event dependence), and incorporates an individual frailty term.

$$\lambda_{ikr}(t) = \nu_i \lambda_0(t) e^{X_i \beta} \tag{3}$$

Results of fitting the described models in scenario 1 are summarized in the Table 2:

Model	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_3$	$SE(\hat{\beta}_3)$
Shared Frailty	0.292	0.043	0.558	0.043	0.843	0.043
SHFMI.CP	0.244	0.034	0.470	0.036	0.706	0.039
SHFMI.GT	0.218	0.030	0.419	0.031	0.632	0.034

Table 2: Average estimates obtained on scenario 1 (10% of subjects at risk prior to the beginning of the cohort).

Below are presented the results for the second scenario, Table 3:
Results for scenario 3, Table 4:

Model	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_3$	$SE(\hat{\beta}_3)$
Shared Frailty	0.287	0.039	0.554	0.039	0.832	0.040
SHFMI.CP	0.242	0.032	0.472	0.035	0.703	0.040
SHFMI.GT	0.217	0.028	0.425	0.030	0.633	0.034

Table 3: Average estimates obtained on scenario 2 (25% of subjects at risk prior to the beginning of the cohort).

Model	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_3$	$SE(\hat{\beta}_3)$
Shared Frailty	0.277	0.033	0.542	0.034	0.818	0.035
SHFMI.CP	0.243	0.030	0.474	0.036	0.710	0.042
SHFMI.GT	0.217	0.025	0.421	0.029	0.632	0.033

Table 4: Average estimates obtained on scenario 3 (50% of subjects at risk prior to the beginning of the cohort).

Finally, Table 5 shows the estimates when 100% of the subjects are at risk prior to the beginning of the follow-up:

Model	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_3$	$SE(\hat{\beta}_3)$
Shared Frailty	0.262	0.026	0.527	0.027	0.797	0.029
SHFMI.CP	0.248	0.029	0.495	0.040	0.742	0.053
SHFMI.GT	0.208	0.022	0.416	0.027	0.626	0.034

Table 5: Average estimates obtained on scenario 4 (100% of subjects at risk prior to the beginning of the cohort).

The average relative bias of the estimates produced by each method is shown in Figure 1. It can be seen that the estimates produced by the *miRecSurv* are less biased than those based on the common baseline hazard model.

4 Conclusion

Left censoring, when analysing recurrent events, is a situation that can occur with certain frequency in cohort studies. For example, it will occur in every cohort that follows subjects that were already at risk of experiencing the outcome of interest before the beginning of the follow-up. If the recurrent event presents event dependence, not knowing the number of episodes that each individual has had prior to the effective initiation of the follow-up, and analysing this through "classical" methods, is an

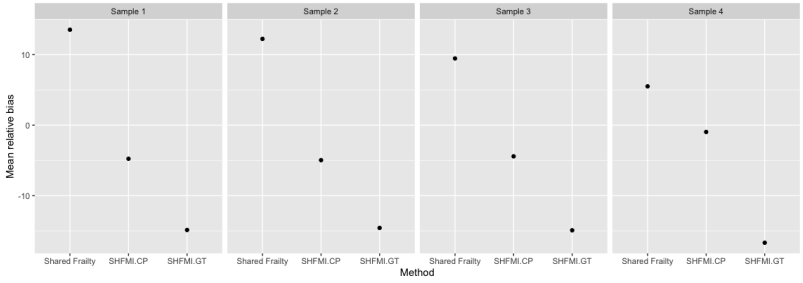


Figure 1: Average relative bias of the estimates produced by the considered methods in each scenario.

important problem, as it will lead to biased and inefficient estimates.

The proposal presented seems to work reasonably well, outperforming the alternative to use a common hazard model. It is important to carry out a comprehensive study to evaluate the performance of the presented proposal. It is also worth to notice that although in this particular example the average relative biases produced by the shared frailty model are relatively low, this is not always the case, as can be seen in a simulation study like [5].

Unavailability of the number of previous episodes that an individual has had should not be a justification for the use of models with a common baseline hazard, that on the large majority of occasions show a higher bias and less coverage than specific baseline hazard models, as shown by the results of this work.

References

[1] Jeffrey Dunn. *compoisson: Conway-Maxwell-Poisson Distribution*, 2012. R package version 0.3.

[2] Amanda Fernández-Fontelo, Alejandra Cabaña, Pedro Puig, and David Moriña. Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, 35(26):4875–4890, nov 2016.

[3] Gilma Hernández-Herrera, Albert Navarro, and David Moriña. Regression-based imputation of explanatory discrete missing data. jul 2020.

[4] David Moriña and Albert Navarro. The R package survsim for the simulation of simple and complex survival data. *Journal of Statistical Software*, 59(2):1–20, 2014.

[5] Albert Navarro, Georgina Casanovas, Sergio Alvarado, and David Moriña. Analyzing recurrent events when the history of previous episodes is unknown or not taken into account: proceed with caution. *Gaceta sanitaria*, 31(3):227–234, 2017.

[6] Albert Navarro, David Moriña, Ricardo Reis, Fúlvio B. Nedel, Miguel Martín, and Sergio Alvarado. Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses. *Scandinavian journal of work, environment & health*, 38(5):447–55, sep 2012.

[7] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley, 1987.

[8] Galit Shmueli, Thomas P. Minka, Joseph B. Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. Technical Report 1, 2005.

11. Discusión

El estudio realizado se ha enfocado principalmente en una propuesta para ayudar a investigadores en el área de la salud, no estadísticos, con métodos que puedan ser aplicados para el análisis de supervivencia en presencia de eventos recurrentes en los que se requiera manejo de datos faltantes.

Si bien cada uno de los artículos sometidos e incluidos en esta monografía tiene su propio apartado de discusión, aquí se resumen los aportes más importantes y se añaden otras de carácter general.

11.1. Sobre la censura a izquierda

En investigación en salud, particularmente en epidemiología y en estudios observacionales de cohorte se presenta frecuentemente el problema de la censura a izquierda, que usualmente es ignorada en los análisis de la supervivencia. Este fenómeno se produce cuando algunos individuos incluidos en la cohorte (o bien todos) han sufrido el evento de interés (o algún episodio del evento de interés) antes del inicio de la observación. La aproximación tradicional a esta problemática ha sido ignorar los casos que presentaban censura a la izquierda y analizar el experimento usando solamente los datos que el investigador recoge a partir del inicio del estudio. Sin embargo, la exclusión de la muestra de los individuos que han sufrido el evento de interés antes del inicio del seguimiento puede llevar a estimadores imprecisos y con notables sesgos. A la vista de los problemas que ignorar el problema de la censura a la izquierda puede producir, es evidente la necesidad de encontrar alternativas para encargarse de este tipo de censura.

En el análisis de supervivencia cuando hay eventos recurrentes, la censura a izquierda juega un papel determinante, ya que la existencia de información previa puede ayudar a definir la dependencia de evento y con esta información determinar el método de análisis más apropiado y preciso. La historia de eventos previos generalmente es desconocida en estudios

epidemiológicos y raras veces se tiene en cuenta para construir modelos de supervivencia. En este caso estamos frente a una situación de censura a la izquierda en la cual la variable censurada es discreta, lo que puede definir diferentes riesgos basales según el episodio del cual el individuo se encuentra a riesgo. Este trabajo muestra como incorporar la información sobre los episodios previos permite obtener estimaciones más precisas y menos sesgadas en un contexto de censura por la izquierda en eventos recurrentes.

11.2. Imputación de datos faltantes cuando la variable es de conteo

En la investigación médica es común que falten parcialmente datos y es necesario afrontar este problema para analizar la información de manera coherente y consistente. Hay muchos métodos disponibles que permiten esto, y el manejo adecuado de la información que falta dependerá de la elección del "mejor" método. Los datos faltantes ocurren en la investigación por una variedad de razones y son motivo de preocupación para el análisis de los datos. En los últimos 20 años se han desarrollado y perfeccionado métodos para el tratamiento de datos faltantes y aún se continua trabajando en ello [Brick y Kalton, 1996].

En general, si no se tiene en cuenta el efecto de los datos faltantes, los resultados de los análisis estadísticos estarán sesgados y la cantidad de variabilidad en los datos no se estimará correctamente, aún así, muchas de las prácticas comunes para el tratamiento de datos faltantes en la investigación en salud siguen siendo, lamentablemente, el uso de eliminación por lista y por pares que son poco recomendadas [Lang y Little, 2018].

Cuando la variable que presenta valores perdidos es una variable de conteo, existen varias formas alternativas de imputar dichos valores perdidos, que dependen de las características de distribución de la variable, particularmente el comportamiento de la media y la varianza. El método más utilizado es asumir que existe equidispersión y que un modelo de Poisson clásico es la mejor alternativa para imputar los valores perdidos, sin embargo, en situaciones reales este supuesto rara vez se cumple y es más frecuente encontrar variables con sobredispersión o infradispersión, o incluso variables de conteo con exceso de ceros para las que el modelo de Poisson no sería el adecuado para llevar a cabo la imputación. En este estudio, se llevó a cabo un subestudio de simulación para identificar la distribución que más se adecuara a la variable de conteo con diferentes porcentajes de datos perdidos y en todos los escenarios de dispersión: equi, infra y sobredispersión, así como cuando se tiene exceso de ceros, que es más frecuente de los que pensamos en la investigación en salud, y usarla en el proceso de imputación múltiple para imputar esta variable que nos interesa. Es de anotar, que en la re-

visión de las distribuciones generalizadas de Poisson, la distribución COMPoisson tiene unas características particulares de flexibilidad para la dispersión, es decir, es una distribución que puede usarse en cualquiera de los escenarios de dispersión y al evaluar los resultados del ajuste de los modelos con las imputaciones, fue la distribución que producía menos sesgos, mayores coberturas y menores longitudes promedio de los intervalos de confianza para los coeficientes de los modelos. Uno de los hallazgos más inesperados en este estudio, fue que el ajuste de modelos utilizando solo los datos disponibles en algunos casos produce un estimador con menos sesgo que realizar la imputación de los valores faltantes , en el caso de una variable de respuesta dicotómica. Aunque se reconoce que ajustar un modelo con solo los datos disponibles puede afectar la potencia del estudio y producir estimaciones imprecisas, es evidente que los efectos sobre la potencia y la precisión pueden ser significativos si el porcentaje de datos faltantes es bajo y el mecanismo de pérdida de los datos es completamente aleatoria (MCAR).

11.3. Sobre los modelos de supervivencia para eventos recurrentes

Cuando se presenta la necesidad de modelar fenómenos recurrentes, los análisis de los datos inician con la pregunta acerca de la existencia de dependencia de ocurrencia y/o heterogeneidad individual; pero la respuesta a estas preguntas no es sencilla y para solucionarla requiere de información detallada de los tiempos en que ocurren los eventos y sí los individuos han presentado o no eventos antes de iniciar el estudio, así como, el número de eventos que han ocurrido a priori. Una estrategia comúnmente empleada en este tipo de estudios es analizar sólo el tiempo hasta el primer evento y considerar éste como representativo de lo que ocurrirá con posterioridad en cada individuo [Glynn y Buring, 1996] y utilizar el modelo de Cox para el análisis, pero la existencia de dependencia de evento y heterogeneidad individual conllevan a la violación del supuesto de independencia en los tiempos entre ocurrencias del modelo de Cox que ha sido el más usado para modelar la supervivencia en investigaciones en salud y por esto, esta estrategia puede llegar a producir resultados sesgados. El análisis de eventos recurrentes parte del supuesto que los investigadores tienen acceso a toda la información requerida por cada modelo. Sin embargo, en la práctica, muchos de estos datos no están disponibles [Amorim y Cai, 2015]. En particular, la información de la historia exhaustiva de cada indviduo generalmente es desconocida, lo que nos deja sin un método para abordar directamente la dependencia de evento y lo que se hace es entonces ajustar modelos con un riesgo de base común que es independiente de los episodios previos en vez de usar modelos con función de base específica que tendría en cuenta la "real naturaleza de los datos". Nuestra propuesta para resolver el problema del desconocimiento de la historia de eventos previos es imputarlos a través de un método "adecuado" que tenga en cuenta que la variable a imputar

es de conteo y que puede caracterizarse por equidispersión, infradispersión, sobredispersión o exceso de ceros, para lo cual usamos modelos de regresión Poisson o generalizaciones de este modelo tales como la distribución COMPoisson, Hermite o Binomial Negativa y a partir de los datos imputados seleccionar el modelo para el ajuste de los datos, teniendo en cuenta además, la heterogeneidad individual que puede ser producida por la variabilidad intra sujeto, presumiblemente debida a una exposición no observada.

Para la elección del método "adecuado" el primer paso recomendado es hacer la descripción de la función de riesgo basal del fenómeno estudiado en cada uno de los episodios y tener en cuenta la verdadera naturaleza de los datos. Si el fenómeno de interés es generado de una función de riesgo constante, el mejor modelo a elegir podría ser un modelo de fragilidad con función de riesgo de base específico con los episodios previos imputados con intervalo de tiempo gap time si se tiene porcentaje de datos faltantes entre 30 y 50 % para el número de episodios previos. Pero para fenómenos con funciones de riesgo no constantes, se debe seleccionar el modelo de fragilidad con función de riesgo de base específico con los episodios previos imputados con intervalos de tiempo counting proccess para porcentajes de hasta el 50 % de datos faltantes de información previa. Ahora, si el número de episodios previos es desconocido para todos los individuos el modelo más adecuado sería modelo de fragilidad con función de riesgo de base específico con los episodios previos imputados con intervalo de tiempo gap time, dado que el modelo con intervalo de tiempo counting proccess en este caso, mostró en los resultados de la simulación, sesgos mayores en las estimaciones de los parámetros y bajas coberturas de los intervalos de confianza (coberturas incluso menores del 80 %). Es importante notar que cuando la dependencia de evento es baja o no hay dependencia de evento, un modelo de fragilidad de base común ajusta necesariamente bien a los datos.

Por otro lado, la poca frecuencia de aplicación de algunos métodos estadísticos para análisis especiales, como el análisis de eventos recurrentes con dependencia de evento y desconocimiento de la historia previa, puede deberse a la falta de software disponible para ello. En este estudio se vio la necesidad de proveer a investigadores e investigadoras de una herramienta que les permita hacer este tipo de análisis y que sea de fácil acceso, para lo cual se desarrolló el paquete *miRecSurv* del R para ponerlo a disposición de los investigadores. En la actualidad, aunque no se cuente con datos reales para identificar el "mejor" modelo, existen estrategias de simulación que cada vez son más usadas, para evaluar el rendimiento de modelos en diferentes escenarios y que pueden ser de ayuda para investigadores que se vean enfrentados a datos de fenómenos recurrentes con características como las que se han simulado en este estudio.

Así pues, el principal aporte de esta tesis es la propuesta de una alternativa de análisis de

fenómenos recurrentes ante la presencia de dependencia de evento y el desconocimiento de los episodios previos para parte de la muestra considerada, así como el desarrollo de una herramienta de software que permita a investigadores replicar estos análisis.

11.4. Limitaciones del estudio

Si bien es cierto que los estudios de simulaciones desarrollados en el marco de este proyecto han permitido comparar el rendimiento de los modelos considerados al ser conocidos los parámetros usados en el proceso de simulación, también es importante contar con datos reales para comprobar el rendimiento de la metodología propuesta en un contexto real. En este estudio, el hecho de que en la literatura están poco descritas las distribuciones que rigen los eventos recurrentes (y la heterogeneidad de las mismas en diferentes contextos, incluso dentro de la Epidemiología o la Salud Pública), y la ausencia de una amplia diversidad de publicaciones que incluyan este tipo de análisis con datos reales limitó los escenarios para las simulaciones y con esto las posibilidades de ampliar el espectro de conocimientos sobre el comportamiento de los fenómenos recurrentes.

12. Conclusiones

- En el estudio de fenómenos recurrentes, y ante la censura a izquierda de la variable "número de episodios previos", es imprescindible encontrar una alternativa analítica a la del ajuste de modelos que no consideran un riesgo basal específico.

- Según los resultados de simulación en este estudio, el modelo COMPoisson se comporta bien para la imputación de variable de conteo ya que es flexible en cuanto al manejo de esta variable con características de sobre y subdispersión, así como con equidispersión.

- La propuesta de modelo realizada en esta tesis parece funcionar razonablemente bien para la mayoría de situaciones estudiadas, en general mucho mejor que la alternativa del uso de modelos con riesgo basal común.

- Para poder elegir con más garantías el modelo de análisis, es imprescindible explorar las funciones que gobiernan el fenómeno de interés, la cantidad de sujetos a riesgo previo al inicio del seguimiento, así como el objetivo concreto de análisis.

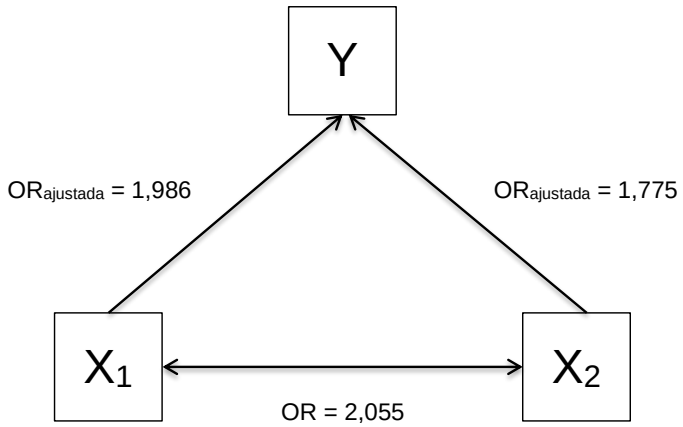
- Si el fenómeno de interés se genera a partir de funciones de riesgo constantes, y cuando todos los sujetos de la muestra han estado a riesgo del evento antes del inicio del seguimiento, el modelo propuesto en su formulación gap time parece ser el más adecuado.

- Si el fenómeno de interés se genera a partir de funciones de riesgo no constantes, el modelo más adecuado es el propuesto con formulación counting process, al menos hasta el 50 % de sujetos a riesgo previo al inicio del seguimiento.

- Para la implementación del modelo propuesto en este estudio se puede utilizar el paquete en R *miRecSurv* desarrollado por los investigadores de esta tesis.

A. Anexo: Imputación

Figura 1S. Parámetros del ejemplo.



Ecuación 1S. Generación de valores ausentes de X_1 según X_2 en el patrón de pérdidas MAR.

$$P(x_1 = \text{valor ausente} | x_2 = 1) = 4{,}5 \times P(x_1 = \text{valor ausente} | x_2 = 0)$$

B. Anexo: Imputación de variable discreta

SUPPLEMENTARY MATERIAL

Accompanying the manuscript:

Handling discrete missing data in longitudinal studies: A simulation approach

Gilma Hernández Herrera
Universidad de Antioquia, Medellín, Colombia. Universidad Autónoma de Barcelona, Barcelona
E-mail: gilma.hernandez@udea.edu.co
Albert Navarro
Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, Barcelona
E-mail: albert.navarro@uab.cat
David Moriña
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Barcelona Graduate School of Mathematics (BGSMath)
E-mail: david.morina@uab.cat

*david.morina@uab.cat

Classical counting models

Poisson regression model

This is the model most widely used in the literature to analyse count data, where the dependent variable corresponds to the number of events which occur per unit of time or space. The Poisson distribution was established based on the limiting case of a binomial distribution when the probability of success is small, in other words, when events are rare and they are assumed to be mutually independent. The fundamental property of the distribution is its equidispersion, from which arise other formulations of models for count data. The probability distribution function for a Poisson random variable **Y** with parameter μ is given by:

$$P(Y = y) = \frac{e^{-\mu}\,\mu^{y}}{y!}$$

where μ > 0 is the mean number of events per unit of time or unit of area.

On the basis of this distribution we construct the Poisson regression model, a Generalised Linear Model. One of the most common applications of Poisson regression models in bio-medical research is the estimation of relative rates or rate ratios, whether crude or adjusted for a series of independent variables, for example in epidemiological studies. Also, in public health the Poisson model has been used to model a wide variety of variables, such as health services demand modeling the number of patients arriving per hour to an emergency department, as a function of covariates which may increase or decrease these numbers, or in occupational health analysing rates of job-related accidents in terms of the characteristics of workers or of the business where employed. In reality, there are many examples of research where the Poisson regression model may be used to estimate the effect of independent variables or factors on a count variable.

However, the Poisson distribution is characterised by being equidisperse (μ=E(Y)= σ^2), where σ^2 is the variance of the distribution, and a problem which frequently arises in this model is that the relationship of μ and σ^2 with the explanatory variables is not an equality, due to a variety of potential factors (unobserved individual heterogeneity, excess of zeros, ...), something which may provoke situations of overdispersion or underdispersion. Overdispersion may be caused by positive correlation between responses or by an excess of variance between response probabilities or counts; this problem can also arise due to violation of assumptions about the distribution of the data. When overdispersion is not taken into account in the models, the result is biased estimators with distorted standard errors, which can lead to erroneous conclusions about the association between response and explanatory variables. Similar consequences result when the problem is of underdispersion.

Negative binomial regression model

It is well known that negative binomial regression is useful for modelling overdisperse count data, in other words, when the conditional variance exceeds the conditional average. This model can be considered a generalisation of Poisson regression, since it has the same

structure, except with an additional parameter to model the overdispersion. If the conditional distribution of the outcome variable is overdisperse, the confidence intervals from negative binomial regression may be more precise, in comparison with those of the Poisson model.

This model is based on the Negative Binomial distribution, obtained from a mixture of a Poisson distribution and a Gamma distribution (Hilbe 2011), and forms part of the exponetial family, allowing modelling using the generalised linear modelling approach. The probability distribution function of a random variable **Y** with this distribution is given by:

$$P(Y = y) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(1/\alpha)\Gamma(y+1)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(1 - \frac{1}{1+\alpha\mu}\right)^{y}$$

Where μ > 0 corresponds to the mean, $(1 + \alpha\mu)$ μ is the variance and α >0 is the dispersion parameter.

Zero-inflated regression models

An excess of zeros occurs when for some reason a count variable has more observations with a value of zero than are to be expected under a Poisson, negative binomial, or other distribution (Lukusa2016). Using a regression model which does not take the excess of zeros into account in this situation can lead to biased estimates of the parameters and to erroneous conclusions.
In general a model with an excess of zeros includes two components: a distribution of counts which may be Poisson, binomial, negative binomial or geometric, and a second component which is the distribution for the zeros. The so-called *zero-inflated models* differ from other count models in the distribution used for the probability function, which is:

$$P(Y = y) = \begin{cases} p + (1-p)f(y;\eta,d) & if \quad y = 0 \\ (1-p)f(y;\eta,d) & if\ y > 0 \end{cases}$$

where *p* is the probability of the excess of zeros, $f(y;\eta,d)$ is the distribution of counts which generally is a distribution with two parameters: η= E(Y) and *d* which is the dispersion parameter (Lukusa 2016)

The inflated zeros are generated by two processes, that which includes the distribution of counts of the component of random zeros, and that for the excess of zeros (structural zeros).

The models used in this case are the zero-inflated Poisson model (ZIP) or the zero-inflated Negative Binomial (ZINB), depending on which distribution is used in the count component of the zero-inflated model (Greene 1994).

Zero-inflated Poisson regression models (ZIP)

The ZIP regression model assumes that observed counts are generated by a mixture of two processes. One process determines the probability of an excess of zeros, i.e. that more observations with value 0 appear in the data set than would be expected under a Poisson or Negative Binomial distribution. A second process models the dependent count variable according to a Poisson distribution. Usually one uses logistic or probit regressions to model the structural zeros, and Poisson or Negative Binomial regression models for the response variable of counts.

Process 1 is a Bernoulli process where z_{ij} is a binary variable which determines whether an excess of zeros is generated for the count variable y_{ij}. Thus, $z_{ij} = 1$ indicates an excess of zeros and $z_{ij} = 0$ the opposite. The model is:

$$logit(\phi_{ij}) = \frac{1}{1 + exp[-(x_{ij}\boldsymbol{\beta})]}$$

where $\phi_{ij} = E(z_{ij}) = P(z_{ij} = 1|x_{ij})$ is the probability of an excess of zeros and $\boldsymbol{\beta}$ are the coefficients of the independent variables (x_{ij}) in the model.

In process 2, given a $z_{ij} = 0$, a y_{ij} is generated on the basis of a Poisson process:

$$log\theta_{ij} = \exp(w_{ij}\alpha)$$

where $\theta = E(y_{ij}^*)$ and y_{ij}^* has a Poisson distribution, denoted by $f(y_{ij}^*)$; and α are the coefficients for each covariable w_{ij} in the model.

The mixture of these two processes results in a ZIP model for the observed counts, y_{ij} The probability function for the ZIP model is:

$$P(Y_i = y_i\ |y_i > 0) = (1 - \phi_{ij})\frac{{\mu_i}^{y_i}e^{-\mu_i}}{y_i!}$$

$$P(Y_i = 0) = \phi_{ij} + (1 - \phi_{ij})e^{-\mu_i}$$

where $f(0)$ is the Poisson probability distribution function for process 2 evaluated in y_{ij}^*; and

$$P(y_{ij} = k) = (1 - \phi_{ij})f(k), \qquad for\ k = 1,2, \ldots$$

It can be shown that the mean and variance of a ZIP model are given by:

$$E(Y_i) = \mu_i(1 - \phi_{ij})$$

$$Var(Y_i) = (1 - \phi_{ij})(\mu_i + \phi_{ij}\mu_i)$$

When the probabiliy $\phi_{ij} = 0$ we obtain the mean and variance of the Poisson regression model. And if $\phi_{ij} > 0$ then the variance is greater than the mean and in this case the excess of zeros causes overdispersion. The values of μ_i and ϕ_{ij} can be modeled in terms of the covariables.

Zero-inflated negative binomial regression model (ZINB)

This model is similar to the previous one, except that in this case the counts follow a negative binomial distribution (Zaninotto 2011).

$$P(y_i = 0) = \phi_{ij} + (1 - \phi_{ij})\left(\frac{k}{\mu_i + k}\right)^k$$

$$P(Y_i = y_i | y_i > 0) = \left(1 - \phi_{ij}\right) f_{NB}(y)$$

where f_{NB} is the negative binomial probability function:

$$f\left(y_{i,}k, \mu_i\right) = \frac{\Gamma(y_i + k)}{\Gamma(k)(y_i + 1)}\left(\frac{k}{\mu_i + k}\right)^k \left(1 - \frac{k}{\mu_i + k}\right)^y$$

\[

1. Simulation results for binary response variable

Table S1. Estimates for the equidispersion scenario β=0.5, n=2000, binary response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,50704	0,04333	0,00704	0,16984	
	lw	0,50697	0,04586	0,00697	0,17975	0,942
	pois	0,50557	0,04645	0,00557	0,18208	0,944
	nb	0,50525	0,04635	0,00525	0,18171	0,945
	hermite	0,48192	0,04618	-0,01808	0,18103	0,859
	comp	0,48227	0,04683	-0,01773	0,18356	0,864
10	fd	0,50632	0,04334	0,00632	0,16991	
	lw	0,50734	0,04914	0,00734	0,19262	0,941
	pois	0,50444	0,05044	0,00444	0,19772	0,957
	nb	0,50418	0,05032	0,00418	0,19726	0,944
	hermite	0,45999	0,04925	-0,04001	0,19305	0,774
	comp	0,46259	0,05154	-0,03741	0,20205	0,801
15	fd	0,50774	0,04340	0,00774	0,17013	
	lw	0,50815	0,05358	0,00815	0,21005	0,946
	pois	0,50301	0,05500	0,00301	0,21560	0,953
	nb	0,50343	0,05528	0,00343	0,21670	0,942
	hermite	0,43984	0,05268	-0,06016	0,20651	0,756
	comp	0,44401	0,05736	-0,05599	0,22484	0,759
20	fd	0,50637	0,04334	0,00637	0,16990	
	lw	0,50953	0,05993	0,00953	0,23491	0,954
	pois	0,50371	0,06201	0,00371	0,24308	0,934
	nb	0,50300	0,06167	0,00300	0,24173	0,931
	hermite	0,42148	0,05827	-0,07852	0,22843	0,748
	comp	0,42515	0,06534	-0,07485	0,25614	0,757
25	fd	0,50613	0,04330	0,00613	0,16972	
	lw	0,50653	0,07012	0,00653	0,27488	0,937
	pois	0,49984	0,07201	-0,00016	0,28227	0,912
	nb	0,50050	0,07359	0,00050	0,28848	0,932
	hermite	0,40015	0,06527	-0,09985	0,25586	0,705
	comp	0,40351	0,07681	-0,09649	0,30110	0,731
30	fd	0,50547	0,04335	0,00547	0,16992	
	lw	0,51033	0,09336	0,01033	0,36599	0,957
	pois	0,50072	0,09693	0,00072	0,37995	0,915
	nb	0,50292	0,09839	0,00292	0,38570	0,917
	hermite	0,38444	0,08348	-0,11556	0,32722	0,726
	comp	0,38498	0,09974	-0,11502	0,39098	0,741

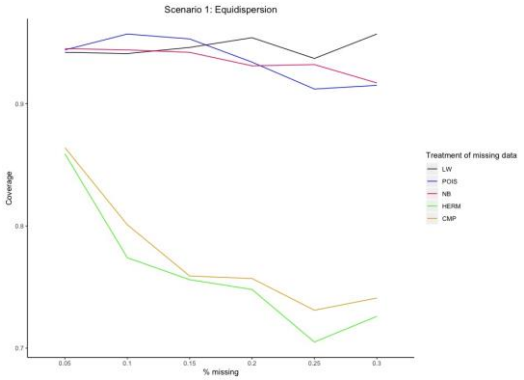


Figure S1. Coverage of confidence intervals for the coefficient in the scenario of equidispersion and binary response variable.

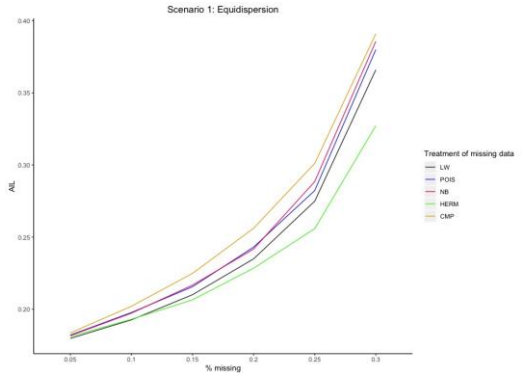


Figure S2. Average length of confidence intervals of the coefficient in the scenario of equidispersion and binary response variable.

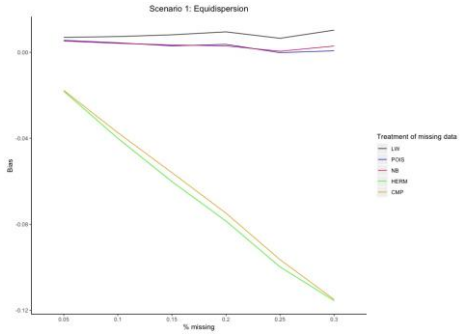


Figure S3. Bias in the coefficient in the scenario of equidispersion and binary response variable.

Table S2. Estimates for the overdispersion scenario β=0.5, n=2000, binary response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,50284	0,03838	0,00284	0,15044	
	lw	0,50301	0,04056	0,00301	0,15900	0,948
	pois	0,50977	0,04040	0,00977	0,15838	0,944

	nb	0,50317	0,04095	0,00317	0,16051	0,948
	hermite	0,47387	0,04045	-0,02613	0,15857	0,811
	comp	0,47374	0,04409	-0,02626	0,17285	0,849
	zpois	0,50789	0,04053	0,00789	0,15887	0,95
	znb	0,50457	0,04106	0,00457	0,16096	0,955
10	fd	0,50282	0,03838	0,00282	0,15044	
	lw	0,50388	0,04334	0,00388	0,16989	0,956
	pois	0,51789	0,04275	0,01789	0,16757	0,933
	nb	0,50426	0,04404	0,00426	0,17262	0,951
	hermite	0,44485	0,04222	-0,05515	0,16549	0,73
	comp	0,44763	0,05257	-0,05237	0,20606	0,785
	zpois	0,51381	0,04339	0,01381	0,17008	0,944
	znb	0,50745	0,04496	0,00745	0,17625	0,952
15	fd	0,50077	0,03830	0,00077	0,15013	
	lw	0,50155	0,04683	0,00155	0,18356	0,956
	pois	0,52199	0,04574	0,02199	0,17929	0,907
	nb	0,50226	0,04787	0,00226	0,18767	0,945
	hermite	0,42477	0,04455	-0,07523	0,17464	0,719
	comp	0,43117	0,05847	-0,06883	0,22922	0,771
	zpois	0,51695	0,04644	0,01695	0,18205	0,933
	znb	0,50531	0,05002	0,00531	0,19607	0,956
20	fd	0,50035	0,03829	0,00035	0,15010	
	lw	0,50144	0,05183	0,00144	0,20318	0,954
	pois	0,52989	0,05084	0,02989	0,19930	0,888
	nb	0,50163	0,05291	0,00163	0,20739	0,926
	hermite	0,40049	0,04776	-0,09951	0,18721	0,676
	comp	0,40918	0,06683	-0,09082	0,26198	0,751
	zpois	0,52250	0,05124	0,02250	0,20085	0,914
	znb	0,50793	0,05615	0,00793	0,22011	0,937
25	fd	0,50100	0,03834	0,00100	0,15030	
	lw	0,50187	0,05969	0,00187	0,23398	0,954
	pois	0,53863	0,05821	0,03863	0,22819	0,873
	nb	0,50244	0,06038	0,00244	0,23670	0,928
	hermite	0,37777	0,05198	-0,12223	0,20374	0,635
	comp	0,39188	0,07692	-0,10812	0,30154	0,729
	zpois	0,52850	0,05832	0,02850	0,22861	0,896
	znb	0,51130	0,06586	0,01130	0,25818	0,947
30	fd	0,50153	0,03832	0,00153	0,15020	
	lw	0,50846	0,07413	0,00846	0,29060	0,958
	pois	0,55166	0,07083	0,05166	0,27764	0,856
	nb	0,51112	0,07674	0,01112	0,30082	0,918
	hermite	0,36595	0,06062	-0,13405	0,23763	0,661
	comp	0,38581	0,08825	-0,11419	0,34595	0,742
	zpois	0,54039	0,07115	0,04039	0,27892	0,872
	znb	0,51860	0,08175	0,01860	0,32045	0,946

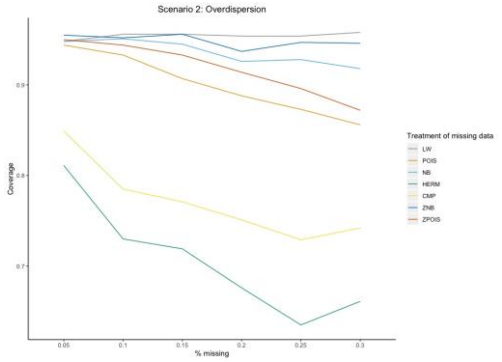


Figure S4. Coverage of confidence intervals for the coefficient in the scenario of overdispersion and binary response variable.

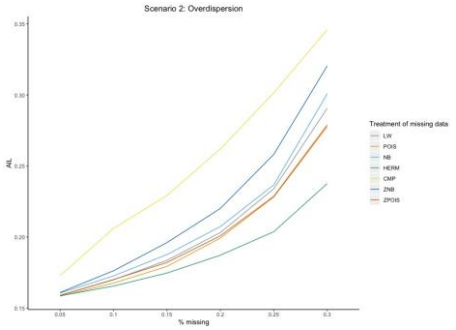


Figure S5. Average length of confidence intervals of the coefficient in the scenario of overdispersion and binary response variable.

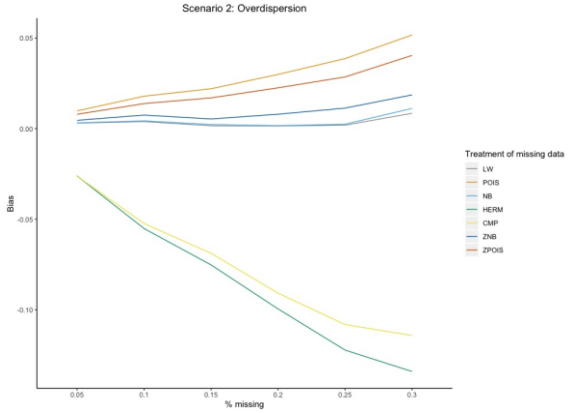


Figure S6. Bias in the coefficient in the scenario of overdispersion and binary response variable.

Table S3. Estimates for the underdispersion scenario β=0.5, n=2000, binary response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,50880	0,05957	0,00880	0,23352	
	lw	0,50873	0,06301	0,00873	0,24700	0,96
	pois	0,48148	0,06652	-0,01852	0,26074	0,959
	nb	0,48053	0,06695	-0,01947	0,26244	0,956
	comp	0,48447	0,06956	-0,01553	0,27267	0,948

10	fd	0,50723	0,05959	0,00723	0,23358	
	lw	0,50666	0,06746	0,00666	0,26445	0,954
	pois	0,45504	0,07444	-0,04496	0,29179	0,915
	nb	0,45518	0,07444	-0,04482	0,29181	0,927
	comp	0,45897	0,08776	-0,04103	0,34402	0,919
15	fd	0,50958	0,05964	0,00958	0,23378	
	lw	0,50780	0,07354	0,00780	0,28829	0,948
	pois	0,43513	0,08192	-0,06487	0,32112	0,894
	nb	0,43347	0,08272	-0,06653	0,32428	0,894
	comp	0,43827	0,11199	-0,06173	0,43901	0,902
20	fd	0,50761	0,05961	0,00761	0,23367	
	lw	0,50711	0,08269	0,00711	0,32414	0,954
	pois	0,41490	0,09411	-0,08510	0,36892	0,855
	nb	0,41413	0,09466	-0,08587	0,37108	0,863
	comp	0,41890	0,14320	-0,08110	0,56134	0,904
25	fd	0,50533	0,05956	0,00533	0,23348	
	lw	0,50763	0,09817	0,00763	0,38482	0,95
	pois	0,39817	0,11219	-0,10183	0,43979	0,846
	nb	0,39668	0,11373	-0,10332	0,44584	0,833
	comp	0,38861	0,19348	-0,11139	0,75843	0,897
30	fd	0,50583	0,05957	0,00583	0,23352	
	lw	0,50878	0,13535	0,00878	0,53055	0,942
	pois	0,38962	0,15684	-0,11038	0,61481	0,877
	nb	0,38508	0,15679	-0,11492	0,61462	0,86
	comp	0,37741	0,28447	-0,12259	0,61452	0,934

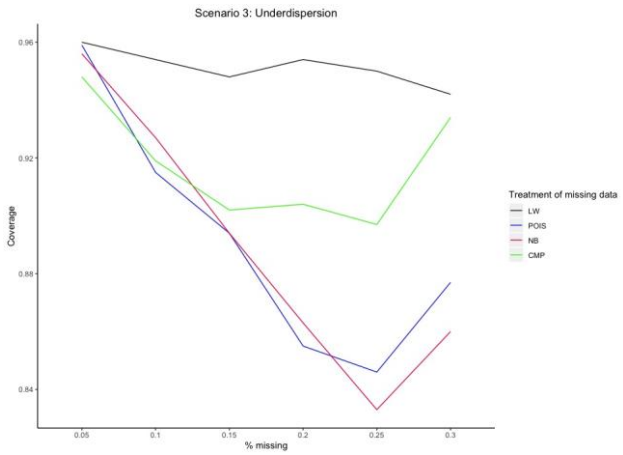


Figure S7. Coverage of confidence intervals for the coefficient in the scenario of underdispersion and binary response variable.

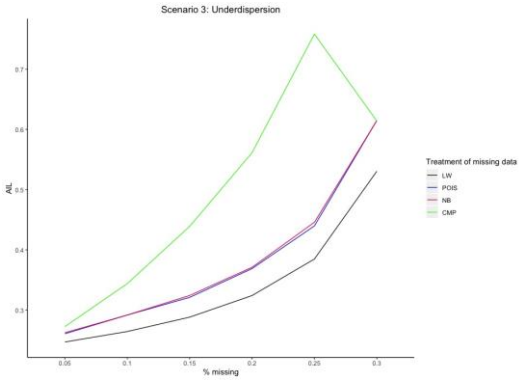


Figure S8. Average length of confidence intervals of the coefficient in the scenario of underdispersion and binary response variable.

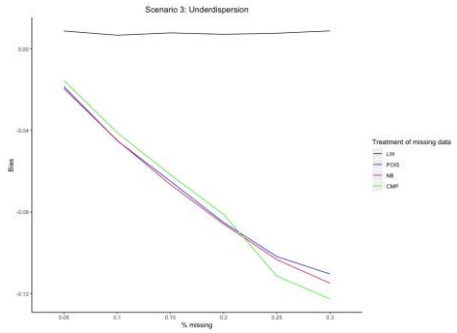


Figure S9. Bias in the coefficient in the scenario of underdispersion and binary response variable.

Table S4. Estimates for the excess zeros scenario β=0.5, n=2000, binary response.

% Missing	Método	$\hat{\beta}$	$se(\hat{\beta})$	Sesgo	LPI	Índice de cobertura
5	fd	0,50119	0,04102	0,00119	0,16081	
	lw	0,50125	0,04322	0,00125	0,16942	0,956
	pois	0,50419	0,04350	0,00419	0,17051	0,949
	nb	0,50292	0,04354	0,00292	0,17069	0,948
	hermite	0,47918	0,04351	-0,02082	0,17055	0,874
	comp	0,47957	0,04422	-0,02043	0,17334	0,887
	zpois	0,50266	0,04356	0,00266	0,17076	0,951
	znb	0,50288	0,04349	0,00288	0,17047	0,951
10	fd	0,50272	0,04104	0,00272	0,16089	
	lw	0,50138	0,04598	0,00138	0,18026	0,962
	pois	0,50783	0,04636	0,00783	0,18172	0,954
	nb	0,50495	0,04672	0,00495	0,18314	0,958
	hermite	0,45767	0,04594	-0,04233	0,18010	0,748
	comp	0,45796	0,04871	-0,04204	0,19093	0,779
	zpois	0,50418	0,04670	0,00418	0,18306	0,955
	znb	0,50422	0,04665	0,00422	0,18286	0,955
15	fd	0,50182	0,04106	0,00182	0,16094	
	lw	0,50315	0,04972	0,00315	0,19492	0,958

	pois	0,51253	0,05010	0,01253	0,19638	0,944
	nb	0,50914	0,05053	0,00914	0,19809	0,944
	hermite	0,44156	0,04917	-0,05844	0,19276	0,752
	comp	0,44277	0,05340	-0,05723	0,20934	0,756
	zpois	0,50762	0,05063	0,00762	0,19848	0,944
	znb	0,50686	0,05069	0,00686	0,19870	0,948
20	fd	0,50296	0,04101	0,00296	0,16078	
	lw	0,50528	0,05473	0,00528	0,21454	0,958
	pois	0,51706	0,05505	0,01706	0,21580	0,933
	nb	0,51226	0,05572	0,01226	0,21844	0,945
	hermite	0,42376	0,05271	-0,07624	0,20664	0,734
	comp	0,42637	0,05946	-0,07363	0,23310	0,736
	zpois	0,51141	0,05584	0,01141	0,21890	0,943
	znb	0,51077	0,05584	0,01077	0,21891	0,945
25	fd	0,50175	0,04102	0,00175	0,16080	
	lw	0,50659	0,06260	0,00659	0,24538	0,947
	pois	0,52351	0,06308	0,02351	0,24726	0,917
	nb	0,51655	0,06423	0,01655	0,25177	0,919
	hermite	0,41334	0,05923	-0,08666	0,23220	0,74
	comp	0,41538	0,06661	-0,08462	0,26112	0,737
	zpois	0,51483	0,06534	0,01483	0,25613	0,921
	znb	0,51462	0,06424	0,01462	0,25183	0,915
30	fd	0,50205	0,04105	0,00205	0,16093	
	lw	0,50565	0,07665	0,00565	0,30048	0,951
	pois	0,52415	0,07905	0,02415	0,30986	0,902
	nb	0,51749	0,07884	0,01749	0,30904	0,916
	hermite	0,39284	0,06768	-0,10716	0,26530	0,714
	comp	0,39607	0,07933	-0,10393	0,31098	0,714
	zpois	0,51492	0,07811	0,01492	0,30618	0,918
	znb	0,51443	0,07902	0,01443	0,30974	0,909

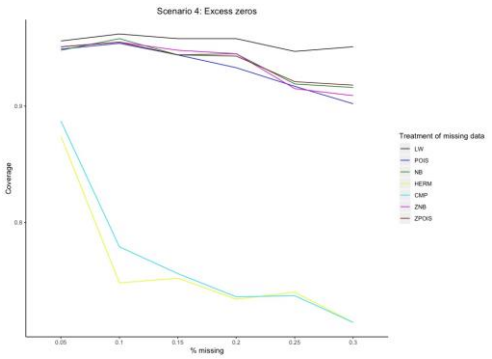


Figure S10. Coverage of confidence intervals for the coefficient in the excess zero scenario and binary response variable.

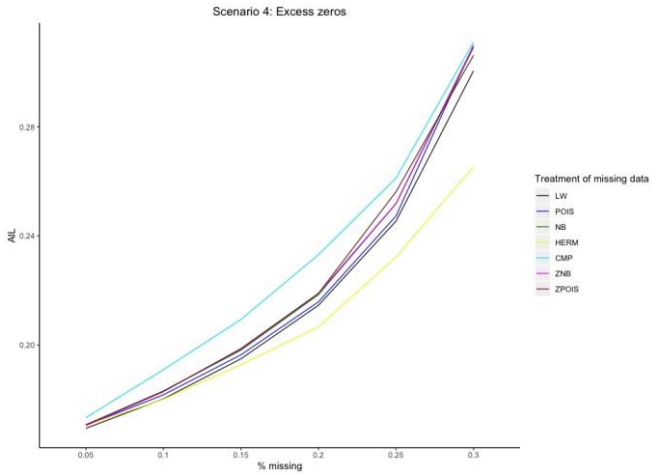


Figure S11. Average length of confidence intervals of the coefficient in the excess zeros scenario and binary response variable.

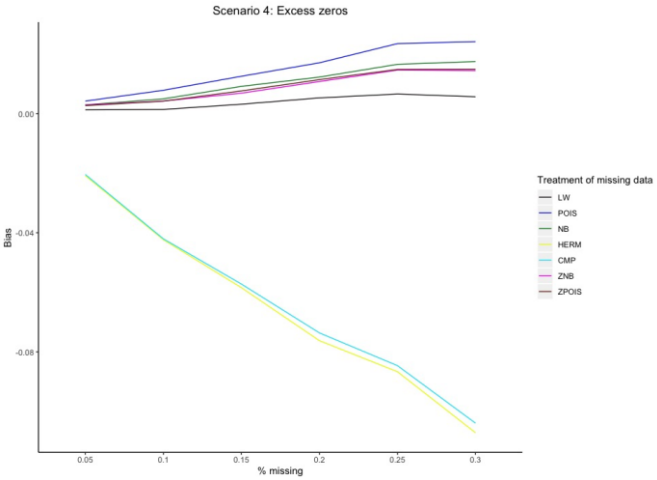


Figure S12. Bias in the coefficient in the excess zeros scenario and binary response variable.

2. Simulation results for discrete response variable

Table S5. Estimates for the equidispersion scenario β=0.5, n=2000, discrete response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,49933	0,02879	-0,00067	0,11287	
	lw	0,49706	0,02931	-0,00294	0,11488	0,953
	pois	0,49133	0,03156	-0,00867	0,12373	0,95
	nb	0,49123	0,03164	-0,00877	0,12404	0,959
	hermite	0,48231	0,03018	-0,01769	0,11829	0,903
	comp	0,48515	0,03048	-0,01485	0,11949	0,922
10	fd	0,50017	0,02881	0,00017	0,11292	
	lw	0,49455	0,02986	-0,00545	0,11706	0,949

	pois	0,48426	0,03341	-0,01574	0,13099	0,955
	nb	0,48446	0,03335	-0,01554	0,13072	0,952
	hermite	0,46959	0,03121	-0,03041	0,12233	0,844
	comp	0,47404	0,03168	-0,02596	0,12420	0,877
15	fd	0,49694	0,02881	-0,00306	0,11295	
	lw	0,48740	0,03046	-0,01260	0,11942	0,941
	pois	0,47477	0,03577	-0,02523	0,14022	0,933
	nb	0,47384	0,03607	-0,02616	0,14141	0,94
	hermite	0,45078	0,03228	-0,04922	0,12652	0,701
	comp	0,45773	0,03310	-0,04227	0,12974	0,771
20	fd	0,49986	0,02892	-0,00014	0,11337	
	lw	0,48724	0,03115	-0,01276	0,12210	0,93
	pois	0,46812	0,03748	-0,03188	0,14690	0,926
	nb	0,46797	0,03796	-0,03203	0,14882	0,919
	hermite	0,43795	0,03328	-0,06205	0,13045	0,631
	comp	0,44678	0,03489	-0,05322	0,13678	0,713
25	fd	0,49955	0,02887	-0,00045	0,11318	
	lw	0,48180	0,03176	-0,01820	0,12448	0,923
	pois	0,46314	0,03854	-0,03686	0,15107	0,91
	nb	0,46347	0,03802	-0,03653	0,14903	0,907
	hermite	0,42588	0,03418	-0,07412	0,13397	0,552
	comp	0,43591	0,03628	-0,06409	0,14223	0,644
30	fd	0,49919	0,02884	-0,00081	0,11303	
	lw	0,47540	0,03244	-0,02460	0,12716	0,906
	pois	0,45311	0,04026	-0,04689	0,15783	0,871
	nb	0,45408	0,03973	-0,04592	0,15573	0,871
	hermite	0,40921	0,03510	-0,09079	0,13761	0,394
	comp	0,42105	0,03776	-0,07895	0,14802	0,526

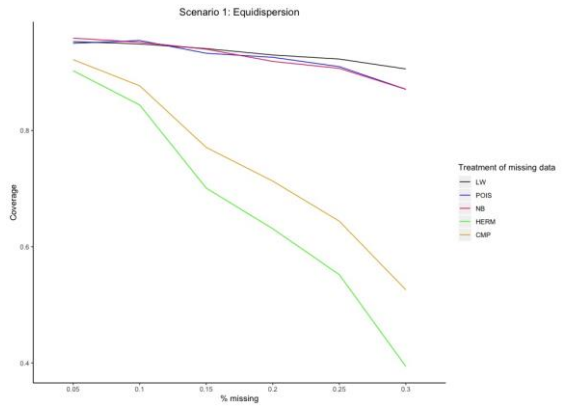


Figure S13. Coverage of confidence intervals for the coefficient in the equidispersion scenario and discrete response variable.

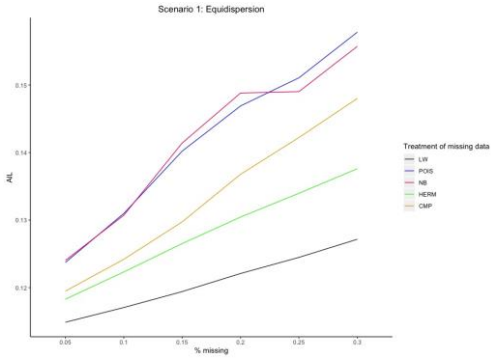


Figure S14. Average length of confidence intervals of the coefficient in the equidispersion scenario and discrete response variable.

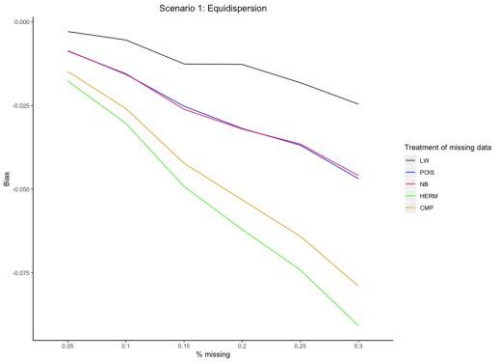


Figure S15. Bias in the coefficient in the equidispersion scenario and discrete response variable.

Table S6. Estimates for the overdispersion scenario β=0.5, n=2000, discrete response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,50182	0,03503	0,00182	0,13732	
	lw	0,49971	0,03567	-0,00029	0,13981	0,955
	pois	0,48563	0,03699	-0,01437	0,14500	0,956
	nb	0,48555	0,03705	-0,01445	0,14524	0,947
	hermite	0,48539	0,03610	-0,01461	0,14151	0,933
	comp	0,50126	0,03495	0,00126	0,13700	0,95
	zpois	0,49586	0,03629	-0,00414	0,14226	0,957
	znb	0,46901	0,03848	-0,03099	0,15085	0,904
10	fd	0,46949	0,03826	-0,03051	0,14997	
	lw	0,46847	0,03715	-0,03153	0,14563	0,876
	pois	0,50166	0,03494	0,00166	0,13698	0,963
	nb	0,49287	0,03699	-0,00713	0,14501	0,957
	hermite	0,45687	0,03905	-0,04313	0,15308	0,836
	comp	0,45660	0,03891	-0,04340	0,15254	0,827

	zpois	0,45256	0,03846	-0,04744	0,15076	0,76
	znb	0,50151	0,03494	0,00151	0,13697	0,945
15	fd	0,48955	0,03776	-0,01045	0,14802	
	lw	0,44324	0,04007	-0,05676	0,15707	0,736
	pois	0,44313	0,03996	-0,05687	0,15663	0,738
	nb	0,43761	0,03935	-0,06239	0,15423	0,652
	hermite	0,50256	0,03496	0,00256	0,13703	0,959
	comp	0,48613	0,03858	-0,01387	0,15122	0,945
	zpois	0,43214	0,04087	-0,06786	0,16019	0,647
	znb	0,43228	0,04091	-0,06772	0,16038	0,665
20	fd	0,42041	0,04107	-0,07959	0,16100	
	lw	0,50299	0,03500	0,00299	0,13721	0,937
	pois	0,48006	0,03952	-0,01994	0,15493	0,928
	nb	0,42029	0,04156	-0,07971	0,16293	0,532
	hermite	0,41974	0,04166	-0,08026	0,16330	0,514
	comp	0,40655	0,04211	-0,09345	0,16509	0,446
	zpois	0,50182	0,03503	0,00182	0,13732	0,954
	znb	0,49971	0,03567	-0,00029	0,13981	0,955
25	fd	0,48563	0,03699	-0,01437	0,14500	
	lw	0,48555	0,03705	-0,01445	0,14524	0,947
	pois	0,48539	0,03610	-0,01461	0,14151	0,933
	nb	0,50126	0,03495	0,00126	0,13700	0,95
	hermite	0,49586	0,03629	-0,00414	0,14226	0,957
	comp	0,46901	0,03848	-0,03099	0,15085	0,904
	zpois	0,46949	0,03826	-0,03051	0,14997	0,914
	znb	0,46847	0,03715	-0,03153	0,14563	0,876
30	fd	0,50166	0,03494	0,00166	0,13698	
	lw	0,49287	0,03699	-0,00713	0,14501	0,957
	pois	0,45687	0,03905	-0,04313	0,15308	0,836
	nb	0,45660	0,03891	-0,04340	0,15254	0,827
	hermite	0,45256	0,03846	-0,04744	0,15076	0,76
	comp	0,50151	0,03494	0,00151	0,13697	0,945
	zpois	0,48955	0,03776	-0,01045	0,14802	0,946
	znb	0,44324	0,04007	-0,05676	0,15707	0,736

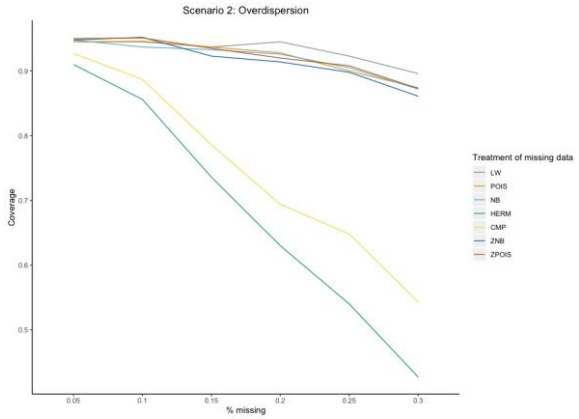


Figure S16. Coverage of confidence intervals for the coefficient in overdispersion and discrete response variable.

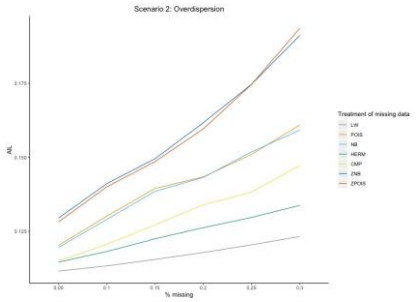


Figure S17. Average length of confidence intervals of the coefficient in the overdispersion scenario and discrete response variable.

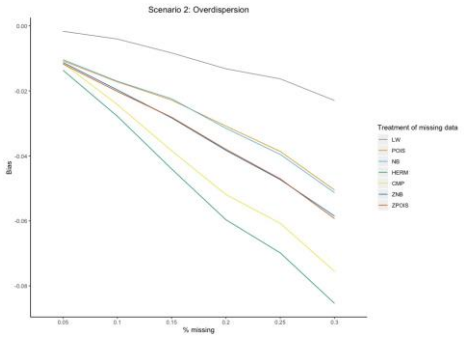


Figure S18. Bias in the coefficient in the overdispersion scenario and discrete response variable.

Table S7. Estimates for the underdispersion scenario β=0.5, n=2000, discrete response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,50182	0,03503	0,00182	0,13732	
	lw	0,49971	0,03567	-0,00029	0,13981	0,955
	pois	0,48563	0,03699	-0,01437	0,14500	0,956
	nb	0,48555	0,03705	-0,01445	0,14524	0,947
	comp	0,48539	0,03610	-0,01461	0,14151	0,933
10	fd	0,50126	0,03495	0,00126	0,13700	
	lw	0,49586	0,03629	-0,00414	0,14226	0,957
	pois	0,46901	0,03848	-0,03099	0,15085	0,904

	nb	0,46949	0,03826	-0,03051	0,14997	0,914
	comp	0,46847	0,03715	-0,03153	0,14563	0,876
15	fd	0,50166	0,03494	0,00166	0,13698	
	lw	0,49287	0,03699	-0,00713	0,14501	0,957
	pois	0,45687	0,03905	-0,04313	0,15308	0,836
	nb	0,45660	0,03891	-0,04340	0,15254	0,827
	comp	0,45256	0,03846	-0,04744	0,15076	0,76
20	fd	0,50151	0,03494	0,00151	0,13697	
	lw	0,48955	0,03776	-0,01045	0,14802	0,946
	pois	0,44324	0,04007	-0,05676	0,15707	0,736
	nb	0,44313	0,03996	-0,05687	0,15663	0,738
	comp	0,43761	0,03935	-0,06239	0,15423	0,652
25	fd	0,50256	0,03496	0,00256	0,13703	
	lw	0,48613	0,03858	-0,01387	0,15122	0,945
	pois	0,43214	0,04087	-0,06786	0,16019	0,647
	nb	0,43228	0,04091	-0,06772	0,16038	0,665
	comp	0,42041	0,04107	-0,07959	0,16100	0,55
30	fd	0,50299	0,03500	0,00299	0,13721	
	lw	0,48006	0,03952	-0,01994	0,15493	0,928
	pois	0,42029	0,04156	-0,07971	0,16293	0,532
	nb	0,41974	0,04166	-0,08026	0,16330	0,514
	comp	0,40655	0,04211	-0,09345	0,16509	0,446

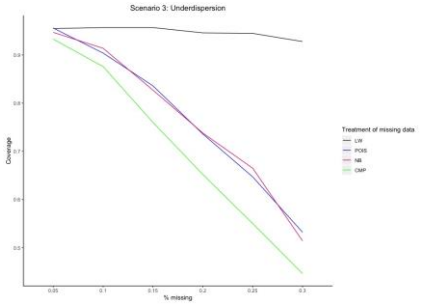


Figure S19. Coverage of confidence intervals for the coefficient in the underdispersion scenario and discrete response variable.

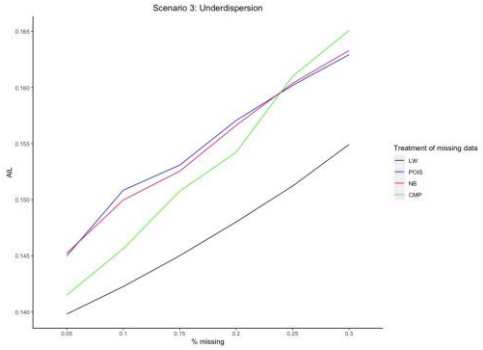


Figure S20. Average length of confidence intervals of the coefficient in the underdispersion scenario and discrete response variable.

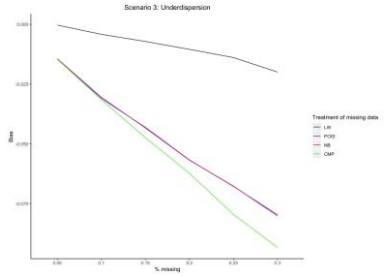


Figure S21. Bias in the coefficient in the underdispersion scenario and discrete response variable.

Table S8. Estimates for the excess zeros scenario β=0.5, n=2000, discrete response.

% Missing	Method	$\hat{\beta}$	$se(\hat{\beta})$	Bias	AIL	Coverage index
5	fd	0,417868	0,056312	-0,082132	0,220741	
	lw	0,439604	0,057893	-0,060396	0,226941	0,674
	pois	0,439207	0,057691	-0,060793	0,226149	0,674
	nb	0,435451	0,057918	-0,064549	0,227039	0,659
	hermite	0,419326	0,058054	-0,080674	0,227573	0,616
	comp	0,420375	0,057978	-0,079625	0,227272	0,617
	zpois	0,435671	0,057822	-0,064329	0,226664	0,674
	znb	0,435284	0,057852	-0,064716	0,226780	0,658
10	fd	0,410762	0,056203	-0,089238	0,220316	
	lw	0,457396	0,059454	-0,042604	0,233061	0,724
	pois	0,456842	0,059245	-0,043158	0,232241	0,722
	nb	0,449187	0,059305	-0,050813	0,232478	0,712
	hermite	0,416632	0,059536	-0,083368	0,233382	0,627
	comp	0,417959	0,059715	-0,082041	0,234084	0,625
	zpois	0,448487	0,059278	-0,051513	0,232368	0,71
	znb	0,448730	0,059322	-0,051270	0,232543	0,712
15	fd	0,414960	0,056164	-0,085040	0,220164	
	lw	0,485387	0,061323	-0,014613	0,240385	0,799
	pois	0,481594	0,060798	-0,018406	0,238326	0,801
	nb	0,470914	0,060924	-0,029086	0,238821	0,788
	hermite	0,421083	0,061082	-0,078917	0,239443	0,661
	comp	0,423798	0,061412	-0,076202	0,240735	0,674
	zpois	0,470732	0,060771	-0,029268	0,238223	0,792
	znb	0,470219	0,060838	-0,029781	0,238483	0,797
20	fd	0,411566	0,056206	-0,088434	0,220328	
	lw	0,514293	0,063303	0,014293	0,248149	0,784
	pois	0,509783	0,062544	0,009783	0,245171	0,793
	nb	0,496761	0,063040	-0,003239	0,247118	0,78
	hermite	0,426925	0,062742	-0,073075	0,245949	0,683
	comp	0,429388	0,063345	-0,070612	0,248311	0,701
	zpois	0,494573	0,062917	-0,005427	0,246633	0,802
	znb	0,494445	0,063019	-0,005555	0,247034	0,803

25	fd	0,413639	0,056172	-0,086361	0,220193	
	lw	0,547039	0,065389	0,047039	0,256326	0,737
	pois	0,540371	0,064442	0,040371	0,252612	0,738
	nb	0,524629	0,064446	0,024629	0,252627	0,763
	hermite	0,437376	0,064386	-0,062624	0,252392	0,715
	comp	0,438931	0,065152	-0,061069	0,255395	0,725
	zpois	0,521102	0,064281	0,021102	0,251983	0,779
	znb	0,520902	0,064687	0,020902	0,253574	0,778
30	fd	0,416089	0,056126	-0,083911	0,220015	
	lw	0,582464	0,067686	0,082464	0,265330	0,667
	pois	0,570962	0,066601	0,070962	0,261076	0,696
	nb	0,556962	0,067046	0,056962	0,262821	0,715
	hermite	0,439203	0,066139	-0,060797	0,259266	0,726
	comp	0,441832	0,067101	-0,058168	0,263034	0,743
	zpois	0,547955	0,066788	0,047955	0,261809	0,747
	znb	0,548860	0,066813	0,048860	0,261907	0,755

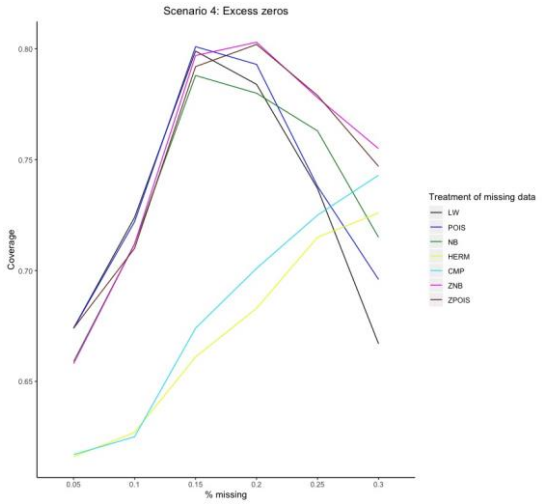


Figure S22. Coverage of confidence intervals for the coefficient in the excess zeros scenario and discrete response variable.

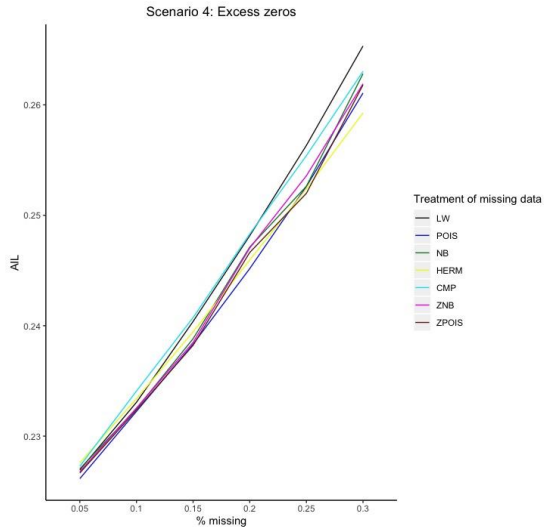


Figure S23. Average length of confidence intervals of the coefficient in the excess zeros scenario and discrete response variable.

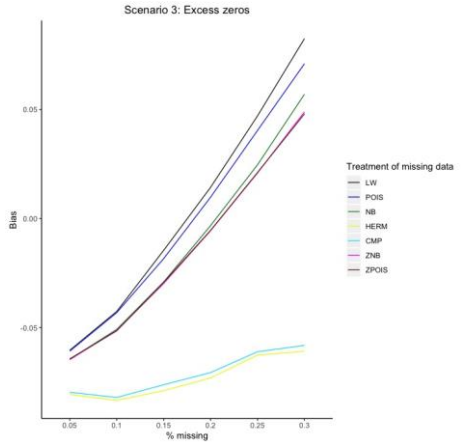


Figure S24. Bias in the coefficient in the excess zeros scenario and discrete response variable.

Referencies

Hilbe, J. (2011). Negative Binomial regression (2nd ed.). New York: Cambridge University Press.

Lukusa, T. Martin, Lee, Shen-Ming and Li, Chin-Shang. (2016, may). Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. Metrika 79(4), 457–483.

Greene, William H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. NYU Working Papers (EC-94-10).

Zaninotto, Paola and Falaschetti, Emanuela. (2011). Comparison of methods for mod- elling a count outcome with excess zeros: application to activities of daily living (adl-s). Journal of Epidemiology & Community Health 65(3), 205–210.

C. Anexo: Propuesta modelos

Supplementary material

Results Graphics for n=500 and n=250

Figure 1s. Bias according to population and time of follow-up, n=500

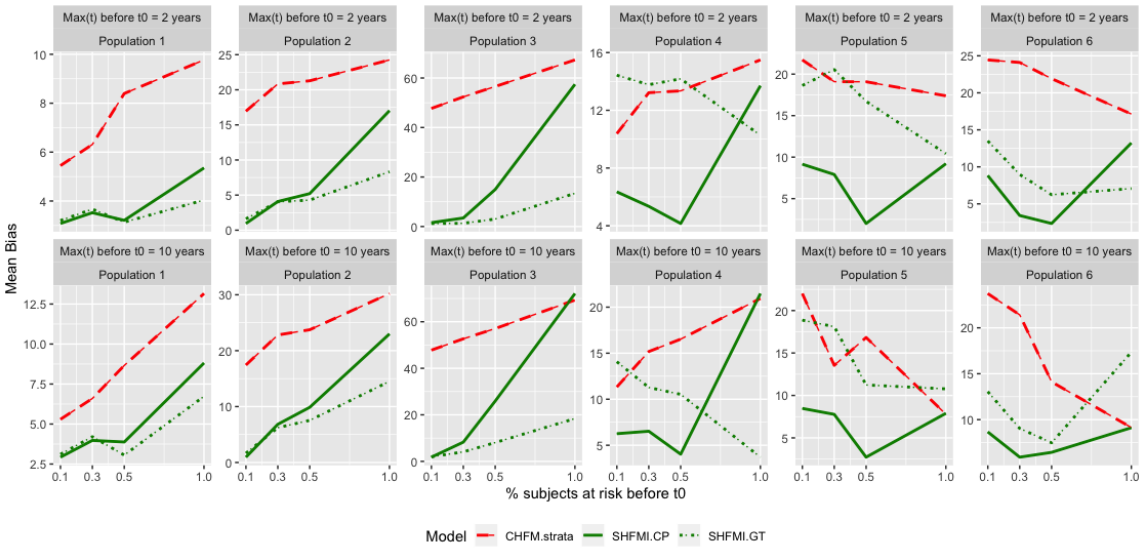


Figure 2s. Average length of the 95% confidence interval according to population and time of follow-up, n=500

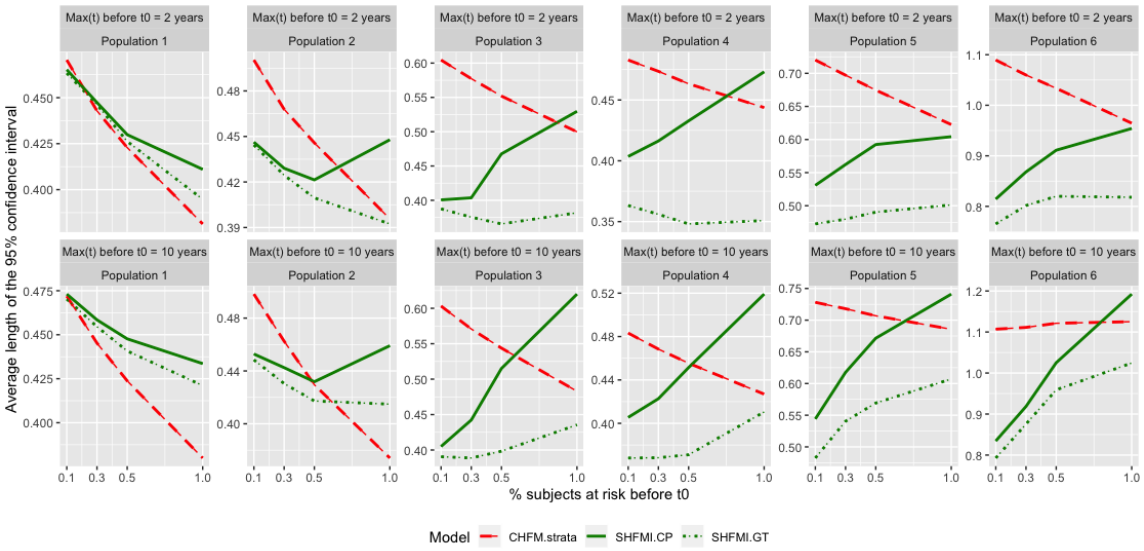


Figure 3s. Coverage of the 95% confidence intervals according to population and time of follow-up, n=500

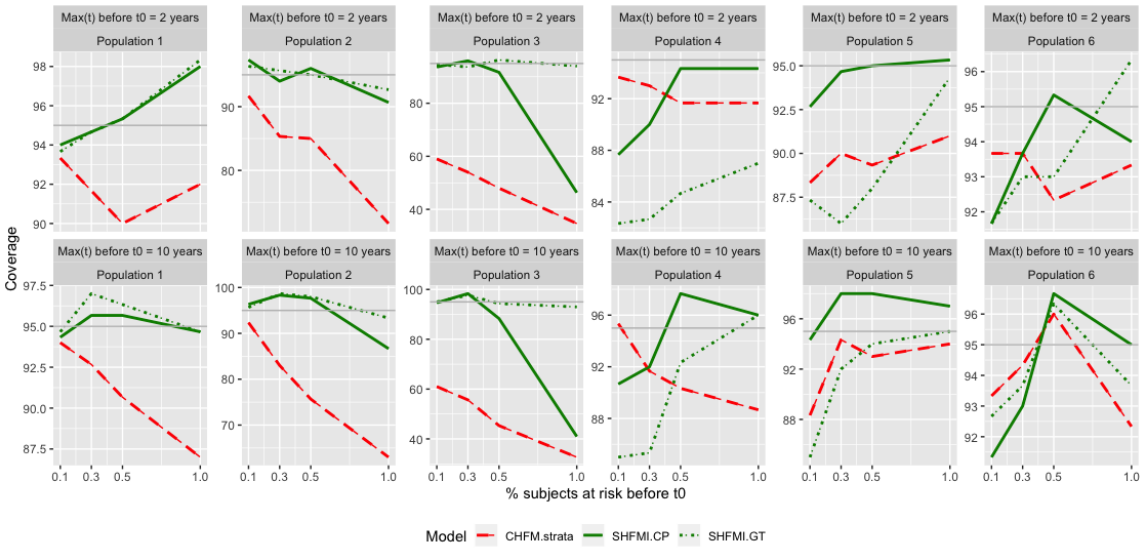


Figure 4s. Bias according to population and time of follow-up, n=250

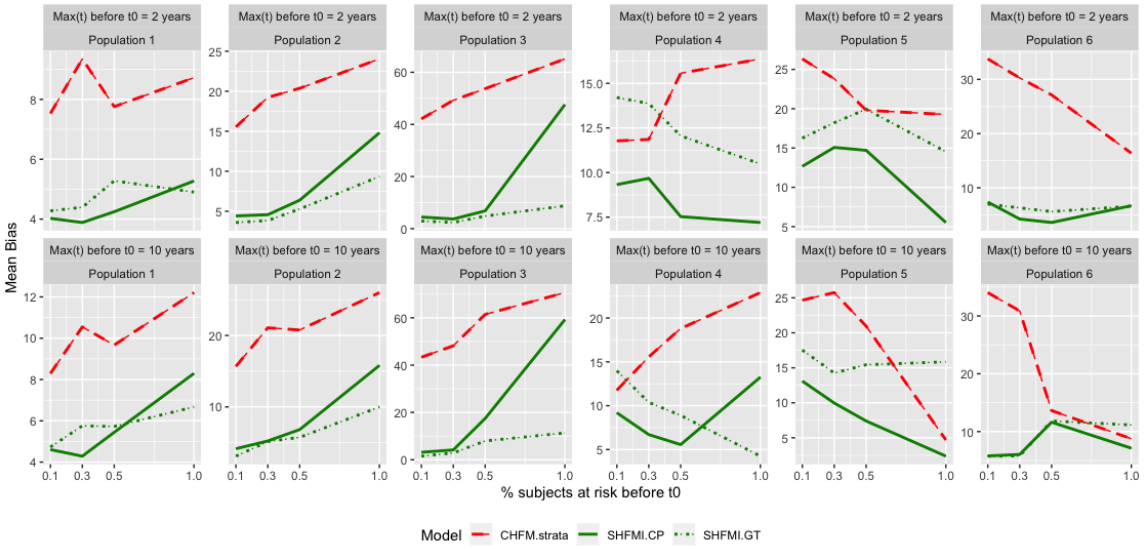


Figure 5s. Average length of the 95% confidence interval according to population and time of follow-up, n=250

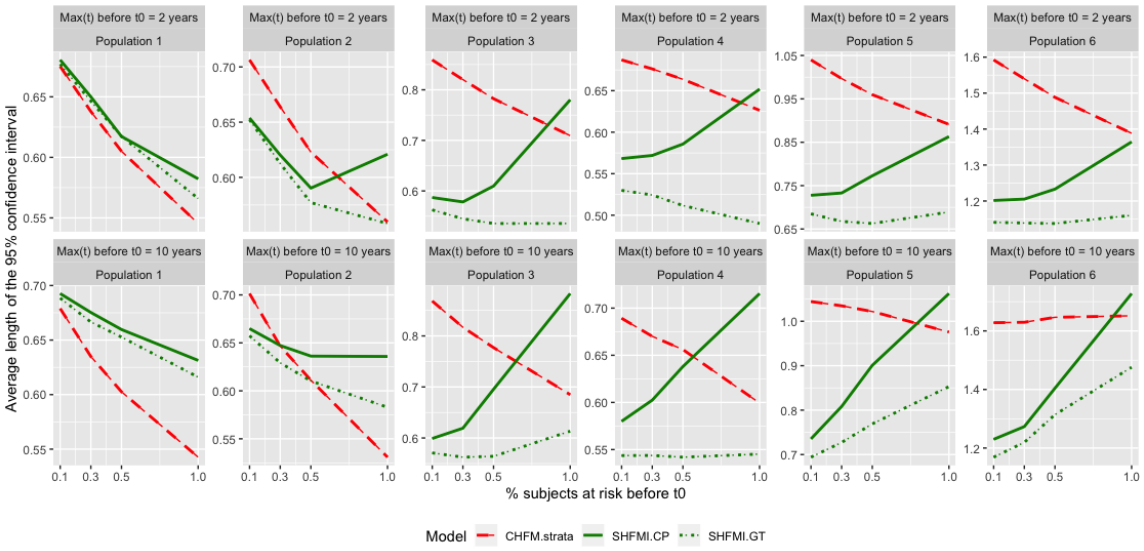
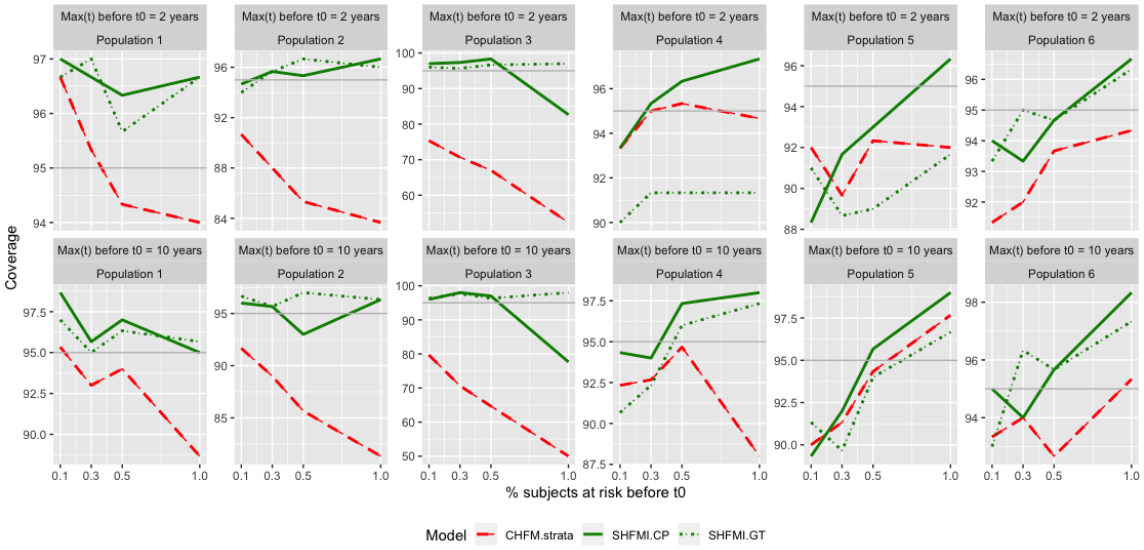


Figure 6s. Coverage of the 95% confidence intervals according to population and time of follow-up, n=250



Bibliografía

[Aalen, 1988] Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in medicine*, 7(11):1121–1137.

[Aalen, 1992] Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound poisson distribution. *The Annals of Applied Probability*, pp. 951–972.

[Aalen, 1994] Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3):227–243.

[Acock, 2012] Acock, A. C. (2012). What to do about missing values.

[Allison, 2010] Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. Sas Institute.

[Amorim y Cai, 2015] Amorim, L. D. y Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, 44(1):324–333.

[Andersen y Gill, 1982] Andersen, P. K. y Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, 10(4):1100–1120.

[Balakrishnan y Peng, 2006] Balakrishnan, N. y Peng, Y. (2006). Generalized gamma frailty model. *Statistics in medicine*, 25(16):2797–2816.

[Balan y Putter, 2020] Balan, T. A. y Putter, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research*, p. 0962280220921889.

[Barceló, 2002] Barceló, M. A. (2002). Modelos marginales y condicionales en el análisis de supervivencia multivariante. *Gaceta sanitaria: Organo oficial de la Sociedad Española de Salud Pública y Administración Sanitaria*, 16(2):59–68.

[Box-Steffensmeier y De Boef, 2006] Box-Steffensmeier, J. M. y De Boef, S. (2006). Repeated events survival models: the conditional frailty model. *Statistics in medicine*, 25(20):3518–3533.

[Breslow, 1974] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, pp. 89–99.

[Brick y Kalton, 1996] Brick, J. M. y Kalton, G. (1996). Handling missing data in survey research. *Statistical methods in medical research*, 5(3):215–238.

[Cai y Schaubel, 2004] Cai, J. y Schaubel, D. E. (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime data analysis*, 10(2):121–138.

[Cain *et al.*, 2011] Cain, K. C., Harlow, S. D., Little, R. J., Nan, B., Yosef, M., Taffe, J. R., y Elliott, M. R. (2011). Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *American journal of epidemiology*, 173(9):1078–1084.

[Cañizares *et al.*, 2004] Cañizares, M., Barroso, I., y Alfonso, K. (2004). Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gaceta sanitaria*, 18(1):58–63.

[Carlin *et al.*, 1999] Carlin, J. B., Wolfe, R., Coffey, C., y Patton, G. C. (1999). Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: prevalence and incidence of smoking in an adolescent cohort. *Statistics in medicine*, 18(19):2655–2679.

[Clayton y Cuzick, 1985] Clayton, D. y Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A (General)*, 148(2):82–108.

[Collett, 1994] Collett, D. (1994). Modelling survival data in medical research 1994 london.

[Cook y Lawless, 2007] Cook, R. J. y Lawless, J. (2007). *The statistical analysis of recurrent events*. Springer Science & Business Media.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[Cox, 1975] Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

[Donders *et al.*, 2006] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., y Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

[Efron, 1977] Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565.

[Gillespie *et al.*, 2010] Gillespie, B. W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., Adriaens, P., Demond, A., Luksemburg, W., y Garabrant, D. H. (2010). Estimating population distributions when some data are below a limit of detection by using a reverse kaplan-meier estimator. *Epidemiology*, pp. S64–S70.

[Glynn y Buring, 1996] Glynn, R. J. y Buring, J. E. (1996). Ways of measuring rates of recurrent events. *Bmj*, 312(7027):364–367.

[Gomez *et al.*, 1992] Gomez, G., Julià, O., Utzet, F., y Moeschberger, M. L. (1992). Survival analysis for left censored data. En *Survival analysis: State of the art*, pp. 269–288. Springer.

[González y Peña, 2004] González, J. R. y Peña, E. A. (2004). Estimación no paramétrica de la función de supervivencia para datos con eventos recurrentes. *Revista Española de Salud Pública*, 78(2):189–199.

[Guo *et al.*, 2008] Guo, Z., Gill, T. M., y Allore, H. G. (2008). Modeling repeated time-to-event health conditions with discontinuous risk intervals. *Methods of information in medicine*, 47(02):107–116.

[Henderson y Oman, 1999] Henderson, R. y Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society: Series*

B (Statistical Methodology), 61(2):367–379.

[Hernández-Herrera *et al.*, 2020] Hernández-Herrera, G., Navarro, A., y Moriña, D. (2020). Regression-based imputation of explanatory discrete missing data. *arXiv e-prints*, p. arXiv:2007.15031.

[Hougaard, 1995] Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis*, 1(3):255–273.

[Islam, 2016] Islam, F. (2016). Parametric reversed hazards model for left censored data with application to hiv.

[Jahn-Eimermacher, 2008] Jahn-Eimermacher, A. (2008). Comparison of the andersen–gill model with poisson and negative binomial regression on recurrent event data. *Computational Statistics & Data Analysis*, 52(11):4989–4997.

[Jung *et al.*, 2018] Jung, T. H., Kyriakides, T., Holodniy, M., Esserman, D., y Peduzzi, P. (2018). A joint frailty model provides for risk stratification of human immunodeficiency virus–infected patients based on unobserved heterogeneity. *Journal of clinical epidemiology*, 98:16–23.

[Karahalios *et al.*, 2012] Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., y Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC medical research methodology*, 12(1):96.

[Keiding *et al.*, 1997] Keiding, N., Andersen, P. K., y Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in medicine*, 16(2):215–224.

[Kelly y Lim, 2000] Kelly, P. J. y Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in medicine*, 19(1):13–33.

[Kleinke *et al.*, 2011] Kleinke, K., de Jong, R., Spiess, M., y Reinecke, J. (2011). Multiple imputation of incomplete ordinary and overdispersed count data.

[Lagakos y Schoenfeld, 1984] Lagakos, S. W. y Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, 40(4):1037–1048.

[Lancaster, 1990] Lancaster, T. (1990). *The econometric analysis of transition data*. Nmero 17. Cambridge university press.

[Landerman *et al.*, 1997] Landerman, L. R., Land, K. C., y Pieper, C. F. (1997). An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. *Sociological Methods & Research*, 26(1):3–33.

[Lang y Little, 2018] Lang, K. M. y Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, 19(3):284–294.

[Lawless, 2011] Lawless, J. F. (2011). Statistical models and methods for lifetime data, vol. 362. *Hoboken: Wiley*.

[Lee *et al.*, 1992] Lee, E. W., Wei, L., Amato, D. A., y Leurgans, S. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. En *Survival analysis: state of the art*, pp. 237–247. Springer.

[Leung *et al.*, 1997] Leung, K.-M., Elashoff, R. M., y Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104.

[Lewsey y Thomson, 2004] Lewsey, J. D. y Thomson, W. M. (2004). The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology*, 32(3):183–189.

[Lin, 1994] Lin, D. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in medicine*, 13(21):2233–2247.

[Navarro y Ancizu, 2009] Navarro, A. y Ancizu, I. (2009). Analyzing the occurrence of falls and its risk factors: Some considerations. *Preventive medicine*, 48(3):298–302.

[Navarro *et al.*, 2017] Navarro, A., Casanovas, G., Alvarado, S., y Moriña, D. (2017). Analy-

zing recurrent events when the history of previous episodes is unknown or not taken into account: proceed with caution. *Gaceta sanitaria*, 31(3):227–234.

[Navarro *et al.*, 2001] Navarro, A., Utzet, F., Puig, P., Caminal, J., y Martín, M. (2001). La distribución binomial negativa frente a la de poisson en el análisis de fenómenos recurrentes. *Gaceta Sanitaria*, 15(5):447–452.

[Nguti, 2003] Nguti, R. W. (2003). *Random effects survival models applied to animal breeding data*. LUC.

[Ødegård y Rossow, 2004] Ødegård, E. y Rossow, I. (2004). Alcohol and non-fatal drug overdoses. *European addiction research*, 10(4):168–172.

[O'Quigley y Stare, 2002] O'Quigley, J. y Stare, J. (2002). Proportional hazards models with frailties and random effects. *Statistics in medicine*, 21(21):3219–3233.

[Pahel *et al.*, 2011] Pahel, B. T., Preisser, J. S., Stearns, S. C., y Rozier, R. G. (2011). Multiple imputation of dental caries data using a zero-inflated Poisson regression model. *Journal of Public Health Dentistry*, 71(1):71–78.

[Petersen, 1998] Petersen, J. H. (1998). An additive frailty model for correlated life times. *Biometrics*, pp. 646–661.

[Petersen, 1991] Petersen, T. (1991). The statistical analysis of event histories. *Sociological Methods & Research*, 19(3):270–323.

[Petersen, 1995] Petersen, T. (1995). Analysis of event histories. En *Handbook of statistical modeling for the social and behavioral sciences*, pp. 453–517. Springer.

[Prentice *et al.*, 1981] Prentice, R. L., Williams, B. J., y Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379.

[Reid y Crépeau, 1985] Reid, N. y Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika*, 72(1):1–9.

[Ripatti y Palmgren, 2000] Ripatti, S. y Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022.

[Rojas *et al.*, 2011] Rojas, L., Achurra, P., Pino, F., Ramírez, P., Lopetegui, M., Sanhueza, L. M., Villarroel, L., y Aizman, A. (2011). Diagnóstico y manejo de la hipoglicemia en adultos diabéticos hospitalizados: evaluación de competencias en un equipo profesional multidisciplinario de salud. *Revista médica de Chile*, 139(7):848–855.

[Rondeau *et al.*, 2012] Rondeau, V., Mazroui, Y., y Gonzalez, J. R. (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw*, 47(4):1–28.

[Rubin, 1988] Rubin, D. B. (1988). An overview of multiple imputation. En *Proceedings of the survey research methods section of the American statistical association*, pp. 79–84. Citeseer.

[Tekindal *et al.*, 2017] Tekindal, M. A., Erdoğan, B. D., y Yavuz, Y. (2017). Evaluating left-censored data through substitution, parametric, semi-parametric, and nonparametric methods: A simulation study. *Interdisciplinary Sciences: Computational Life Sciences*, 9(2):153–172.

[Therneau y Grambsch, 2000] Therneau, T. M. y Grambsch, P. M. (2000). Modeling survival data: extending the cox model. dietz k, gail m, krickeberg k, samet j, tsiatis a, editors.

[Therneau y Hamilton, 1997] Therneau, T. M. y Hamilton, S. A. (1997). rhdnase as an example of recurrent event analysis. *Statistics in medicine*, 16(18):2029–2047.

[Turnbull, 1974] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American statistical association*, 69(345):169–173.

[Ullah *et al.*, 2014] Ullah, S., Gabbett, T. J., y Finch, C. F. (2014). Statistical modelling for recurrent events: an application to sports injuries. *Br J Sports Med*, 48(17):1287–1293.

[Vaida y Xu, 2000] Vaida, F. y Xu, R. (2000). Proportional hazards model with random effects. *Statistics in medicine*, 19(24):3309–3324.

[Vaupel *et al.*, 1979] Vaupel, J. W., Manton, K. G., y Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

[Vermunt, 1996] Vermunt, J. K. (1996). Log-linear event history analysis. *Series on Work and Organization*.

[Wei *et al.*, 1989] Wei, L.-J., Lin, D. Y., y Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408):1065–1073.

[White, 1980] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pp. 817–838.

[Wienke *et al.*, 2010] Wienke, A., Ripatti, S., Palmgren, J., y Yashin, A. (2010). A bivariate survival model with compound poisson frailty. *Statistics in Medicine*, 29(2):275–283.

[Xu y Cheung, 2015] Xu, Y. y Cheung, Y. B. (2015). Frailty models and frailty-mixture models for recurrent event times. *The Stata Journal*, 15(1):135–154.

[Xu *et al.*, 2014] Xu, Y., Lam, K. F., y Cheung, Y. B. (2014). Estimation of intervention effects using recurrent event time data in the presence of event dependence and a cured fraction. *Statistics in medicine*, 33(13):2263–2274.

[Yadav *et al.*, 2018] Yadav, C. P., Sreenivas, V., Khan, M., y Pandey, R. (2018). Epidemiology: Open access.

[Yang *et al.*, 2017] Yang, Wei and Jepson, Christopher and Xie, Dawei and Roy, Jason A and Shou, Haochang and Hsu, Jesse Yenchih and Anderson, Amanda Hyre and Landis, J Richard and He, Jiang and Feldman, Harold I and others (2017). Statistical methods for recurrent event analysis in cohort studies of ckd. *Clinical Journal of the American Society of Nephrology*, 12(12):2066–2073.