Gabriela Boggio
Leticia Hachuel
Guillermina Harvey
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

MODELOS DE REGRESIÓN PARA DATOS DE DURACIÓN EN ESTUDIOS MULTICÉNTRICOS[1]

Resumen

Es sabido que el análisis de los datos provenientes de estudios multicéntricos de carácter observacional presenta desafíos metodológicos debido a la correlación previsible entre las respuestas de individuos de un mismo centro sumado a la falta de rigurosidad en comparación con la de los diseños experimentales. Tal es el caso de un estudio prospectivo sobre lupus llevado adelante por el Grupo Latinoamericano de Estudio del Lupus. En este trabajo se realiza la puesta a prueba de diferentes alternativas de modelización al clásico modelo de regresión de Cox para el análisis del tiempo hasta la ocurrencia de un problema renal en pacientes con lupus. El riesgo en común que pueden presentar los individuos de un mismo centro se tiene en cuenta a través de la consideración de un efecto centro bajo un enfoque condicional estratificando por centro, o bien asignando una distribución de probabilidad a esa susceptibilidad compartida por los individuos de un mismo centro mediante la incorporación de efectos aleatorios al modelo. El análisis de los resultados permitió evaluar las ventajas y desventajas de cada una de las alternativas.

Abstract

It is known that data from observational multicenter studies have methodological challenges because of correlation between individuals from the same center and the lack of accuracy in observational designs. Such is the case of a prospective study about lupus performed by the Latin American Group for the Study of Lupus. Different alternative models from the conventional Cox regression model are evaluated for time to renal disease analysis in lupus patients. The common risk in patients from the same center is taken into account by stratification through a conditional approach or assigning a distribution to the frailty by the inclusion of random effects in the model. The advantages and disadvantages of these approaches were assessed.

Introducción

Los estudios multicéntricos se llevan a cabo simultáneamente en varios centros siguiendo un protocolo acordado. Entre sus ventajas se puede mencionar la posibilidad de incluir un mayor número de participantes, diferentes ubicaciones geográficas, una gama más amplia de grupos poblacionales y la posibilidad de comparar los resultados entre los centros, lo cual permite ampliar la generalización de los resultados (Worthington, 2004). Sin embargo, estos

[1] En este trabajo participó como auxiliar de investigación la alumna de la carrera Licenciatura en Estadística Karen Roberts.

estudios traen aparejados desafíos metodológicos debido a la posible correlación entre mediciones de pacientes de un mismo centro ya que se supone que las respuestas de dichos individuos tenderán a parecerse más entre ellas que con las de otros centros o grupos. Se suma a ello la falta de rigurosidad de los diseños observacionales en contraste con los de carácter experimental. Es en este contexto donde el analista se enfrenta a una variedad de modelos estadísticos con el objeto de dar respuesta a ciertos interrogantes científicos. Entre ellos, los modelos grupo-específicos permiten que el modelo difiera para cada centro mediante la inclusión de parámetros asociados a cada uno de ellos los cuales pueden considerarse como efectos fijos o aleatorios. Debido a que el número de estos parámetros crece a medida que aumenta el número de centros, el enfoque más popular consiste en suponer que los mismos constituyen una muestra al azar de una población de centros según una determinada distribución de probabilidad. Grizzle (1987) argumentó que a pesar de que los centros no sean elegidos al azar, este enfoque proporciona tests e intervalos de confianza que capturan adecuadamente la variabilidad inherente al sistema. Otra alternativa dentro del enfoque grupo-específico es considerar los centros como efectos fijos donde cada parámetro representa el nivel de riesgo de cada centro, pero resignar la estimación de su valor particular para cada grupo condicionando sobre ellos en el proceso de estimación. Una alternativa quizás menos rigurosa para manejar la correlación intra-centro consiste simplemente en modificar la matriz de variancias y covariancias a través de un estimador robusto de la misma. Esta opción se puede encuadrar dentro del denominado enfoque marginal.

El objetivo de este trabajo es poner a prueba estas alternativas de modelización a datos registrados en un estudio observacional multicéntrico sobre lupus que lleva adelante el Grupo Latinoamericano de Estudio del Lupus (GLADEL). Uno de los compromisos orgánicos que más condicionan el pronóstico de un paciente con lupus es el compromiso renal post-diagnóstico. De ahí el interés por investigar el efecto de características sociodemográficas y clínicas de los pacientes al momento del diagnóstico de lupus sobre el tiempo transcurrido desde ese momento hasta la aparición del primer evento de compromiso renal durante el seguimiento. Este objetivo da lugar al ajuste de los denominados modelos de datos de duración o supervivencia.

En la sección siguiente se describen brevemente los diferentes modelos que tienen en cuenta de alguna manera la estructura jerárquica de la información para luego presentar los resultados.

Metodología

La correlación presente entre las respuestas de pacientes de un mismo centro hace pensar que la aplicación de modelos de supervivencia clásicos que suponen independencia de las observaciones, bajo un enfoque ingenuo, no proveen los mejores resultados. De ahí que se presentan variantes al modelo ampliamente utilizado para datos de duración denominado modelo de regresión de Cox, el cual se presenta a continuación.

Modelo de Cox

Un modelo clásico para el análisis de los tiempos de duración o supervivencia es el modelo de hazards proporcionales, también denominado modelo de regresión de Cox, el cual no requiere supuestos distribucionales sobre el tiempo que transcurre hasta la presencia del evento de interés, en este caso la complicación renal (Hosmer *et al.* 2008). El mismo se puede formalizar de la siguiente forma:

$$h(t, \boldsymbol{x}_i, \boldsymbol{\beta}) = h_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

siendo $h(.)$ la tasa instantánea de que se produzca el evento, $\boldsymbol{\beta}$ el vector de coeficientes asociados a las variables explicativas, $\boldsymbol{x}_i$ el vector de covariables para el i-ésimo individuo y $h_0(.)$, la tasa instantánea basal, es decir, para el valor cero en todas las covariables.

La razón de las funciones hazards para individuos con valor a y b en la k-ésima variable explicativa y los mismos valores en el resto de las variables constituye una medida del riesgo asociado al cambiar el valor de la k-ésima variable de a a b. Se la denomina razón de hazards o razón de riesgos: $RR(a,b) = \exp[\beta_k(a-b)]$.

Modelo de Cox marginal

Dado que la dependencia entre las observaciones de un mismo centro provoca mayor heterogeneidad que la supuesta bajo independencia, Lin y Wei (1989) proponen utilizar un estimador de la variancia robusto para la regresión de Cox adoptando un enfoque marginal basado en las ecuaciones de estimación generalizadas (GEE). De esta manera los coeficientes del modelo y las razones de hazards son iguales a los del modelo ingenuo ya que la estimación robusta de la variancia afecta sólo a los errores estándares de los coeficientes del modelo y a las estadísticas asociadas, pudiendo modificar por ende las conclusiones acerca de la significación de las variables explicativas.

Modelo estratificado

El modelo de hazards proporcionales estratificado asume que los individuos en cada grupo o estrato tienen una función hazard basal diferente y todas las variables explicativas satisfacen el supuesto de hazards proporcionales en cada estrato (Allison, 2005; 2010). Se denota, entonces esa función hazard basal para cada estrato como $h_{0j}(t),\ j = 1, \dots, J$, donde J es la cantidad de estratos. El modelo se puede expresar entonces de la siguiente forma:

$$h_j(t, \boldsymbol{x}_{ij}, \boldsymbol{\beta}) = h_{0j}(t)\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}), \quad j = 1, \dots, J$$

Se puede notar que el efecto de las variables explicativas sobre el hazard es el mismo para todos los estratos.

Estos modelos no son completamente eficientes debido a la pérdida de información proveniente de utilizar sólo la variación intra-grupo**.** Su principal ventaja es que controla todas las características estables medidas y no medidas de los centros si bien no permite examinar el efecto particular de ellas.

Modelo con efectos aleatorios

El modelo con efectos aleatorios incluye en la función hazard el valor de una covariable adicional no medida, a veces denominada susceptibilidad compartida por todos los individuos del j-ésimo grupo denotada por z_j, lo que conduce a la siguiente expresión del modelo:

$$h(t, \boldsymbol{x}_{ij}, \boldsymbol{\beta}) = h_0(t) z_j \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}).$$

Este modelo con susceptibilidad compartida, común a los individuos de un mismo grupo y responsable de generar la dependencia entre las observaciones, supone que dado el valor de la susceptibilidad las observaciones son independientes (Hosmer *et al.*, 2008; Racca, 2012; Therneau y Grambsch, 2010).

Notar que puede expresarse como un modelo de hazards proporcionales estratificado como sigue:

$$h\left(t, \boldsymbol{x}_{ij}, \boldsymbol{\beta}\right) = h_{0j}(t)\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta})$$

donde $h_{0j}(t) = h_0(t)z_j$ es la función hazard basal estrato-específica bajo un enfoque de susceptibilidad compartida, sólo que en lugar de estar completamente no especificada como en el modelo estratificado, se realiza un supuesto distribucional para z_j. Las distribuciones de probabilidad más comunes para las susceptibilidades son la gamma, la estable positiva y la log-normal. La elección correcta de esta distribución es importante para tratar de captar la estructura de dependencia presente en los datos. Sin embargo, teniendo en cuenta que el interés principal radica en la estimación de los efectos de las variables explicativas teniendo en cuenta la heterogeneidad de los datos y no la estructura de dependencia en sí misma, se elige la distribución log-normal en función de su disponibilidad en el paquete estadístico SAS.

Los datos

El lupus eritematoso sistémico (LES) es una enfermedad crónica e inflamatoria de origen autoinmune que puede afectar varias partes del cuerpo tales como piel, articulaciones y riñones. En general se comprometen diferentes sistemas orgánicos siendo la lesión renal una de las de mayor incidencia.

El estudio de carácter prospectivo llevado adelante por GLADEL está conformado por 1480 pacientes de 34 centros de atención médica distribuidos en 9 países de Latinoamérica que cuentan con expertos en el diagnóstico y seguimiento de pacientes con LES. Se registraron características sociodemográficas, clínicas, de laboratorio y de actividad de la enfermedad como así también órganos dañados y otras características consideradas importantes para el pronóstico de la enfermedad.

En este estudio se incluyeron los 945 pacientes de la cohorte libres de complicación renal al momento del diagnóstico de LES.

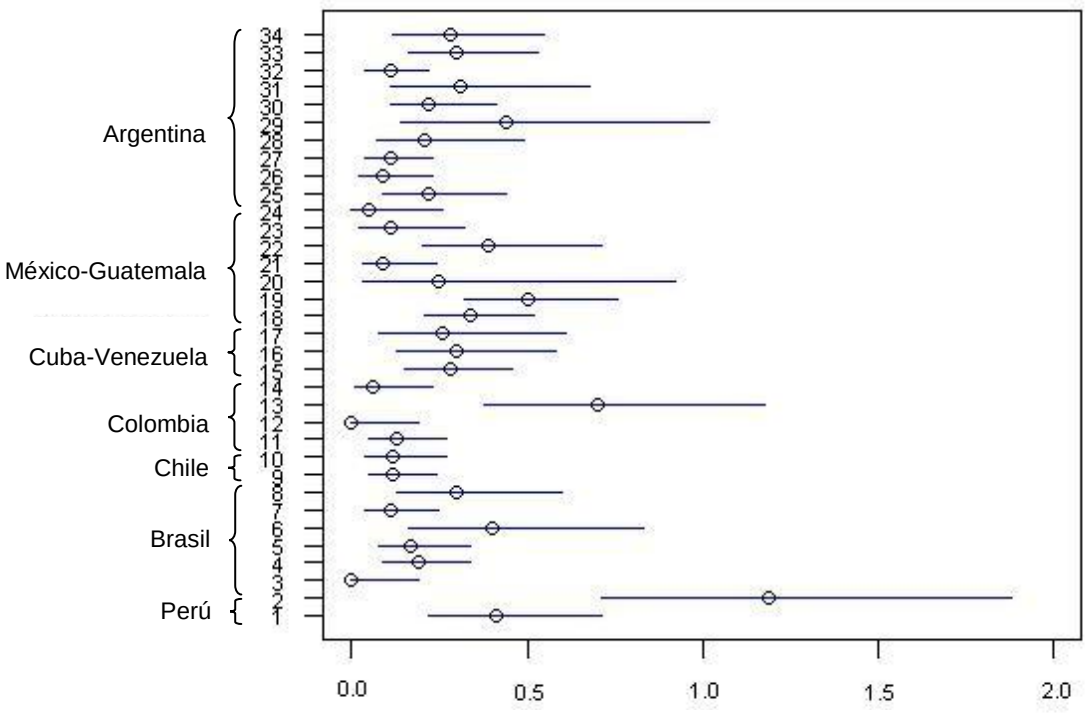
Se considera como variable respuesta el tiempo (medido en meses) desde el diagnóstico de LES hasta el desarrollo de la primera complicación renal. Se tienen en cuenta como posibles factores explicativos de esta complicación renal los considerados en un trabajo anterior (Pons-Estel *et al*, 2013), a saber: edad al comienzo de síntomas de LES, etnia, hipertensión, consumo de antimaláricos, fotosensibilidad e índice de actividad de LES al diagnóstico (SLEDAI). Se tiene también en cuenta el país de residencia de los pacientes (Argentina, Perú, Brasil, Chile, Colombia, Cuba, Venezuela, México y Guatemala).

Resultados

En primer lugar se muestran las tasas de incidencia de complicación renal y los respectivos intervalos de confianza para cada uno de los centros identificando el país de pertenencia (Figura 1).

Es notable la variabilidad de estas tasas a través de los centros de los diferentes países. Se observan tasas de incidencia pequeñas e incluso nulas, para un centro de Colombia y otro de Brasil, como así también tasas de incidencia mayores a uno por cada mil pacientes/meses en Perú. El gráfico incluso muestra variabilidades distintas en centros de un mismo país. Estos resultados orientan a elegir los centros como posibles responsables de heterogeneidad en la información a analizar.

Figura 1 – Tasas de incidencia de enfermedad renal en pacientes con LES según centro



Con fines comparativos se presentan los resultados del ajuste del modelo de regresión de Cox clásico (enfoque ingenuo) y de modificaciones del mismo para tener en cuenta la correlación de pacientes de un mismo centro: modelo de efectos aleatorios, estratificado y marginal (Tabla 1).

Como era de esperar, el tratamiento ingenuo de los datos conduce en general a probabilidades asociadas menores que bajo los otros enfoques. En el ajuste del modelo marginal, si bien hay variaciones en las probabilidades asociadas, no alcanzan a alterar prácticamente la significación de las variables si se toma como criterio un nivel del 5%. Sólo se observa que la probabilidad asociada al efecto de los países México y Guatemala, en comparación con Argentina, pasa de 0.0836 a 0.0527, valor muy cercano al nivel de significación elegido. En cambio, en los modelos estratificado y el de efectos aleatorios, *presencia de fotosensibilidad* pierde su significación estadística, de manera más contundente en el enfoque condicional. El efecto de las variables *edad al comienzo de síntomas, hipertensión, SLEDAI al diagnóstico* y *consumo de antimaláricos* permanece significativo en todos los modelos.

La inclusión de la covariable *etnia* se origina en el conocimiento de su importancia en el estudio de LES. Sin embargo en estos modelos no resulta significativa, por lo que es razonable pensar que su influencia podría estar absorbida por el *país de residencia*. El efecto de *país de residencia* se refleja sólo en las diferencias entre Perú, México y Guatemala con Argentina en los modelos ingenuo y marginal y en forma más atenuada en el modelo de efectos aleatorios sólo la comparación entre Perú y Argentina. Por otro lado, al ser *país de residencia* una variable estable dentro de los centros, no es posible incluirla en el modelo estratificado si bien el mismo tiene la habilidad de controlar todas aquellas variables medidas y no medidas que resultan constantes dentro de los centros.

Tabla 1 – Estimaciones de los modelos propuestos

Variable	Modelo de efectos aleatorios			Modelo estratificado			Modelo marginal			Modelo ingenuo		
	Estim.	**SE**	**p**	**Estim.**	**SE**	**p**	**Estim.**	**SE**	**p**	**Estim.**	**SE**	**p**
Edad al comienzo de los síntomas	-0.02	0.01	0.0021	-0.01	0.01	0.0382	-0.02	0.01	<.0001	-0,02	0,01	0,0003
Etnia(*)												
Mestizo	0.16	0.24	0.5063	0.10	0.27	0.7133	0.20	0.23	0.3795	0.20	0.22	0.3494
ALA	0.23	0.28	0.4135	0.16	0.30	0.6011	0.21	0.40	0.5954	0.21	0.26	0.4158
Hipertensión	1.40	0.15	<.0001	1.39	0.16	<.0001	1.35	0.13	<.0001	1.35	0.14	<.0001
Presencia de fotosensibilidad	-0.20	0.15	0.1856	-0.10	0.16	0.5305	-0.28	0.15	0.0668	-0.28	0.15	0.0555
SLEDAI al diagnóstico	0.05	0.01	<.0001	0.05	0.01	<.0001	0.04	0.01	0.0063	0.04	0.01	0.0002
Consumo de antimaláricos	-0.48	0.15	0.0020	-0.44	0.16	0.0055	-0.54	0.18	0.0029	-0.54	0.15	0.0004
País de residencia(*)												
Perú	0.84	0.47	0.0767	-	-	-	0.92	0.28	0.0012	0.92	0.31	0.0029
Brasil	0.15	0.36	0.6842	-	-	-	0.17	0.23	0.4518	0.17	0.26	0.5153
Chile	-0.36	0.49	0.4611	-	-	-	-0.26	0.17	0.1361	-0.26	0.35	0.4559
Colombia	-0.12	0.42	0.7788	-	-	-	0.10	0.57	0.8670	0.10	0.30	0.7527
Cuba-Venezuela	-0.03	0.45	0.9407	-	-	-	0.10	0.30	0.7259	0.10	0.31	0.7392
México-Guatemala	0.25	0.36	0.4755	-	-	-	0.44	0.23	0.0527	0.45	0.26	0.0836

(*) Categoría de referencia: etnia blanca; país de residencia Argentina.

La débil significación de *país de residencia* en el modelo de efectos aleatorios, se podría pensar que se debe a que la heterogeneidad entre las respuestas de los pacientes de estos países queda absorbida por la susceptibilidad compartida por los pacientes de un mismo centro médico ya que los efectos aleatorios asociados a ellos resultan significativos (p=0,0042). Este resultado confirma la existencia de la heterogeneidad no observable entre los centros ya detectada en el análisis descriptivo.

Con datos observacionales, la variación entre centros es muy probable que esté contaminada por características no medidas que pueden estar correlacionadas con las variables explicativas consideradas. Ello puede conducir a estimaciones sesgadas, lo cual se aprecia solamente en el coeficiente asociado a fotosensibilidad (-0,10 en el modelo estratificado y -0,20 en el modelo con efectos aleatorios).

En el enfoque condicional hay una pérdida inevitable en el tamaño muestral debido a que los centros en los cuales ningún paciente tuvo enfermedad renal no contribuyen a la verosimilitud del modelo. En este caso los 26 pacientes de un centro de atención de Colombia no presentaron enfermedad renal y 12 pacientes de un centro de Brasil. Como consecuencia de esta restricción los errores estándares de los coeficientes del modelo resultan mayores que los del modelo con efectos aleatorios, pero en este caso el aumento resultó insignificante.

Ante la falta de un método formal para elegir entre uno y otro modelo se podría hacer un balance entre la ventaja del modelo estratificado respecto de su habilidad para controlar variables no medidas estables en los centros y la ventaja del de efectos aleatorios de poder incluir ese tipo de variables en el predictor. Como en este caso sólo se dispone de información de una única característica estable dentro de los centros -*país de residencia*- y con escasa significación, parecería apropiado elegir el modelo estratificado, más aún cuando la pérdida de información que éste conlleva no ha sido de importancia. La cualidad de controlar las variables estables del enfoque condicional resulta relevante ya que no se dispone de información respecto a características de los centros de salud como por ejemplo dimensión del establecimiento, capacidad médica, cantidad de especialistas, etc., no pudiendo, en consecuencia, considerarlas en forma explícita en el modelo.

Se comparan de cualquier forma las estimaciones de ambos modelos en términos de razones de hazards para evaluar las diferencias interpretativas según cual fuera la elección (Tabla 2).

Para las variables que resultaron significativas, las razones de riesgos calculadas a partir de ambos modelos resultan muy semejantes y los intervalos de confianza presentan amplitudes sólo levemente mayores en el modelo estratificado en virtud del escaso aumento de los errores estándares de sus coeficientes.

En definitiva, tanto a partir del modelo estratificado como del de efectos aleatorios, se desprende que el riesgo de desarrollar enfermedad renal aumenta en la medida que el índice de actividad de LES sea mayor en el momento del diagnóstico de la enfermedad, aumenta también ante la presencia de hipertensión y cuando los pacientes comienzan con los síntomas del lupus a una edad más temprana. Es importante el resultado acerca de que el consumo de antimaláricos reduce el riesgo de desarrollar enfermedad renal, mostrando el efecto protector del consumo de esta droga para evitar complicaciones frecuentes en el lupus.

Tabla 2 – Razones de riesgo estimadas en los modelos de efectos aleatorios y estratificado

Variable	Modelo de efectos aleatorios		Modelo estratificado	
	$\widehat{RR}$	$IC_{95\%}$	$\widehat{RR}$	$IC_{95\%}$
Edad al comienzo de los síntomas	0.98	0.97-0.99	0.99	0.97-0.99
Etnia(*)				
Mestizo	1.18	0.73-1.89	1.10	0.65-1.86
ALA	1.26	0.73-2.18	1.17	0.65-2.10
Hipertensión	4.07	3.03-5.44	4.02	2.48-5.47
Presencia de fotosensibilidad	0.82	0.61-1.10	0.91	0.67-1.23
SLEDAI al diagnóstico	1.05	1.03-1.07	1.05	1.03-1.08
Consumo de antimaláricos	0.62	0.46-0.84	0.64	0.47-0.88
País de residencia(*)				
Perú	2.31	0.91-5.81	-	-
Brasil	1.16	0.57-2.36	-	-
Chile	0.70	0.26-1.83	-	-
Colombia	0.89	0.39-2.01	-	-
Cuba-Venezuela	0.97	0.40-2.36	-	-
México-Guatemala	1.29	0.64-2.59	-	-

(*) Categoría de referencia: etnia blanca; país de residencia Argentina.

Consideraciones finales

Las alternativas de modelización consideradas hacen posible tener en cuenta de una u otra forma la correlación entre las observaciones relativas a individuos que se origina en el hecho de compartir, en este caso particular, su atención en un mismo centro de salud. Desde un punto de vista metodológico, el objetivo es procurar obtener errores estándares de las estimaciones más realísticos que aquéllos obtenidos por los métodos convencionales que ignoran la falta de independencia.

El modelo con efectos aleatorios tiene en cuenta tanto la variación intra-centro como entre los centros, aunque se basa en el fuerte supuesto de independencia entre las variables explicativas y los efectos aleatorios, supuesto que puede no cumplirse particularmente en estudios observacionales. Permite incluir como variables explicativas aquellas comunes o estables en los individuos de un mismo centro, por lo que es posible examinar si éstas son las que originan las diferencias entre ellos.

Por otro lado, el enfoque estratificado se destaca por su habilidad para controlar todas las características estables de los centros en estudio, las medidas y las no medidas, eliminando de esta manera fuentes de sesgo potencialmente importantes. Su principal desventaja es la imposibilidad de considerar variables estables dentro de los centros como predictores en el modelo.

En el problema presentado, se obtuvieron resultados semejantes en ambos enfoques por lo que a pesar de las restricciones que impone uno u otro procedimiento se detectan los principales factores que influyen sobre el desarrollo de compromiso renal en pacientes con LES.

REFERENCIAS BIBLIOGRÁFICAS

Allison, P. D. 2005. *Fixed Effects Regression Methods for Longitudinal Data Using SAS®*. Cary, NC: SAS Institute Inc.

Allison, P. D. 2010. *Survival Analysis using SAS: a Practical Guide*, 2º edition. Cary, NC: SAS Institute Inc.

Hosmer, D.W.; Lemeshow, S.; May, S. 2008. *Applied Survival Analysis. Regression Modeling of Time to Event Data*, 2º edition. John Wiley & Sons.

Pons-Estel, G.; Alarcón, G.; Burgos, P.; Hachuel, L.; Boggio, G.; Wojdyla, D.; Nieto, R.; Alvarellos, A.; Catoggio, L.; Guibert-Toledano, M.; Sarano, J.; Massardo, L.; Vásquez, G.; Iglesias-Gamarra, A.; Costallat, L. L.; Da Silva, N.; Alfaro, J.; Abadi, I.; Segami, M.; Huerta, G.; Cardiel, M.; Pons-Estel, B.; GLADEL. 2013. Mestizos with systemic lupus erythematosus develop renal disease early while antimalarials retard its appearance: Data from a Latin American cohort. *Lupus,* 22(9): 899-907.

Racca, L. 2012. *Modelos para datos de duración correlacionados* (Tesis de Maestría). Universidad Nacional de Rosario.

Worthington, J., 2004. Methods for pooling results from multicenter studies. *Journal of Dental Research*, 83: C119-C121.

Grizzle, J. E., 1987. Letter to the Editor. *Controlled Clinical Trials*, 8: 392-393.

Lin, D. Y.; Wei, L. J. The robust inference for proportional hazards model. *Journal of the American Statistical Association*, 84:1074-1078.

Therneau, T. M.; Grambsch, P. M. 2000. *Modeling Survival Data. Extending the Cox model.* Springer-Verlag.