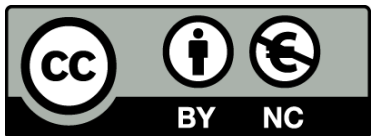
UNIVERSITAT DE BARCELONA

Aplicació de models d'efectes aleatoris en l'epidemiologia quantitativa

Martí Casals i Toquero



Departament de Salut Pública
Unitat de Bioestadística
Facultat de Medicina

Aplicació de models d'efectes aleatoris en l'epidemiologia quantitativa

Memòria de Tesi per optar al grau de Doctor
Programa de Doctorat en Estadística

Doctorand: Martí Casals i Toquero
Directors: Dr. Josep Lluís Carrasco i el Dr. Klaus Langohr

Barcelona, 2015

Agraïments

Tot va començar amb el projecte fi de màster que vaig realitzar l'any 2009 amb els tutors Joan Caylà i Klaus Langohr. El treball que va ser un repte personal i divertit que aplicava l'estadística a malalties infeccioses en un escenari multidisciplinar em va fer adonar que l'estadística, la recerca i posteriorment la docència eren la meva passió. Agrair en especial al Dr. Joan Caylà per donar-me la confiança de liderar el treball fi de màster amb unes dades reals del Camp de la Bota, i també per mostrar-me el camí dia a dia que el llenguatge mèdic (o aplicat) i l'estadístic es poden entendre i fer grans projectes.

En primer lloc vull donar les gràcies als meus directors, en Josep Lluís Carrasco i Klaus Langohr, que m'han permès madurar com a professional de l'estadística i com a persona. Tot i passar els dos per greus problemes de salut, sempre heu estat al meu costat i això mai ho oblidaré. L'experiència d'en Josep Lluís m'ha transmès saviesa, ser més pràctic, estratègic i constància. Amb en Klaus, després de molts cafès, simulacions computacionals, edicions de text amb LaTex, m'ha ensenyat els valors, serenor i paciència que cal d'un estadístic, i sobretot ser un bon amic. En general, heu representat per mi un gran tàndem, el més semblant a un primer i segon entrenador d'elit d'un gran equip (no diré de quin esport, jeje). Espero continuar gaudint de la vostra companyia en el terreny professional i personal. Gràcies.

Agrair la direcció d'en Lars Ronnegard, estadístic que em va supervisar durant la meva estada a Suècia durant tres mesos. En Lars em va mostrar una manera de treballar, una filosofia que no era ni freqüentista ni bayesiana i sobretot la humilitat i senzillesa que cal tenir com a professional.

Donar les gràcies a la Montse (la Dra. Montserrat Girabent) per obrir-me la porta a la docència i convertir-se en una de les meves millors amigues en el terreny professional i personal. La teva visió, estratègia i manera de veure les coses m'ha ajudat a tirar endavant, a madurar i sobretot a poder compartir inquietuds professionals com si fossis part de

la codirecció de la tesi en tot moment. Gràcies.

Aquesta tesi no ves estat possible si no ves llegit probablement un dels meus articles preferits *Generalized linear mixed models: a practical guide for ecology and evolution* d'en Ben Bolker. Els seus escrits i el fòrum 'user-r-mixed-models'em van fer motivar més que mai amb el meu tema principal de la tesi, els models mixtes.

Vull agrair en Vicente Martin de la Universitat de León pel seu suport i mostrar-me les possibilitats d'aplicació estadística que hi pot haver en dades reals pel que fa a la prevenció i control de lesions en el món de l'esport. Gracias Vicente, ya conozco la Lucha Leonesa:-)

També donar les gràcies a bons amics estadístics (Leila Luján, Josep Anton Sánchez, Marc Saez, Erik Cobo, Lupe Gomez, Isaac Subirana, Moisés Gomez, Altea Lorenzo, Carmen Cadarso, etc), a la SCE, grup GRASS de l'UPC, que han estat, són i seran grans motivadors per fer aquest projecte i altres recerques. En especial vull agrair en Josep Anton Sánchez, que el vaig conèixer en un dels cursos de la UAB sobre models mixtes, un dels meus estadístics preferits, un apassionat dels models estadístics i optimització i computació, i sobretot una gran font de coneixement i gran persona. Gràcies.

Al Servei d'Epidemiologia de l'Agència de Salut Pública de Barcelona, on hi porto treballant fa molts anys amb diferents companys. Agrair en especial en Joan Asensio, un dels meus companys de feina que a part d'esmorzar plegats i parlar de temes tan diversos, és un gran amic amb qui tenim un *feeling* especial i de qui sempre aprenc alguna cosa en les nostres converses. Gràcies.

Al Centre d' Investigació Biomèdica d'Epidemiología i Salut Pública (CIBERESP)i a la convocatoria del programa d'estades breus de doctorat europeu que em va permetre anar a un centre de recerca a Suècia.

A la UIC, en especial, al seu departament de fisioteràpia, amb qui darrerament hem establert més amistat. Gràcies pel vostre bon rotllo.

Als meus professors preferits durant la meva formació acadèmica, en especial, a la meva professora 'madre Pepa' de les Escolàpies d'El Masnou, en Joan Vila (professor de Matemàtiques) de l'institut Badalonès, i al

meu amic professor, matemàtic i filòsof d'El Masnou, en Miquel Àngel Perelló, que em va impulsar i motivar a fer Estadística i a fer la tesi doctoral. Gràcies Miquel Àngel, sempre has estat un mirall a seguir.

També agrair a investigadors que he conegut recentment (Jose Antonio Martínez de la Universitat Politècnica de Cartagena i en Javier Peña de la Univeristat de Vic), amants de l'esport i de la ciència, amb qui espero anar fent projectes, i que m'han fet valorar que és possible combinar les meves dues passions (estadística i bàsquet) amb rigor científic. 'Gracias Jose por darme la oportunidad de poder trabajar juntos sobre mi deporte favorito ('basketball'), ciència y estadística.'

Als meus amics que veig cada dia, de qui aprenc a ser millor persona i em fa ser com sóc. A entorns que he passat o hi passo més estona, El Masnou, Girona, Teià, Serra de Daró, Sant Julia de Ramis, Menorca..., i futurs encara per conèixer. I sobretot a les estones d'oci que hi dedico (anar al palau Blaugrana, jugar a bàsquet o pàdel, ioga, cinema, excursions....)

I un punt a part mereix la família. Sense ells no ves estat possible aquesta tesi. En especial, vull agrair als meus pares i la meva germana. De ben petit fins ara m'han ensenyat el camí, i que els valors de la constància, voluntat i creure en un mateix fan que puguis aconseguir allò que et proposes. Moltes gràcies.

I, per últim i molt molt important, vull agrair a la Iona. Des de que t'he conegut, he après a aplicar l'energia positiva que cal en tot moment a un treball tan llarg com és la tesi. La teva paciència, el que compartim dia a dia, i les passions que tenim, m'han donat l'impuls necessari per fer realitat aquest somni. Moltes gràcies.

Índex

Índex de taules

Índex de figures

1 Introducció

L'aplicació de models estadístics sempre ha estat una qüestió que ha generat un gran interès en qualsevol àmbit científic.

Per entendre què és un model estadístic en primer lloc definirem un model des d'un punt de vista matemàtic. Un model matemàtic és un patró teòric o experimental que permet interpretar mitjançant mètodes matemàtics fenòmens reals o problemes tècnics per fer inferència i prendre decisions. Els models matemàtics poden ser determinístics quan es té la certesa sobre el funcionament i els resultats del problema, o poden ser aleatoris o estocàstics quan només es té la certesa parcial sobre el funcionament i es coneixen els resultats probables del model. Per exemple, les lleis de la dinàmica clàssica de partícules poden considerar-se un model matemàtic determinístic, i el model de regressió lineal estadístic (que veurem més endavant amb detall) és un model matemàtic aleatori [1].

Un model estadístic és un tipus especial de model matemàtic. El que distingeix a un model estadístic d'altres models matemàtics és doncs que un model estadístic no és determinista. Un model estadístic s'especifica a través d'equacions matemàtiques on algunes de les variables no tenen valors específics, sinó que tenen distribucions de probabilitat, és a dir, algunes de les variables són estocàstiques. Herman Adèr va citar a Kenneth Bollen tot dient: 'A model is a formal representation of a theory' [2].

Hi ha tres propòsits per a un model estadístic, segons Konishi i Kitagawa [3]:

- Descripció de les estructures estocàstiques
- Extracció d'informació
- Prediccions

En el Capítol 1 d'aquesta tesi s'introduiran els models clàssics més utilitzats en l'àmbit de ciències de la salut. Aquests models presenten certes limitacions quan ens trobem amb dissenys més complexes on les dades presenten estructures jeràrquiques i on s'assumeix que les observacions no són independents. Per exemple, els subjectes poden ser observats dins d'unitats més grans com escoles, famílies, hospitals o àreas geogràfiques. És per això que seran necessaris models més sofisticats. A més a més, farem una breu introducció de l'epidemiologia, ciència que ha motivat els casos reals presentats al llarg de la tesi.

1.1 Motivació de l'estudi: Aplicació a l'epidemiologia

L'epidemiologia és la ciència que estudia la distribució de malalties i els determinants de les seves freqüències en els diferents grups d'humans o animals. Aquesta definició engloba dues areas separades d'investigació: l'estudi de la distribució de la malaltia i la cerca dels determinants d'aquesta malaltia en funció de la seva distribució.

A més a més, l'epidemiologia es considera la ciència bàsica per a la medicina preventiva i una font d'informació per a la formulació de politiques de salut pública.

Existeixen diferents definicions d'epidemiologia: en els seus inicis era "la doctrina de les epidemies", però aquesta definició és clarament insuficient a l'actualitat. Amb el temps, la definició de l'epidemiologia s'ha vist enriquida per conceptes més propis de la sociologia, la demografia i l'estadística. La paraula "epidemiologia" ve de les paraules gregues "epi" (sobre), "demos" (la població, la gent) i "logos" (estudi de, doctrina).

La definició següent és oficial de l'Associació Epidemiològica Internacional [4]:

> "L'epidemiologia és l'estudi dels factors que determinen la freqüència i distribució de malalties en poblacions humanes."

El científic John Snow, considerat un dels pares de l'epidemiologia va ser un dels primers en dur a terme un estudi sobre l'epidèmia de la còlera a Londres el setembre de l'any 1854. L'explicació d'aquesta epidèmia a Londres i que alguns barris estiguessin més afectats que altres, va ser que existien diferents companyies d'aigua fent-se la competència. La qualitat de l'aigua havia estat dolenta durant anys, ja que, es recollia l'aigua del Tàmesis, on també hi havia el clavegueram de Londres. No es realitzaven depuracions ni filtracions abans de distribuir l'aigua a les seves respectives fonts, tot i que en 1853 la companyia Lambeth va canviar de lloc de captura de les seves aigües, obtenint així un aigua lliure de contaminació fecal i determinant als seus clients que patissin menys còlera que els clients de la competència. L'estudi va portar a Snow a la conclusió de que la causa de la malaltia era un microorganisme paràsit que molts anys més tard es va arribar a identificar com el *Vibrio cholerae* [5].

Objectius i observació en epidemiologia

El coneixement de la distribució de la malaltia pot ser utilitzat per comprendre els mecanismes causals, explicar les característiques locals de la presència de la malaltia, descriure la historia natural d'una malaltia o servir de guia durant l'administració dels serveis de salut. Un dels objectius principals de l'epidemiologia consisteix en identificar causes alterables de la malaltia. Per això, és necessari comprendre, en primer lloc, què entenem per causa i, en segon lloc, quines són les bases per a la formació de categories d'individus dels quals es diu que tenen una malaltia.

L'epidemiologia persegueix el propòsit pràctic de descobrir relacions que ofereixin possibilitats per a la prevenció de la malaltia. En les últimes dècades hi ha un extraordinari creixement de l'epidemiologia obser-

vacional en general, i d'aquella referida a grans poblacions. L'estímul que va provocar aquest gran creixement va ser la demostració, als voltants dels anys 50, de la varietat i gravetat dels efectes del consum de tabac sobre la salut de les persones. Tot i que existeix una extensa bibliografia sobre aquest tema, els estudis de Doll i Hill, en 1952 i 1954, van conscienciar a la població sobre l'utilitat de l'epidemiologia en l'estudi de malalties cròniques. Aquest últim treball, publicat al *British Medical Journal*, és un estudi de casos i controls, on els casos el constituïen els pacients que ingressaven en certs hospitals amb diagnòstic de càncer de pulmó, mentre que els controls eren pacients en què el seu ingrés es devia a altres causes. Ambdós tipus de pacients se'ls interrogava sobre els seus hàbits tabàquics, inhalació d'altres gasos, i altres possibles agents etiològics diferents. Les enquestes varen ser per personal "ceg", en el sentit de que desconeixien el propòsit del treball. El resultat va ser que els casos i controls tenien una exposició similar a tots els possibles factors de risc, excepte el tabac.

El tipus d'anàlisi realitzada va ser el de la comparació de proporcions mitjançant la prova χ^2, proposada per Pearson a l'any 1900 i la prova exacte de Fisher en la dècada dels 20. Aquest treball però va rebre moltes crítiques de personalitats tan respectades com Joseph Berkson, estadístic principal de la Clínica Mayo que ja a l'any 1938 l'havia descrit, tot i que el treball va ser ampliat i publicat al 1946. Aquest treball va ser titulat com les "Limitacions de l'aplicació de les taules dos per dos per dades hospitalàries", i és on segons Sander Greenland, és la primera anàlisi algebraic d'un biaix de selecció descrit en epidemiologia. També van criticar aquest treball Jerzy Neyman, i sobretot en R.A. Fisher, fumador empedreït i impulsor de l'anàlisi de la variancia (ANOVA) i de mètodes de màxima versemblança a la dècada dels 20, qui a l'any 1958 va publicar un article titulat *"Cigarettes, Cancer and Statistics"* al Centennial Review, i dos articles en la prestigiosa revista Nature titulats *"Lung Cancer and Cigarettes"* i *"Cancer and smoking"* [6–9]. Una de les principals critiques es basava en que no tots els fumadors desenvolupaven càncer de pulmó ni tots els malalts de càncer de pulmó fumaven. Això, entrava en clara contradicció amb el primer postulat "L'agent (tabac) ha d'estar present

en cada cas de la malaltia sota les condicions apropiades", al que els autors van respondre afirmant que el que volien dir era que el tabac no era la causa del càncer, sinó una de les causes. Com que pràcticament ningú es creia els resultats de l'informe, Doll i Hill varen decidir canviar de mètode i es van plantejar dur a terme un estudi de seguiment en un grup poblacional.

A l'any 1954 Doll i Hill van començar un estudi prospectiu, de cohorts, en el que s'efectuava un seguiment de metges britànics i s'estudiava la possible associació entre les taxes de mortalitat i l'hàbit tabàquic, que va corroborar no solament els resultats anteriors sinó també una mortalitat més ràpida també per altres causes, fonamentalment malalties coronàries, entre els fumadors. A mesura que l'evidència es va anar acumulant tan Berkson com Neyman van canviar d'opinió, tot i que Fisher va mantenir-se amb la mateixa opinió.

1.2 Model Lineal General

El model lineal general és un model de regressió que permet analitzar relacions lineals entre una variable resposta (continua) i un conjunt de covariables de naturalesa tant quantitativa com qualitativa. Casos particulars del model lineal general són també la generalització del model de regressió, anàlisi de la variància i anàlisi de la covariància. El model lineal (LM) és per exemple un cas particular del model lineal general, conegut també com a model lineal de regressió.

El model lineal general és una família de models on els errors són independents i idènticament distribuïts i segueixen una distribució normal amb una variància comú. Aquests models són també referits com a models lineals normals i han estat una de les aplicacions principals de l'estadística durant molts anys.

Donada una mostra aleatoria $(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, la relació entre les observacions Y_i de la variable resposta i les covariables X_{ij} es formula segons l'equació (1.1).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i \tag{1.1}$$

on $i = 1, \ldots, n$. Les quantitats ϵ_i són variables aleatòries que representen errors amb la propietat que $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$, és a dir, el model suposa homocedasticitat i independència. La part lineal està constituïda pels coeficients de regressió β_i.

Pel que fa a l'estimació, l'assumpció d'homocedasticitat es realitza quan es fa l'estimació mitjançant els mínims quadrats ordinaris (OLS) on les estimacions dels paràmetres desconeguts β_i es determinen al minimitzar una funció de suma de quadrats. El model lineal general també es podrà estimar mitjançant els mínims quadrats generalitzats (GLS) quan es presenti una variància heterocedàstica.

Per tant, en el LM la variable resposta Y_i segueix una distribució gaussiana i existeix una relació lineal entre l'esperança de Y_i i les covariables (predictor lineal).

1.3 Model Lineal Generalitzat

Els models lineals generalitzats (GLM) generalitzen els LM en el sentit que permeten utilitzar distribucions de probabilitat més enllà del model Normal de la resposta condicionada a les covariables (Binomials, Poisson, Gamma, etc) i variàncies no constants [10]. Nelder i Wedderburn a l'any 1972 van ser els primers en definir els GLM de manera que permetien que la distribució de probabilitat de la variable resposta i per tan dels errors, fos qualsevol de la família exponencial [11]. D'aquesta manera la distribució pot ser escollida dintre de distribucions que pertanyen a la família exponencial com són la Binomial, Poisson, Binomial Negativa, Gamma, Beta, Inversa Gaussiana i on la distribució normal n'és un cas particular. Això va permetre desenvolupar un algoritme general per l'estimació de màxima versemblança en aquests models com es veurà a continuació.

L'objectiu dels GLM és descriure la relació entre l'esperança de Y i les covariables:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

on $i = 1, \ldots, n$, μ_i és el valor esperat de la variable resposta i $g(\mu_i)$ és la funció d'enllaç que relaciona l'esperança $E(Y_i)$ i les covariables i pot ser no lineal. La funció d'enllaç acostuma a eliminar les restriccions sobre els paràmetres.

Nelder i Wedderburn (1972) van definir els GLM de manera que:

1. La funció de densitat de la variable resposta Y_i en els GLM [10] és definida com:

$$f_Y(Y_i; \theta_i, \phi_i) = \exp\left\{\frac{Y_i \cdot \theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi_i)\right\}$$

   on ϕ_i representa un paràmetre extra d'escala, conegut com a paràmetre de dispersió, i θ_i és el paràmetre 'natural' de la distribució, conegut com a paràmetre canònic.

2. La relació entre la variable resposta i les covariables es dóna mitjançant una funció d'enllaç $g(\mu_i)$, que és monòtona i diferenciable. Aquesta funció defineix la relació entre l'esperança de la variable resposta, $E(Y_i) = \mu_i$, i les covariables. Per tant, la funció d'enllaç connecta els components sistemàtics i aleatoris. Els dos primers moments de la funció de distribució donada anteriorment venen donats com:

$$E(Y_i; \theta_i, \phi_i) = \mu_i = b'(\theta_i)$$

$$Var(Y_i; \theta_i, \phi_i) = b''(\theta_i) a_i(\phi)$$

$$Cov(Y_i; \theta_i, \phi_i) = 0$$

   Matricialment l' especificació del model GLM quedarà definit com:

$$E(Y) = \mu$$

$$g(\mu) = X\beta$$

$$Var(Y) = Var(\mu) \cdot \phi$$

on μ és el valor esperat de la variable resposta i $g(\cdot)$ és la funció d'enllaç que especifica la relació entre μ i les covariables X. La funció d'enllaç que transforma la mitjana cap al paràmetre natural és l'anomena't enllaç canònic. Pel que fa a la $Var(Y)$ hi trobem dos termes: $Var(\mu)$, és la funció de variància associada a la distribució de les dades, i ϕ que permet recollir la possible sobredispersió o sotadispersió no recollida per a la funció de la variància.

En resum, un GLM s'especifica a partir de tres components:

- Un component aleatori que identifica la variable resposta Y i la seva distribució de probabilitat que pertany a la família exponencial.
- Un component sistemàtic que identifica les covariables utilitzades en una funció predictor lineal $\eta = X\beta$.
- Una funció d'enllaç que connecta $\mu = E(Y)$ amb el component sistemàtic.

En la següent taula es mostren les funcions anteriors en el cas que la variable resposta segueixi una distribució Binomial o de Poisson.

Taula 1.1: Model lineal generalitzat per a les distribucions Binomial i Poisson

	Binomial	Poisson
Mitjana	π	λ
$g(\mu)$	$\log[\pi(1-\pi)]$	$\log(\lambda)$
$a(\phi)$	$1/n$	1
$Var(\mu)$	$\pi(1-\pi)$	λ
$Var[Y]$	$\pi(1-\pi)/n$	λ
Enllaç canònic	logit	log

Per a dades normals l'enllaç canònic és l'identitat, obtenint el model lineal general clàssic. Per a dades Poisson tenim el logaritme, donant lloc als models log-lineals o regressió Poisson, i per a dades binàries l'enllaç canònic és la funció logit, que dóna lloc a la regressió logística.

L'estimació dels GLM

L'estimació dels paràmetres de regressió β en els GLM es pot obtenir mitjançant el mètode de la màxima versemblança. El valor de la log versemblança de l'individu i és: :

$$\ell(\theta_i; \phi_i, y_i) = \left\{ \frac{y_i \cdot \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i \phi_i) \right\}$$

I l'estimació dels paràmetres s'obtindrà igualant a zero la derivada de la log versemblança:

$$\ell(\theta) = \sum_{i=1}^{n} \frac{y_i \cdot \theta_i - b(\theta_i)}{\alpha} + c(y_i \alpha) = \sum_{i=1}^{n} \ell_i(\theta),$$

on $\theta = [\theta_1(\beta), \ldots, \theta_n(\beta)]$ és el vector de paràmetres canònics.

Degut a que la derivada de ℓ_i i β_i no és lineal, el logaritme de la funció de versemblança habitualment no es pot maximitzar directament. Aquesta derivada es pot solucionar mitjançant les relacions parcials que existeixen entre ℓ_i i β_i:

$$S(\beta) = \frac{\ell}{\beta} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\theta_i} \frac{d\theta_i}{\mu_i} \frac{d\mu_i}{\beta} = \sum_{i=1}^{n} \frac{Y_i - b'(\theta_i)}{\alpha} \frac{1}{V_i} \frac{d\mu_i}{d\beta}$$

on $V_i = Var(Y|\beta)/\alpha$ i $\frac{d^2 b}{d\theta_i^2} = \frac{d}{d\theta_i}\mu_i = V_i$

Per tant l'expressió final que s'obté,

$$S(\beta) = \sum_{i=1}^{n} (\frac{\partial \mu_i}{\partial \beta})^T \frac{Y_i - E[Y_i|\mu_i]}{var(Y_i|\mu_i)} = \mathbf{D}^T \mathbf{V}^{-1} \{\mathbf{Y} - \mu(\beta)\},$$

on D és la matriu de dimensió $\frac{\partial \mu_i}{\partial \beta_j}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ i V és la matriu diagonal amb ièssim elements $Var(Y_i|\mu_i)$.

Degut a que la mitjana $\mu = g^{-1}(X\beta)$ apareix sovint en l'expressió de la variància, per a l'estimació s'utilitzen mètodes iteratius. Alternativament es poden utilitzar aproximacions a partir de l'algoritme de Newton-Raphson o es pot utilitzar el mètode de mínims quadrats ponderats iteratiu (IRLS) [11].

En resum, un GLM és un model lineal per al valor esperat d'una variable resposta que té una distribució que pertany a una família exponencial.

1.4 Anàlisi de la supervivència

L'origen de l'anàlisi de la supervivència el trobem en les taules de vida en què es tabulaven els temps de mort d'una població en funció de l'edat. Les primeres aplicacions es troben en estudis actuarials [12], trobant més endavant per exemple en medicina que la majoria d'aplicacions són en oncologia tant per avaluar mortalitat com per altres esdeveniments com la recidiva d'un tumor [13, 14]. L'anàlisi de la supervivència, també coneguda com anàlisi del temps de fallada, té aplicacions més enllà de l'anàlisi de la mortalitat. Per exemple, en l'àmbit industrial es coneix com anàlisi de fiabilitat (*reliability analysis*), entre els economistes i els sociòlegs es coneix com anàlisi de durades i els demografs de vegades es refereixen a l'anàlisi de la història dels successos (*event history analysis*).

L'anàlisi de la supervivència és aquella part de l'estadística dissenyada per a l'anàlisi de dades que representen el temps (des d'un origen ben definit) fins a l'ocurrència d'un determinat esdeveniment. Hi ha models de probabilitat que poden recollir aquest comportament asimètric de les dades de temps com, per exemple, el models Log-Normal, Exponencial, Gamma i Weibull. No obstant, hi ha casos en que les dades tampoc s'ajusten prou bé a aquests models de probabilitat, i com a alternativa han estat proposats procediments no paramètrics. El tret més rellevant és però la censura que sovint presenten les dades de supervivència, és a dir, el temps fins a l'esdeveniment d'alguns individus s'observa incomplet. Aquestes dues característiques fan que per exemple la mitjana sigui un

mal descriptor de tendència central degut a l'asimetria i la presència de valors extrems. En canvi, la mediana és una mesura de tendència central més robusta.

A continuació explicarem els conceptes més bàsics i els models més utilitzats a l'anàlisi de la supervivència. A més a més es tractarà més amb detall les situacions amb truncament per l'esquerra i dades correlacionades, respectivament.

1.4.1 Conceptes bàsics

Es defineix T com el temps fins un succés d'interès, $\mathcal{E}$. El succés $\mathcal{E}$ generalment serà la mort, però també pot ser qualsevol altre com per exemple la curació d'una malaltia o recaiguda. Formalment T és una variable aleatòria no negativa corresponent a una població homogènia.

La distribució per a T queda caracteritzada per exemple per les següents funcions:

- Funció de supervivència (*survival function*): $S(t)$,
- Funció de risc (*hazard function*): $\lambda(t)$,

A l'hora de treballar amb dades de supervivència normalment els dos objectius que es volen dur a terme són:

- Tenir resums en taules o gràfics dels temps de supervivència pels individus de diferents grups,
- Conèixer l'efecte que tenen certes variables sobre el temps de supervivència.

Aquests dos objectius i el propi procés aleatori es descriu habitualment amb la funció de supervivència $S(t)$ i la funció de risc $\lambda(t)$.

La funció de supervivència

Aquesta funció es denota per S i correspon a la probabilitat que un individu sobrevisqui el temps t, per exemple t anys, és a dir, la probabilitat que el succés $\mathcal{E}$ succeeixi després de t anys: $S(t) = P(T > t)$, i està definida per $t \geq 0$. Les seves propietats bàsiques són:

- $S(0) = 1$ i $S(\infty) = 0$,
- $S(t)$ és una funció monòtona decreixent,
- Si T és continua, $S(t)$ és continua i estrictament decreixent.

La funció de risc

Quan T és una variable aleatòria continua, la funció de risc es defineix formalment com

$$\lambda(t) = \lim \frac{1}{\Delta t} \mathrm{P}[t \leq T < t + \Delta t | T \geq t].$$

Aquesta funció, que no dóna cap probabilitat ni és una funció de densitat, descriu el risc instantani de morir al temps t quan s'ha sobreviscut fins a aquest temps. Les seves propietats bàsiques són:

- $\lambda(t)$ és una funció no negativa,
- Es pot expressar en funció de la funció de densitat $f(t)$,

$$\lambda(t) = \frac{f(t)}{S(t)}$$

1.4.2 Censura

El temps de supervivència és censurat quan l'esdeveniment d'interès no s'ha pogut observar. Ens podem trobar amb diferents tipus de censura:

Censura per la dreta: La censura per la dreta és la més comú i ens la podem trobar en diferents estudis clínics, mèdics o epidemiologics on s'aplica estudis de supervivència. Es produeix quan el succés $\mathcal{E}$ no s'ha produït durant el temps d'observació de l'estudi, i per tant el temps observat és inferior al temps real de supervivència.

Les circumstàncies poden ser les següents:

1. Finalització de l'estudi en un moment predeterminat o finalització de l'estudi en un moment determinat (censura administrativa).
2. Es perd el seguiment d'alguns dels individus (*lost to follow-up*). Aquests individus són observats només durant part del període d'observació.
3. Pèrdua de seguiment d'alguns individus (*dropout*) degut a la no tolerància del tractament, abandonament, etc. El comportament d'aquests individus s'han de mirar amb molta cura, ja que, poden estar relacionats amb l'evolució de la malaltia o $\mathcal{E}$.
4. Es produeix l'event per una altra causa diferent a la d'interès. L'anàlisi per aquest tipus de situacions es coneix amb el nom d'anàlisis de riscos competitius (*competing risk analysis*).

Per formalitzar la versemblança, s'han de definir noves variables: sigui C la variable aleatòria dels temps de censura, llavors

$$Y = \min(T, C) \qquad \text{i} \qquad \delta = \begin{cases} 1 & Y = T \\ 0 & Y = C \end{cases}. \tag{1.2}$$

És a dir, Y és la variable aleatòria del temps observat i δ la variable indicadora de que es tracta d'un temps censurat o no.

Donada una mostra de dades independents $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ i suposant la independència entre T i C, la funció de versemblança és igual a

$$L(F) = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}. \tag{1.3}$$

Censura per l'esquerra: Censura que es produeix quan el temps fins l'esdeveniment $\mathcal{E}$ s'ha produït abans de l'inici del temps de seguiment, i per tant, el temps observat és superior al temps real de supervivència.

Censura en un interval: Censura que es produeix quan l'esdeveniment d'interès $\mathcal{E}$ no es pot observar exactament i solament se sap que ha succeït en un cert interval de temps, per exemple, entre visites mèdiques consecutives. El temps transcorregut entre dues visites consecutives depèn no solament del protocol establert pel pacient, sinó també del possible no compliment de les mateixes per causes diverses: el pacient viu lluny de l'hospital i decideix saltar visites establertes o per exemple el pacient es troba malament i decideix demanar una visita extra.

Censura doble: Censura que es produeix quan en un mateix estudi hi ha dades censurades per la dreta, per l'esquerra i no censurades. També es parla de censura doble quan també el punt d'origen està censurat.

1.4.3 Estimador de Kaplan i Meier

L'estimador de Kaplan i Meier és utilitzat per a l'estimació no paramètrica de la funció de supervivència $S(t)$ [15]. També és coneix com l'estimador del límit del producte i està basat en la mateixa idea que l'estimador actuarial, és a dir, en la descomposició de la funció de supervivència en un producte de probabilitats condicionades. La diferència essencial amb el mètode actuarial està en el fet que els intervals $I_1, \ldots, I_k$ són ara al.leatoris.

Sigui $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ una mostra de les variables definides en (1.2) i suposem que els temps de censura C no estan relacionats amb els temps potencials de fallada T, és a dir que la censura és no informativa [16]. A més a més, siguin $Y_1, \ldots, Y_n$ ordenats del temps menor al temps màxim. Llavors, la supervivència estimada en un instant t (tenint en compte que

hi ha censura) es pot expressar com:

$$\hat{S}(t) = \prod_{j:y_j \leq t} \hat{p}_j$$

on $\hat{p_j} = P(T > y_j \mid T > y_{j-1})$ és la supervivència en el moment y_j d'aquells individus vius en el moment y_{j-1}.

Quan y_j correspon a un temps de mort, la probabilitat p_j s'estima mitjançant la proporció d'individus amb una vida superior a y_j sabent que en el moment y_{j-1} estaven vius:

$$\hat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \frac{d_j}{n_j}$$

on n_j és el nombre d'individus vius just abans de y_j i on d_j és el nombre d'individus que moren en el moment y_j.

Quan en el temps y_j únicament hi ha hagut una censura el valor de d_j és igual a zero i per tan la probabilitat condicionada de superar l'interval j és igual a 1.

Definició: Es defineix l'**estimador de Kaplan-Meier** com:

$$\hat{S}(t) = \prod_{j:y_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{j:y_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

amb variància (segons formula de Greenwood)

$$\mathrm{Var}_G\big(S(t)\big) = \big(S(t)\big)^2 \sum_{i:y_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

Les propietats més importants de l'estimador de Kaplan i Meier són les següents:

1. Si no hi ha censura, es redueix a la funció de supervivència empírica.

2. Sota certes condicions de regularitat l'estimador de Kaplan i Meier és l'estimador de màxima versemblança no paramètric generalitzat.

3. L'estimador de Kaplan i Meier és consistent, és a dir, asintòticament

no té biaix i la variància tendeix a zero.

Una manera habitual de representar l'estimador Kaplan i Meier és gràficament tal com es mostra en la Figura 1.1, que fa referència a la supervivència en pacients amb càncer de pulmó avançat de l'estudi de Loprinzi CL et al a l'any 1994 [17].

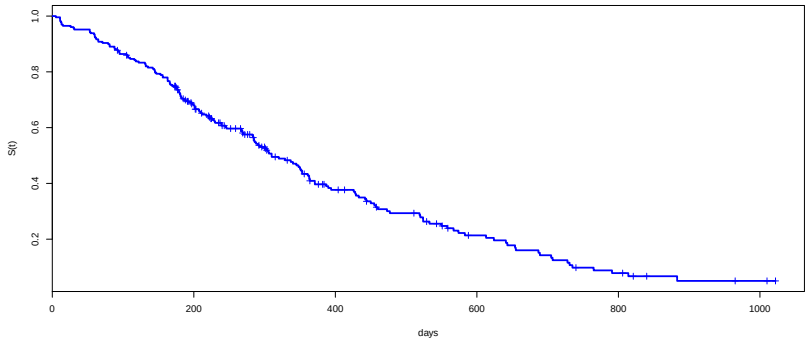


Figura 1.1: Estimació de la Supervivència en pacients de càncer de pulmó de l'estudi de Loprinzi (1994) mitjançant l'estimador de Kaplan i Meier

.

Gràficament l'estimador de Kaplan i Meier és una funció esglaonada que només salta quan es produeix un esdeveniment. Quan la darrera dada està censurada l'estimador de Kaplan i Meier no baixa fins a zero i $S(t)$ no està definit més enllà d'aquesta observació. Cal comentar que l'estimador de la mediana, M, és el valor més petit M pel qual l'estimador Kaplan i Meier és menor o igual a 0.5.

Per a la comparació de dues funcions de supervivència, és a dir per contrastar l'hipòtesi de la igualtat de la supervivència de dues o més mostres existeixen diferents proves no-paramètriques com la prova de log-rank, la prova de Breslow o la de Gehan.

$$H_0 : S_1(t) = S_2(t), \forall t \quad \text{vs.} \quad H_1 : S_1 \neq S_2(t) \text{ per algun } t \leq T \qquad (1.4)$$

Aquesta última prova de Gehan per exemple, posa més pes a les observacions més petites i degut això és més potent per detectar els efectes

de riscos a curt termini. La prova de log-rank posa el mateix pes en cada observació, i per tant és més sensible a exposicions amb un risc relatiu constant. La prova de log-rank té una potencia òptima per detectar diferències quan les funcions de risc són proporcionals. Aquestes dues proves permeten concloure no solament si l'exposició té algun efecte sinó també la naturalesa d'aquest efecte: a curt o llarg termini.

1.4.4 El model de Cox o riscos proporcionals

En molts treballs d'investigació mèdica és molt comú que en cada pacient a més a més de ser observat el temps de supervivència siguin observades les característiques clíniques dels pacients, anomenades covariables. Si l'interès és determinar l'efecte de les covariables en el temps de supervivència, l'objectiu de l'estudi es centra en l'anàlisi de les relacions entre el temps de supervivència i les covariables mitjançant un model de regressió.

Els models paramètrics basats en diferents distribucions (Weibull, Exponencial, Log-Normal) són eficients quan es té informació del model de distribució subjacent a les variables i només falta per determinar un nombre finit de paràmetres. Tot i així, una font d'error pot ser triar una família paramètrica no adequada. En aquests casos podem utilitzar els models semiparamètrics o no paramètrics que a més a més de permetre graduar les probabilitats brutes que no segueixen un model paramètric establert, poden utilitzar-se per proporcionar una prova de diagnòstic dels models paramètrics o simplement per explorar les dades.

En l'anàlisi de supervivència, el model de regressió semiparamètric més freqüentment utilitzat és el model de riscos proporcionals de Cox [18], anomenat generalment com a model de Cox.

El model de Cox permet investigar la relació entre un conjunt de variables pronòstiques i el temps de supervivència. Els treballs de Cox en general, i en particular el treball **Regression Models and Life Tables** publicat l'any 1972 en el *Journal of the Royal Statistical Society* (Series B), són els pioners d'aquesta metodologia [19].

Recordem que s'indica per T el temps fins a produir-se el succés $\mathcal{E}$ i per $X_1, \ldots, X_n$ les variables explicatives, o valors pronòstics, recollides a l'origen de l'estudi.

Definició: El model de Cox estableix la següent relació entre la funció de risc $\lambda(t|X)$ en el moment t d'un individu amb perfil $X = (X_1, \ldots, X_n)$ i la funció de risc basal en el mateix moment t:

$$\lambda(t|X) \equiv \lambda_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n\right). \tag{1.5}$$

En aquest model se suposa que la raó entre les funcions de risc es manté constant al llarg del temps, ja que es verifica:

$$\frac{\lambda(t|X)}{\lambda_0(t)} = \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n\right),$$

on el terme de la dreta solament depèn dels valors de les covariants i no del temps t. Aquest terme correspon al ***hazard rati*** en el moment t d'un individu amb perfil X respecte d'un individu amb $X = 0$.

El logaritme del *hazard rati* ve expressat per la coneguda equació de regressió lineal múltiple, i és un valor constant, independentment d'en quin moment del temps es calcula, és a dir:

$$\ln\left(\frac{\lambda(t|X)}{\lambda_0(t)}\right) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n.$$

A la quantitat $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$ se la coneix com l'índex pronòstic o *risk score*. Aquest índex pot utilitzar-se per comparar per exemple dos tractaments o comparar el pronòstic de pacients amb diferents nivells de les variables explicatives. En aquest model cal destacar que no hi ha terme constant, ja que, aquest ho absorbeix $\lambda_0(t)$.

Algunes propietats del model de Cox són:

1. El logaritme del quocient de les funcions de risc es relaciona linealment amb els factors pronòstics, mentre que en una regressió lineal és la mesura de la variable resposta la que es relaciona linealment amb els factors pronòstics.

2. El logaritme del quocient de les funcions de risc, així com el quocient de les funcions de risc, se suposa constant independentment del moment del temps en que es calcula. Aquest motiu és pel que es coneix com a **model de riscos proporcionals**.

3. La funció de risc basal, λ_0, pot prendre qualsevol forma; per aquest motiu es diu que el model de Cox es de naturalesa **semiparamètrica**.

4. Si $p \doteq \lambda(t|X')$ representa la probabilitat de morir en el interval $(t, t+\triangle t)$ donat que està viu a t, i suposem que p és molt petit $p/1-p \simeq p = \lambda(t|X')$, aleshores el model de Cox podria interpretarse com un model de regressió logística:

$$\ln \frac{p}{1-p} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n.$$

L'estimació dels paràmetres del model de riscos proporcionals requereix els següents dos passos:

1. L'estimació dels coeficients (paràmetres) de les variables explicatives $\beta_j, j = 1, \ldots, p$, mitjançant la maximització de la versemblança parcial. Els estimadors s'escriuen $\hat{\beta}_j, j = 1, \ldots, p$.

2. L'estimació de la funció de risc basal $\lambda_0(t)$ mitjançant la funció de versemblança condicionada als valors estimats $\hat{\beta}_j, j = 1, \ldots, p$.

La manera estàndard d'estimar els coeficients β_j és mitjançant la maximització de l'anomenada funció de versemblança parcial $L(\beta)$, que només te en compte l'ordre dels valors de supervivència observats, no el seu valor en sí.

Finalment destacar que la interpretació del model de Cox no es fa directament mitjançant els seus coeficients estimats sinó del seu exponencial: $\exp(\hat{\beta}_j), j = 1, \ldots, n$. L'interpretació per a covariables dicotòmiques és diferent que per a variables continues. Per a cada covariable dicotòmica, $\exp(\hat{\beta})$ és un estimador del *hazard* rati i s'interpreta com a la quantitat de risc que es té amb la presència de la covariable amb relació a l'absència de la covariable. Per al cas de covariables continues, $\exp(\hat{\beta})$

representa la raó de riscos (el *hazard rati*) al incrementar en una unitat la covariable.

1.4.5 Truncament

El truncament succeeix quan només els pacients amb el temps de fallada dins un cert interval de temps són observats. Els pacients els quals el seu esdeveniment s'ha produït fora d'aquest temps (Y_L, Y_R) no són observats i l'investigador no sap de la seva existència ni té dades sobre aquests. En aquest cas, tota la inferència haurà de fer-se condicionada a aquesta condició. Aquest fet obliga a analitzar anàlisis diferents dels que es fan amb presència de censura, donat que un individu censurat informa sobre el temps de fallada, i en canvi no coneixem ni tan sols l'existència dels individus truncats.

Truncament per l'esquerra: El truncament pot ser per l'esquerra quan solament els individus amb un cert temps de supervivència entren en l'estudi i per la dreta quan solament els individus que han experimentat l'esdeveniment abans un cert temps són observats.

Quan Y_R és infinit, parlem de truncament per l'esquerra, aquí els únics individus observats són aquells amb temps de l'esdeveniment més gran que Y_L. És a dir, X és observat si, i només si, $Y_L < X$.

El truncament per l'esquerra es dóna quan únicament els individus que compleixen el requisit determinat, amb anterioritat a l'esdeveniment d'interès, s'inclouen en l'estudi. També ens referim a dades truncades per l'esquerra com a dades amb entrades retardades. Si per exemple Y és el temps de truncament, només els individus amb $T \geq Y$ són observats. El truncament per l'esquerra no s'ha de confondre amb la censura per l'esquerra. És important mencionar que la variable del truncament ha de ser igual que la variable del temps d'interès.

Per a l'estimació amb dades truncades per l'esquerra per a cada individu es disposa d'un temps aleatori d'entrada a l'estudi L_j i d'un temps T_j on a l'individu es produeix l'esdeveniment o és censurat. S'entén els diferents temps de fallada per $t_1 < \ldots < t_D$ i per d_i el nombre d'individus

que els hi succeeix l'esdeveniment en el moment t.

El nombre d'individus a risc en el moment t_i es defineix com el número d'individus que han entrat a l'estudi abans de t_i i el seu temps de que es produeix l'esdeveniment o censura és posterior a t_i. L'individu no es considera a risc en aquells temps anteriors al temps en què entra. Per tant, al contrari que en els estudis sense truncament per l'esquerra, el nombre d'individus a risc és una quantitat dinàmica que no té perquè disminuir amb el pas del temps.

Una de les principals diferencies amb el procediment estàndard és la seva interpretació, ja que, en realitat tot queda condicionat a $T \geq L$ sent L el més petit dels temps d'entrada. La supervivència que s'estima és en realitat un estimador de $P(T > t|T \geq L)$.

L'estimador de Kaplan i Meier per a la supervivència condicionat al haver viscut L anys ve donat per

$$\hat{S}_L(t) = \prod_{i:L \leq t_i \leq t} \Big(1 - \frac{d_i}{n_i}\Big).$$

Hi ha certes precaucions que s'han de prendre abans de calcular els estimadors [20]:

- Pot succeir que en un moment donat el nombre d'individus a risc sigui zero (o molt petit), aleshores l'estimador de Kaplan i Meier serà zero després d'aquest moment encara quan hi hagi individus supervivents en aquesta data. En aquests casos el més sensat seria estimar la supervivència condicionada a un temps en el que això ja no passi.

- Si es considera que el temps és l'eix temporal, es tindrà en compte directament l'efecte del temps sobre la mortalitat sense necessitat d'ajustar pel mateix.

1.5 Limitacions dels models descrits i la necessitat de l'ús de models més sofisticats

Fins ara, hem vist els models més utilitzats coneguts en l'àmbit de les ciències de la salut, com són els models LM i GLM. En el model lineal general se suposen les premisses següents: linealitat, normalitat, homocedasticitat (variància constant) i independència. A mesura que no es compleix alguna d'aquestes premisses ens trobem amb diferents models. (Figura 1.2).

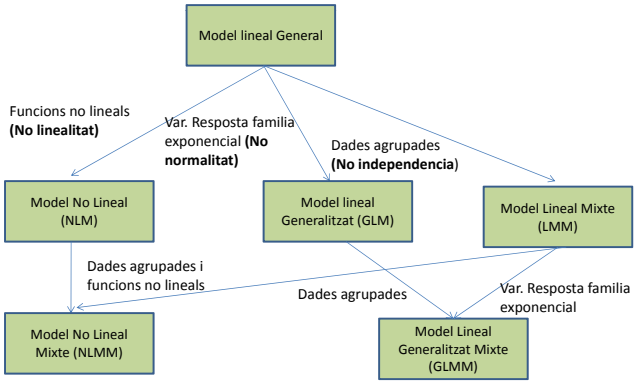


Figura 1.2: Model lineal General i les seves premises. Font: Josep Anton Sánchez Espigares

Els LM no són una metodologia apropiada per a variables amb resposta no continua (per exemple binària o recompte), degut a que les dades no compleixen les assumpcions comunes (residus homocedastics, normalitat de la distribució de la probabilitat de la resposta, linealitat de la relació entre l'esperança de la resposta i covariables). En dissenys experimentals habituals, quan la variable resposta és discreta i s'assumeix

que les observacions són independents, estaríem davant dels coneguts models GLM [10]. En ciències de la salut els GLM més utilitzats són aquells que modelen recomptes o dicotomies, assumint una distribució de la família exponencial com la de Poisson, Binomial o Bernoulli [10, 21]. Els GLM assumeixen habitualment que les observacions condicionades als predictors són independents i idènticament distribuïdes. No obstant, aquestes assumpcions poden ser violades en algunes situacions, com en els estudis longitudinals on hi ha mesures repetides i les dades poden aparèixer correlacionades. Això, comporta que els mètodes inferencials per estimar i resoldre contrasts d'hipòtesis sobre els paràmetres poden donar resultats erronis [22].

Com a alternativa davant aquestes assumpcions, hi ha els models mixtes o models de regressió amb efectes aleatoris que són models que tracten aquesta correlació entre les observacions. Els models mixtes són models que tenen tan efectes fixes com aleatoris, i habitualment es poden trobar exemples o casos particulars amb diferents noms (models amb coeficients aleatoris, *clustered models*, models jeràrquics (*hierarchical models*), models multinivell (*multilevel models*), *cluster-specific models*, models amb mesures repetides, models de component de la variància (*variance component models*), models de dades de panell (*panel data*) segons l'àmbit d'aplicació) [23–26].

En la present tesi descriurem diferents tipus de models amb efectes aleatoris: els models lineals mixtes (LMM), els models no lineals mixtes (NLMM), els models lineals generalitzats mixtes (GLMM) i els models de Cox amb efectes aleatoris, també anomenats *frailty models* [22]. Els LMM són una extensió dels models LM on s'afegeix l'efecte aleatori en el predictor lineal. Els NLMM són models semblants als LMM però la funció d'enllaç de la variable resposta no és lineal, hi ha una dependència no lineal en els paràmetres, entre la part fixa i part aleatòria. Per últim, estaríem davant dels GLMM que serien una extensió que combina els models GLM i els LMM.

1.6 Hipòtesis, objectius i estructura de la tesi

L'objectiu general d'aquesta tesi doctoral es centra en l'estudi i descripció de models d'efectes aleatoris en problemes d'investigació quantitativa, i majoritàriament aplicats a l'epidemiologia.

En el Capítol 2 s'introdueixen els models amb efectes aleatoris. Als darrers anys s'ha observat un increment de l'ús dels models lineals generalitzats mixts (GLMM) en articles de medicina clínica. A l'apartat de mètodes d'aquests articles moltes vegades hi ha manca d'informació rellevant (mètode d'estimació, algoritme, *software*,...) respecte els models GLMM al no haver-hi una pauta o guia del què cal informar sobre aquests models. En aquest capítol s'inclou un article on es recull una revisió del grau d'aplicació dels GLMM i s'avalua també la qualitat de la informació aportada en relació amb l'anàlisi amb GLMM en articles originals en el camp de la medicina clínica. L'objectiu de l'article és comprovar l'ús i la qualitat de la informació proporcionada dels GLMM en la recerca clínica.

El Capítol 3 inclou un article on es descriu l'estimació dels paràmetres del model Poisson GLMM en el cas de recomptes a través de tres filosofies estadístiques. La filosofia, mètode d'estimació, algoritme o *software* utilitzats en un estudi poden influir en l'estimació dels paràmetres. A més a més, cal tenir present alhora d'estimar els paràmetres el grau de sobredispersió, el tamany mostral i la mitjana marginal. L'objectiu de l'article ha estat comparar el rendiment de tres filosofies diferents d'un GLMM mitjançant estudis de simulació i amb dades reals sobre lesions d'un esport de contacte. En aquest capítol s'inclou també com a exemple de la utilitat dels GLMM un altre article on s'ajusten les lesions d'un esport de contacte. D'aquesta manera es pretén conèixer les associacions amb factors potencials de risc de forma que es pugui ampliar el coneixement sobre les possibles causes de lesions.

El Capítol 4 introdueix els frailty models centrant-se en els models de supervivència semiparamètrics amb efectes aleatoris. També s'inclou un article on s'investiga la incidència de malalties infeccioses i la supervivència en una cohort d'ètnia gitana en un barri marginal durant 23 anys

de seguiment. Donada la insalubritat i marginalitat del barri, és molt probable que les malalties infeccioses s'haguessin seguit succeint. Pel fet de treballar amb una població molt determinada, jove i amb unes condicions de marginalitat pronunciades, es preveu que utilitzant l'edat com a variable objectiu es pot fer una interpretació més acurada des del punt de vista metodològic. A més a més, el fet que convivien diferents famílies a cada barraca fa pensar que les estimacions són més acurades tenint en compte la correlació en les dades a nivell familiar al ajustar un model *frailty model* en un cas real d'epidemiologia amb malalties infeccioses i truncament per a l'esquerra. L'objectiu de l'article ha estat analitzar la incidència de malalties infeccioses i la supervivència ajustada per edat i sexe tenint en compte que es tractava d'una població molt determinada. En aquest mateix capítol s'exposen línies de recerca futura fent ús del cas real mencionat anteriorment.

Finalment, en el Capítol 5 es troben les principals conclusions derivades dels resultats de la tesi.

2 Models de regressió amb efectes aleatoris

2.1 Introducció als efectes aleatoris, dades agrupades i mesures repetides

La modelització estadística és una eina que rep una gran atenció per predir resultats i avaluar també l'associació entre ells i els seus factors de risc. Per tant, un dels reptes importants que ens trobem en els models estadístics és testar eficientment les hipòtesis d'investigació eliminant el biaix de les estimacions. A més a més, cal tenir en compte les fonts de variabilitat de les dades que freqüentment ens condueixen a trobar dades que presenten estructures jeràrquiques o també conegudes com a multinivell o tipus cluster. En aquests models és comú l'ús de conceptes com els efectes aleatoris, dades correlacionades i mesures repetides que tot seguit s'explica amb més detall.

2.1.1 Efectes aleatoris

Els efectes aleatoris són atribuïbles a un conjunt infinit de nivells dels factors, dels quals només una mostra aleatoria són considerats en les dades utilitzades [27]. D'aquesta manera no ens interessen els nivells

concrets dels factors utilitzats en l'anàlisi sinó l'efecte genèric d'aquests factors.

En el model d'efectes fixes l'experimentador decideix quins nivells concrets es van a considerar i les conclusions que s'obtinguin només són aplicables a aquests nivells, no podent fer extensives a altres nivells no inclosos en l'estudi. En canvi, en el model d'efectes aleatoris, els nivells del factor se seleccionen a l'atzar i les conclusions que s'obtinguin es generalitzen a tota la població de tots els possibles nivells del factor, hagin estat explícitament considerats en l'estudi o no.

En resum, per tal de decidir l'ús d'efectes fixes o aleatoris els criteris principals han de ser l'estructura de les dades i la finalitat de l'anàlisi. Quan les dades són heterogènies és degut a la variabilitat natural associada als nivells del factor aleatori, i dins de cada nivell les observacions normalment tenen certa homogeneïtat. Les observacions que apareixen dins d'un mateix nivell mostrarien una lleugera variabilitat a causa de l'error pròpiament experimental o variabilitat entre les subrèpliques.

Els efectes aleatoris recullen la variabilitat entre individus però poden no tenir un interès directe per a l'investigador i aleshores en aquest cas, podrien no ser reportats. Tot i així, els efectes aleatoris han de ser examinats. Per exemple, si l'efecte aleatori és molt petit hi hauria l'opció de tenir-lo en compte, ja que, podria millorar l'estimació dels paràmetres del model. Els efectes aleatoris són utilitzats majoritàriament per a diferents raons. Per exemple, poden ser una de les causes de variació en les dades o pot haver un interès directe per a l'investigador per obtenir resultats empírics i un model més parsimoniós. Una altra raó d'utilitzar els efectes aleatoris seria quan l'atenció es centra en estimar la quantitat de variació. Un clar exemple d'això el trobem en estudis de genètica, en què la variància genètica additiva és important (per exemple, en l'estimació de la velocitat de canvi fenotípic degut a la selecció), i els valors estimats per als individus (els valors d'una cria, 'breeding values') sovint són menys importants. L'enfocament es basa doncs en estimar on succeeix la variació en les dades [28].

2.1.2 Dades agrupades

Definim dades agrupades a las observacions que s'agrupen en grups o conglomerats (*clusters*). Cada *cluster* representa un nivell del factor aleatori. Per exemple, pacients agrupats en hospitals, en els quals interessa conèixer quines característiques del pacient afecten la variable analitzada o també quines característiques del grup (en aquest cas l'hospital) afecten també a aquest resultat. Un exemple podria ser un estudi per analitzar quins factors de risc s'associen amb hipertensió en pacients diabètics en atenció primària, en el qual s'inclouen centres amb diferents característiques. Cada centre aporta inicialment el mateix nombre de pacients a l'estudi, seleccionats de forma aleatòria. És raonable pensar que amb dades agrupades les observacions pertanyents al mateix grup són en general més similars entre si que respecte a les d'altres grups, el que violaria la condició d'independència entre les observacions. Així, en aquest exemple pot passar que els pacients que atén un dels centres siguin tots ancians, un altre centre bàsicament subjectes aturats per trobar-se en una zona de població amb alta taxa d'atur i altres centres amb diferents tipus de pacients.

Identificada la covariable causant de dades correlacionades, una possible solució a l'incompliment de les suposicions és afegir la covariable dins del model d'efectes fixes. Una altre alternativa davant la correlació d'observacions seria incorporar la informació associada a la variable latent a través de la modificació de l'estructura de correlacions. En aquestes situacions és important tenir present la utilització de models amb efectes aleatoris.

2.1.3 Mesures Repetides

Entenem com a mesura repetida quan una mateixa variable es mesura en diferents ocasions ja sigui al llarg del temps o en diferents condicions de mesura.

Els estudis en què apareixen mesures repetides normalment com-

porten només un nivell d'agrupació (*cluster*) on les mesures repetides són intercanviables. En canvi, en dissenys multinivell (explicat amb més detall en el següent apartat) es presenten varis nivells de grups (*clusters*), donant lloc a una estructura jeràrquica en cada grup, com es pot observar en estudis longitudinals o mesures repetides. Cal tenir present que les mesures repetides, i en concret el concepte de rèpliques, són transversals i no hi ha efecte temps o ordre.

S'han descrit diferents estratègies per analitzar mesures repetides si ens trobem davant dades longitudinals, com per exemple anàlisis separades per a cada moment en el temps o procediments per a efectes aleatoris. El mètode més senzill per analitzar mesures repetides seria incloure simplement el temps com a un factor, i ignorar la dependència entre dues observacions sobre el mateix individu. Tan atractiu ens pot semblar aquest enfocament que pot dur a conclusions completament errònies. L'essència del problema és que es com si es pretengués tenir més observacions de les que hi ha disponibles. Dues observacions correlacionades contenen menys informació que dues observacions independents, perquè un s'explica en part per l'altra. En els dissenys de mesures repetides sembla just esperar que dues mesures d'un mateix (*cluster*) tingui una correlació positiva, el que donaria lloc a mesures més similars de dues mesures de diferents individus. A més, dues mesures realitzades en el mateix individu poden estar altament correlacionades si es mesura en dos punts del temps a prop un de l'altre, però menys correlacionades (o potser independentment) si es mesuren lluny.

2.2 Consideracions prèvies

Hi ha alguns conceptes relacionats amb la literatura dels models amb efectes aleatoris, que es fa necessari exposar en aquest punt del treball, a fi d'aclarir algunes nomenclatures utilitzades que més endavant, al llarg d'aquesta tesi, en farem referència.

- Factor o efecte fix: Factor en què els seus nivells experimentals

són tots els possibles nivells observables o que han estat fixats per l'investigador. Dins dels efectes fixes hem de distingir dos tipus de covariables, les covariables tipus *outer* i les tipus *inner*. Les covariables *outer* que no són covariables principals és un tipus de variable en què el seu valor és constant en cada unitat experimental. En contrapartida, a les covariables en què el seu valor varia dins de cada grup són covariables tipus *inner*. Per tant, les covariables tipus 'outer' són independents del temps, afecten només a la component de la variància de l'efecte aleatori.

- Factor o efecte aleatori: Com s'ha dit abans és un factor en què els nivells són una mostra de tots els possibles nivells. El factor té molts nivells possibles (idealment infinits), l'interès està en tots els nivells possibles, però només una mostra aleatòria dels nivells estarà inclòs en les dades. La variable resposta s'observa en més d'una ocasió en cada nivell del factor aleatori. Per exemple, en un assaig clínic, el nombre de visites seria un factor aleatori.

- *Crossed random effects*: Múltiples efectes aleatoris que s'apliquen de forma independent a un individu, com per exemple els efectes temporal i espacial en el mateix disseny, on la variabilitat temporal actua igualment sobre tots els grups espacials.

- *Individual random effect*: els efectes que s'apliquen a nivell de cada individu (és a dir, blocs de mida 1).

- *Nested random effects*: múltiples efectes aleatoris que són jeràrquicament estructurats, com per exemple pacients en diferents hospitals, participants en diferents països, individus dins la mateixa família, etc. En aquests dissenys s'utilitzen també els anomenats *Nested models*, coneguts majoritàriament com a models multinivell que són subconjunts d'un model més complex.

- Sobredispersió: l'ocurrència de més variància en les dades que en les predites mitjançant un model estadístic. Aquest fenomen serà explicat detalladament al proper capítol.

- Mesures repetides, dades agrupades i dades longitudinals: És important diferenciar els tres conceptes. En **dades de mesures repetides**, la variable dependent es mesura més d'una vegada per a cada subjecte. Un cas particular de dades de mesures repetides són les dades longitudinals. En les **dades longitudinals**, hi ha mesures repetides al llarg del temps, ja sigui, modificant o no les condicions experimentals. Les mesures repetides són identificades per l'instant o moment (temps) de mesura (visita, data, etc). En dades longitudinals, la variable dependent es mesura en diferents moments del temps per a cada subjecte, normalment més d'un període relativament llarg de temps. Per últim, en **dades agrupades** o *clustered data* la variable dependent es mesura una vegada per a cada característica, però els mateixos subjectes d'alguna manera estan agrupats (ex: pacients agrupats en hospitals). No hi ha ordre als subjectes dins del grup, per la qual cosa les seves respostes han de ser igualment correlacionades.

- Dades espacials: Mesures repetides en diferents ubicacions de l'espai. Les mesures repetides són identificades per la seva ubicació física o geogràfica.

- Dissenys balancejats i no balancejats: Quan el nombre de mesures és el mateix per a tots els individus o subjectes i aquestes mesures són equidistants al llarg del temps en un disseny longitudinal, podem dir que tenim dades balancejades. En altres casos, parlaríem de dades no balancejades. Les dades no balancejades no han de ser un problema, excepte si són extremes (algunes combinacions 'missings' en els efectes fixes, o tots zeros o tots uns en alguns efectes aleatoris, etc). El fet de treballar amb dissenys balancejats o dissenys amb efectes ortogonals comporta un anàlisi i interpretació més simple. Tot i així, en dades reals i en els dissenys d'experiments comuns ens trobem amb molts dissenys no balancejats que depenen de l'estructura de covariància que millor s'ajusta a les dades, i que comporta treballar amb metodologies estadístiques més complexes.

A continuació es mostra l'especificació i característiques d'alguns models mixtes.

2.3 Models Lineals Mixtes

Els models LM i GLM vistos fins ara, són models amb efectes fixes. L'especificació clàssica dels models amb efectes fixes és $g(\mu) = X\boldsymbol{\beta}$ on s'incorpora només el vector β dels efectes fixes [25, 29]. Els models lineals mixtes (LMM) són models en els quals tant la part fixa com la part aleatòria tenen una relació lineal entre els paràmetres i la resposta. En els LMM la interpretació de l'estimació dels components de la variància és senzilla, ja que, els efectes aleatoris operen en la mateixa escala que els valors de la variable resposta. En general, els efectes aleatoris representen a diferents famílies o grups/estrats (centres hospitalaris, les regions geogràfiques com comarca, província, municipi, districte, secció censal) i aquests s'assumeix que són independents. Imaginem el model amb un únic efecte aleatori.

L'especificació del LMM és definit com:

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i \tag{2.1}$$

on $i = 1, \ldots, n$ és l'identificador de l'observació i, $\epsilon_i \sim \mathcal{N}(0, R)$ correspon als errors aleatoris o terme residual, i $u_i \sim \mathcal{N}(0, G)$ és el vector que correspon als efectes aleatoris de llargada q. Els dos termes se suposen independents i les característiques dels altres components són:

- β són els efectes fixes (paràmetres) de llargada p
- X_i (dimensió $n_i \times p$) Z_i (dimensió $n_i \times q$) són les matrius de disseny dels efectes dixes i aleatoris
- G és la matriu de covariàncies de l'efecte aleatori de dimensió $q \times q$
- R és la matriu de covariàncies dels errors

Estimació del model

El LMM és un model condicional representat com $Y_i|u \sim \mathcal{N}(X_i\beta + Z_iu, R)$ on s'assumeix normalment que $u \sim \mathcal{N}(0, G)$. Els paràmetres relacionats en la modelització del valor esperat correspondria als efectes fixes, i els paràmetres relacionats en la modelització de l'estructura i propietats de la variancia correspondria als efectes aleatoris.

La distribució marginal d'Y_i a l'equació anterior ve donada per $Y_i \sim \mathcal{N}(X_i\beta, V = Z_i^T G Z_i + R)$ on la matriu V descriu la variància marginal de la resposta. Podem comprovar per exemple que si $G = 0$, $R = I_n\sigma_\epsilon^2$, aleshores estem davant d'un LM d'efectes fixes. Els paràmetres d'interès que es volen estimar són els coeficients β i els components de la variància són representats per $\alpha = (\sigma_u^2, \sigma_\epsilon^2)$, a partir de la funció de versemblança marginal $\ell(\beta, \alpha)$.

La funció de versemblança és la funció de densitat conjunta de les dades observades en funció dels paràmetres del model. La funció de versemblança cobreix models multiparamètrics donant lloc a versemblances multidimensionals sofisticades que de vegades poden ser difícils de descriure o tractar. Els LMM avaluats en aquesta secció estarien en aquest tipus de situació, ja que, inclouen paràmetres d'efectes fixes o aleatoris. Com que els efectes aleatoris no són observats directament, apareixen certes dificultats en el maneig de la funció de versemblança conjunta i per tant és més complexa integrar la funció de versemblança conjunta. El fet que l'interès principal es centra en els paràmetres fixes s'utilitza la versemblança marginal o integrada, de forma separada (integrant) els efectes aleatoris. Quan la variable resposta segueix una distribució normal es pot obtenir la versemblança integrada de forma exacte, però no és així en el cas amb respostes no normals, que s'explicarà amb més detenciment en subapartats posteriors.

La versemblança marginal es definida amb la següent expressió:

$$L(\beta, \alpha|y) = f_Y(y|\beta, \alpha) = \int_u f_{Y|u}(y|u, \beta, \alpha) f_u(u|\alpha) du \qquad (2.2)$$

on $f_Y(y|\beta,\alpha)$ és la funció de densitat conjunta de Y, $f_{Y|u}(y|u,\beta,\alpha)$ la funció de densitat condicionada d'Y donada u, i $f_u(u|\alpha)$ la funció de densitat de l'efecte aleatori u.

Els paràmetres β són estimats mitjançant el mètode de màxima versemblança (ML) [29]. La log versemblança ve expressada com:

$$\ell_{ML} = \frac{-1}{2}\log|V(\alpha)| - \frac{n}{2}\log(2\Pi)\frac{-1}{2}(Y - X\hat{\beta}_{ML}(\alpha))V^{-1}(Y - X\hat{\beta}_{ML}(\alpha)) \tag{2.3}$$

Els components de la variància poden ser estimats mitjançant el mètode de ML però aquest proporciona resultats amb biaix, ja que, no es contempla la pèrdua d'informació (graus de llibertat) atribuïble al fet d'estimar β. Per avaluar la versemblança no es necessita conèixer ni els efectes aleatoris ni β. Només es necessita conèixer la norma dels residus del problema dels mínims quadrats que s'obté utilitzant una descomposició QR. L'estimació màxim versemblant té el problema de sotsestimar la variància residual i com alternativa, aquest biaix és corregit mitjançant el mètode de la màxima versemblança restringida (REML), que no modifica el mètode d'estimació dels β i consisteix en maximitzar per α la següent log-versemblança:

$$\begin{aligned}\ell_{REML} =& \frac{-1}{2}\log|V(\alpha)| - \frac{n-p}{2}\log(2\Pi)\frac{-1}{2} \\ & \cdot \log|X'V^{-1}(\alpha)X|\frac{-1}{2}(Y - X\hat{\beta}_{ML}(\alpha)'V^{-1}(\alpha)(Y - X\beta_{ML}(\alpha)).\end{aligned}$$

Aquesta maximització es dóna a terme sovint mitjançant el mètode de Newton-Raphson [30]. Un cop realitzada la maximització, ja es pot procedir a estimar els paràmetres β i u. L'estimació de β per la ML ve donada per la següent expressió:

$$\hat{\beta}_{ML} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y$$

Si G i R són conegudes, $\hat{\beta}$ és el millor estimador no esbiaixat (BLUE) de β, i $\hat{u}$ és el millor predictor lineal no esbiaixat (BLUP) de u. Com que G i R són estimades, aleshores es diu que β és un estimador EBLUE i u EBLUP, on la E inicial correspon al terme empíric.

A la pràctica de l'anàlisi de dades experimentals V normalment és desconeguda i es reemplaça pel seu estimador $\hat{V} = Z\hat{G}Z' + R$. Si es pot assumir que u i ϵ tenen distribució normal, la millor aproximació per a la estimació s'aconsegueix amb mètodes basats en ML. Els mètodes d'estimació més utilitzats són ML i REML.

Per a l'estimació de la matriu de covariàncies del vector $(\hat{\beta} - \beta, \hat{u} - u)$ tenim:

$$\begin{pmatrix} (X'\hat{R}^{-1}X) & (X'\hat{R}^{-1}Z)^{-1} \\ (Z'\hat{R}^{-1}X) & (Z'\hat{R}^{-1}Z)^{+}\hat{G}^{-1} \end{pmatrix}^{-1}$$

que sovint subestima la variabilitat mostral de $(\hat{\beta}, \hat{u})$ i no incorpora la correcció associada a la incertessa extra provocada per l'estimació dels paràmetres en G i R. Com alternativa s'han proposat algunes aproximacions com la de Satttherthwaite o l'aproximació de Kenward-Roger (casos on existeixen efectes aleatoris i modelització de covariància residual).

Inferència

Un cop estimat un model es voldrà saber la precisió de l'estimació mitjançant els errors estàndard, i si els paràmetres són significativament diferents de zero mitjançant contrast d'hipòtesis que habitualment es resoldran amb aquests tests: el test de raó de versemblances (LR) i de Wald.

El LR test serveix per comparar models estimats amb màxima versemblança i l'especificació dels efectes fixes ha de ser comuna a tots els models. Els models que comparem han d'estar aniuats. Per a la inferència del component de la variància ($H_0 : \sigma_u^2 = 0$) s'utilitza el LR test (en aquest cas, vàlid tan amb ML o REML només si el nombre els efectes fixes són els mateixos, és a dir, els models són jeràrquics), tenint en compte, que

ara la distribució de l'estadístic és una mixtura de distribucions χ^2, al considerar un valor frontera.

Si L_2 és la versemblança del model més general i L_1 és la versemblança del model restringit, es satisfà que $L_2 > L_1$. L'estadístic de LR test es expressat com:

$$LR = 2\log(L_2/L_1) = 2[\log(L_2) - \log(L_1)]$$

i serà positiu. Si K_1 és el nombre de paràmetres del model restringit, la distribució de l'estadístic LR sota la hipòtesis nul.la serà $\chi^2_{K_2-K_1}$.

Per a models no aniuats, s'utilitzen mesures d'entropia com el *Akaike Information Criterion* (AIC), *Bayesian Information Criterion* (BIC), *Consistent Akaike Information Criterion* (CAIC). Respecte a la validació del model, s'haurà de tenir en compte que el model i l'estructura de variàncies del model estigui ben especificat, i comprovar i diagnosticar la validesa del model mitjançant els gràfics dels residus, i la detecció de les observacions influents.

2.4 Models no Lineals Mixtes

Els models no lineals mixtes (NLMM) són models mixtes on hi ha una relació no lineal entre els paràmetres (efectes fixes i aleatoris) del model respecte la resposta.

L'especificació del model NLMM és definit com:

$$Y_i = f(X_i; \beta; Z_i; u_i) \tag{2.4}$$

on $f(X_i; \beta; Z_i; u_i)$ és una funció no lineal dels paràmetres β i u.

Es poden trobar diferents famílies o funcions no lineals f, el problema és l'inestabilitat en l'estimació i problemes de convergència. En la pràctica, algunes funcions ja estan definides en la literatura, com per exemple en l'àmbit de la farmacocinètica i farmacodinàmica. [31–33].

Per defecte el mètode d'estimació és el de la ML. Degut a la complexitat en l'estimació de la funció de versemblança, s'han proposat diferents aproximacions basades en l'aproximació de Laplace, i quadratura gaussiana, entre altres [34,35].

2.5 Models Lineals Generalitzats mixtes

Actualment, la complexitat dels dissenys per poder contrastar adequadament les hipòtesis d'investigació d'interès ha provocat que les dades habitualment presentin una estructura jeràrquica o multinivell. En la pràctica i sobretot en ciències de la salut, ens trobem amb pacients o subjectes que són observats dins d'unitats més grans, com famílies, hospitals, barris, àrees geogràfiques. Aquesta estructura de dades apareix també en els estudis longitudinals, on les mesures estan agrupades dins dels propis subjectes.

Generalized Linear Mixed Models (GLMM) són una extensió dels GLM que incorpora efectes aleatoris en el predictor lineal [36]. Els GLMM proporcionen un procediment d'anàlisi més flexible quan la variable resposta no és continua, i on no es compleix l'assumpció d'independència donat que permet modelar dades agrupades mitjançant els efectes aleatoris. A més a més, els GLMM són útils per modelar la sobredispersió [37] i autocorrelació en models amb distribució de Poisson o Binomial.

2.5.1 Definició i especificació del model GLMM

El GLMM és especificat com:

$$\eta = g(E(Y_i) = g(\mu_i) = X_i\beta + Z_i u_i \tag{2.5}$$

on Y_i és el vector de m observacions de la variable resposta corresponent al subjecte i, $i = 1, \ldots, n$.

Les característiques dels altres components són:

- u_i és el vector d'efectes aleatoris d'un mateix subjecte i. Normalment s'assumeix que $u_i \sim \mathcal{N}(0, \Sigma)$ on Σ és la matriu de covariàncies definida positiva. La seva funció de densitat és definida com $f(u; \Sigma)$.
- S'assumeix que $(Y_i|u_i)$ segueix una distribució dins de la família exponencial amb la funció de densitat $f(Y_i|u_i; \cdot)$.
- La mitjana i variància concidionals són definides com $\mu_i = \mathrm{E}(Y_i|u_i)$ i $\mathrm{Var}(Y_i|u_i) = \Phi \mathrm{Var}(\mu_i)$ respectivament, on Φ és el paràmetre de dispersió i $\mathrm{Var}(\mu_i)$ és la funció de la variància del GLMM.
- $g(\cdot)$ és la funció monòtona diferenciable, coneguda com a funció d'enllaç
- η és el predictor lineal
- X_i i Z_i són les matrius de disseny
- β és el vector amb els paràmetres corresponents dels efectes fixes

La funció de densitat condicionada de Y donat u té la forma següent:

$$f(Y|u; \beta) = \prod_{i=1}^{n} f(Y_i|u_i; \beta)$$

I la funció de densitat multivariant de u és expressada com:

$$f(u; \Sigma) = \prod_{i=1}^{n} f(u_i; \Sigma).$$

2.5.2 Estimació dels paràmetres dels GLMM

L'estimació dels paràmetres (efectes fixes, aleatoris i components de la variància) en els GLMM es basen en la maximització de la versemblança marginal.

$$l(\beta, u|Y) = f(Y; u, \beta) = \int f(Y|u; \beta) f(u; \Sigma)\, du \tag{2.6}$$

L'estimació dels paràmetres en els GLMM no és una qüestió trivial, perquè la derivació de la versemblança sovint no es pot aconseguir d'una manera analítica. El primer mètode d'estimació dels GLMM es va introduir en la dècada dels 90 [38] i actualment estan disponibles la majoria dels mètodes en els diferents *softwares* estadístics, ja que, els mètodes d'estimació i també les llibreries dels paquets estadístics estan avui dia encara en desenvolupament [39]. Donat que els efectes aleatoris no són observats directament, apareixen dificultats pel que fa a l'ús de la funció de versemblança conjunta. Hem comentat que en els LMM la versemblança integrada pot avaluar-se de forma exacte. No és així en els models GLMM.

Aquest fet ha provocat que en els darrers anys hagin sorgit diversos mètodes d'estimació basats en aproximacions o en simulació. Actualment, és possible trobar diferents mètodes d'estimació per als GLMM com el mètode 'Penalized quasi-likelihood' (PQL) [36], el mètode de 'Laplace', 'Gauss-Hermite quadrature' (GHQ) [40,41], procediments jeràrquics [42] i procediments bayesians basats en les tècniques de 'Markov Chain Monte Carlo'(MCMC) [43] i recentment el 'Integrated Nested Laplace Approximation' (INLA) [44].

Cal destacar que els problemes amb l'obtenció de l'estimació dels paràmetres en els GLMM, ha donat lloc a altres alternatives, com l'utilització dels models marginals com el *Generalized Estimating Equations*(GEE), o també l'ús dels efectes aleatoris assumint que no segueixen una distribució normal com els models lineals generalitzats jeràrquics (HGLM), on és possible l'ús de distribucions conjugades.

2.5.3 Procediments inferencials dels GLMM

Un cop els models han estat parametritzats, s'utilitzen els mètodes d'inferència per tal d'obtenir conclusions sobre les variables d'interès. A part dels llibres clàssics de la inferència dels GLMM, també hi ha algunes guies que expliquen amb més deteniment l'actualitat de la inferència estadística i l'estratègia per a la selecció dels millors models. Aquesta

informació es complementada i sobretot actualitzada per Ben Bolker en la web http://glmm.wikidot.com/.

Una de les preguntes clàssiques en inferència respecte els GLMM és quin és el millor camí per testar els seus paràmetres. Podem contrastar hipòtesis del tipus:

$$H_0 : \beta = \beta^* \quad \text{vs.} \quad H_1 : \beta \neq \beta^* \tag{2.7}$$

Les hipòtesis sobre els efectes fixes i aleatoris (o les seves variàncies) són testades de forma separada. Per testar les hipòtesis dels efectes fixes s'utilitza normalment el test de Wald (Z, χ^2, t, F) i per testar els efectes aleatoris s'utilitza la LR o bé comparant la bondat d'ajust dels models utilitzant el AIC o el BIC. I pel que fa al test de Z Wald i χ^2 són més apropiats per als models que no presenten sobredispersió, mentre que els test de Wald t i F test té en compte la incertesa en l'estimació de la sobredispersió [39].

Finalment, destacar que com tot model de regressió és important seguir una estratègia per a la construcció del model, un cop especificat. Per obtenir el millor model, dins dels candidats es farà servir els criteris d'informació (AIC, BIC, , ..), tenint en compte també l'estratègia per a la selecció de les variables [39]. Recentment ha estat definida també una expressió pel coeficient de determinació pels GLMM [45] que permet avaluar la capacitat predictiva dels models.

2.6 Article 1: Methodological quality and reporting of Generalized Linear Mixed Models in Clinical Medicine (2000 – 2012): a systematic review (PLOS ONE)

Les següents pàgines mostren el treball publicat a la revista Plos One, revista de primer cuartil amb factor d'impacte de 3.23 (JCR, 2014), i que és citat a continuació:

Casals M, Girabent-Farrés M, Carrasco JL (2014). Methodological Quality and Reporting of Generalized Linear Mixed Models in Clinical Medicine (2000–2012): A Systematic Review. PLoS ONE 9(11), e112653.

OPEN ACCESS Freely available online

PLOS ONE

Methodological Quality and Reporting of Generalized Linear Mixed Models in Clinical Medicine (2000–2012): A Systematic Review

Martí Casals[1,2,3,4]*, Montserrat Girabent-Farrés[5], Josep L. Carrasco[2]

1 CIBER de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, **2** Bioestadística, Departament de Salut Pública, Universitat de Barcelona, Barcelona, Spain, **3** Departament de Ciencies Basiques, Universitat Internacional de Catalunya, Barcelona, Spain, **4** Servei d'Epidemiologia, Agència de Salut Pública de Barcelona, Barcelona, Spain, **5** Departament de Fisioteràpia (unitat de Bioestadística), Universitat Internacional de Catalunya, Barcelona, Spain

Abstract

Background: Modeling count and binary data collected in hierarchical designs have increased the use of *Generalized Linear Mixed Models (GLMMs)* in medicine. This article presents a systematic review of the application and quality of results and information reported from GLMMs in the field of clinical medicine.

Methods: A search using the Web of Science database was performed for published original articles in medical journals from 2000 to 2012. The search strategy included the topic "*generalized linear mixed models*","*hierarchical generalized linear models*", "*multilevel generalized linear model*" and as a research domain we refined by science technology. Papers reporting methodological considerations without application, and those that were not involved in clinical medicine or written in English were excluded.

Results: A total of 443 articles were detected, with an increase over time in the number of articles. In total, 108 articles fit the inclusion criteria. Of these, 54.6% were declared to be longitudinal studies, whereas 58.3% and 26.9% were defined as repeated measurements and multilevel design, respectively. Twenty-two articles belonged to environmental and occupational public health, 10 articles to clinical neurology, 8 to oncology, and 7 to infectious diseases and pediatrics. The distribution of the response variable was reported in 88% of the articles, predominantly Binomial (n = 64) or Poisson (n = 22). Most of the useful information about GLMMs was not reported in most cases. Variance estimates of random effects were described in only 8 articles (9.2%). The model validation, the method of covariate selection and the method of goodness of fit were only reported in 8.0%, 36.8% and 14.9% of the articles, respectively.

Conclusions: During recent years, the use of GLMMs in medical literature has increased to take into account the correlation of data when modeling qualitative data or counts. According to the current recommendations, the quality of reporting has room for improvement regarding the characteristics of the analysis, estimation method, validation, and selection of the model.

Citation: Casals M, Girabent-Farrés M, Carrasco JL (2014) Methodological Quality and Reporting of Generalized Linear Mixed Models in Clinical Medicine (2000–2012): A Systematic Review. PLoS ONE 9(11): e112653. doi:10.1371/journal.pone.0112653

Editor: Antonio Guilherme Pacheco, FIOCRUZ, Brazil

Received June 25, 2014; **Accepted** October 10, 2014; **Published** November 18, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

* Email: marticasals@gmail.com

Introduction

Statistical modeling is a highly important tool that receives a lot of attention in any scientific field. In health sciences, statistical models arise as an important methodology to predict outcomes and assess association between outcomes and risk factors as well. Thus, one important aspect is to efficiently test the investigational hypothesis by avoiding biases and accounting for all the sources of variability present in data. This usually leads to complex designs where data is hierarchically structured. Multilevel, longitudinal or cluster designs are examples of such structure. In health sciences, longitudinal studies probably are more common, where measurements are grouped in subjects who are followed over time. Furthermore, other possibilities are studies where measurements are hierarchically grouped in subgroups such as schools, hospitals, neighborhoods, families, geographical areas or place of employment.

In the classic linear model (linear regression analysis, ANOVA, ANCOVA), the variable response is continuous and it is assumed that the response conditioned to covariates follows a normal distribution with maximum likelihood based approaches as the principal estimation methods [1–3]. However, the general linear model is not appropriate for non-continuous responses (e.g. binary, counts) because the underlying assumptions of the model do not hold.

Generalized linear models (GLMs) arose as an extension of the classic linear model that allowed for the accommodation of non-normal responses as well as a non-linear relationship between the expectation of the response and the covariates [2,4,5]. GLMs are most often applied to count or binary responses in health sciences [6], assuming Poisson, Binomial or Bernoulli as probability distributions for the response.

Similar to the classic linear model (which is indeed a particular type of GLM), GLMs also assume that the observations (conditioned to covariates) are independent and identically distributed. Regarding study designs with hierarchical structure, the assumption of independence is usually violated because measurements within the same cluster are correlated. The main disadvantage of ignoring within-cluster correlation is the bias in point estimates and standard errors. These biases might cause a loss of statistical power and efficiency of hypothesis testing on fixed effects [7,8]. Thus, the statistical significance could be wrongly assessed [9] and the type I error rate could be different than that *a priori* determined in hypothesis testing.

Generalized linear mixed models (GLMMs) are a methodology based on GLMs that permit data analysis with hierarchical GLMs structure through the inclusion of random effects in the model. The GLMMs are also known in the literature as hierarchical generalized linear models (HGLMs) and multilevel generalized linear models (MGLMs) depending on the field [10–12]. For the sake of simplicity we will use the term GLMMs throughout the text. The first estimation method of GLMMs was introduced in the early 1990 s [13]. Nowadays various estimation methods can be found for GLMMs, such as the penalized quasi-likelihood method (PQL) [14], the Laplace method [14], Gauss-Hermite quadrature [15], hierarchical-likelihood methods [11], and Bayesian methods based on the Markov chain Monte Carlo technique [16,17], and, recently also based on the integrated nested Laplace approximation [18].

Furthermore, GLMM methodology is now available in the main statistical packages, though estimation methods as well as statistical packages are still under development [19,20].

The increasing interest in GLMMs is reflected by the publication of tutorials in various fields, such as ecology [19], psychology [21], biology [22], and medicine [23–26]. Nowadays, original articles, academic work and reports which utilize GLMMs exist, and methodological guidelines and revisions are also available for the analysis of GLMMs in each field [19,27–29].

However, it is not possible to find guidelines that specifically address the appropriate reporting of population modeling studies [30]. In addition, no reviews of the use and quality of reported information by GLMMs exist despite an important increase in quantitative analyses in the academic and professional science settings.

Reporting guidelines are evidence-based tools that employ expert consensus to help authors to report their research such that readers can both critically appraise and interpret study findings [30–34]. Recently, minimal rules that can serve as standardized guidelines should be established to improve the quality of information and presentation of data in medical scientific articles [35]. Only Thiele [22] has made reference to GLMMs in the field of biology and still no standardized guidelines indicate what information is relevant to present in medical articles.

For this reason, the objective of the present study is to review the application of GLMMs and to evaluate the quality of reported information in original articles in the field of clinical medicine during a 13-year period (2000–2012), while analyzing the evolution over time, journals, and areas of publication.

Methods

This review was conducted according to the Preferred Reporting Items for Systematic Reviews and Metanalyses (PRISMA) Statement [36,37]. We also report the review in accordance with PRISMA guidelines (Checklist S1).

With the objective to obtain and analyze the existing scientific literature related to the use of GLMMs in clinical medicine, a strategic search of original published articles in this field from 2000 to 2012 was performed using the Web of Science database.

The search strategy included the topic "*generalized linear mixed models*", "*hierarchical generalized linear models*", "*multilevel generalized linear model*" and as a research domain we refined by science technology (Appendix S1).

The following fields of clinical medicine were included in the search:

Endocrinology Metabolism, Urology Nephrology, Public environmental occupational health, Orthopedics, Respiratory system, Entomology, Health care sciences services, Medical laboratory technology, Pediatrics, Pathology, Life sciences biomedicine other topics, Hematology, Geriatrics gerontology, Gastroenterology hepatology, Rheumatology, Critical care medicine, Medical informatics, Emergency medicine, Integrative complementary medicine, Obstetrics gynecology, Neurosciences neurology, Cardiovascular system cardiology, Infectious diseases, Radiology nuclear medicine medical imaging, Transplantation, Tropical medicine, Allergy, Anesthesiology, Anatomy morphology, General internal medicine, Immunology, Research experimental medicine, Dermatology, Oncology, Surgery.

Selection of the studies included in the review

Articles were eligible for inclusion if they were original research articles written in English in peer-reviewed journals reporting an application of GLMM. We excluded articles of statistical methodology development and those that were not entirely involved in clinical medicine (biology, psychology, genetics, sports, dentistry, air pollution, education, economy, family and health politics, computer science, ecology, nutrition, veterinary and nursing).

Identification of studies

The information from Appendix S1 (Table) was extracted from the selected articles. Data were collected and stored in a database. Then, data were checked to find discrepancies between the two reviewers. Discrepancies were solved by consensus after reviewing again the conflictive articles.

Figure 1 uses the PRISMA flowchart to summarize all stages of the paper selection process [37]. In the first review phase, 462 articles were identified, nineteen of which were duplicates.

After inspection of the abstracts, we excluded the articles that were non-original articles (reviews, short articles or conferences) and those articles that did not have a GLMM as a key word in the abstract or in the title of the article.

In the second review phase, of the 428 articles, only 129 pertained to the aforementioned medical fields. Thus, 299 articles were excluded because they belonged to other fields, such as ecology, computer science, air pollution or statistical methodology. In the third review phase, we obtained full text versions of potentially eligible articles. Two articles were excluded due to inconsistency in the specification of the model applied because in the full text version they were not a GLMM as it was stated in the abstract. We then conducted a detailed review of the 127 articles and we excluded 19 articles because they were not published in an indexed journal included in Journal Citation Reports (JCR).

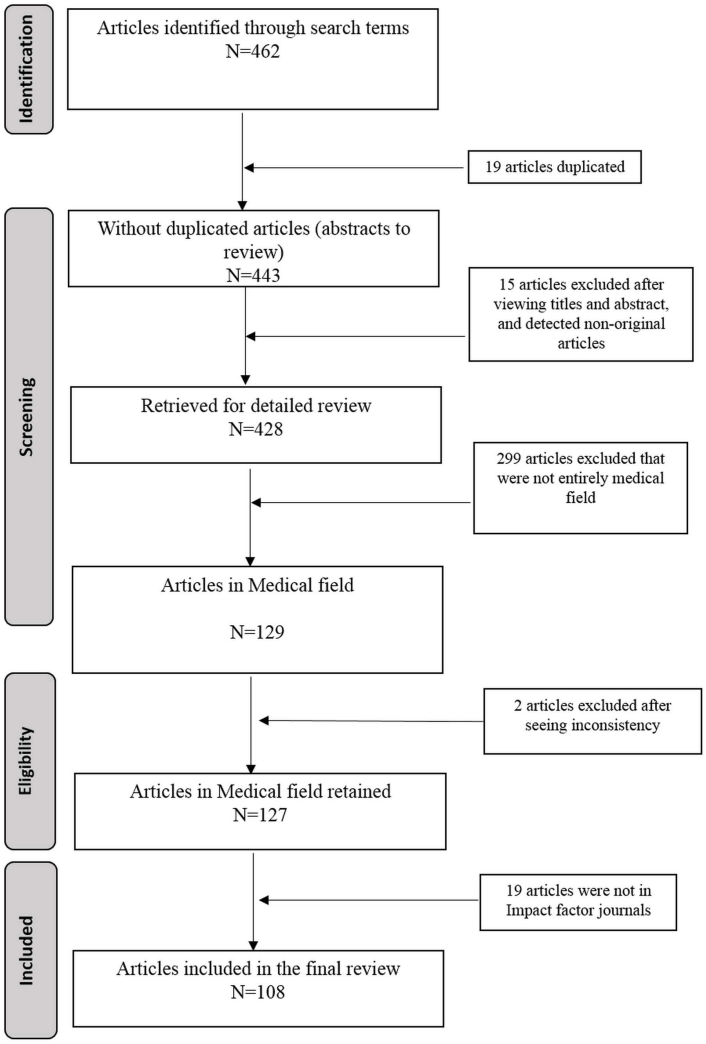


Figure 1. Flow chart of the selection of reviewed articles.
doi:10.1371/journal.pone.0112653.g001

Finally, 108 articles were included in the final review (Appendix S2). Figure 1 summarizes the numbers of articles identified and the reasons for exclusion at each stage.

Information collected from the selected articles

Based on Thiele's and Bolker's works [22,38], a list of relevant information and basic characteristics of the study that should be reported in an article with GLMM analysis was suggested (Appendix S1).

Study characteristics

Regarding the study design, we refer to different aspects of each study, such as hierarchical structure of data and sample size. The hierarchical structure was used to differentiate between the different study designs that are not mutually exclusive, such as longitudinal, repeated measurements, and multilevel studies. Longitudinal data consist of outcome measurements repeatedly taken on each experimental unit over time. Longitudinal analysis is distinct from cross-sectional analysis as it addresses dependency among measurements taken on the same experimental unit [39]. The studies with repeated measurements usually involve only one level of clustering, where the repeated measurements are interchangeable (replicates).

Finally, multilevel studies present various levels of clusters, potentially providing hierarchical structure in each cluster, as seen in longitudinal or repeated measurement studies. We also took note of whether the probability distribution of the variable response was mentioned or easily deducible. Regarding sample size, the number of clusters, individuals or experimental units were collected.

Inferential issues

This section includes information regarding the GLMM model, as seen in Appendix S1 (Table).

The mixed models are characterized by including fixed and random effects in the linear predictor. Random effects are usually related to the cluster variable. Therefore, it is important to provide information about the cluster variable in the model.

It is also important to report the estimation method of the study and the software applied because they can influence the validity of the GLMM estimates [6,20,38]. Furthermore, the software implementations differ considerably in flexibility, computation time and usability [20].

Concerning the computational issues, the macro GLIMMIX from SAS (1992) was the first available software to fit GLMMs using penalized quasilikelihood (PQL) estimation method. The first production version of PROC GLIMMIX for SAS was first released in 2005 and became the standard procedure in version 9.2 in 2008 [40]. Nowadays, there are other available softwares to fit GLMMs. Among them the lme4 package was first implemented for R in 2003 [41]. Moreover, in R software, we can find other packages to fit GLMMs such as glmmML [42], MASS (with the glmmPQL function) [43] or gar (with the repeated function) [44,45]. Concerning SAS software besides the aforementioned PROC GLIMMIX, the PROC NLMIXED is also able to fit GLMMs [46]. Additionally, it is also possible to use ASReml [47], MLwiN [48] and STATA software (which uses the functions xtmixed and gllamm [22,28,49,50]) [22,28,49,50]. The SPSS (starting with SPSS 19) software now also includes a GLMM obtained in the GENLINMIXED procedure [51,52].

With respect to statistical inference, the hypotheses concerning fixed and random effects (or their variances) are tested in separated form. Thus, testing the hypotheses for fixed effects is commonly assessed by the Wald score tests. On the other hand, hypotheses concerning random effects variances can be tested using the likelihood ratio test [19] or by comparing the goodness of fit of the models using the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) [19].

Validation model

Similar to GLMs, validation of GLMMs is commonly based on the inspection of residuals to determine if the model assumptions are fulfilled.

An important point is related to the so-called scale parameter when it is fixed to a specific value because of the probability model assumed. For example, the scale parameter for Poisson and Binomial distribution should be equal to 1. A parameter different from 1 implies that the probability distribution of the responses conditioned to covariates is not correctly specified and the model is not valid. This phenomenon is known as over or underdispersion and causes incorrect standard errors that can produce different clinical conclusions [53]. Thus, it is relevant to evaluate the presence of over- or underdispersion and report the results of this analysis.

Finally, information on the use of a concrete strategy to select the variables in the model and its criterion was obtained. Variable selection strategy usually consist of stepwise selection of variables (forward or backward) [19]. Concerning the criterion, it can be based on entropy as the aforementioned AIC and BIC, or hypotheses testing (likelihood ratio test or Wald test). However, it is possible to find studies with no need of variable selection, for example confirmatory analysis where a particular hypothesized model is fit. This hypothesized model may be based on theory and/or previous analytic research [54,55]. In this latter case, the selection variable strategy was considered appropriately reported.

Results

The evolution of the use of GLMMs in medical journals of the 443 articles selected in the first phase is described in Figure 2. The remaining results (Tables 1, 2, 3 and Appendix S3 and S4) make reference to the 108 articles included in the final in-depth review. Of these, 92 (85.2%) were defined as GLMMs, 14 (13.0%) as HGLMs, and 2 (1.9%) as MGLMs.

Most of these articles were found in the following journals: *American Journal of Public Health*, which had 7 publications; *PLoS ONE, Cancer Causes & Control*, *BMC Public Health, Annals of Surgery,* and *Headache*, which had 3 publications each. Twenty-two articles pertained to environmental and occupational public health area, 10 articles pertained to clinical neurology, 8 to oncology, and 7 to infectious diseases and pediatrics (Appendix S3).

Forty-five articles (41.7%) were written by an author who was part of a biometric or statistical department and some co-authors (53.3%) were affiliated with a public health department.

Of the 108 selected articles, 59 (54.6%) declared to be longitudinal studies, whereas 56 (58.3%) and 29 (26.9%) were defined as repeated measurements and multilevel design, respectively (Table 1). It is important to note that over 8% of the articles were unclear when reporting the cluster design. Twenty-seven articles (25%) involved confirmatory analysis whereas 81 (75%) were declared as exploratory analysis. Ninety-five of the articles stated their sample size, which ranged from 20–785,385 with a median of 2,201 (Q1 = 408; Q3 = 25000). One random effect in the *intercept* was used in 61 articles, and two or more random effects were used in 36 articles. Of these, 61.1% of the articles had a random effect that pertained to a multilevel model. The size of the random effect or *cluster*, as the number of levels of random effects or the number of clusters, was clearly described in only 33 articles, which ranged from 9–16,230 clusters with a median of 167 (Q1 = 55; Q3 = 1187). The cluster was principally the individual (subject, patient, participant, etc) (n = 46), hospital (n = 15), center (n = 10), geographical area (n = 9) and family (n = 3).

The type of study design was described as cross-sectional (n = 31), cohort (n = 26), clinical trial (n = 18), case-control (n = 2) and *cross-over* (n = 1). Eight articles did not mention study design

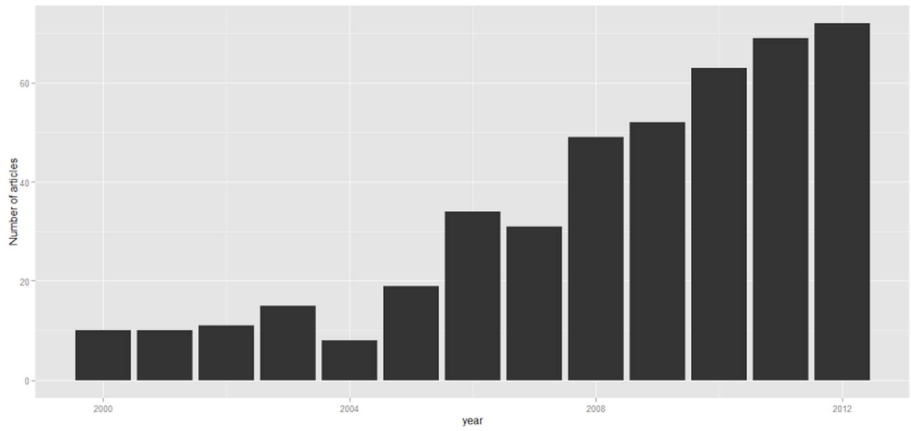


Figure 2. Number of reviewed articles by year of publication.
doi:10.1371/journal.pone.0112653.g002

and 18 articles only described the characteristics of the study design (i.e. experimental, prospective, multicenter, etc) without specifying which study design was used (Table 1).

The response variable ('clinical') of the study differed in each of the reviewed articles, and thus there was no common illness or pathology. Available software can fit different response variables

Table 1. Characteristics of the study design in the reviewed articles.

	N = 108
Longitudinal study:	
NO	40 (37.0%)
Unclear	9 (8.30%)
YES	59 (54.6%)
Repeated measures:	
NO	34 (31.5%)
Unclear	11 (10.2%)
YES	56 (58.3%)
Multilevel (nested design):	
NO	79 (73.1%)
YES	29 (26.9%)
Type of analysis	
Exploratory	81 (75.0%)
Confirmatory	27 (25.0%)
Design	
Case-control	2 (2.30%)
Case-crossover	1 (1.10%)
Cluster Random Trial	18 (16.7%)
Cohorts	26 (24.1%)
Cross-sectional	31 (28.7%)
NR	8 (7.40%)
Unclear	22 (20.4%)

NR: Not reported.
doi:10.1371/journal.pone.0112653.t001

for exponential family, such as Poisson, binomial, Gamma, and Inverse Gaussian, though Poisson and Binomial (or binary) are the most used in medicine. The distribution of the response variable was reported in 88% of the articles, and the most common was binomial (n = 64), Poisson (n = 22), negative binomial (n = 1) and multinomial (n = 2).

Furthermore, the estimation method for each model was reported in only 21 articles (19.4%), and the following estimation methods were used: *maximum likelihood* (n = 3), *penalized quasi-likelihood* (n = 8), pseudo-likelihood (n = 2), restricted maximum likelihood (n = 2), *adaptative quadrature likelihood approximation* (n = 1), and Markov chain Monte Carlo (MCMC; n = 5). It is important to mention that over 90% of the articles did not report the test used for the fixed nor random effects, which implies that the section on statistical methods was insufficiently described (Table 2).

The most used statistical software packages were SAS (n = 57), R (n = 13), Stata (n = 12), and HLM (n = 6). For SAS, the use of macro GLIMMIX was reported in 24 articles and the macro NLMIXED with PROC MIXED to fit the GLMM was used in five articles. For R, different packages were used to fit the GLMM, such as lme4 (n = 2), glmmPQL (n = 4), glmmML(n = 1), BayesX (n = 2) or repeated (n = 1). For Stata, the gllamm (n = 2) and xtmixed functions were also used (n = 1).

Overdispersion for models with counts or binary response which assume a Poisson or Binomial distribution was evaluated in 10 articles. Of these, different approaches were proposed to fit as alternatives (GEE, Negative Binomial, Quasi-Poisson, Zero-Inflated). For the articles that used Poisson or Binomial distribution of probability, 90.7% did not state if under-over-dispersion was evaluated, 99.1% did not report the magnitude of the scale parameter, and 92.6% did not suggest alternatives for possible under-overdispersion. Variance estimates of random effects were described in only 10 articles (9.3%). With respect to the fixed effects, the standard error and confidence interval were reported in 20% and 71.3%, respectively, whereas in the variance components, they were reported in 3.7% and 2.8%, respectively. The model validation, the method of covariate selection and the method of goodness of fit were reported in 6.5%, 35.2% and 15.7% of the articles, respectively (Table 3).

Discussion

The articles selected in this review showed that the number of bibliographical references that use GLMMs in medical journals increased from the year 2000 to 2012.

Our review also indicated that there is room for improvement in quality when basic characteristics about the GLMMs are reported in medical journals.

A predominance of the articles reviewed were in the fields of environmental and occupational public health. Furthermore, for 45 of the articles (41.7%) at least one of the co-authors was associated with a biometrics or statistical department. This result is consistent with the systematic review of Diaz-Ordaz that showed that trials having a statistician as co-author was associated with a increase in the methodological quality of the analyses [56].

In any scientific paper, the validity of the conclusions is linked to the adequacy of the methods used to generate the results. Thus, it is important to adequately describe the statistical methods used in the analysis. Hence, the reader is able to judge whether the methods used are appropriate, and by extension whether the conclusions are correct.

In the case of GLMM's, as we observed in the results section, the majority of the useful and relevant information about GLMMs that is proposed by Bolker [19] and Thiele [22] was not reported. Therefore, the main consequence is the difficulty to assess the reliability of the results and the validity of the conclusions. For example, the majority of the articles did not mention the estimation method or software that was used. The inferential issues (hypothesis testing, confidence interval estimation) and model validation are closely linked to the estimation method (for instance, bayesian or frequentist). As a consequence, the lack of reporting of the estimation method (or software) used makes it complicated to evaluate the adequacy of the approaches used to inference purposes. Furthermore, the estimation method may have important flaws depending on the situation. For example, PQL yields biased parameter estimates if the standard deviations of the random effects are large, especially with binary data [19].

Additionally, an important deficit regarding the inference of fixed and random effects was observed. Such inference may consist of : 1) hypothesis testing of a set of parameters; 2) competing models using entropy measures; 3) confidence interval of parameters. Here again the validity of the conclusions drawn from the analysis depends on the appropriateness of the procedures used in the inference. For example, the likelihood ratio test is only applicable to nested models. Another example arises when testing the existence of a random effect. This question could be solved by a common hypothesis testing using a null hypothesis whose variance is zero. However, the null hypothesis is set to the boundary of the parameter domain (variance must be positive). Therefore, it is necessary to modify the probability distribution function under the null hypothesis otherwise the p-value obtained is incorrect [57]. Additionally, as we mentioned above, the inferential procedures must be coherent with the estimation technique used.

Furthermore, the validity and model selection as proposed by Bolker and Thiele [19,22] were also not reported in most cases. Once again, the results of the inference and the conclusions of the study will be valid when the assumptions made on the model and estimation method are fulfilled. This is the aim of the validation and, thus, it is essential that the researchers report the results of such a validation and how it was made.

Therefore, in our opinion the methodological information reported in articles using GLMMs could be improved.

We also think that standardized guidelines to report GLMM characteristics in medicine could be beneficial, even though they would not imply by themselves a direct improvement on quality of the articles. As stated by Cobo [35] and Moher [58], it is necessary that both authors and reviewers are aware of recommendations to improve the quality of the manuscripts.

Limitations of the study

One of the limitations of our study could be that the number of identified articles was not high, despite the 13-years review. Nonetheless, the only similar existing review by Thiele [22] in the field of "invasion biology" included only 50 articles. One possible explanation for this number of articles that use GLMMs in health sciences is that medical literature frequently uses models with fixed effects in a hierarchical structure, even though the use of GLMMs is well known in statistical literature [6,59].

Another possible limitation of our review is the potential bias to disregard articles that use a GLMM but do not specify the term as a topic. However, we could assume that articles that use GLMM as topic are more sensitive to this methodology. Thus, it is expected that if this bias existed, the reporting quality would be even better in those potential articles that applied GLMM and used it as a topic.

Table 2. Characteristics of inference and estimation methods reported in the review articles.

	N = 108
Test for fixed effects:	
NR	103 (95.4%)
t-value	1 (0.90%)
Wald F test	4 (3.7%)
Test for random effects:	
LRT	3 (2.80%)
NR	105 (97.2%)
Variance estimates of random effects:	
NR	98 (90.7%)
YES	10 (9.30%)
Statistical software:	
SAS	57 (52.8%)
R	13 (12.0%)
Stata	12 (11.1%)
WinBugs	2 (1.90%)
S-plus	3 (2.80%)
HLM	6 (5.60%)
Statistical Analysis System	1 (0.90%)
SPSS	2 (1.90%)
SEER*Stat	1 (0.90%)
MLwiN	1 (0.90%)
NR	10 (9.30%)
Estimation method:	
Adaptative Quadrature likelihood Approximation	1 (0.90%)
Maximum Likelihood	3 (2.80%)
NR	87 (80.6%)
Penalized Quasi- likelihood	8 (7.50%)
Posterior mean	5 (4.60%)
Restricted Maximum Likelihood	2 (1.90%)
Pseudo likelihood	2 (1.90%)
Statistical software function or macro:	
PROC GLIMMIX	24 (22.2%)
glmmPQL	4 (3.70%)
Gllamm	2 (1.90%)
BayesX	2 (1.90%)
Xtmixed	1 (0.90%)
PROC MIXED/NLMIXED	5 (4.70%)
lme4	2 (1.90%)
glmmML	1 (0.90%)
Repeated	1 (0.90%)
NR	66 (61.1%)

NR: No reported; MCMC: Markov chain Monte Carlo.
doi:10.1371/journal.pone.0112653.t002

There could be also a trend on the estimation methods according to the names given to GLMMs in the articles. Bayesians usually prefer the term hierarchical models instead of mixed effects models whereas frequentists are more likely to use mixed models, which seems to be consistent with our results (Appendix S4).

Conclusions

During recent years, the use of GLMMs in medical literature has increased to take into account the correlation of data when modeling binary or count data. Our review included articles from indexed medical journals included in JCR that mainly consisted of longitudinal studies in a medical setting.

Table 3. Characteristics of the specification, validation and construction of the model for the reviewed articles.

	N = 108
Variable response distribution:	
2 distributions: Binomial, Poisson	1 (0.90%)
2 distributions: Binomial, Multinomial	1 (0.90%)
Binomial	64 (59.2%)
Binomial count	1 (0.90%)
Negative Binomial with offset	1 (0.90%)
NR	11 (10.2%)
Poisson	22 (20.4%)
Poisson with offset	2 (1.90%)
Multinomial	2 (1.90%)
Ordinal	1 (0.90%)
Unclear	2 (1.90%)
Overdispersion evaluation:	
NR	98 (90.7%)
YES	10 (9.20%)
Overdispersion measurement:	
NR	107 (99.1%)
Pearson residuals	1 (0.90%)
Proposed alternative for overdispersion:	
GEE	2 (1.90%)
Negative Binomial	2 (1.90%)
NR	100 (92.6%)
Quasi-Poisson	1 (0.90%)
Variogram	1 (0.90%)
Dscale-adjusted	1 (0.90%)
Zero-inflated	1 (0.90%)
Method of variable selection:	
Backward	3 (2.80%)
Forward	1 (0.90%)
Forward stepwise	1 (0.90%)
NR	70 (64.8%)
Unnecessary (Confirmatory analysis)	27 (25.0%)
Stepwise	6 (5.60%)
Method of goodness of fit comparison model:	
AIC	12 (11.1%)
BIC	3 (2.80%)
DIC	1 (0.90%)
NR	91 (84.3%)
Pseudo R-squared	1 (0.90%)
GLMM Validation:	
NR	101 (93.5%)
YES	7 (6.50%)

NR: No reported; MCMC: Markov chain Monte Carlo; GEE: Generalized estimating equation;
DIC: Deviance information criterion; AIC: Akaike information criterion; BIC: Bayesian information criterion; df: freedom degree.
doi:10.1371/journal.pone.0112653.t003

According to the current recommendations, the quality of reporting has room for improvement regarding the characteristics of the analysis, estimation method, validation and selection of the model.

After analyzing and reviewing the quality of the publications, we believe it is important to consider the use of minimal rules as standardized guidelines when presenting GLMM results in medical journals.

Supporting Information

Appendix S1 Search strategy protocol.
(DOCX)

Appendix S2 Articles included in our study.
(DOC)

Appendix S3 Journals according to field of knowledge.
(DOC)

Appendix S4 Estimation methods according to the name used (GLMM, HGLM, MGLM).
(DOC)

Checklist S1 PRISMA Checklist.
(DOC)

Acknowledgments

We thank LLuís Jover and Klaus Langohr for helpful comments.

Author Contributions

Conceived and designed the experiments: MC MGF JLC. Performed the experiments: MC MGF JLC. Analyzed the data: MC MGF. Contributed reagents/materials/analysis tools: MC MGF JLC. Contributed to the writing of the manuscript: MC MGF JLC.

References

1. Davidson R, MacKinnon JG (1993) Estimation and inference in econometrics. OUP Catalogue.
2. Nelder JA, Wedderburn RW (1972) Generalized linear models. Journal of the Royal Statistical Society. Series A (General) : 370–384.
3. Diggle P, Heagerty P, Liang K, Zeger S (2013) Analysis of longitudinal data. : Oxford University Press.
4. MacCullagh P, Nelder JA (1989) Generalized linear models. : CRC press.
5. Draper N, Smith H (1998)Applied regression analysis new york.
6. Austin PC (2010) Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. The international journal of biostatistics 6.
7. Littell RC, Pendergast J, Natarajan R (2000) Tutorial in Biostatistics modelling covariance structure in the analysis of repeated measures data.
8. Wang LA, Goonewardene Z (2004) The use of MIXED models in the analysis of animal experiments with repeated measures data. Canadian Journal of Animal Science 84: 1–11.
9. Campbell MJ (2008) Statistics at square two: Understanding modern statistical applications in medicine. : Wiley. com.
10. Garson GGD (2012) Hierarchical linear modeling: Guide and applications. : Sage Publications.
11. Lee Y, Nelder JA, Pawitan Y (2006) Generalized linear models with random effects: Unified analysis via H-likelihood. : CRC Press.
12. Stryhn H, Christensen J (2014) The analysis-Hierarchical models: Past, present and future. Prev Vet Med 113: 304–312.
13. Schall R (1991) Estimation in generalized linear models with random effects. Biometrika 78: 719–727.
14. Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88: 9–25.
15. Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and computing 6: 251–262.
16. Zeger SL, Karim MR (1991) Generalized linear models with random effects; a gibbs sampling approach. Journal of the American statistical association 86: 79–86.
17. Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. : Cambridge University Press.
18. Rue H, Martino S, Chopin N (2009) Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71: 319–392.
19. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, et al. (2009) Generalized linear mixed models: A practical guide for ecology and evolution. Trends in ecology \& evolution 24: 127–135.
20. Li B, Lingsma HF, Steyerberg EW, Lesaffre E (2011) Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes. BMC medical research methodology 11: 77.
21. Moscatelli A, Mezzetti M, Lacquaniti F (2012) Modeling psychophysical data at the population-level: The generalized linear mixed model. Journal of vision 12.
22. Thiele J, Markussen B (2012) Potential of GLMM in modelling invasive spread. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources 7.
23. Brown H, Prescott R (2006) Applied mixed models in medicine. : Wiley Chichester.
24. Platt RW, Leroux BG, Breslow N (1999) Generalized linear mixed models for meta-analysis. Stat Med 18: 643–654.
25. Cnaan A, Laird N, Slasor P (1997) Tutorial in biostatistics: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med 16: 2349–2380.
26. Skrondal A, Rabe-Hesketh S (2003) Some applications of generalized linear latent and mixed models in epidemiology: Repeated measures, measurement error and multilevel modeling. Norsk epidemiologi 13.
27. Dean C, Nielsen JD (2007) Generalized linear mixed models: A review and some extensions. Lifetime Data Anal 13: 497–512.
28. Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language 59: 390–412.
29. Tuerlinckx F, Rijmen F, Verbeke G, Boeck P (2006) Statistical inference in generalized linear mixed models: A review. Br J Math Stat Psychol 59: 225–255.
30. Bennett C, Manuel DG (2012) Reporting guidelines for modelling studies. BMC medical research methodology 12: 168.
31. Weinstein MC, Toy EL, Sandberg EA, Neumann PJ, Evans JS, et al. (2001) Modeling for health care and other policy decisions: Uses, roles, and validity. Value in Health 4: 348–361.
32. Kopec JA, Finès P, Manuel DG, Buckeridge DL, Flanagan WM, et al. (2010) Validation of population-based disease simulation models: A review of concepts and methods. BMC Public Health 10: 710.
33. Bagley SC, White H, Golomb BA (2001) Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 54: 979–985. 10.1016/S0895-4356(01)00372-9.
34. Lang TA, Altman DG (2013) Basic statistical reporting for articles published in biomedical journals: The ''Statistical analyses and methods in the published literature'' or the SAMPL guidelines''. Science Editors' Handbook, European Association of Science Editors.
35. Cobo E, Cortés J, Ribera J, Cardellach F, Selva-O'Callaghan A, et al. (2011) Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: Masked randomised trial. BMJ: British Medical Journal 343.
36. Hutton B, Salanti G, Chaimani A, Caldwell DM, Schmid C, et al. (2014) The quality of reporting methods and results in network meta-analyses: An overview of reviews and suggestions for improvement. PloS one 9: e92508. 10.1371/journal.pone.0092508.
37. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group (2010) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International journal of surgery (London, England) 8: 336–341. 10.1016/j.ijsu.2010.02.007.
38. Bolker BM (2008) Ecological models and data in R. : Princeton Univ Pr.
39. Liu C, Cripe TP, Kim M (2010) Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. Molecular Therapy 18: 1724–1730.
40. Institute Inc S (2008) SAS/STAT 9.2. user's guide.
41. Bates D., Sarkar D (2004) 1. lme4: Linear mixed-effects models using S4 classes. R package version 0.6–9.
42. Bröstrom G, Holmberg H. Glmmml: Generalized linear models with clustering. 2011; r package version 0.82–1. Available: http://CRAN. R-project.org/package = glmmML.
43. Venables WN, Ripley BD (2002) Modern applied statistics with S. : Springer.
44. Lindsey JK (1999) Models for repeated measurements. : Oxford University Press, UK.
45. Lindsey JK (2001) Nonlinear models in medical statistics. : Oxford University Press, Oxford, UK.
46. Wolfinger RD (1999) Fitting nonlinear mixed models with the new NLMIXED procedure. : 278–284.
47. Gilmour A, Gogel B, Cullis B, Welham S, Thompson R (2002) ASReml user guide release 1.0 VSN international ltd, hempstead, HP1 1ES, UK. Online im Internet unter.
48. Rowe K (2007) Practical multilevel analysis with MLwiN & LISREL: An integrated course.
49. Rabe-Hesketh S, Skrondal A, Pickles A (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal 2: 1–21.
50. Rabe-Hesketh S, Skrondal A (2008) Estimation using xtmixed. In: Anonymous Multilevel and longitudinal modeling using Stata. : STATA press. 433–436.
51. Garson GD (2013) Fundamentals of hierarchical linear and multilevel modeling. GD Garson, Hierarchical linear modeling guide and applications. Raleigh, NC: North Carolina State University. Sage.
52. Heck RH, Thomas SL, Tabata LN (2013) Multilevel and longitudinal modeling with IBM SPSS. : Routledge.
53. Milanzi E, Alonso A, Molenberghs G (2012) Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions. Stat Med 31: 1475–1482.

54. Preedy VR, Watson RR (2010) Handbook of disease burdens and quality of life measures. : Springer New York.
55. Bender R, Lange S (2001) Adjusting for multiple testing–when and how? J Clin Epidemiol 54: 343–349.
56. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S (2013) A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. BMC medical research methodology 13: 127.
57. Molenberghs G, Verbeke G, Demetrio CGB (2007) An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Anal 13: 513–531. 10.1007/s10985-007-9064-y.
58. Moher D, Schulz KF, Simera I, Altman DG (2010) Guidance for developers of health research reporting guidelines. PLoS medicine 7: e1000217.
59. Austi PC, Alte DA (2003) Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: Should we be analyzing cardiovascular outcomes data differently? Am Heart J 145: 27–35.

El material complementari de l'article és presentat a l'apartat de l'annex.

Systematic review of GLMM

Databases

Web of Science database was used.

Search Terms

TOPIC: ("generalized linear mixed models" OR "generalized linear mixed-effects models" OR "generalized linear models with random effects" OR " hierarchical generalized linear models" OR "multilevel generalized linear model")
Refinedby: RESEARCH DOMAINS=(SCIENCE TECHNOLOGY) AND RESEARCH AREAS=(UROLOGY NEPHROLOGY OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR ONCOLOGY OR GENERAL INTERNAL MEDICINE OR MEDICAL INFORMATICS OR HEALTH CARE SCIENCES SERVICES OR INFECTIOUS DISEASES OR SURGERY OR ORTHOPEDICS OR LIFE SCIENCES BIOMEDICINE OTHER TOPICS OR RESPIRATORY SYSTEM OR ENDOCRINOLOGY METABOLISM OR MEDICAL LABORATORY TECHNOLOGY OR HEMATOLOGY OR PEDIATRICS OR GASTROENTEROLOGY HEPATOLOGY OR ANATOMY MORPHOLOGY OR RHEUMATOLOGY OR OBSTETRICS GYNECOLOGY OR TRANSPLANTATION OR RADIOLOGY NUCLEAR MEDICINE MEDICAL IMAGING OR TROPICAL MEDICINE OR PATHOLOGY OR GERIATRICS GERONTOLOGY OR DERMATOLOGY OR CRITICAL CARE MEDICINE OR CARDIOVASCULAR SYSTEM CARDIOLOGY OR ENTOMOLOGY OR INTEGRATIVE COMPLEMENTARY MEDICINE OR NEUROSCIENCES NEUROLOGY OR RESEARCH EXPERIMENTAL MEDICINE OR EMERGENCY MEDICINE OR ALLERGY OR ANESTHESIOLOGY OR IMMUNOLOGY) AND PUBLICATION YEARS=(2011 OR 2006 OR 2004 OR 2009 OR 2007 OR 2000 OR 2005 OR 2008 OR 2001 OR 2012 OR 2002 OR 2010 OR 2003) AND DOCUMENT TYPES=(ARTICLE) AND LANGUAGES=(ENGLISH)

Date

2000 to 2012

Language

English

Publication type/status

Published original articles in peer reviewed journals.

Study selection

Stage 1: Screening of titles/ abstracts against inclusion criteria.

Stage 2 and 3: Full papers obtained and assessed against inclusion criteria. Papers will be either accepted or rejected due to failure to meet inclusion criteria and the reason will be specified.

Inclusion Criteria

- Original articles written in English that were entirely clinical medical field.

Exclusion criteria

- Articles of statistical methodology development and those which were not entirely involved in clinical medicine (biology, psychology, genetics, sports, dentistry, air pollution, education, economy, family and health politics, computer science, ecology, nutrition, veterinary and nursing).
- Inconsistency in the specification of the model applied because in the full text version they were not a GLMM.

- Articles were not in Impact factor journals.

Data Extraction

Table: Information collected from the selected articles.

Characteristics of the study:
– Study outcome – Study design – Sample size – Number of clusters – Journals – Field – Authors' affiliation to a biostatistics or biometric department – Type of analysis (confirmatory/exploratory)
Inferential issues:
– Information about the cluster variable – Estimation method – Statistical software – Statistical software function or macro – Test for fixed effects – Test for random effects – Variance estimates of random effects
Model validation:
– Overdispersion (if assessed and reported) – Method of goodness of fit for model comparison (if necessary) – Method of covariate selection

Table: Articles included in our study.

References included
1. Abellana R, Ascaso C, Carrasco JL, Castell C, Tresserras R: **Geographical variability of the incidence of Type 1 diabetes in subjects younger than 30 years in Catalonia, Spain.** Med Clin 2009, **132**(12):454-458.
2. Ahmed A, Allman RM, Kiefe CI, Person SD, Shaneyfelt TM, Sims RV, Howard G, DeLong JF: **Association of consultation between generalists and cardiologists with quality and outcomes of heart failure care.** Am Heart J 2003, **145**(6):1086-1093.
3. Andreozzi VL, Bailey TC, Nobre FF, Struchiner CJ, Barreto ML, Assis AMO, Santos LMP: **Random-effects models in investigating the effect of vitamin A in childhood diarrhea.** Ann Epidemiol 2006, **16**(4):241-247.
4. Assassi S, Leyva AL, Mayes MD, Sharif R, Nair DK, Fischbach M, Ngan Nguyen, Reveille JD, Gonzalez EB, McNearney TA, GENISOS Study Grp: **Predictors of Fatigue Severity in Early Systemic Sclerosis: A Prospective Longitudinal Study of the GENISOS Cohort.** Plos One 2011, **6**(10):e26061.
5. Assassi S, Sharif R, Lasky RE, McNearney TA, Estrada-Y-Martin RM, Draeger H, Nair DK, Fritzler MJ, Reveille JD, Arnett FC, Mayes MD, GENISOS Study Grp: **Predictors of interstitial lung disease in early systemic sclerosis: a prospective longitudinal study of the GENISOS cohort.** Arthritis Research & Therapy 2010, **12**(5):R166.
6. Baechle C, Haastert B, Holl RW, Beyer P, Grabert M, Giani G, Icks A, DPV Initiative: **Inpatient and Outpatient Health Care Utilization of Children and Adolescents with Type 1 Diabetes before and after Introduction of DRGs.** Experimental and Clinical Endocrinology & Diabetes 2010, **118**(9):644-648.
7. Bennett KE, Hopper JE, Stuart MA, West M, Drolet BS: **Blood-feeding behavior of vesicular stomatitis virus infected Culicoides sonorensis (Diptera : Ceratopogonidae).** J Med Entomol 2008, **45**(5):921-926.
8. Berdahl TA: **Racial/Ethnic and Gender Differences in Individual Workplace Injury Risk Trajectories: 1988-1998.** Am J Public Health 2008, **98**(12):2258-2263.
9. Boudourakis LD, Wang TS, Roman SA, Desai R, Sosa JA: **Evolution of the Surgeon-Volume, Patient-Outcome Relationship.** Ann Surg 2009, **250**(1):159-165.
10. Boyd HA, Flanders WD, Addiss DG, Waller LA: **Residual spatial correlation between geographically referenced observations - A Bayesian hierarchical modeling approach.** Epidemiology 2005, **16**(4):532-541.
11. Boyd HA, Waller LA, Flanders WD, Beach MJ, Sivilus JS, Lovince R, Lammie PJ, Addiss DG: **Community- and individual-level determinants of Wuchereria bancrofti infection in Leogane Commune, Haiti.** Am J Trop Med Hyg 2004, **70**(3):266-272.
12. Bradley CJ, Dahman B, Bear HD: **Insurance and Inpatient Care Differences in Length of Stay and Costs Between Surgically Treated Cancer Patients.** Cancer 2012, **118**(20):5084-5091.
13. Bradley EH, Herrin J, Wang Y, Barton BA, Webster TR, Mattera JA, Roumanis SA, Curtis JP, Nallamothu BK, Magid DJ, McNamara RL, Parkosewich J, Loeb JM, Krumholz HM: **Strategies for reducing the door-to-balloon time in acute myocardial infarction.** N Engl J Med 2006, **355**(22):2308-2320.
14. Bucher BT, Guth RM, Saito JM, Najaf T, Warner BW: **Impact of Hospital Volume on In-Hospital Mortality of Infants Undergoing Repair of Congenital Diaphragmatic Hernia.** Ann Surg 2010, **252**(4):635-641.
15. Caffrey AR, LaPlante KL: **Changing epidemiology of methicillin-resistant Staphylococcus aureus in the Veterans Affairs Healthcare System, 2002-2009.** Infection 2012, **40**(3):291-297.
16. Campitelli MA, Inoue M, Calzavara AJ, Kwong JC, Guttmann A: **Low Rates of Influenza Immunization in Young Children Under Ontario's Universal Influenza Immunization Program.** Pediatrics 2012, **129**(6):E1421-E1430.
17. Cardo MV, Vezzani D, Carbajo AE: **Environmental Predictors of the Occurrence of Ground Water Mosquito Immatures in the Parana Lower Delta, Argentina.** J Med Entomol 2011, **48**(5):991-998.
18. Cherpitel CJ, Ye Y, Bond J, Rehm J, Poznyak V, Macdonald S, Stafstrom M, Hao W: **Multi-level analysis of alcohol-related injury among emergency department patients: a cross-national study.** Addiction 2005, **100**(12):1840-1850.
19. Childs JD, Teyhen DS, Van Wyngaarden JJ, Dougherty BF, Ladislas BJ, Helton GL, Robinson ME, Wu SS, George SZ: **Predictors of web-based follow-up response in the Prevention of Low Back Pain in the Military Trial (POLM).** Bmc Musculoskeletal

References included

Disorders 2011, **12**:132.

20. Chung JH, Phibbs CS, Boscardin WJ, Kominski GF, Ortega AN, Gregory KD, Needleman J: **Examining the effect of hospital-level factors on mortality of very low birth weight infants using multilevel modeling.** Journal of Perinatology 2011, **31**(12):770-775.

21. Cleveland MJ, Feinberg ME, Bontempo DE, Greenberg MT: **The role of risk and protective factors in substance use across adolescence.** Journal of Adolescent Health 2008, **43**(2):157-164.

22. Colford JM,Jr., Hilton JF, Wright CC, Arnold BF, Saha S, Wade TJ, Scott J, Eisenberg JNS: **The Sonoma Water Evaluation Trial: A Randomized Drinking Water Intervention Trial to Reduce Gastrointestinal Illness in Older Adults.** Am J Public Health 2009, **99**(11):1988-1995.

23. Comber H, Sharp L, Timmons A, Keane FBV: **Quality of rectal cancer surgery and its relationship to surgeon and hospital caseload: a population-based study.** Colorectal Disease 2012, **14**(10):E692-E700.

24. Cooper HLF, Des Jarlais DC, Ross Z, Tempalski B, Bossak B, Friedman SR: **Spatial Access to Syringe Exchange Programs and Pharmacies Selling Over-the-Counter Syringes as Predictors of Drug Injectors' Use of Sterile Syringes.** Am J Public Health 2011, **101**(6):1118-1125.

25. Cram P, Bayman L, Popescu I, Vaughan-Sarrazin MS, Cai X, Rosenthal GE: **Uncompensated care provided by for-profit, not-for-profit, and government owned hospitals.** Bmc Health Services Research 2010, **10**:90.

26. Cuadros DF, Branscum AJ, Garcia-Ramos G: **No evidence of association between HIV-1 and malaria in populations with low HIV-1 prevalence.** PloS one 2011, **6**(8):e23458-e23458.

27. Daniak CN, Peretz D, Fine JM, Wang Y, Meinke AK, Hale WB: **Factors associated with time to laparoscopic cholecystectomy for acute cholecystitis.** World Journal of Gastroenterology 2008, **14**(7):1084-1090.

28. Debbink MP, Bader MDM: **Racial Residential Segregation and Low Birth Weight in Michigan's Metropolitan Areas.** Am J Public Health 2011, **101**(9):1714-1720.

29. Dodd CC, Renter DG, Shi X, Alam MJ, Nagaraja TG, Sanderson MW: **Prevalence and Persistence of Salmonella in Cohorts of Feedlot Cattle.** Foodborne Pathogens and Disease 2011, **8**(7):781-789.

30. Dumpa V, Katz K, Northrup V, Bhandari V: **SNIPPV vs NIPPV: does synchronization matter?** Journal of Perinatology 2012, **32**(6):438-442.

31. Egede LE, Gebregziabher M, Lynch CP, Gilbert GE, Echols C: **Longitudinal ethnic differences in multiple cardiovascular risk factor control in a cohort of US adults with diabetes.** Diabetes Res Clin Pract 2011, **94**(3):385-394.

32. Fabio A, Tu L, Loeber R, Cohen J: **Neighborhood Socioeconomic Disadvantage and the Shape of the Age-Crime Curve.** Am J Public Health 2011, **101**:S325-S332.

33. Fernandez de Larrea-Baz N, Alvarez-Martin E, Morant-Ginestar C, Genova-Maleras R, Gil A, Perez-Gomez B, Lopez-Abente G: **Burden of disease due to cancer in Spain.** BMC Public Health 2009, **9**:42-42.

34. Figueiras A, Carracedo-Martinez E, Saez M, Taracido M: **Analysis of case-crossover designs using longitudinal approaches - A simulotion study.** Epidemiology 2005, **16**(2):239-246.

35. Figueiras A, Herdeiro MT, Polonia J, Jesus Gestal-Otero J: **An educational intervention to improve physician reporting of adverse drug reactions - A cluster-randomized controlled trial.** Jama-Journal of the American Medical Association 2006, **296**(9):1086-1093.

36. Filion KB, Steffen LM, Duval S, Jacobs DR,Jr., Blackburn H, Luepker RV: **Trends in Smoking Among Adults From 1980 to 2009: The Minnesota Heart Survey.** Am J Public Health 2012, **102**(4):705-713.

37. Finkelstein JA, Stille C, Nordin J, Davis R, Raebel MA, Roblin D, Go AS, Smith D, Johnson CC, Kleinman K, Chan KA, Platt R: **Reduction in antibiotic use among US children, 1996-2000.** Pediatrics 2003, **112**(3):620-627.

38. Flexeder C, Thiering E, Brueske I, Koletzko S, Bauer C-, Wichmann H-, Mansmann U, von Berg A, Berdel D, Kraemer U, Schaaf B, Lehmann I, Herbarth O, Heinrich J, GINIplus, LISAplus Study Grp: **Growth velocity during infancy and onset of asthma in school-aged children.** Allergy 2012, **67**(2):257-264.

39. Foraker RE, Rose KM, Kucharska-Newton AM, Ni H, Suchindran CM, Whitsel EA: **Variation in Rates of Fatal Coronary Heart Disease by Neighborhood Socioeconomic Status: The Atherosclerosis Risk in Communities Surveillance (1992-2002).** Ann

References included

Epidemiol 2011, **21**(8):580-588.

40. Forte ML, Virnig BA, Eberly LE, Swiontkowski MF, Feldman R, Bhandari M, Kane RL: **Provider Factors Associated with Intramedullary Nail Use for Intertrochanteric Hip Fractures.** Journal of Bone and Joint Surgery-American Volume 2010, **92A**(5):1105-1114.

41. Froeschke G, Sommer S: **Insights into the complex associations between MHC class II DRB polymorphism and multiple gastrointestinal parasite infestations in the striped mouse.** PloS one 2012, **7**(2):e31820-e31820.

42. Goodman ER, Platt R, Bass R, Onderdonk AB, Yokoe DS, Huang SS: **Impact of an environmental cleaning intervention on the presence of methicillin-resistant Staphylococcus aureus and vancomycin-resistant enterococci on surfaces in intensive care unit rooms.** Infection Control and Hospital Epidemiology 2008, **29**(7):593-599.

43. Gumpertz ML, Pickle LW, Miller BA, Bell BS: **Geographic patterns of advanced breast cancer in Los Angeles: Associations with biological and sociodemographic factors (United States).** Cancer Causes & Control 2006, **17**(3):325-339.

44. Hall CB, Lipton RB, Tennen H, Haut SR: **Early follow-up data from seizure diaries can be used to predict subsequent seizures in same cohort by borrowing strength across participants.** Epilepsy & Behavior 2009, **14**(3):472-475.

45. Hollingsworth JM, Krein SL, Dunn RL, Wolf JS,Jr., Hollenbeck BK: **Understanding variation in the adoption of a new technology in surgery.** Med Care 2008, **46**(4):366-371.

46. Holmboe ES, Wang Y, Meehan TP, Tate JP, Ho S, Starkey KS, Lipner RS: **Association between maintenance of certification examination scores and quality of care for medicare beneficiaries.** Arch Intern Med 2008, **168**(13):1396-1403.

47. Holmboe ES, Wang Y, Tate JP, Meehan TP: **The effects of patient volume on the quality of diabetic care for Medicare beneficiaries.** Med Care 2006, **44**(12):1073-7.

48. Hsia RY, Kanzaria HK, Srebotnjak T, Maselli J, McCulloch C, Auerbach AD: **Is Emergency Department Closure Resulting in Increased Distance to the Nearest Emergency Department Associated With Increased Inpatient Mortality?** Ann Emerg Med 2012, **60**(6):707-715.

49. Hunter S, Love-Jackson K, Abdulla R, Zhu W, Lee J, Wells KJ, Roetzheim R: **Sun Protection at Elementary Schools: A Cluster Randomized Trial.** J Natl Cancer Inst 2010, **102**(7):484-492.

50. Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD: **A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis - Does the effect change over time?** Arthritis Rheum 2007, **56**(3):840-849.

51. Janjua NZ, Skowronski DM, Hottes TS, Osei W, Adams E, Petric M, Lem M, Tang P, De Serres G, Patrick DM, Bowering D: **Transmission dynamics and risk factors for pandemic H1N1-related illness: outbreak investigation in a rural community of British Columbia, Canada.** Influenza and Other Respiratory Viruses 2012, **6**(3):e54-e62.

52. Janjua NZ, Skowronski DM, Hottes TS, Osei W, Adams E, Petric M, Sabaiduc S, Chan T, Mak A, Lem M, Tang P, Patrick DM, De Serres G, Bowering D: **Seasonal Influenza Vaccine and Increased Risk of Pandemic A/H1N1-Related Illness: First Detection of the Association in British Columbia, Canada.** Clinical Infectious Diseases 2010, **51**(9):1017-1027.

53. Jia H, Feng H, Wang X, Wu SS, Chumbler N: **A longitudinal study of health service utilization for diabetes patients in a care coordination home-telehealth programme.** J Telemed Telecare 2011, **17**(3):123-126.

54. Johnson DS, Hoeting JA: **Bayesian multimodel inference for geostatistical regression models.** PloS one 2011, **6**(11):e25677-e25677.

55. Kelley ME, Haas GL, van Kammen DP: **Longitudinal progression of negative symptoms in schizophrenia: A new look at an old problem.** Schizophr Res 2008, **105**(1-3):188-196.

56. Kleinman K, Lazarus R, Platt R: **A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism.** Am J Epidemiol 2004, **159**(3):217-224.

57. Kleinschmidt I, Sharp BL, Clarke GPY, Curtis B, Fraser C: **Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa.** Am J Epidemiol 2001, **153**(12):1213-1221.

58. Konety SH, Rosenthal GE, Vaughan-Sarrazin MS: **Surgical volume and outcomes of off-pump coronary artery bypass graft surgery: Does it matter?** J Thorac Cardiovasc

References included

Surg 2009, **137**(5):1116-U98.

59. Kravitz RL, Epstein RM, Feldman MD, Franz CE, Azari R, Wilkes MS, Hinton L, Franks P: **Influence of patients' requests for direct-to-consumer advertised antidepressants - A randomized controlled trial.** Jama-Journal of the American Medical Association 2005, **293**(16):1995-2002.

60. Lambertz CK, Johnson CJ, Montgomery PG, Maxwell JR: **Premedication to reduce discomfort during screening mammography.** Radiology 2008, **248**(3):765-772.

61. Lau DT, Mercaldo ND, Harris AT, Trittschuh E, Shega J, Weintraub S: **Polypharmacy and Potentially Inappropriate Medication Use Among Community-dwelling Elders With Dementia.** Alzheimer Disease & Associated Disorders 2010, **24**(1):56-63.

62. Lederer DJ, Kawut SM, Wickersham N, Winterbottom C, Bhorade S, Palmer SM, Lee J, Diamond JM, Wille KM, Weinacker A, Lama VN, Crespo M, Orens JB, Sonett JR, Arcasoy SM, Ware LB, Christie JD, Lung Transplant Outcomes Grp: **Obesity and Primary Graft Dysfunction after Lung Transplantation The Lung Transplant Outcomes Group Obesity Study.** American Journal of Respiratory and Critical Care Medicine 2011, **184**(9):1055-1061.

63. Lewis EC, Mayer JA, Slymen D: **Postal workers' occupational and leisure-time sun safety behaviors (United States).** Cancer Causes & Control 2006, **17**(2):181-186.

64. Lichtman JH, Leifheit-Limson EC, Jones SB, Wang Y, Goldstein LB: **30-Day Risk-Standardized Mortality and Readmission Rates After Ischemic Stroke in Critical Access Hospitals.** Stroke 2012, **43**(10):2741-2747.

65. Lynch BM, Cerin E, Owen N, Aitken JF: **Associations of leisure-time physical activity with quality of life in a large, population-based sample of colorectal cancer survivors.** Cancer Causes & Control 2007, **18**(7):735-742.

66. Lynch BM, Cerin E, Owen N, Hawkes AL, Aitken JF: **Television viewing time of colorectal cancer survivors is associated prospectively with quality of life.** Cancer Causes & Control 2011, **22**(8):1111-1120.

67. Lynch BM, Cerin E, Owen N, Hawkes AL, Aitken JF: **Prospective relationships of physical activity with quality of life among colorectal cancer survivors.** Journal of Clinical Oncology 2008, **26**(27):4480-4487.

68. Mather FJ, Chen VW, Morgan LH, Correa CN, Shaffer JG, Srivastav SK, Rice JC, Blount G, Swalm CM, Wu XC, Scribner RA: **Hierarchical modeling and other spatial analyses in prostate cancer incidence data.** Am J Prev Med 2006, **30**(2):S88-S100.

69. Mayer JA, Woodruff SI, Slymen DJ, Sallis JF, Forster JL, Clapp EJ, Hoerster KD, Pichon LC, Weeks JR, Belch GE, Weinstock MA, Gilmer T: **Adolescents' Use of Indoor Tanning: A Large-Scale Evaluation of Psychosocial, Environmental, and Policy-Level Correlates.** Am J Public Health 2011, **101**(5):930-938.

70. McCall WV, Blocker JN, D'Agostino R,Jr., Kimball J, Boggs N, Lasater B, Rosenquist PB: **Insomnia severity is an indicator of suicidal ideation during a depression clinical trial.** Sleep Med 2010, **11**(9):822-827.

71. McQueen A, Vernon SW, Myers RE, Watts BG, Lee ES, Tilley BC: **Correlates and predictors of colorectal cancer screening among male automotive workers.** Cancer Epidemiology Biomarkers & Prevention 2007, **16**(3):500-509.

72. Molloy SF, Tanner CJ, Kirwan P, Asaolu SO, Smith HV, Nichols RAB, Connelly L, Holland CV: **Sporadic Cryptosporidium infection in Nigerian children: risk factors with species identification.** Epidemiol Infect 2011, **139**(6):946-954.

73. Mueller S, Polley M, Lee B, Kunwar S, Pedain C, Wembacher-Schroeder E, Mittermeyer S, Westphal M, Sampson JH, Vogelbaum MA, Croteau D, Chang SM: **Effect of imaging and catheter characteristics on clinical outcome for patients in the PRECISE study.** J Neurooncol 2011, **101**(2):267-277.

74. Murphy HR, Rayman G, Duffield K, Lewis KS, Kelly S, Johal B, Fowler D, Temple RC: **Changes in the glycemic profiles of women with type 1 and type 2 diabetes during pregnancy.** Diabetes Care 2007, **30**(11):2785-2791.

75. O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, Ekstrom HL, Gilmer TP: **Impact of Electronic Health Record Clinical Decision Support on Diabetes Care: A Randomized Trial.** Annals of Family Medicine 2011, **9**(1):12-21.

76. Paintsil E, Ghebremichael M, Romano S, Andiman WA: **Absolute CD4(+) T-lymphocyte count as a surrogate marker of pediatric human immunodeficiency virus disease progression.** Pediatr Infect Dis J 2008, **27**(7):629-635.

77. Partovian C, Gleim SR, Mody PS, Li S, Wang H, Strait KM, Allen LA, Lagu T, Normand ST, Krumholz HM: **Hospital Patterns of Use of Positive Inotropic Agents in Patients With Heart Failure.** J Am Coll Cardiol 2012, **60**(15):1402-1409.

References included

78. Patel MM, Chillrud SN, Correa JC, Hazi Y, Feinberg M, Deepti KC, Prakash S, Ross JM, Levy D, Kinney PL: **Traffic-Related Particulate Matter and Acute Respiratory Symptoms among New York City Area Adolescents.** Environ Health Perspect 2010, **118**(9):1338-1343.

79. Polgreen PM, Bohnett LC, Yang M, Pentella MA, Cavanaugh JE: **A spatial analysis of the spread of mumps: the importance of college students and their spring-break-associated travel.** Epidemiol Infect 2010, **138**(3):434-441.

80. Polgreen PM, Sparks JD, Polgreen LA, Yang M, Harris ML, Pentella MA, Cavanaugh JE: **A statewide outbreak of Cryptosporidium and its association with the distribution of public swimming pools.** Epidemiol Infect 2012, **140**(8):1439-1445.

81. Rao S, Van Donkersgoed J, Bohaychuk V, Besser T, Song X, Wagner B, Hancock D, Renter D, Dargatz D, Morley PS: **Antimicrobial Drug Use and Antimicrobial Resistance in Enteric Bacteria Among Cattle from Alberta Feedlots.** Foodborne Pathogens and Disease 2010, **7**(4):449-457.

82. Regenbogen SE, Gawande AA, Lipsitz SR, Greenberg CC, Jha AK: **Do Differences in Hospital and Surgeon Quality Explain Racial Disparities in Lower-Extremity Vascular Amputations?** Ann Surg 2009, **250**(3):424-431.

83. Roe CM, Xiong C, Miller JP, Cairns NJ, Morris JC: **Interaction of neuritic plaques and education predicts dementia.** Alzheimer Disease & Associated Disorders 2008, **22**(2):188-193.

84. Roe CM, Xiong C, Miller JP, Morris JC: **Education and Alzheimer disease without dementia - Support for the cognitive reserve hypothesis.** Neurology 2007, **68**(3):223-228.

85. Ross JS, Maynard C, Krumholz HM, Sun H, Rumsfeld JS, Normand ST, Wang Y, Fihn SD: **Use of Administrative Claims Models to Assess 30-Day Mortality Among Veterans Health Administration Hospitals.** Med Care 2010, **48**(7):652-658.

86. Rusconi PG, Ludwig DA, Ratnasamy C, Mas R, Harmon WG, Colan SD, Lipshultz SE: **Serial measurements of serum NT-proBNP as markers of left ventricular systolic function and remodeling in children with heart failure.** Am Heart J 2010, **160**(4):776-783.

87. Salhofer-Polanyi S, Frantal S, Brannath W, Seidel S, Woeber-Bingoel C, Woeber C, PAMINA Study Grp: **Prospective Analysis of Factors Related to Migraine Aura - The PAMINA Study.** Headache 2012, **52**(8):1236-1245.

88. Sarnat SE, Raysoni AU, Li W, Holguin F, Johnson BA, Luevano SF, Garcia JH, Sarnat JA: **Air Pollution and Acute Respiratory Response in a Panel of Asthmatic Children along the U.S.-Mexico Border.** Environ Health Perspect 2012, **120**(3):437-444.

89. Schelbert EB, Rosenthal GE, Welke KF, Vaughan-Sarrazin MS: **Treatment variation in older black and white patients undergoing aortic valve replacement.** Circulation 2005, **112**(15):2347-2353.

90. Seetharamaiaha R, West BT, Ignash SJ, Pakarinen MP, Koivusalo A, Rintala RJ, Liu DC, Spencer AU, Skipton K, Geiger JD, Hirschl RB, Coran AG, Teitelbaum DH: **Outcomes in pediatric patients undergoing straight vs J pouch ileoanal anastomosis: a multicenter analysis.** J Pediatr Surg 2009, **44**(7):1410-1417.

91. Sikkema KJ, Wilson PA, Hansen NB, Kochman A, Neyfeld S, Ghebremichael MS, Kershaw T: **Effects of a coping intervention on transmission risk behavior among people living with HIV/AIDS and a history of childhood sexual abuse.** Jaids-Journal of Acquired Immune Deficiency Syndromes 2008, **47**(4):506-513.

92. Svensson J, Johannesen J, Mortensen HB, Nordly S, Danish Childhood Diabet Registry: **Improved metabolic outcome in a Danish diabetic paediatric population aged 0-18 yr: results from a nationwide continuous Registration.** Pediatric Diabetes 2009, **10**(7):461-467.

93. Szyszkowicz M: **Ambient air pollution and daily emergency department visits for headache in Ottawa, Canada.** Headache 2008, **48**(7):1076-1081.

94. Szyszkowicz M: **Air pollution and daily emergency department visits for headache in Montreal, Canada.** Headache 2008, **48**(3):417-423.

95. Thomson BKA, MacRae JM, Barnieh L, Zhang J, MacKay E, Manning MA, Hemmelgarn BR: **Evaluation of an electronic warfarin nomogram for anticoagulation of hemodialysis patients.** Bmc Nephrology 2011, **12**:46.

96. Tran AT, Diep LM, Cooper JG, Claudi T, Straand J, Birkeland K, Ingskog W, Jenum AK: **Quality of care for patients with type 2 diabetes in general practice according to patients' ethnic background: a cross-sectional study from Oslo, Norway.** Bmc Health Services Research 2010, **10**:145.

References included

97. van Baal PH, Engelfriet PM, Hoogenveen RT, Poos MJ, van den Dungen C, Boshuizen HC: **Estimating and comparing incidence and prevalence of chronic diseases by combining GP registry data: the role of uncertainty.** BMC Public Health 2011, **11**:163.

98. Vithiananthan S, Gero D, Zhang JY, Machan JT: **A case-controlled matched-pair cohort study of single-incision and conventional laparoscopic gastric band patients in a single US center with 1-year follow-up.** Surgical Endoscopy and Other Interventional Techniques 2012, **26**(12):3467-3475.

99. Wagner A, Simon C, Oujaa M, Platat C, Schweitzer B, Arveiler D: **Adiponectin is associated with lipid profile and insulin sensitivity in French adolescents.** Diabetes Metab 2008, **34**(5):465-471.

100. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, Agusti A, Anderson WH, Lomas DA, DeMeo DL: **Systemic Steroid Exposure Is Associated with Differential Methylation in Chronic Obstructive Pulmonary Disease.** American Journal of Respiratory and Critical Care Medicine 2012, **186**(12):1248-1255.

101. Williams AL, Khattak AZ, Garza CN, Lasky RE: **The behavioral pain response to heelstick in preterm neonates studied longitudinally: Description, development, determinants, and components.** Early Hum Dev 2009, **85**(6):369-374.

102. Williams ED, Magliano DJ, Zimmet PZ, Kavanagh AM, Stevenson CE, Oldenburg BF, Shaw JE: **Area-Level Socioeconomic Status and Incidence of Abnormal Glucose Metabolism The Australian Diabetes, Obesity and Lifestyle (AusDiab) study.** Diabetes Care 2012, **35**(7):1455-1461.

103. Wright KC, Ravoori MK, Dixon KA, Han L, Singh SP, Liu P, Gupta S, Johnson VE, Kan Z, Kundra V: **Perfusion CT Assessment of Tissue Hemodynamics Following Hepatic Arterial Infusion of Increasing Doses of Angiotensin II in a Rabbit Liver Tumor Model.** Radiology 2011, **260**(3):718-726.

104. Yih WK, Lieu TA, Rego VH, O'Brien MA, Shay DK, Yokoe DS, Platt R: **Attitudes of healthcare workers in US hospitals regarding smallpox vaccination.** BMC Public Health 2003, **3**:20.

105. Zafar AM, Harris TJ, Murphy TP, Machan JT: **Patients' Perspective about Risks and Benefits of Treatment for Peripheral Arterial Disease.** Journal of Vascular and Interventional Radiology 2011, **22**(12):1657-1661.

106. Zeltzer LK, Lu Q, Leisenring W, Tsao JCI, Recklitis C, Armstrong G, Mertens AC, Robison LL, Ness KK: **Psychosocial outcomes and health-related quality of life in adult childhood cancer survivors: A report from the Childhood Cancer Survivor Study.** Cancer Epidemiology Biomarkers & Prevention 2008, **17**(2):435-446.

107. Zhang J, Himes JH, Hannan PJ, Arcan C, Smyth M, Rock BH, Story M: **Summer effects on body mass index (BMI) gain and growth patterns of American Indian children from kindergarten to first grade: a prospective study.** BMC Public Health 2011, **11**:951.

108. Zhu CW, Scarmeas N, Torgan R, Albert M, Brandt J, Blacker D, Sano M, Stern Y: **Clinical characteristics and longitudinal changes of informal cost of Alzheimer's disease in the community.** J Am Geriatr Soc 2006, **54**(10):1596-1602.

Table: Journals according to field of knowledge.

AREA	N
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	22
CLINICAL NEUROLOGY	10
ONCOLOGY	8
INFECTIOUS DISEASES	7
PEDIATRICS	7
MEDICINE, GENERAL & INTERNAL	6
CARDIAC & CARDIOVASCULAR SYSTEMS	5
ENDOCRINOLOGY & METABOLISM	5
SURGERY	5
BIOLOGY	4
HEALTH CARE SCIENCES & SERVICES	4
RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING	3
CRITICAL CARE MEDICINE	2
ENTOMOLOGY	2
ENVIRONMENTAL SCIENCES	2
FOOD SCIENCE & TECHNOLOGY	2
GASTROENTEROLOGY & HEPATOLOGY	2
OBSTETRICS & GYNECOLOGY	2
RHEUMATOLOGY	2
ALLERGY	1
EMERGENCY MEDICINE	1
GERIATRICS & GERONTOLOGY	1
ORTHOPEDICS	1
PSYCHIATRY	1
SUBSTANCE ABUSE	1
TROPICAL MEDICINE	1
UROLOGY & NEPHROLOGY	1
Total	**108**

Table: Estimation methods according to the name used (GLMM, HGLM, MGLM)

Estimation method	**GLMM N=92**	**HGLM N=14**	**MLGM N=2**	**Total**
Adaptative Quadrature likelihood Approximation	1 (1.1%)	0 (0.0%)	0 (0.0%)	1 (0,9%)
Maximum Likelihood	3 (3.3%)	0 (0.0%)	0 (0.0%)	3 (2.8%)
NR	74 (80.4%)	11 (78.6%)	2 (100%)	87 (80.6%)
Penalized Quasi- likelihood	8 (8.7%)	0 (0.0%)	0 (0.0%)	8 (7.4%)
Posterior mean	3 (3.3%)	2 (14.3%)	0 (0.0%)	5 (4.6%)
Pseudo-likelihood	1 (1.1%)	1 (7.1%)	0 (0.0%)	2 (1.9%)
Restricted Maximum Likelihood	2 (2,2%)	0 (0.0%)	0 (0.0%)	2 (1.9%)

3 Estimació dels paràmetres en els 'Generalized Linear Mixed Models'

Els principals detalls d'aquest capítol s'expliquen en l'article corresponent a la revista SORT que apareixerà a la Secció 3.4. En aquest treball es descriu i es compara l'estimació dels paràmetres del model Poisson GLMM en el cas de recomptes a través de tres filosofies estadístiques.

3.1 Filosofies, mètodes i algoritmes d'estimació

La filosofia o 'principle', el mètode d'estimació, l'algoritme i el *software* utilitzat en un estudi poden influir en la validació i fiabilitat de l'estimació dels paràmetres d'un 'Generalized Linear Mixed Model'(GLMM) [39, 46, 47].

El 'principle' consta tant de la filosofia com dels procediments específics que utilitza la versemblança. La filosofia clàssica ('likelihood principle') dóna lloc a models que inclouen efectes fixes i aleatoris [48, 49], i la inferència clàssica es basa en la versemblança marginal on els efectes aleatoris s'integren a part [27]. En la filosofia jeràrquica ('extended like-

lihood') [42,50] tota la informació de les dades (efectes aleatoris i fixes) s'inclouen en una probabilitat conjunta ('joint likelihood'). I pel que fa a la filosofia Bayesiana ('Bayesian principle') es basa en un marc probabilistic que combina la versemblança i la informació a priori [51].

El mètode d'estimació és una manera de resoldre la integral (2.6) i així obtenir la versemblança per estimar els paràmetres en el model. Diferents mètodes d'estimació basats en aproximacions o de simulació s'han desenvolupat en els últims anys [10,39].

Les estimacions es poden obtenir mitjançant l'ús d'algoritmes (tècniques iteratives) tal com s'aplica en paquets de *software*. Les implementacions de *software* difereixen considerablement en la flexibilitat, el temps de càlcul i la facilitat d'ús [46]. I pel que fa a nivell de *software* o computacional, s'ha especificat un apartat en aquest capítol.

En la literatura dels GLMMs, moltes vegades, per esmentar els mètodes d'estimació, s'han utilitzat diferents termes com *approaches*, *techniques*, *algorithms*, *paradigmes*, *principles*, *criterion*, *packages*, etc que poden donar lloc a confussió. Per aquesta raó, una distinció entre els enfocaments en termes de filosofia, mètode d'estimació i algoritme es mostra a la Taula1 del primer article publicat en aquest capítol.

Taula 3.1: Resum de filosofies estadístiques

Filosofies	**Mètodes**	**Algoritmes**
Marginal Likelihood	Maximum likelihood	Newton-Raphson (N-R), Fisher scoring, Penalized iteratively reweighted least squares (PIRLS) Adaptative Gauss Hermite Quadrature (GHQ)
Extended likelihood	h-likelihood	N-R, Iterative weighted least squares (IRWLS)
Bayesian	Posterior mean	MCMC, Integrated Nested Laplace Approximations (INLA)

Per exemple en els llibres de GLMM com el de McCulloch i Searle

(2001) [27] es mostra la filosofia que es basa amb la inferència clàssica, el llibre de Lee i Nelder (2006) [42] mostra la filosofia que es basa amb els models jeràrquics amb efectes aleatoris, o el de Havard Rue que mostra la filosofia que [44] que es basa amb els models *Latent Gausssian Fields* i amb la metodologia INLA bayesiana.

L'existència de diferents mètodes d'estimació dels paràmetres d'un model GLMM implica que abans de començar el modelatge s'hagi de seleccionar un d'aquests mètodes. En aquest capítol han estat proposades diferents filosofies per estimar els paràmetres, i és en el primer article que s'ha comparat el rendiment de tres filosofies estadístiques, la versemblança marginal clàssica, la jeràrquica, i la filosofia bayesiana.

Dintre dels mètodes freqüentistes, una primera solució seria la de màxima versemblança. Però, com ja s'ha dit, moltes vegades no es viable obtenir una solució analítica. Altres aproximacions, a part de PQL, com Laplace i GHQ varen ser proposades [27]. Lee i Nelder van proposar incloure efectes aleatoris als GLM a través de la utilització de la versemblança jeràrquica o 'h-likelihood' [42, 52]. A diferència de la inferència clàssica, on s'utilitza la versemblança marginal i es separen els efectes aleatoris de la integral, en la versemblança jeràrquica tota la informació dels efectes fixes i aleatoris queden inclosos en una mateixa versemblança conjunta. A més a més, en aquesta filosofia jeràrquica no hi ha la restricció que la distribució de l'efecte aleatori sigui una Normal, i permet així l'ús de distribucions conjugades. I pel que fa als mètodes bayesians, difereixen de la inferència clàssica i jeràrquica, tan en la seva pròpia filosofia com en l'especificació dels procediments utilitzats. En la filosofia bayesiana, tots els paràmetres són tractats com a variables aleatòries i es basa en l'estimació de la distribució de probabilitat posterior dels paràmetres, obtinguda de combinar la versemblança de les dades amb les creences a 'priori'. L'ús de l'algoritme MCMC és el més popular i més utilitzat, però recentment s'ha desenvolupat un nou algoritme, l'INLA, que ofereix avantatges [39, 44, 53].

3.2 Sobredispersió en els GLMM

En els models GLM o GLMM, un dels problemes que sorgeixen en l'anàlisi de dades amb distribucions Binomials i de Poisson (els casos que trobem majoritàriament en l'àmbit de medicina) és que el paràmetre d' escala o dispersió ϕ és diferent a la unitat. Aquest fenomen és conegut com a sobredispersió quan $\phi > 1$ o sotsdispersió quan $\phi < 1$. Ignorar la sobredispersió i sotsdispersió podria provocar l'obtenció de conclusions clínicament diferents. En aquest capítol ens centrarem en el cas de sobredispersió. Models amb el fenomen de la sotsdispersió es tenen en compte també amb més detall en el llibre de Hilbe (2007).

La situació de sobredispersió es produeix quan la variància $Var(Y_i)$ és superior a la mitjana. En el cas de la distribució de Poisson, concretament en un model GLMM, què és el que ens hem centrat en el capitol 3, seria quan $Var(Y_i) > \mu_i$.

Si definim el model Poisson generalitzat lineal mixt com:

$$\log(\mu_i) = \log(\lambda_i) + X_i\beta + u_i,$$

on β són els paràmetres que representen els efectes fixes, u_i l'efecte aleatori en l'intercept pel subjecte i, i λ_i és l'*offset*. Els efectes aleatoris són assumits com independents i distribuïts com una Normal amb mitjana 0 i variància σ^2. La variància marginal es pot expressar com:

$$\mathrm{Var}(Y_i) = \phi \cdot \mu_i,$$

on Φ és el paràmetre de dispersió. En el cas de la distribució de Poisson tenim $\phi = 1$, per això si $\phi > 1$, ens referim al fenomen de sobredispersió. En aquest cas, les dades tenen la variància més gran que l'esperada.

Si es vol conèixer la sobredispersió generada assumint que $u_i \sim \mathcal{N}(0, \sigma^2)$, ens fixarem abans amb els moments marginals (la mitjana i la variància) descrits en el següent sistema d'equacions [27].

Sabent que l'esperança marginal es defineix:

$$\mu_i = \lambda_i \cdot \exp(\beta' X_i + u_i)$$
$$\mathrm{E}(\mu_i) = \mu = \bar{\lambda}_i \cdot \big(\exp(\beta' \bar{X} + (\sigma_u^2/2))\big)$$

o

$$\log(\mathrm{E}(\mu_i)) = \log(\mu) = \log(\bar{\lambda}_i) + (\beta' \bar{X} + (\sigma_u^2/2))\big)$$

on σ_u^2 és la variància de l'efecte aleatori, $\bar{\lambda}_i$ és l'offset marginal i μ és la mitjana marginal. L'escala ϕ depèn de σ_u^2 i μ.

La variància marginal es definida:

$$\mathrm{Var}(Y_i) = \mathrm{E}(Y_i)\big(\exp(\beta' X_i)(\exp(3\sigma_u^2/2) - \exp(\sigma_u^2/2)) + 1\big)$$

La variància de l'efecte aleatori σ_u^2 expressa la sobredispersió, ja que, si σ_u^2 és igual a zero la variància teòrica i l'empírica coincidirien.

Podem expressar doncs el terme de dispersió generat a partir de:

$$\phi = \frac{\mathrm{Var}(Y_i)}{\mathrm{E}(Y_i)} = \exp(\beta' X_i(\exp(3\sigma_u^2/2) - \exp(\sigma_u^2/2)) + 1$$
$$= \exp(\beta' X_i + \sigma_u^2/2)(\exp(\sigma_u^2) - 1) + 1 = \mu \cdot (\exp(\sigma_u^2) - 1) + 1 \quad (3.1)$$

Si el terme dispersió és més gran que 1, podem veure com la variància és més gran que la mitjana [27].

3.2.1 Validació dels models i obtenció del paràmetre de dispersió

La validació dels models GLM o GLMM es pot realitzar mitjançant la deviància, una generalització de la suma de quadrats de l'error (SSE) en els models lineals. La deviància del model es calcula com:

$$D(\hat{\mu}; y) = 2[\ell(y; y) - \ell(\hat{\mu}; y)]$$

on $\ell(y;y)$ és la versemblança avaluada en $\mu = y$ representant doncs la versemblança màxima i $\ell(\hat{\mu};y)$ és la versemblança segons el model considerat. Quan el paràmetre d'escala ϕ és conegut, s'ha de calcular la deviància escalada:

$$D(\hat{\mu};y) = \frac{D(\hat{\mu};y)}{\phi}$$

que es distribueix aproximadament com una distribució χ^2_{n-p}.

Una punt important al validar el model és conèixer si hi ha sobredispersió o sotsdispersió. Cal comentar que hi ha altres aspectes alhora de validar el model GLMM tal com s'especifica a l'article de l'apartat 2.6.

Per conèixer si el model té sobredispersió o no, hi ha dues opcions [27,54].

1. Quan el paràmetre d'escala és desconegut i és vol estimar, es pot utilitzar la següent expressió:

$$\hat{\phi} = D(\hat{\mu};y)/(n-p) \tag{3.2}$$

2. Una altra opció és a partir dels anomenats residus de Pearson. Un índex amb propietats similars a la deviància és l'estadístic χ^2 de Pearson que donarà lloc a la prova de bondat d'ajust:

$$\hat{\phi} = \frac{1}{n-r}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{V(\mu_i)}.$$

En dividir la funció de versemblança per un paràmetre d'escala no s'obté una funció de versemblança pròpiament dita. S'obté el que es diu una funció de quasi-versemblança. En aquest cas però no estaríem

parlant de realitzar inferència pròpiament dita (per exemple en el cas de distribució Poisson estaríem davant de la distribució Quasi-Poisson), i els índexs de deviància no seran apropiats.

És possible també definir els residus de Pearson estudantitzats al dividir els residus de Pearson per l'arrel quadrada de $1 - h_i$, on h_i és el *leverage* de l'observació i.

3.2.2 Causes i solucions per a problemes de sobredispersió

És important preguntar-se per què es produeix la sobredispersió. Aquest fenomen es pot produir per la presència de *missings* en les covariables o interaccions, observacions anòmales (*outliers*) en la variable resposta, una situació de *zero inflation* (una distribució que permet observacions freqüents de valor zero), correlació, gran variació de les dades, patrons no lineals modelats com a lineals o elecció no apropiada d'una funció d'enllaç.

La sobredispersió real però existeix quan no identifiquem cap d'aquestes possibles causes esmentades anteriorment. Així, aquest fenomen es produeix perquè la variació de les dades és realment més gran que la mitjana.

El fet d'ajustar models més complexes no ajuda, quan per exemple s'afegeixen moltes covariables i també la presència d'interaccions.

Una de les causes més recurrents en la distribució Binomial i de Poisson és la presencia anòmala de valors zero. En aquesta situació, en què les distribucions de Poisson, Binomial, Negativa Binomial no són adequades, les dades tenen un excés de zeros que donaran lloc a una sobredispersió. Alternatives com la utilització de models 'zero inflated poisson'(ZIP) o Zero Inflated Negative Binomial (ZINB) serien una solució [55].

Davant la incertesa de com resoldre aquest problema, diferents estratègies s'han proposat:

- Fer una correcció, concretament multiplicar els errors estàndard dels paràmetres per l'arrel quadrada del valor del paràmetre de sobredispersió.

- Ajustar una distribució Quasi-Poisson o Quasi-Binomial, Binomial Negativa o Beta-Binomial, o altres procediments com el GEE.
- Examinar les dades i intentar trobar la causa del problema de sobredispersió.

Cal destacar que normalment la sobredispersió s'ignora quan $\phi < 1.5$ [54]. En canvi, quan ens trobem amb el fenomen de $\phi > 1.5$ necessitarem corregir els errors estàndard per als paràmetres estimats. Segons Zuur [54], és a partir d'un valor de paràmetre de dispersió més gran de 15 o 20, quan aquesta correcció no seria útil i s'haurien de considerar altres mètodes o distribucions alternatives com GEE, Poisson-Lognormal, Binomial Negativa o ZIP.

3.3 *Software* per a l'estimació dels GLMM

Actualment els GLMM estan disponibles en diferents *softwares* estadístics.

Com s'ha comentat anteriorment, els mètodes d'estimació més recents dels GLMM són de la dècada dels 90. A nivell computacional, la macro GLIMMIX de SAS (1992) va ser la primera que es va disposar per ajustar GLMM amb el mètode d'estimació PQL, mètode que produiria estimacions esbiaixades en els paràmetres [56]. De totes maneres, la macro GLIMMIX és en la versió 9.2 la que permetria ajustar els GLMM amb les aproximacions de mètodes d'estimació de Laplace i GHQ de manera eficient [57,58]. En l'actualitat, hi ha diferents *softwares* disponibles per ajustar els GLMM, però és a partir de l'any 2000 que es van implementar en la seva majoria, com ara ASReml [59], R (lme4; primera actualització del paquet va ser a l'any 2003), ADMB (exemple de Poisson GLMM; 2006), MLwiN [60], SAS (PROC GLIMMIX va esdevenir el procediment estàndard amb la versió 9.2, 2008), entre d'altres [57,61].

Els *softwares* disponibles poden ajustar diferents variables resposta de la família exponencial, com la distribució de Poisson, Binomial, Gamma, Inversa Gaussiana, encara que les més utilitzades i que la majoria té per

defecte són la distribució Poisson i Binomial. Actualment, s'estan realitzant diferents estudis que comparen els paquets estadístics disponibles i mètodes d'estimació [46,47,62,63] . Darrerament hi ha disponible diferents llibres acadèmics sobre els GLMM, la seva aplicació, amb diferents filosofies i paquets d'R [54]. Tot i així, és a través d'algun llibre com 'Mixed-effects Modeling with R' de Douglas Bates, dels diferents tutorials dels diferents paquets (MCMC, hglm) i sobretot de la web `http://glmm.wikidot.com/faq` revisada per Ben Bolker que es pot obtenir informació actualitzada sobre el *software* estadístic disponible dels GLMM. A més a més, es pot trobar gran informació respecte els GLMM i aplicacions amb el seu *software* (principalment amb R) en el forum de R, 'R-sig-mixed models' (`https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`). En la present tesi ens hem centrat en el *software* lliure R i amb tres dels seus paquets (INLA, hglm, i lme4) que representen tres filosofies per ajustar els models GLMM. El paquet lme4 és dels paquets més estables que segueix la filosofia clàssica dels GLMM amb les seves possibles extensions. L'INLA és una metodologia recent que permet fer inferència bayesiana per a models anomenats *Latent Gaussian Models*. Tot i així, és un paquet especialitzat amb dades espacials que poden ser modelades utilitzant models jeràrquics, molt similars als GLMM [53]. Una altre paquet és el hglm que es basa en una altra filosofia poderosa per a models jeràrquics, coneguts com a HGLM, els 'Hierarchical Generalized Linear Models' (Lee et al 2006). Aquests, es construeixen basats en la idea dels GLMM, però proporcionen més flexibilitat. Per exemple, aquesta filosofia proporciona un procediment per modelar la variància de la mateixa manera que la mitjana, utilitzant un model lineal doblement jeràrquic, anomenat 'doubly hierarchical linear model' [28].

Actualment, segons Bolker [39], hi ha diferents paquets d'R disponibles per ajustar els GLMM com glmmML [64], glmmPQL o lme4 [65]. Una de les funcions més populars i estables per GLMM es diu 'glmer 'disponible en el paquet lme4. Aquest paquet implementa el GHQ per aproximar la log-versemblança utilitzant integració numèrica. Per defecte, s'utilitza

l'aproximació de Laplace quan només s'utilitza un punt de quadratura. Hem de tenir en compte que la versió del paquet lme4.0 (versió 0) ha estat substituida per una de nova, la versió lme4.1 que ha ajudat a tenir menys problemes amb la falsa convergència i una major flexibilitat per a l'exploració i la solució de problemes de convergència.

Pel que fa a la filosofia o principi jeràrquic, dos dels paquets disponibles per ajustar GLMM són hglm i HGLMM [66, 67]. Pel que fa als paquets per realitzar inferència bayesiana trobem també un ampli ventall disponible en els models GLMM, paquets generals com glmmBUGS [68] i R2WinBugs [69] i els paquets especialitzats com glmmAK [70], MCMCglmm [71], i INLA [72].

L'INLA és una eina recent d'inferència bayesiana que es basa en tres ingredients: 'Gaussian Markov random fields', 'Latent Gaussian models', i aproximacions de Laplace. Aquest mètode combina aproximacions de Laplace i integració numèrica de forma eficient com a gran alternativa de l'estimació via MCMC. Aquest marc analític és flexible a manipular models i versemblances però està dirigit sobretot a aplicacions complexes amb models on es pot incloure estructura espacial o suavitzat temporal i on l'estimació MCMC pot ser massa difícil d'aplicar. L'INLA substitueix les simulacions MCMC amb rapidesa a nivell computacional, i també en quant a precisió i qualitat.

Darrerament han sorgit noves versions en aquests paquets esmentats anteriorment i és per això que descrivim algunes de les noves característiques que s'inclouen.

lme4 Des de la versió 0.99 (la més coneguda i utilitzada com *lme*4.0) del paquet s'han anat creant noves versions i discutint noves característiques del paquet a través del fòrum 'r-sig-mixed models'. Actualment la nova versió és la $1.1-7$ i ofereix algunes avantatges respecte la 0.99, sobretot a nivell de de convergència i d'optimització. Aquestes novetats es poden trobar al següent enllaç: `http://cran.r-project.org/web/packages/lme4/news.html`

A continuació es destaquen les principals avantatges en aquest paquet.

- El més estable dels paquets estudiats. Permet ajustar la versemblança GHQ, Laplace o PQL.
- Per a dissenys balancejats, no balancejats i *nested models*.
- Permet incloure més d'un efecte aleatori.
- S'acaba d'establir una versió nova, $1.1-7$ que permet més rapidesa, flexibilitat per als models GLMM.

A continuació es destaquen els principals inconvenients d'aquest paquet.

- L'efecte aleatori s'assumeix només que segueix una distribució normal.
- No reporta l'error estàndard per defecte de l'estimació de la variància de l'efecte aleatori. Amb la nova versió del paquet només és possible via *bootstrapping*.
- Encara en desenvolupament.

hglm La nova versió 2.0 del paquet permet ajustar models Generalitzats amb efectes aleatoris amb les següents novetats respecte la versió 1.2 anterior:

- Ajustar varis efectes aleatoris de diferents distribucions.
- Ajustar un predictor lineal pel component de la variància o 'dispersion of the random effects'.
- Ajustar un model espacial CAR per efectes aleatoris introduït per Besag, York i Mollié.
- Utilitza la correcció HL11 de Lee i Nelder.
- Utilitza la prova de la raó de versemblança per al paràmetre de la dispersió dels efectes aleatoris (funció LRT).

A continuació es destaquen les principals avantatges en aquest paquet.

- Més ràpid que el paquet HGLMM (paquet amb la mateixa filosofia).
- L'efecte aleatori pot tenir distribucions diferents a la Normal.
- Permet distribucions conjugades com Beta-Binomial, Poisson-Gamma.
- Més especialitzat en l'àmbit de genètica.
- Fàcil migració i interface flexible, i computació eficient.
- És possible utilitzar quasi-versemblança en casos de sots o sobre-dispersió.

A continuació es destaquen els principals inconvenients d'aquest paquet.

- Quan la variable resposta segueix una distribució Binomial presenta més problemes de convergència.
- Encara en desenvolupament.
- No permet ajustar efectes aleatoris aniuats.

Més novetats es poden trobar en el tutorial: http://cran.r-project.org/web/packages/hglm/vignettes/hglm.pdf

INLA Hi ha un gran creixement d'articles amb l'ús d'aquest paquet. Un dels més recents i que tracta sobre aspectes més novedosos des de la filosofia bayesiana és el de Martins i Rue [73] que explica una novetat en aquesta filosofia que és com construir les priors. Més informació respecte les novetats d'aquest paquet és poden trobar a la web principal d'INLA http://www.r-inla.org/.

A continuació es destaquen les principals avantatges en aquest paquet.

- Camp d'aplicació molt ampli: Més especialitzat en models 'spatio-temporal', però pot ajustar els 'Latent gaussian models' que cobreix els models més comuns i utilitzats (GLM, *Generalized additive models, smoothing spline models, state space models, semi-parametric regression, spatial and spatiotemporal models, log-Gaussian Cox process and geostatistical* i *geoadditive models*).

- Molt ràpid si es compara amb altres paquets bayesians.
- Paquet d'R disponible, i amb fòrum per discussió sobre qualsevol problema de codi.
- Permet incloure més d'un efecte aleatori, i amb component geoespacial.

A continuació es destaquen els principals inconvenients d'aquest paquet.

- Caldrien eines més intuïtives de validació dels models.
- Encara en desenvolupament.

3.4 Article 2: Parameter Estimation of Poisson Generalized Linear Mixed Models Based on Three Different Statistical Principles: a Simulation Study

Les següents pagines mostren un dels treballs acceptat per a la publicació a la revista SORT, revista situada al segon cuartil i amb factor d'impacte 1.33 (JCR,2014).

El primer treball acceptat a la revista SORT present en aquest capítol es descriu amb més detall i és citat a continuació:

Casals M, Langohr K, Carrasco JLL, Rönnegård L. Parameter estimation of Poisson generalized linear mixed models based on three different statistical principles: a Simulation Study

El problema de seleccionar el 'millor' procediment per a l'estimació de paràmetres en els GLMM és molt complexa. Per això, nosaltres hem proposat com a estratègia la implementació d'una simulació amb diferents escenaris aplicat a dades reals d'un esport de contacte, la 'Lucha leonesa'.

En aquestes dades s'ha treballat amb una cohort de 213 lluitadors (només professionals d'aquest esport) que han competit durant les temporades 2005-2010. La variable d'interès és la incidència de lesions, concretament, el nombre de lesions per combat, que segueixen una distribució

de Poisson amb un 'offset', i és per això que s'ha emprat un model Poisson generalitzat lineal mixt.

Més

[SORT] Editor Decision

Safata d'entrada x

RACO no-reply@csuc.cat mitjançant cesca.cat
per a usuari

Dear Dr. Casals,

I have now received the report from the referees assigned with the evaluation of this revised version. They are very positive and then, I am pleased to inform you that your submission "Parameter Estimation of Poisson Generalized Linear Mixed Models Based on Three Different Statistical Principles: a Simulation Study" has been accepted for publication, congratulations for this nice piece of work.

Your paper has been placed in the queue to appear in one of the next issues of SORT after the editorial grammar/language revision.

Let me comment you that the Supplementary Material of your paper will be published in the journal website.

Thanks again for considering SORT for this work.

Best regards,

Pere Puig

[Message sent from RACO do not reply directly. If you need to contact the person who sent this message, log in www.raco.cat and do it from there.
For any questions, write to raco@csuc.cat.]
The Editorial Committee

Parameter Estimation of Poisson Generalized Linear Mixed Models Based on Three Different Statistical Principles: a Simulation Study

Martí Casals[1,2,3,4], Klaus Langohr[5], Josep Lluís Carrasco[1], Lars Rönnegård[6]

[1] Department of Public Health, Universitat de Barcelona, Barcelona, Spain

[2] Epidemiology Service, Public Health Agency of Barcelona, Barcelona, Spain

[3] Area of Biostatistics, Universitat Internacional de Catalunya, Barcelona, Spain

[4] CIBER de Epidemiología y Salud Pública (CIBERESP), Spain

[5] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya/ BARCELONATECH, Barcelona, Spain

[6] Statistics Unit, Dalarna University, Falun, Sweden

Abstract

Generalized linear mixed models are flexible tools for modeling non-normal data and are useful for accommodating overdispersion in Poisson regression models with random effects. Their main difficulty resides in the parameter estimation because there is no analytic solution for the maximization of the marginal likelihood. Many methods have been proposed for this purpose and many of them are implemented in software packages. The purpose of this study is to compare the performance of three different statistical principles –Marginal likelihood, Extended likelihood, Bayesian analysis– in R via simulation studies. Real data on contact wrestling are used for illustration.

Keywords: Estimation methods; Overdispersion; Poisson Generalized Linear Mixed Models; Simulation study; Statistical principles; Sport injuries.

MSC2010: 62J12; 62P99; 62F99

1 Introduction

One of the methodologies used to study disease incidence in medicine or injuries in sport research is the generalized linear model (GLM). This methodology is able to model counts and proportions besides normally distributed variables (McCullagh and Nelder, 1989). Furthermore, GLMs assume that the observations conditioned on the predictors are independent and identically distributed. However, these assumptions may be violated in some situations, such as longitudinal studies, where there are repeated measures and, hence, correlated data. Ignoring correlation of data when fitting the model may lead to biased estimates and misinterpretation of results (Bolker et al., 2009).

Generalized linear mixed models (GLMMs) are an extension of GLMs adding random effects in the linear predictor term in a regression setting (Breslow and Clayton, 1993). The GLMM is a more flexible analysis approach for analyzing non-normal data and it is known to be useful for accommodating the overdispersion in Binomial or Poisson regression models, and modelling the dependence structure among outcome variables for longitudinal or repeated measures data (Williams, 1982; Breslow, 1984).

The main difficulty of these models is the estimation of their parameters,

as it is often not viable to obtain an analytic solution that allows maximizing the marginal likelihood of the data. Due to this fact, different estimation methods based on approximation or simulation have been developed in recent years. One approximation using numerical integration is the Gauss-Hermite quadrature (GHQ) (McCulloch and Searle, 2001). However, there are alternatives to the marginal likelihood principle including Bayesian statistics and the extended likelihood principle. For example, the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) is a Bayesian implementation and the hierarchical (h-)likelihood is an implementation of the extended likelihood principle (Lee and Nelder, 1996, 2001). It is worth mentioning that the comparison between Bayesian and non-Bayesian methods is difficult to perform given that they are different principles.

Nowadays, GLMMs are implemented in most statistical software packages and several researchers have published and updated different guides and reviews of different software packages for fitting a GLMM. West et al. (2014) introduce the fitting and interpretation of several types of linear mixed models using the statistical software packages SAS, SPSS, Stata, R, and HLM. Dean and Nielsen (2007) review the theoretical background of generalized linear mixed models and the inferential techniques that have been developed for SAS, S-Plus, and contributed R packages. Bolker et al.

(2009) describe the use of generalized linear mixed models for ecology and evolution and give information on available functions and packages in SAS or R. For further comparisons of statistical software for GLMMs for binary responses and frailty models, see, for instance, Zhang et al. (2011); Li et al. (2011); Hirsch and Wienke (2012); Kim et al. (2013); or Grilli et al. (2014).

The aim of this work is to compare three different statistical principles —Marginal likelihood, Extended likelihood, and Bayesian analysis; see, Table 1— to estimate the parameters of a Poisson Mixed Model in R using both real and simulated data. It is structured as follows: in Section 2, we briefly review the definition of the GLMM and highlight the problem of deriving and maximizing the likelihood. In Sections 3 and 4, we give a theoretical description according to the statistical principle used. Several contributed R packages for the fit of GLMMs are presented in Section 5 and three of them are used in Section 6 for the analysis of the motivating real data set on Leonese Wrestling. These data are then used to define the settings of the simulation study presented in Section 7. In Sections 8 and 9, the results of the simulation are presented and discussed, and recommendations are given on which statistical principle is, preferably, to be used in each of the settings under study.

2 Generalized Linear Mixed Models

The GLMM extends the GLM by adding normally distributed random effects to the linear predictor. As Bolker et al. (2009) point out, GLMMs combine the properties of linear mixed models (LMMs) and GLMs by using link functions and exponential family distributions such as Binomial or Poisson distributions.

Let $Y_i = (Y_{i1}, \ldots, Y_{im})'$ be a vector of m observations of the response variable of interest corresponding to subject $i, i = 1, \ldots, n$ and $u_i, i = 1, \ldots, n$, be the random effects vector of the same subject. Conditional on u_i, the distribution of Y_i is assumed to be from the exponential family type with density function $f(Y_i|u_i;\cdot)$ and with conditional mean $\mu_i = \mathrm{E}(Y_i|u_i)$ and conditional variance $\mathrm{Var}(Y_i|u_i) = \Phi V(\mu_i)$, where Φ is the dispersion parameter and $V(\mu_i)$ is the variance function of the GLMM.

The definition of the GLMM is completed by introducing a monotone and differentiable function $g(\cdot)$ known as the link function (McCullagh and Nelder, 1989) and a linear predictor η as follows:

$$\eta_i = g(\mu_i) = X_i\beta + Z_i u_i, i = 1, \ldots, n,$$

where X_i (of dimension $m \times k$) and Z_i ($m \times l$) are subject i's design matrices associated with fixed and random effects, respectively. Vector β ($k \times 1$) is

the fixed effects vector and u $(l \times 1)$ is the random effects vector assumed to follow a multivariate Gaussian distribution with mean vector 0 and unknown positive definite covariance matrix Σ. Its density function is denoted by $f(u;\Sigma)$.

Estimation in the GLMM and theoretical description of the likelihood principle

By the local independence assumption, the conditional density of Y given u has the form

$$f(Y|u;\beta) = \prod_{i=1}^{n} f(Y_i|u_i;\beta)$$

and the multivariate density function of u is given by

$$f(u;\Sigma) = \prod_{i=1}^{n} f(u_i;\Sigma).$$

The likelihood principle involves two kinds of objects: observed random variables (the data) and (unknown) fixed parameters. In the case of models with random effects, the estimation is based on the marginal likelihood where the random effects are integrated out (Birnbaum, 1962; Pawitan, 2001). Hence, the following likelihood function needs to be maximized in order to obtain the maximum likelihood (ML) estimates for β and the vari-

ance components in Σ:

$$l(\beta,\Sigma|Y) = f(Y;\beta) = \int f(Y|u;\beta)f(u;\Sigma)\,du. \quad (1)$$

The classical method that uses ML estimation and in which u is integrated out does not present problems with linear mixed models. The problem exists with GLMMs because of the more complicated integral (McCulloch and Searle, 2001). For this reason, one of the main interests of the research on the GLMM is to develop more efficient estimation methods for the fixed effects vector and the variance components.

Several ways to solve the integration in (1) and to obtain the marginal likelihood to estimate the parameters of a GLMM have been proposed. The Laplace method for integral approximation is considered to be a possible solution, which can be used to estimate the parameters of interests (Breslow and Clayton, 1993). Alternatives are the GHQ method or pseudo and penalized quasilikelihood methods (Aitkin, 1996). The GHQ method presents better estimation properties than the other methods because the GHQ estimates are maximum likelihood estimates. However, it is not feasible for analyses with more than two or three random effects because the speed of the GHQ decreases rapidly when increasing the number of random effects (Bolker et al., 2009).

3 Theoretical Description of the Extended Likelihood Principle

Lee and Nelder (1996, 2001) extended generalized linear models to include random effects by using their hierarchical (h-)likelihood method. This method is based on the extended likelihood principle (Bjørnstad, 1996) and is an implementation of the extended likelihood restricted by a weak canonical link for the random effects (Lee et al., 2006).

The h-likelihood is given by the log joint likelihood, that is, the extended likelihood L_E:

$$h = \log\left(L_E(y;\beta,v)\right) = \log\left(f(y;\beta|v)\right) + \log\left(f(v)\right)$$

where $\log(f(y;\beta|v)$ denotes the log of the density function with β as parameter and conditional on $v = v(u)$, where u is a vector of random effects and $v(\cdot)$ is an appropriate link function defining the h-likelihood. Unlike the GLMMs, the random effect is not restricted to be normal and can follow other distributions (e.g., gamma, beta, or inverse gamma).

A fundamental difference compared to classical marginal likelihood theory is that estimation and inference based on the h-likelihood includes random effects, whereas in classical likelihood theory the random effects are integrated out and a marginal likelihood is used. Hence, the use of the h-

likelihood avoids the integration required for a classical marginal likelihood.

To estimate parameters (β, v), the fixed and random effects are estimated from the score functions of the h-likelihood:

$$\frac{\partial h}{\partial \beta} = 0, \quad \frac{\partial h}{\partial v} = 0.$$

The variance components are estimated by maximizing the adjusted profile h-likelihood defined as

$$p_{\beta,u} = \left(h + \frac{1}{2}\log(2\pi H^{-1})\right)\Big|_{\beta=\hat{\beta},u=\hat{u}},$$

where H is a Hessian matrix of the h-likelihood.

The estimates can be obtained by using iterative weighted least squares (IRWLS) as implemented in the `hglm` package (Rönnegård et al., 2010). The variance components are then estimated iteratively by applying a gamma GLM to the estimated deviances and with an intercept term included in the linear predictor and appropriate weights (Lee et al., 2006).

4 Theoretical Description of the Bayesian Principle

The Bayesian methods differ from the likelihood and extended likelihood principles in their philosophy as well as in the specific procedures used.

In order to implement a Bayesian principle, prior distributions are required for all parameters in the model, since under the Bayesian paradigm, all parameters are treated as random variables rather than fixed unknowns.

The Bayesian principle is based on assigning prior distributions to the parameters of the model. Thus, following the model defined in Section 2, the following prior distributions must be specified: $f(\boldsymbol{\beta}|\cdot)$, $f(\boldsymbol{u}|\cdot)$, and $f(\boldsymbol{\Sigma}|\cdot)$. These prior distributions express the beliefs on the parameters and these beliefs are modified by the data to obtain the posterior distribution of the parameters, $f(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\Sigma}|\boldsymbol{Y})$, which is defined as to be proportional to the product of the prior distributions and the likelihood of the data. The posterior distribution is therefore used for inference purposes.

Here, a non-informative normal distribution is assumed as prior distribution for $\boldsymbol{\beta}$, that is, a normal distribution with a huge variance. Let $\boldsymbol{\gamma} = (\boldsymbol{u},\boldsymbol{\beta})'$ denote the G x 1 vector of Gaussian parameters. Concerning the random effects, we assume $\boldsymbol{u}$ to follow a multivariate normal distribution, $\boldsymbol{u}|\boldsymbol{\Gamma} \sim \mathcal{N}(0,\boldsymbol{\Gamma}^{-1})$, where the precision matrix $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\phi})$ depends on parameters $\boldsymbol{\phi}$. Let $\boldsymbol{\phi}$ also be the vector of the variance components for which the prior $\boldsymbol{\Pi}(\boldsymbol{\phi})$ is assigned. However, often it is not possible to obtain an explicit expression of the posterior distribution and algorithms such as Markov chain Monte Carlo (MCMC) methods are used to generate the

posterior distribution by simulation.

The Bayesian principle is attractive because it offers several advantages over the likelihood principle (e.g., it can increase the stability in small samples or in clustered binary data), but it has the difficulty of specifying prior distributions with variance components (Fong et al., 2010).

The use of MCMC methods for GLMMs is the most popular approach, but has problems in terms of convergence and computational time. These problems with Bayesian estimation have been greatly improved by Integrated Nested Laplace Approximations (Rue et al., 2009).

Integrated Nested Laplace Approximation (INLA)

INLA is a new tool for Bayesian inference based on latent Gaussian models introduced by Rue et al. (2009). The method combines Laplace approximations and numerical integration in a very efficient manner. For the GLMM described in Section 2 and using γ and ϕ as defined in the previous paragraphs, the posterior density is given by

$$\pi(\gamma,\phi|Y) \propto \pi(\gamma|\phi)\pi(\phi)\prod_{i=1}^{m} p(Y_i|\gamma,\phi).$$

It is computed via numerical integration as

$$\pi(\gamma|Y) = \int \pi(\gamma|\phi,Y)\pi(\phi|Y)d\phi,$$

where Laplace approximation is applied to carry out the integrations required for the evaluation of $\pi(\gamma|\phi,Y)$. For more details we refer the readers to Rue et al. (2009).

5 Contributed R Packages for GLMMs

For the likelihood principle, there exist different packages in R such as `glmmML` (Broström and Holmberg, 2013), `lme4` (Bates et al., 2015), or the function `glmmPQL` in the `MASS` package (Venables and Ripley, 2002). One of the most popular and stable functions for fitting GLMMs is called `glmer` and is found within the package `lme4`. This package implements the GHQ to approximate the log-likelihood using numerical integration. By default, it uses the Laplace approximation with one quadrature point.

For the extended likelihood principle, two packages are available for fitting Hierarchical Generalized Linear Models with random effects: `hglm` (Rönnegård et al., 2010) and `HGLMMM` (Molas and Lesaffre, 2011).

Concerning packages for performing Bayesian inference on GLMM, general packages such as `glmmBUGS` (Brown and Zhou, 2010) and `R2WinBugs` (Sturtz et al., 2005), and specialized packages such as `glmmAK` (Komárek and Lesaffre, 2008), `MCMCglmm` (Hadfield, 2010), and `INLA` (Lindgren and Rue, 2015) exist. INLA substitutes MCMC simulations with accuracy and

the quality of such approximations is extremely high.

For both the analysis of the wrestling data (Section 6) and the simulation (Section 7), the R packages `lme4`, `hglm`, and `INLA` were used. Note that the `lme4` package does not report standard errors for variance components. The reason is provided by the developer of the package in his book (Bates, 2010) stating that the sampling distribution of the variance is highly skewed, which makes the standard error nonsensical (Li et al., 2011). Regarding the GHQ method, we used 5 quadrature points since it was indicated that this method can give a poor approximation to the integrated likelihood when the number of quadrature points is low (Lesaffre and Spiessens, 2001). This method can be made arbitrarily accurate by increasing the number of quadrature points. We adopted a strategy of increasing the number of quadrature points until there was a negligible difference in the values of the estimators (Ormerod and Wand, 2012).

Recent changes of the packages used

The three packages studied in this paper have implemented some new features in their latest versions.

Regarding the `lme4` package, the authors of the package have been discussing new features and new versions of the package through the fo-

rum "R-sig-mixed models" (`https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`) since its version 0.99 (better known and used as lme4.0). Currently, the present version is 1.1-7, which offers some advantages with respect to the version 0.99, especially in terms of convergence and optimization. These developments can be found at the following link: `https://github.com/lme4/lme4`.

Concerning the `hglm` package, since version 2.0 it is possible to fit several random effects from different distributions (e.g., gamma or Gaussian), to fit a linear predictor for the dispersion of the random effects, to fit spatial conditional autoregressive (CAR) and spatial autoregressive (SAR) models for the random effects, and to perform a likelihood-ratio test for the dispersion parameter of the random effects (Alam et al., 2014). The method options have also been extended to include the EQL1 method which is a "HL(1,1) correction" (Lee and Lee, 2012; Noh and Lee, 2007) applied on the default EQL method. These developments can be found at the following link: `http://cran.r-project.org/web/packages/hglm/vignettes/hglm.pdf`.

Regarding the INLA package, this analytical framework includes normally distributed latent variables and thus allows for hierarchical data structure, but it is targeted towards complex applications involving temporal and spatial smoothing where MCMC estimation may be too difficult to apply

(requiring specialized MCMC samplers) or prohibitively slow. On the one hand, now there is an increase of articles using this package in several application fields such as fishering (Cosandey-Godin et al., 2014) or ecology (Quiroz et al., 2014). Few applications of the INLA methodology in injury epidemiology or sport science have been published (Cervone et al., 2014), and it would be of interest to study in more detail the medical impact in terms of understanding sport injuries with this methodology. On the other hand, one of the most recent papers of Martins et al. (2014) shows a new perspective on the selection of default priors.

6 Real Data Example: Folk Wrestling Data

Leonese Wrestling (LW) or Aluche is a traditional and popular sport of the province of León, in Northwestern Spain. It is registered with and recognized by three international associations: *Fédération Internationale des Luttes Associées* (FILA), *Asociacion Española de luchas tradicionales* (AELT), and International Belt Wrestling Association (IBWA), respectively. Like with all styles of wrestling, the risk of injury is always present.

The main variable of interest in epidemiological investigation of sports injuries is the incidence of injury, which is generally expressed as the propor-

tion of injuries per fight (Ayán et al., 2010; Hägglund et al., 2010). There are few studies in the international scientific literature about the incidence of injury in combat sports and its associated factors (Klügl et al., 2010; Hewett et al., 2005). Nonetheless, in published papers, it has been found that the incidence of injury in these sports is higher than in other sports activities (Hägglund et al., 2005; Junge et al., 2009).

Concerning factors associated with the incidence of injury, it is known that this incidence is higher during wrestling matches than in training. However, there is not much information on the frequency of injuries, their incidence, and their causes to carry out prevention and control programs in this sport. This lack of knowledge has motivated this analysis of the impact and risk factors of injuries.

Data on matches and injuries of the LW summer seasons were available for 213 wrestlers during the summer seasons from 2005 through 2010. The response variable of interest was the frequency of injuries which was assumed to follow a Poisson distribution. The study design was unbalanced with different numbers of repeated measures given that not all wrestlers participated in official competitions in each of the six years from 2005 to 2010. The possible risk factors for injuries considered were: i) Winner: This variable is defined as a function of the falls during a match. It is set to

'Yes' if the wrestler had more falls in his favor than against him; otherwise, the value of Winner is set to 'No'; and ii) Weight category: a categorical variable with levels Light, Medium (chosen as reference), Semi-heavy, and Heavy. It has been taken into account that these variables could change from one season to another.

6.1 The model under study

Let $Y_i, i = 1,\ldots,n$, be the vector of the number of injuries per season of wrestler i. The length of Y_i depends on the number of seasons the wrestler took part in official competitions. It is assumed that the distribution of Y_i follows a Poisson distribution: $Y_i \sim Po(\mu_i)$. Usually, the counts are considered in relation to some differential or offset (λ) in order to obtain rates. Here, the offset is the number of the wrestler's matches per season. The Poisson generalized linear mixed model used to analyze the data links the mean of Y_i with both covariates X_i of interest —Winner and Weight category— by means of the following equation

$$\log(\mu_i) = \log(\lambda_i) + X_i\beta + u_i, i = 1,\ldots,n, \tag{2}$$

where the vector β contains the fixed effects parameters and u_i stands for the random effect intercept for wrestler i. Random effects are assumed to be independent and normally distributed with mean 0 and variance σ^2.

Random effects for the slope parameters were not considered in order to keep the model simpler. In addition, posterior model fits including such random effects did not improve the model fit significantly at a 0.05 significance level.

The model's marginal variance (over subjects) can be expressed as

$$\mathrm{Var}(Y_i) = \Phi \cdot \mu_i, i = 1, \ldots, n,$$

where Φ is the dispersion parameter. The Poisson distribution assumes $\Phi = 1$ and if $\Phi > 1$, overdispersion is present. In that case, the data have larger variance than expected under the assumption of a Poisson distribution.

The dispersion parameter can be estimated based on the χ^2 approximation of the residual deviance or Pearson residuals. The dispersion parameter is estimated by dividing the χ^2 statistic by the residual degrees of freedom, $n-r$:

$$\hat{\Phi} = \frac{1}{n-r} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \quad (3)$$

If there is high overdispersion, the negative binomial distribution is an alternative to the Poisson distribution for count data because the negative binomial distribution allows for a variance greater than mean. Apparent overdispersion is normally due to missing covariates or interactions, outliers in the response variable, non-linear effects of covariates entered as linear terms in the systematic part of the model, or the choice of a wrong link function. These are mainly model misspecifications. Real overdispersion

exists when none of the previous causes can be identified. The reason for this might be that the variation in the data is, actually, larger than the mean. Or, there may be many zero observations, clustering of observations, or correlation between observations (Hardin and Hilbe, 2007; Zuur et al., 2009).

6.2 Results

We fitted Model (2) with packages `lme4`, `hglm`, and `INLA`, and for the sake of comparison, we also analysed the data as if they were not correlated. That is, we fitted a GLM using function `glm` of the R package `stats`. All analyses were carried out with R, version 3.1.1, and the estimates obtained, together with the 95% confidence and (in the case of the `INLA` package) credible intervals, are presented in Table 2. The confidence intervals of the `lme4` package were calculated using the function `confint.merMod`, which computes a likelihood profile and finds the appropriate cutoffs based on the likelihood ratio test (by choosing `method = "profile"`).

We calculated the dispersion term based on the Pearson residuals using Equation (3) for function `glm` and packages `lme4` and `hglm`. In addition, we checked the possible overdispersion using an individual-level random variable (translating to a lognormal-Poisson model, which is qualitatively

similar to a negative binomial (Elston et al., 2001)).

According to the results obtained, the estimates of the coefficients of the linear predictor are quite similar with only slight differences in the second decimal digit. The same is true for the confidence and credible intervals obtained with packages `lme4`, `hglm`, and `INLA`, whereas function `glm` provides smaller confidence intervals because it treats the data as if they were independent observations. Moreover, it can be seen that the estimate of the random effect variance and the dispersion term differ a little from each other.

Concerning both variables of interest, the positive signs of the parameter estimates corresponding to the weight category indicate a higher risk for injuries among all three weight categories as compared with the medium weight category. Nonetheless, all 95% confidence and credible intervals include 0, that is, the differences are not statistically significant at a 0.05 significance level. In the case of variable Winner, the model indicates a lower injury risk —$\hat{\boldsymbol{\beta}} < \mathbf{0}$ and the 95% confidence and credible intervals do not include 0— among wrestlers with more falls in their favor.

7 Simulation Study

In this section, we present the simulation study designed to assess the performance of the three statistical principles using different scenarios based on the wrestling data. The aim is to compare the three methods in different settings defined by overdispersion or sample size with respect to measures of model accuracy, precision, empirical bias, and empirical coverage of the estimators as well as computation time and possible problems of convergence. A total of 40 different simulation settings were used that can be classified into two main simulation scenarios.

7.1 Simulation scenario 1

For the first simulation scenario, we used the structure of the real data set introduced in Section 6. The values of the response variable —number of injuries— were generated as a function of the observed values of the independent variables and the number of matches of each wrestler in each of the years under study following the model expression in (2). The aim of this scenario was to closely represent the structure of this real data set.

We simulated the number of injuries using the parameter values given in Table 2 obtained with the `lme4` package ($\beta_1 = 0.24, \beta_2 = 0.1, \beta_3 = 0.4, \beta_4 =$

-0.46). Concerning the model intercept (β_0) and the variance of the random effects (σ_u^2), we used combinations of both parameters that lead to three different values of overdispersion ($\Phi \in \{1.5, 3, 10\}$) that we identify as low, moderate, and high overdispersion settings; see McCulloch and Searle (2001) for technical details. Furthermore, we added the value of $\Phi = 1$, i.e., no overdispersion, to also assess how the GLMM behaves in this situation. In addition, the values of β_0 were chosen such that two different marginal means of the injury numbers were obtained ($\mu = 1$ and $\mu = 10$). As can be seen in the R code of the first simulation study in the Supplemental Material, the values of β_0 ranged from -4.8 to -1.7, those of σ_u^2 from 0 to $2.1^2 = 4.41$.

In total, with four different values of dispersion and two of the marginal mean, the number of simulation settings for the Simulation scenario 1 was $4 \cdot 2 = 8$.

7.2 Simulation scenario 2

The second simulation scenario was motivated by the goal to study the effect of different sample sizes on the parameter estimation. For this purpose, we considered two different sample sizes of $n = 30$ and $n = 100$ wrestlers and for each wrestler, a random number of seasons was generated using a

discrete uniform distribution ranging from 1 through 6. In the sequel, the match numbers for each wrestler and season were generated using a Poisson distribution with parameters 60 and 100, respectively. These were the offset terms λ_i in Model (2). The remaining parameters were chosen similar to the first simulation scenario resulting in four dispersion parameters ($\Phi \in \{1, 1.5, 3, 10\}$) and two marginal means for the number of injuries ($\mu = 1$ and $\mu = 10$) so that the number of simulation settings was $2 \cdot 2 \cdot 4 \cdot 2 = 32$ for Simulation scenario 2.

The values of both independent variables —weight category and winner— were generated using equal probabilities for all categories and, as in Simulation scenario 1, the injury numbers were generated using the expression of the model in (2).

7.3 Evaluation criteria

For each of the $8 + 32 = 40$ simulation settings, we simulated 1000 data sets of the model under study and used the three methods to estimate the model parameters (Marginal Likelihood, Extended Likelihood, and Bayesian Analysis). In addition, we used R function `glm` treating the data as if we dealt with a GLM. Measures for the comparison of the different estimation methods were the empirical mean squared error (MSE) as a measure of model

accuracy, the ratio of precision, the empirical bias, and the empirical coverage of the confidence and credible intervals, respectively. Moreover, we recorded the computation times and studied possible problems of convergence.

For each simulation setting and estimation method, the empirical bias was calculated as the mean bias over the 1000 data sets and its squared value was used together with the empirical variance to compute the empirical MSE. The rate of precision was computed as the ratio between the estimator's empirical variance and the mean of the squared standard errors. In order to calculate and compare the empirical bias and the empirical MSE in the case of the `INLA` package, the distribution of the parameters provided by `INLA` were reduced to only one value (the posterior mean).

For the likelihood and the extended likelihood principles, we used the 95% confidence interval and for the Bayesian principle, we used the 95% credible interval of parameter given by the 0.025 and 0.975 sample quantiles of the posterior parameter distributions. In the case of the `lme4` package, we used 5 quadrature points for the GHQ method and non-informative priors were assumed for the Bayesian analysis. Moreover, the random intercepts were assumed to have a normal distribution. Regarding the prior distribution for the precision, a half-normal distribution with mean 0 and precision

0.0001 was assigned to the standard deviations (Gelman et al., 2006).

The comparison was done for the two main parameters of interest: the parameter β_4, which corresponds to the variable Winner, and the variance of the random effects (σ_u^2). The former was chosen, since the analysis of the wrestling data showed a statistically significant association between the number of injuries and this variable. Whereas the value of σ_u^2 varied across the simulation settings, the value of β_4 was kept constant in all settings.

8 Results of the Simulation Study

For the simulation scenarios, results are presented only for the intercept (β_0), the slope (parameter β_4, corresponding to the covariate Winner), and variance of the random effect (σ_u^2). The performance of the estimation methods in terms of empirical bias, empirical MSE, precision ratio, and empirical coverage of $\hat{\beta}_4$ and $\hat{\sigma}_u^2$ is summarized in Figures 1a, 1b, 2a, and 2b. The corresponding figures for $\hat{\beta}_0$ are provided in the Supplemental Material.

In order not to mix up statistical principles, methods, and algorithms as presented in Table 1, following, we only refer to these by the names of the R packages used. Note that the results are yielded by package `INLA` with the prior distribution selected, a half-normal distribution with mean 0 and

precision 0.0001.

Concerning the percentage of convergence of the estimation methods, a model was considered as "not convergent" if either the estimation process did not converge or if the estimate or its standard error was not provided. For example, in some cases the parameter can be estimated but the estimation process may be unable to provide a positive definite variance-covariance matrix of the parameters (for problems with the Hessian matrix), mainly due to the instability of the model. Convergence was checked and obtained using the criteria offered in each software package. In the case of the first simulation scenario, the convergence percentages were always equal to 100% with only one exception: package `hglm` achieved convergence in 99.2% of all data sets in the case of $\mu = 1$ and $\Phi = 10$. The results for Simulation scenario 2 are shown in Table 3. The rate of convergence of all estimation methods was close to 100% for most of the settings. However, in the case of packages `hglm` and `lme4`, the percentage of convergence slightly decreased in some settings with $\mu = 1$ and $n = 30$ as overdispersion increased.

Regarding the empirical bias of the slope, all packages provided mostly unbiased and similar estimates. In terms of accuracy, the highest empirical MSE for the slope for all GLMM packages is given when $\Phi = 10$, $\mu = 1$, and $n = 30$. In this case, function `glm` is the one that presents highest values.

The empirical MSE value obtained with `INLA` is higher than with `hglm` and `lme4`, which are both similar; see Figure 1a.

In terms of precision (upper panel of Figure 1b), we calculated the ratio of the estimator's empirical variance and the mean of the squared standard errors as a precision measure. In general, we found that almost all estimation methods presented an underestimation in the case of $\Phi = 1$ and $\Phi = 1.5$ together with $\mu = 10$. More differences between the packages were observed with sample size equal to 30. In the case of the `lme4` package, the ratio was slightly larger than 1 (equivalent to 100%) especially with moderate and high overdispersion. By contrast, the values of the `hglm` and `INLA` packages were close to 100% for that sample size independently of the offset, the marginal mean, and the dispersion term. The function `glm`, in general, showed values far larger than 100% (and, hence, out of the range of the corresponding plots) from $\Phi = 1.5$.

The empirical coverage of the confidence intervals for the GLMM packages were close to 95% in all settings; see the lower panel of Figure 1b. The GLM appeared to have bad coverage, only acceptable for those combinations with low overdispersion. It suffered from substantial undercoverage (down to 75%) when $\Phi = 3$ and $\Phi = 10$. This result may be expected for the GLM since it does not include random effects and therefore can not

assume any overdispersion. As overdispersion increased, the empirical coverage behavior became worse.

Regarding the empirical bias of the variance component (upper panel of Figure 2a), the three packages performed similarly for $\Phi = 1$ and $\Phi = 1.5$. For $\Phi = 3$, the `INLA` package showed the largest empirical bias with $n = 30$, whereas with $n = 100$, the differences among the packages were small for $\Phi = 3$. For $\Phi = 10$ and $n = 30$, package `lme4` had the smallest empirical bias in terms of the absolute value, `INLA` the largest (with values out of the range of the plot). By contrast, for $\Phi = 10$ and $n = 100$ the absolute values of the empirical bias of `lme4` and `INLA` were roughly the same. It was largest for `hglm` in that setting. Values for function `glm` do not appear in that figure since a GLM does not consider any random effects.

In terms of the accuracy of the variance component (lower panel of Figure 2a), the empirical MSEs were very similar except when $\Phi > 1.5$ and $\mu = 1$. The `INLA` package had the largest empirical MSE when $\Phi = 3$, $\mu = 1$, and $n = 30$. By contrast, in this same setting, the `lme4` and `hglm` packages had very similar values. None of the packages showed a satisfactory behavior when $\Phi = 10$, $\mu = 1$, and $n = 30$: all empirical MSEs are excessively high and, hence, out of the range of the corresponding plot. With $n = 100$, the empirical MSE of package `hglm` was still out of the plot's range, whereas

that of `lme4` and `INLA` were close to 1.

Concerning the precision of the estimation of the variance component (Figure 2b), the ratio obtained with `hglm` and `INLA` was close to zero when $\Phi \neq 10$ and $\mu = 10$. For $n = 30$, in `INLA`, the values were close to 150 when $\mu = 10$, and they were close to 100 when $\Phi > 1.5$ and $\mu = 1$. For `hglm` and when $\Phi > 1$ and $\mu = 1$, the values were excessively high in most of the simulation settings. It could not be computed with `lme4` since this package does not provide the standard error of the estimation of the random effect's variance.

Contrary to overdispersion, sample size, and marginal mean, the choice of the offset did not seem to have any effect on the estimators' performance.

Finally, we also compared the computational times measured with R function `system.time`. Among the three packages that consider random effects, the average computing times of packages `lme4` and `hglm` were very similar in each setting. On average, they were four times faster than package `INLA`.

In summary, the approaches involving a random effect (`lme4`, `INLA`, and `hglm`) showed good performance on estimating the model parameters except for the estimation of the random effect's variance in the case of combinations of huge overdispersion, a small marginal mean, and a small sample size.

Given that the empirical MSE and the empirical bias of the `lme4` package are close to zero for most of the simulation settings, it seems that this package, generally, outperforms the other packages, even though often only slightly.

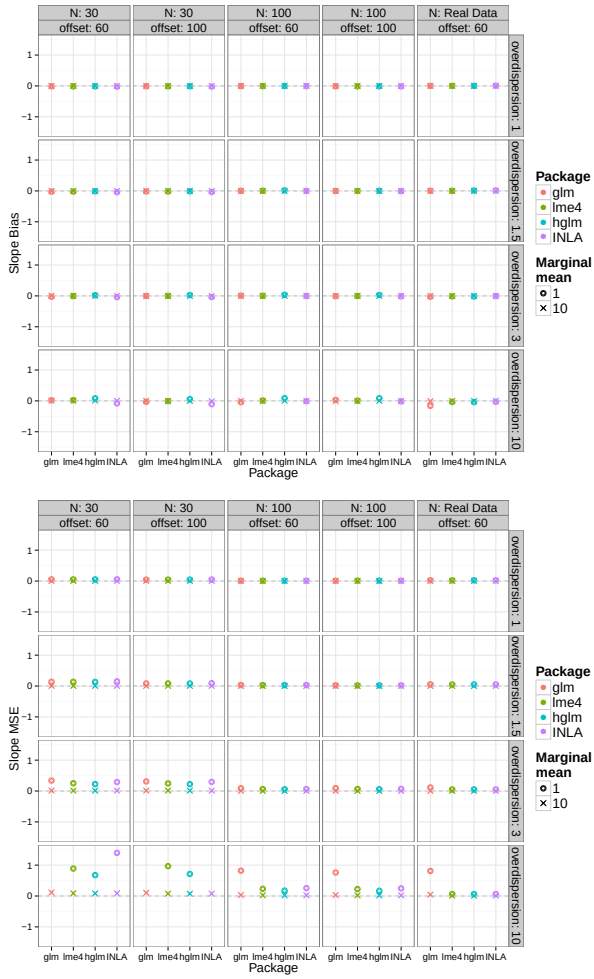


Figure 1a: Empirical bias (upper panel) and empirical MSE of the slope estimate ($\hat{\beta}_4$) as a function of overdispersion, marginal mean, offset, and sample size.

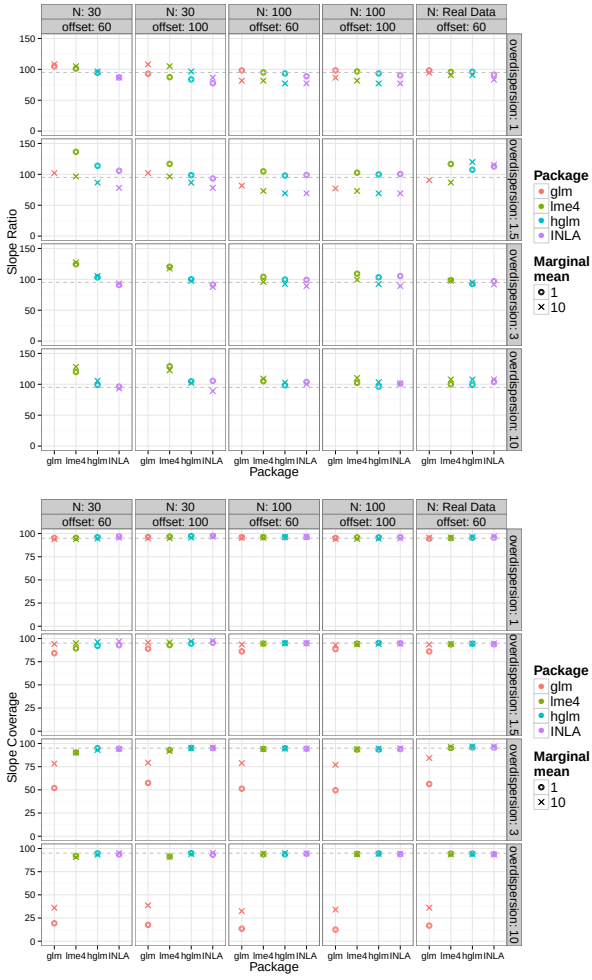


Figure 1b: Precision (upper panel) and empirical coverage of the slope estimate ($\hat{\beta}_4$) as a function of overdispersion, marginal mean, offset, and sample size. Precision is measured as the ratio of the estimator's empirical variance divided by the average of the squared standard errors.

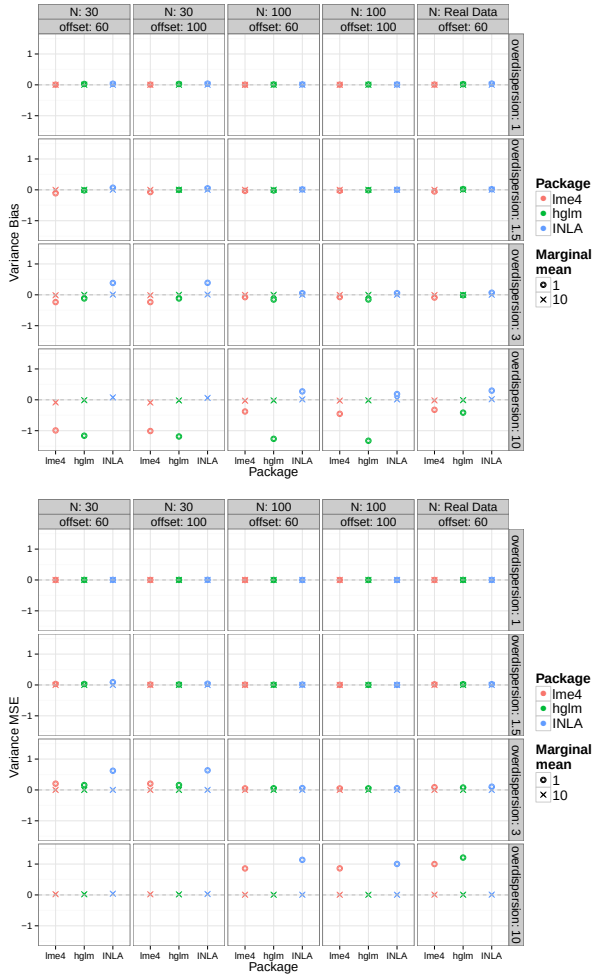


Figure 2a: Empirical bias (upper panel) and empirical MSE of the variance component estimate ($\hat{\sigma}_u^2$) as a function of overdispersion, marginal mean, offset, and sample size.

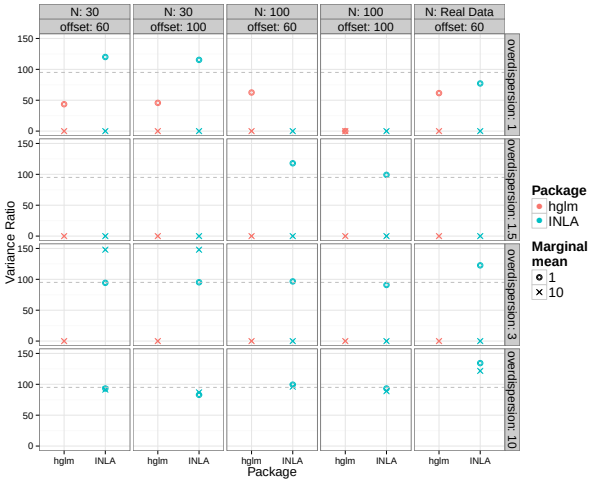


Figure 2b: Precision of the variance component estimate ($(\hat{\sigma}_u^2)$) as a function of overdispersion, marginal mean, offset, and sample size. Precision is measured as the ratio of the estimator's empirical variance divided by the average of the squared standard errors.

9 Discussion

An overview of statistical principles for GLMMs is presented in this paper. The problem of selecting the best approach for estimation and inference within Poisson Mixed Models is very complex and too difficult to solve analytically. For this reason, we have carried out a simulation study that has evaluated the impact of overdispersion, marginal mean, offset, and sample size assuming Poisson Mixed Models using different statistical principles.

The fact that the bias and mean square error improve with larger sample size can be interpreted as that the estimators are consistent when overdispersion is taken into account in the model by means of random effects. By contrast, the empirical bias and MSE were larger with the GLM since this type of model does not include random effects and, hence, ignores overdispersion (Bolker et al., 2009; Milanzi et al., 2012). The results for `INLA`, in general, did hardly differ with the values obtained with both `lme4` and `hglm`; at least not with the prior distribution used.

We have found that for small sample sizes, the random effects variances are difficult to estimate, which has also been described in other studies (Li et al., 2011). In addition, the results are worse in the case of a moderate dispersion term ($\Phi = 3$) and especially with high overdispersion ($\Phi = 10$), in

which none of the methods provide satisfactory results. As Zuur et al. (2009) explain, with a dispersion term up to 1.5 there are no important overdispersion problems. To solve overdispersion problems, other methods and distributions may be used, for instance, the Poisson-lognormal distribution, GEE models, or quasi-Poisson distributions (Booth et al., 2003; Bolker et al., 2009). In the case of high overdispersion with real data, it would be more reasonable to change the Poisson distribution for another distribution that does not have the restriction of the variance equalling the mean, for instance, the Negative Binomial distribution (Czado et al., 2007). Recently, Aregay et al. (2013) suggested a Hierarchical Poisson-Normal overdispersed model (HPNOD) as an alternative using the Bayesian principle. The HPNOD performs better than a Hierarchical Poisson-Normal model for data with low, moderate, and high overdispersion.

In the simulation study, for most combinations, we observed similar performance in terms of the empirical bias and the empirical MSE whatever was the package applied. On the other hand, differences arise with respect to the precision of estimates. This fact may indicate a problem of underestimation because the methods do not capture well all the variability present in data. A bootstrap approach may be a solution to solve the standard error problem.

Regarding computational time and convergence, the `glm` function requires less time because it does not capture the presence of the random effect. The `hglm` and `lme4` packages need similar computational times, whereas the `INLA` package takes more time. Nonetheless, at least for the simulation settings under study, computation time was not excessive and results were obtained in less than 5 seconds in most of the cases. In some combinations of the simulation with small sample size ($n = 30$), huge overdispersion ($\Phi = 10$), and small marginal mean ($\mu = 1$), we found problems of convergence in the `hglm` package. To solve a convergence problem we recommend specifying other starting values.

Several studies carried out until now have compared estimation methods only for the Bayesian principle, the marginal likelihood principle or both in the GLMMs (Zhang et al., 2011; Ormerod and Wand, 2012; Li et al., 2011; Kim et al., 2013). For example, Li et al. (2011) recommend the use of the `lme4` package. The authors highlight that in case a Bayesian package is chosen, the parameter estimates might be influenced by the priors for the variances of the random effects. They also mention that when the data set is small, the random effects' variances are difficult to estimate with both frequentist and Bayesian methods. Kim et al. (2013) recommend the use of the GHQ method given that it performs well in terms of accuracy, precision,

convergence rates, and computing speed. According to the authors, this is also valid with small sample sizes and for longitudinal studies with a few time points. On the other hand, there are some studies that compare the extended likelihood approach with the Bayesian principle (Pan and Thompson, 2007; Collins, 2008). However, they do not take into account the `INLA` method, which is a recently proposed approximate Bayesian approach for latent Gaussian models. According to Collins (2008), in some case studies both Bayesian and extended likelihood approach estimators of the variance component were positively biased, whereas GHQ method based estimators were not.

Our work is different from these previous studies given that we have focused our work on different estimation methods, principles, and more commonly used free software packages. Regarding the real data example, most GLMMs are used in applications of ecology, epidemiology, genetics, clinical medicine, and other applications but these models are now gaining attention in sports sciences, too (Avalos et al., 2003; Bullock and Hopkins, 2009; Sampaio et al., 2010; Casals and Martínez, 2013; Casals et al., 2014). However, there are only a few studies in epidemiology of sports injuries.

Concerning the Bayesian principle, when the sample size is small, the posterior distribution may be more influenced by the choice of prior dis-

tributions than when the sample size is moderate or large. Note that the posterior means depend on the choice of the non-informative prior of the variance component (Li et al., 2011). Bayesian algorithms such as MCMC offer different advantages over frequentist algorithms but they have problems in terms of convergence and computational time. These aspects have improved with `INLA`. However, for some combinations, the default initial value is not adequate for the data because of the use of very weak priors. For example, it is known that the inverse Gamma$(0.001, 0.001)$ prior is not always a good choice (Fong et al., 2010). Frühwirth-Schnatter and Wagner (2010b) and Frühwirth-Schnatter and Wagner (2010a) demonstrate overfitting due to Gamma-priors and suggest using a (half) Gaussian prior for the standard deviation to overcome this problem, as suggested by Gelman et al. (2006). Another interesting possibility that could be considered in the future is the use of Penalized Complexity (PC) priors that have heavier tails than the half Gaussian, but lighter tails than the half-Cauchy (Martins et al., 2014). According to the developers of the `INLA` package, the use of PC priors works pretty much identically to the half-Cauchy in practice.

There are some limitations of the present work. First, in this study we have worked with a Poisson mixed model with a random intercept, but we have not considered models with random slopes. The reasons for this de-

cision were twofold: On one hand, the inclusion of random slopes in the real data example did not significantly improve the model fit at a 0.05 significance level. On the other hand, to study the impact of overdispersion, marginal mean, sample size, and offset, we preferred to analyze a Poisson mixed model with a random intercept due to its frequency in sports medicine research. Given the importance of more complex mixed models, future research should investigate the performances of these principles in mixed Poisson models including random slopes, cross-classified random effects and multiple membership structures that may be analyzed in future simulation studies. Second, we only examined three packages corresponding to the three estimation methods and principles. There are other R packages such as `glmmML` and `MCMCglmm` as well as the function `glmmPQL` of the `MASS` package that could be included in further simulation studies. In addition, such simulation studies could include other software packages such as SAS, STATA, or SPSS, which were not considered for the present work. We decided to focus our work on R because of its great popularity among statisticians (Muenchen, 2015) and the constant development of new packages and functions to deal with GLMMs.

In addition to the two limitations mentioned, the simulation study could have also studied other parameters of interest such as the cluster size, which,

in the real data example in Section 6, is equivalent to the number of seasons of a wrestler. Indeed, several simulation studies point out that for binary responses, the performance of the estimators is influenced by the cluster size, e.g., that clusters of size two usually entail problems (Breslow and Clayton, 1993; Diaz, 2007; Kim et al., 2013; Grilli et al., 2014). Following the suggestion of one of the reviewers, we decided to assess the role of the cluster size on empirical bias and MSE with a small additional simulation study with eight different settings defined by two values of cluster size ($m \in \{2, 5\}$), two values of overdispersion ($\Phi \in \{1.5, 10\}$), and two different marginal means ($\mu \in \{1, 10\}$). The offset ($\lambda = 60$), the sample size ($n = 100$), and the slope parameter ($\beta_4 = -0.46$) were kept constant. Regarding the results of the empirical bias and the MSE, which are shown in Tables 4 and 5 in the Supplemental Material, there were hardly any differences between cluster sizes two and five. However, the estimates obtained by the three packages had larger empirical bias and MSE when $\Phi = 10$ and $\mu = 1$. Under this scenario, the values of package `hglm` were larger than those of the `lme4` and `INLA` packages. Moreover, with respect to computing time under the settings of this additional simulation study, the `hglm` package was somewhat faster than `lme4`. The computing times of the `INLA` package were, again, much larger. Problems of convergence were not detected.

It is important to highlight that the concepts of estimation and standard error are different in the three statistical principles used. In the case of the marginal likelihood, estimation is the value that maximizes the likelihood and the standard error reflects the sample variation of the estimator. In Bayesian analysis, estimation is a summary of the posterior distribution and the standard error is a measure of the variability of this distribution. As for the classical likelihood, the extended likelihood advocates the use of standard errors computed from the Fisher information matrix of the marginal likelihood. In practice, the standard errors are computed from the matrix of second derivatives for the adjusted profile h-likelihood, which is an approximation of the marginal likelihood (Lee et al., 2006).

Parameter estimation of GLMM is nowadays possible using different statistical principles, though estimation methods as well as statistical packages are still under development (Bolker et al., 2009). The problem of selecting the most adequate approach for the estimation and inference within GLMM is very complex. In addition, the software implementations can differ considerably in flexibility, computation time and usability (Austin, 2010; Li et al., 2011). A strategy could be to carry out a simulation study that emulated the data design and to apply the different estimation methods. Although one may think that this strategy is not very practical, it would be indeed

worse to use an estimation method that could provide biased and inefficient estimates.

We have shown through simulations that ignoring overdispersion in Poisson Mixed Models can have serious consequences on the parameter estimates. Available R packages can handle this problem very satisfactorily; however, care must be taken in situations with small sample size, large overdispersion, and small marginal mean. In such situations, the `lme4` package seems to have a slightly better performance than packages `hglm` and the `INLA`, which also depends on the choice of the prior (Grilli et al., 2014), especially concerning the estimation of the random effect's variance (Figure 2a). This observation coincides with the recommendation of Kim et al. (2013) to use the GHQ method under such settings and the active discussions of the package's authors (Douglas Bates, Martin Maechler, and Ben Bolker) and members of the R sig-mixed-models mailing list (`https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`). All packages under study have recently been improved in terms of convergence and optimization.

Acknowledgements

We would like to thank Vicente Martin, Moudud Alam, Lesly Acosta, and members of the R sig-mixed-models mailing list and the INLA group discussion forum for useful comments. We are also grateful for the thorough revision of the manuscript which has helped to improve it. The work was partially supported by the grants MTM2012-38067-C02-01 of the Spanish Ministry of Economy and Competitiveness and 2014 SGR 464 from the *Departament d'Economia i Coneixement* of the *Generalitat de Catalunya* and by the *Diputación de León y la Federación Territorial de Castilla y Leon de Lucha*.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing 6*(3), 251–262.

Alam, M., L. Ronnegard, and X. Shen (2014). Fitting conditional and simultaneous autoregressive spatial models in hglm. Submitted.

Aregay, M., Z. Shkedy, and G. Molenberghs (2013). A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion:

A simulation study. *Computational Statistics & Data Analysis 57*(1), 233–245.

Austin, P. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics 6*(1), Article 16. doi: 10.2202/1557–4679.1195.

Avalos, M., P. Hellard, and J. C. Chatard (2003). Modeling the training-performance relationship using a mixed model in elite swimmers. *Medicine and Science in Sports and Exercise 35*(5), 838–846.

Ayán, C., A. J. Molina, H. García, G. González, M. J. Álvarez, T. Fernández, and V. Martín (2010). Rules modification effect's in incidence of injuries in Lucha Leonesa (Leonesa wrestling). *Apunts. Medicina de l'Esport 45*(165), 17–22.

Bates, D. (2010, February). lme4: Mixed-effects modeling with R. URL: `http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*(298), 269–306.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association 91*(434), 791–806.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. White (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution 24*(3), 127–135.

Booth, J. G., G. Casella, H. Friedl, and J. P. Hobert (2003). Negative binomial loglinear mixed models. *Statistical Modelling 3*(3), 179–191.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Applied Statistics 33*(1), 38–44.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Broström, G. and H. Holmberg (2013). *glmmML: Generalized linear models with clustering*. R package version 1.0.

Brown, P. E. and L. Zhou (2010). Mcmc for generalized linear mixed models with glmmbugs. *The R Journal 2*, 13–17.

Bullock, N. and W. G. Hopkins (2009). Methods for tracking athletes' competitive performance in skeleton. *Journal of Sports Sciences 27*(9), 937–940.

Casals, M., M. Girabent-Farrés, and J. L. Carrasco (2014). Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): A systematic review. *PloS one 9*(11), e112653.

Casals, M. and J. A. Martínez (2013). Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sports 13*, 64–82.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry (2014). A multiresolution stochastic process model for predicting basketball possession outcomes. Submitted. arXiv:1408.0777v1.

Collins, D. (2008). *The performance of estimation methods for generalized linear mixed models*. Ph. D. thesis, School of Mathematics & Applied Statistics - Faculty of Informatics, University of Wollongong.

Cosandey-Godin, A., E. Teixeira Krainski, B. Worm, and J. Mills Flem-

ming (2014). Applying Bayesian spatio-temporal models to fisheries bycatch in the canadian arctic. *Canadian Journal of Fisheries and Aquatic Sciences 72*(999), 1–12.

Czado, C., V. Erhardt, A. Min, and S. Wagner (2007). Zero-inflated generalized poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling 7*(2), 125–153.

Dean, C. B. and J. D. Nielsen (2007). Generalized linear mixed models: A review and some extensions. *Lifetime Data Analysis 13*(4), 497–512.

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics & Data Analysis 51*(6), 2871–2888.

Elston, D., R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin (2001). Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology 122*(5), 563–569.

Fong, Y., H. Rue, and J. Wakefield (2010). Bayesian inference for generalized linear mixed models. *Biostatistics 11*(3), 397–412.

Frühwirth-Schnatter, S. and H. Wagner (2010a). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 9*, pp. 1–21. Oxford University Press.

Frühwirth-Schnatter, S. and H. Wagner (2010b). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics 154*(1), 85–100.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis 1*(3), 515–534.

Grilli, L., S. Metelli, and C. Rampichini (2014). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation 85*(13), 2718–2726.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software 33*(2), 1–22.

Hardin, J. W. and J. Hilbe (2007). *Generalized Linear Models and Extensions*. Stata Press.

Hewett, T., C. Pasque, R. Heyl, and R. Wroble (2005). Wrestling injuries. *Medicine and Sport Science 48*, 152–178.

Hägglund, M., M. Waldén, R. Bahr, and J. Ekstrand (2005). Methods for epidemiological study of injuries to professional football players: developing the UEFA model. *British Journal of Sports Medicine 39*(6), 340–346.

Hägglund, M., M. Waldén, L. Til, and R. Pruna (2010). The importance of epidemiological research in sports medicine. *Apunts. Medicina de l'Esport 45*(166), 57–59.

Hirsch, K. and A. Wienke (2012). Software for semiparametric shared gamma and log-normal frailty models: An overview. *Computer methods and programs in biomedicine 107*(3), 582–597.

Junge, A., L. Engebretsen, M. L. Mountjoy, J. M. Alonso, P. A. F. H. Renström, M. J. Aubry, and J. Dvorak (2009). Sports injuries during the Summer Olympic Games 2008. *The American Journal of Sports Medicine 37*(11), 2165–2172.

Kim, Y., Y.-K. Choi, and S. Emery (2013). Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. *The American Statistician 67*(3), 171–182.

Klügl, M., I. Shrier, K. McBain, R. Shultz, W. H. Meeuwisse, D. Garza, and G. O. Matheson (2010). The prevention of sport injury: an analysis of 12 000 published manuscripts. *Clinical Journal of Sport Medicine 20*(6), 407.

Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random-effects distribution. *Computational Statistics and Data Analysis 52*(7), 3441–3458.

Lee, W. and Y. Lee (2012). Modifications of REML algorithm for HGLMs. *Statistics and Computing 22*(4), 959–966.

Lee, Y., J. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton: Chapman & Hall/CRC.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B (Methodological) 58*(4), 619–678.

Lee, Y. and J. A. Nelder (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika 88*(4), 987–1006.

Lesaffre, E. and B. Spiessens (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 325–335.

Li, B., H. Lingsma, E. Steyerberg, and E. Lesaffre (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology 11*(1), 77–87.

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software 63*(19), 1–25.

Martins, T. G., D. P. Simpson, A. Riebler, H. Rue, and S. H. Sørbye (2014). Penalising model component complexity: A principled, practical approach to constructing priors. Submitted. arXiv:1403.4630v3.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons.

Milanzi, E., A. Alonso, and G. Molenberghs (2012). Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions. *Statistics in Medicine 31*(14), 1475–1482.

Molas, M. and E. Lesaffre (2011). Hierarchical generalized linear models: The R package HGLMMM. *Journal of Statistical Software 39*(13), 1–20.

Muenchen, R. A. (2015). The popularity of data analysis software. URL: `http://r4stats.com/articles/popularity/`.

Noh, M. and Y. Lee (2007). Reml estimation for binary data in GLMMs. *Journal of Multivariate Analysis 98*(5), 896–915.

Ormerod, J. and M. Wand (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics 21*(1), 2–17.

Pan, J. and R. Thompson (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis 51*(12), 5765–5775.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. OUP Oxford.

Quiroz, Z. C., M. O. Prates, and H. Rue (2014). A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics 71*(1), 208–217.

Rönnegård, L., X. Shen, and M. Alam (2010). hglm: A package for fitting hierarchical generalized linear models. *The R Journal 2*(2), 20–28.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Methodological) 71*(2), 319–392.

Sampaio, J., E. J. Drinkwater, and N. M. Leite (2010). Effects of season period, team quality, and playing time on basketball players' game-related statistics. *European Journal of Sport Science 10*(2), 141–149.

Sturtz, S., U. Ligges, and A. Gelman (2005). R2winbugs: A package for running WinBUGS from R. *Journal of Statistical Software 12*(3), 1–16.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

West, B. T., K. B. Welch, and A. T. Galecki (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software* (2nd ed.). CRC Press.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics 31*(2), 144–148.

Zhang, H., N. Lu, C. Feng, S. W. Thurston, Y. Xia, L. Zhu, and X. M. Tu

(2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine 30*(20), 2562–2572.

Zuur, A. F., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

A Tables

Table 1: Overview of statistical principles

Principle	**Method**	**Algorithms**
Marginal Likelihood	Maximum likelihood	Newton-Raphson (N-R), Fisher scoring,
		Penalized iteratively reweighted
		least squares (PIRLS)
		Adaptative Gauss Hermite Quadrature (GHQ)
Extended likelihood	h-likelihood	N-R, Iterative weighted least squares (IRWLS)
Bayesian	Posterior mean	MCMC,
		Integrated Nested Laplace
		Approximations (INLA)

Table 2: Results from the Poisson mixed model applied to the folk wrestling data. CI stands for confidence interval and credible interval (in the case of INLA), respectively.

	Function `glm`		Package `lme4`		Package `hglm`		Package `INLA`	
	$\hat{\beta}$	95% CI	$\hat{\beta}$	95% CI	$\hat{\beta}$	95% CI	$\hat{\beta}$	95% CI
Intercept	−4.34	[-4.73, -4.0]	−4.37	[-4.82, -3.99]	−4.37	[-4.79, -3.95]	−4.41	[-4.85, -4.01]
Weight category[1]								
Light	0.25	[-0.2, 0.71]	0.24	[-0.26, 0.76]	0.25	[-0.28, 0.78]	0.25	[-0.26, 0.76]
Semiheavy	0.1	[-0.36, 0.57]	0.1	[-0.41, 0.63]	0.11	[-0.43, 0.65]	0.12	[-0.4, 0.64]
Heavy	0.39	[-0.1, 0.87]	0.4	[-0.14, 0.96]	0.4	[-0.16, 0.97]	0.41	[-0.14, 0.97]
Winner	−0.48	[-0.82, -0.15]	−0.46	[-0.82, -0.07]	−0.46	[-0.85, -0.07]	−0.44	[-0.82, -0.06]
σ_u^2		–	0.08	[0.0, 0.39]	0.09		0.12	[0.01, 0.31]
Dispersion (Φ)[2]	1.45		1.29		1.35		–	

[1] The reference category is Medium

[2] Obtained by means of equation (3)

Table 3: Percentages of convergence in Simulation scenario 2 as a function of the marginal mean (μ), the average match number per season (Offset), overdispersion (OD), and the sample size

			glm		**lme4**		**hglm**		**INLA**	
	Offset	**OD**	n=30	n=100	n=30	n=100	n=30	n=100	n=30	n=100
μ=1	60	Φ=1	100	100	99.7	100	99.7	100	100	100
		Φ=1.5	100	100	99.4	100	99.7	100	100	100
		Φ=3	100	100	99.4	100	99.6	100	100	100
		Φ=10	100	100	98.1	99.9	98	99.8	100	100
	100	Φ=1	100	100	100	100	100	100	100	100
		Φ=1.5	100	100	99.8	99.9	99.9	100	100	100
		Φ=3	100	100	99.4	100	99.7	100	100	100
		Φ=10	100	100	97.2	100	97.9	99.9	100	100
μ=10	60	Φ=1	100	100	100	100	100	100	100	100
		Φ=1.5	100	100	100	100	100	100	100	100
		Φ=3	100	100	100	100	100	100	100	100
		Φ=10	100	100	100	100	100	100	100	100
	100	Φ=1	100	100	100	100	100	100	100	100
		Φ=1.5	100	100	100	100	100	100	100	100
		Φ=3	100	100	100	100	100	100	100	100
		Φ=10	100	100	100	100	100	100	100	100

A Supplemental Material: Plots

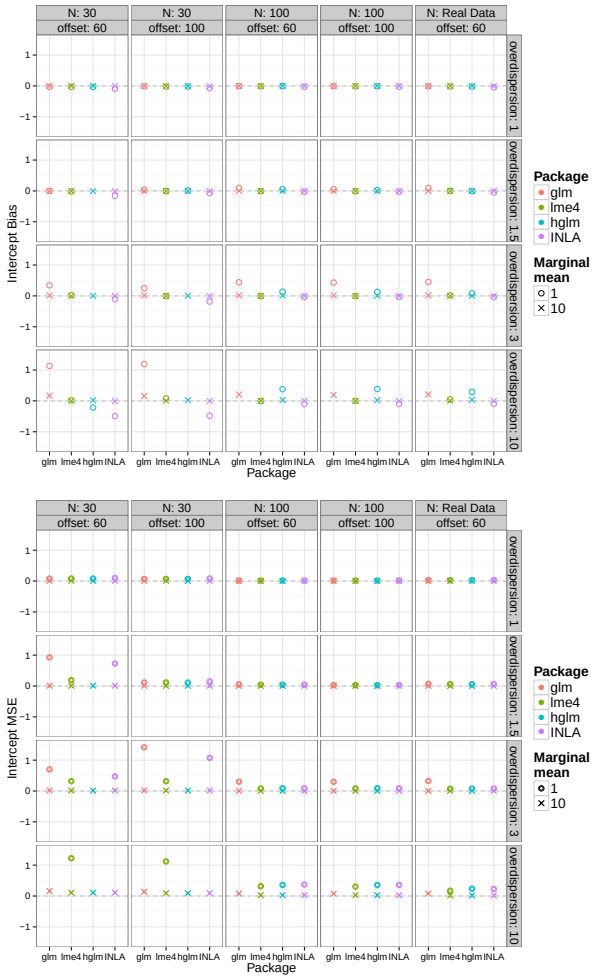


Figure 3: Empirical bias (upper panel) and empirical MSE of the intercept estimate ($\hat{\beta}_0$) as a function of Φ, μ, offset, and N

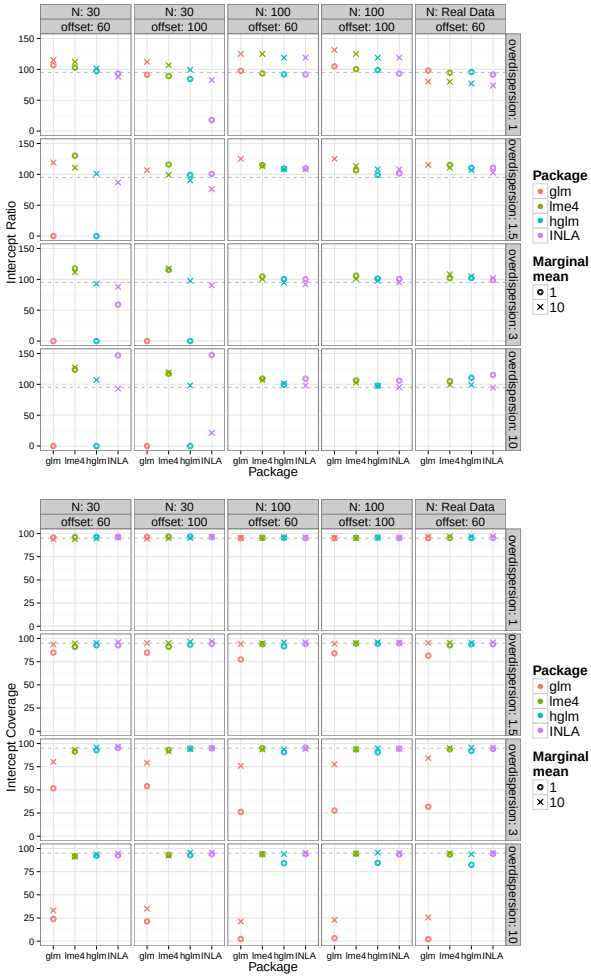


Figure 4: Precision (upper panel) and empirical coverage of the intercept estimate ($\hat{\beta}_0$) as a function of Φ, μ, offset, and N. Precision is measured as the ratio of the estimator's empirical variance divided by the average of the squared standard errors.

B Supplemental Material: Tables

Table 4: Empirical Bias of the slope ($\hat{\beta}_4$) and variance component ($\hat{\sigma}^2_u$) as a function of the marginal mean (μ), overdispersion (OD), and the cluster size (m). Sample size is $N = 100$, the offset is equal to 60.

			lme4		**hglm**		**INLA**	
		OD	m=2	m=5	m=2	m=5	m=2	m=5
μ=1	$\hat{\beta}_4$	Φ=1.5	0	−0.012	0.01	0.001	−0.008	−0.017
		Φ=10	−0.01	−0.006	0.083	0.064	−0.042	−0.025
	$\hat{\sigma}^2_u$	Φ=1.5	−0.045	−0.029	−0.027	−0.023	0.009	0.009
		Φ=10	−0.492	−0.345	1.522	−1.13	0.221	0.217
μ=10	$\hat{\beta}_4$	Φ=1.5	−0.002	−0.001	−0.002	−0.001	−0.002	−0.001
		Φ=10	0.003	0.003	0.014	0.009	0.001	0.002
	$\hat{\sigma}^2_u$	Φ=1.5	0	−0.001	0.002	0.001	0.002	0
		Φ=10	−0.027	−0.025	−0.023	−0.013	0.002	0.015

Table 5: Empirical MSE of the slope ($\hat{\beta}_4$) and variance component ($\hat{\sigma}_u^2$) as a function of the marginal mean (μ), overdispersion (OD), and the cluster size (m). Sample size is $N = 100$, the offset is equal to 60.

			lme4		**hglm**		**INLA**	
		OD	m=2	m=5	m=2	m=5	m=2	m=5
μ=1	$\hat{\beta}_4$	Φ=1.5	0.041	0.025	0.04	0.024	0.043	0.026
		Φ=10	0.274	0.231	0.189	0.175	0.315	0.25
	$\hat{\sigma}_u^2$	Φ=1.5	0.017	0.006	0.013	0.005	0.02	0.007
		Φ=10	0.968	0.696	2.624	1.58	1.179	0.836
μ=10	$\hat{\beta}_4$	Φ=1.5	0.002	0.001	0.002	0.001	0.002	0.001
		Φ=10	0.023	0.022	0.022	0.022	0.023	0.022
	$\hat{\sigma}_u^2$	Φ=1.5	0	0	0	0	0	0
		Φ=10	0.007	0.005	0.006	0.005	0.008	0.006

C Supplemental Material: R code for simulations

C.1 Simulation study 1

```
## Note: Simulation study 1 uses the data set of Folk wrestling data (Section 6)
##       This data cannot be provided. Contrary to that, the code for
##       simulation study 2 does not require any external data.
################################################################################


################################################
## Simulation 1: On Real Data (Poisson response)
################################################

rm(list=ls(all=TRUE))
setwd("C:\\Users\\klangohr\\Documents\\A_UPC\\Marti\\PhD")
load("Aluche2.RData")


### Global settings and needed packages
#######################################
library(lme4)
library(hglm)
library(INLA)
library(R.utils)

### The real data set (Subset of "veran" containing data of regular fighters)
################################################################################

# Regular fighters only
dd.reg <- subset(veran, Regular=="Yes")
dd.reg <- with(dd.reg, dd.reg[order(Nwrestler, Season), ])
# Discard variables that are no longer needed
dd.reg <- dd.reg[c(2, 3, 9, 17, 28)]
rownames(dd.reg) <- 1:nrow(dd.reg)
names(dd.reg)[1:2] <- tolower(names(dd.reg)[1:2])
names(dd.reg)[4] <- 'ncombat'

formula1 <- with(dd.reg, LT~category+fallswinner)
X <- model.matrix(formula1)
formula2 <- ~0+factor(dd.reg$nwrestler)
```

```
Z <- model.matrix(formula2)
rm(formula1, formula2)

### Function to generate data based on data set dd.reg
### (adds new variable Y to existing data set dd.reg)
#########################################################
## Info:
## b0: intercept;
## b1--b4 correspond to 'category': b1, heavy; b2, light; b3, semi-heavy; b4, Winner

simfun <- function(X, Z, b0=-2.09, b1=0.40, b2=0.24, b3=0.10, b4=-0.46, sigmaZu){
  b <- c(b0, b1, b2, b3, b4)
  u=rnorm(length(unique(dd.reg$nwrestler)), 0, sigmaZu)
  eta=exp(X%*%b + Z%*%u + log(dd.reg$ncombat))
  dd.reg$Y <- rpois(length(eta), eta)
  dd.reg
}


### Simulation settings
#######################
# Fixed values
b1 <- 0.40
b2 <- 0.24
b3 <- 0.10
b4 <- -0.46

# Varying parameters
bet0 <- c(-4.1, -1.7, -4.5, -1.75, -4.35, -1.85, -4.8, -2.09)
sigmaZu <- c(0, 0, 0.5, 0.05, 1, 0.2, 2.1, 0.67)
# Save in data frame
SimSettings1to8 <- data.frame(Combination=1:8, Beta0=bet0, Sigma=sigmaZu)
comment(SimSettings1to8) <- 'Simulation settings for real data simulations'
rm(bet0, sigmaZu)

# Number of data sets to be generated
nrep <- 1000


### Preparation of lists with results from simulation settings 1 to 8
#####################################################################
SimDetails1to8 <- vector('list', 8)
SimComparison1to8 <- vector('list', 8)

### SimDetails1to8 contains 8 lists, each with all results of each method
##########################################################################
names(SimDetails1to8) <- paste("SimSetting", 1:8, sep='')
```

```
for (i in 1:8){
  SimDetails1to8[[i]] <- vector('list', 4)
  names(SimDetails1to8[[i]]) <- c('GLM', 'GLMM', 'HGLM', 'INLA')
}
rm(i)

### SimComparison1to8 contains 8 lists, each with the comparison criteria
#########################################################################
names(SimComparison1to8) <- paste("SimSetting", 1:8, sep='')
for (i in 1:8){
  SimComparison1to8[[i]] <- vector('list', 3)
  names(SimComparison1to8[[i]]) <- c('ConvergenceRates', 'EstimationResults',
                                     'ComputationTimes')
}
rm(i)


### La simulació
################
for (s in 1:8)){
  set.seed(111114)
  b0    <- SimSettings1to8$Beta0[s]
  sigZu <- SimSettings1to8$Sigma[s]

  # Generate nrep data sets for each setting
  alldata <- vector('list', nrep)
  for (i in 1:nrep){
    alldata[[i]] <- simfun(X=X, Z=Z, b0=b0, sigmaZu=sigZu)
  }
  rm(i)

  ### Application of all packages to nrep data sets:
  ### * Temporary lists for each package are created.
  ### * 3 data frames will contain results for each setting
  ############################################################
  # List for function glm
  estglm <- vector('list', 8)
  names(estglm) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4', 'time',
                     'coverBet0glm', 'coverBet4glm')

  # List for package lm4
  estglmer <- vector('list', 9)
  names(estglmer) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4', 'var_id',
                       'time', 'coverBet0lmer', 'coverBet4lmer')

  # List for package hglm
  esthglm <- vector('list', 10)
```

```
names(esthglm) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4', 'var_id',
                    'se_var_id', 'time', 'coverBet0hglm', 'coverBet4hglm')

# List for package INLA
estinla <- vector('list', 10)
names(estinla) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4', 'var_id',
                    'se_var_id', 'time', 'coverBet0inla', 'coverBet4inla')

# Default values (FALSE and NA)
estglm$conv   <- rep(FALSE, nrep)
estglmer$conv <- rep(FALSE, nrep)
esthglm$conv  <- rep(FALSE, nrep)
estinla$conv  <- rep(FALSE, nrep)

for (l in 2:8){
  estglm[[l]]   <- rep(NA, nrep)
  estglmer[[l]] <- rep(NA, nrep)
  esthglm[[l]]  <- rep(NA, nrep)
  estinla[[l]]  <- rep(NA, nrep)
}
estglmer[[9]] <- rep(NA, nrep)
for (l in 9:10){
  esthglm[[l]] <- rep(NA, nrep)
  estinla[[l]] <- rep(NA, nrep)
}
rm(l)
for (i in 1:nrep){
  # Data set to be used
  dd <- alldata[[i]]

  # Fit of an GLM
  ################
  estglm$time[i] <- system.time(mod0 <- try(glm(Y~offset(log(ncombat))+category
                                +fallswinner, family='poisson', data=dd)))[3]
  if (is.list(mod0)){
    estglm$conv[i] <- mod0$converged
    if (mod0$converged){
      estglm$beta0[i] <- coef(summary(mod0))[1, 1]
      estglm$beta4[i] <- coef(summary(mod0))[5, 1]
      estglm$se_beta0[i] <- summary(mod0)$coef[1, 2]
      estglm$se_beta4[i] <- summary(mod0)$coef[5, 2]
      n.glm <- mod0$df.residual
      t05 <- qt(0.975, n.glm)
      estglm$coverBet0glm[i] <-
        coef(summary(mod0))[1, 1]-t05*summary(mod0)$coef[1, 2] <= b0 &
        coef(summary(mod0))[1, 1]+t05*summary(mod0)$coef[1, 2] >= b0
      estglm$coverBet4glm[i] <-
```

```
      coef(summary(mod0))[5, 1]-t05*summary(mod0)$coef[5, 2] <= b4 &
      coef(summary(mod0))[5, 1]+t05*summary(mod0)$coef[5, 2] >= b4
    rm(n.glm, t05)
  } else{
      estglm$time[i]
  }
} else{
    estglm$time[i] <- NA
}
      rm(mod0)

# Fit of an GLMM
################
estglmer$time[i] <- system.time(mod1 <- try(glmer(Y~category+fallswinner+
  (1|nwrestler)+offset(log(ncombat)), family="poisson", data=dd)))[3]
mod1 <- update(mod1, nAGQ=5)
if (class(mod1)=="glmerMod"){
  estglmer$conv[i] <- is.null(mod1@optinfo$conv$lme4$messages)
  if (is.null(mod1@optinfo$conv$lme4$messages)){
    row.fw <- which(rownames(summary(mod1)$coefficient) == 'fallswinnerYes')
    estglmer$beta0[i] <- mod1@beta[1]
    estglmer$beta4[i] <- mod1@beta[row.fw]
    estglmer$se_beta0[i] <-coef(summary(mod1))[1, 2]
    estglmer$se_beta4[i] <- coef(summary(mod1))[row.fw, 2]
    estglmer$var_id[i] <- attr(lme4::VarCorr(mod1)$nwrestler, "stddev")^2

    n <- length(fitted(mod1))
    k <- attr(logLik(mod1), "df")
    t05 <- qt(0.975, n-k)
    estglmer$coverBet0lmer[i] <-
      mod1@beta[1]-t05*coef(summary(mod1))[1, 2] <= b0 &
      mod1@beta[1]+t05*coef(summary(mod1))[1, 2] >= b0
    estglmer$coverBet4lmer[i] <-
      mod1@beta[5]-t05*coef(summary(mod1))[row.fw, 2] <= b4 &
      mod1@beta[5]+t05*coef(summary(mod1))[row.fw, 2] >= b4
    rm(n, k, t05, row.fw)
  } else{
      estglmer$time[i] <- NA
  }
} else{
    estglmer$time[i] <- NA
}
rm(mod1)

# Fit of an HGLM
################
esthglm$time[i] <- system.time(mod2 <- try(hglm(fixed=Y~category+fallswinner,
```

```
  random=~1|nwrestler, offset=(log(ncombat)), fix.disp=1, family=poisson(),
  method="HL11", data=dd, maxit=200, conv=1e-8)))[3]
if (is.list(mod2)){
  esthglm$conv[i] <- mod2$Converge=='converged'
  if (mod2$Converge=='converged'){
    esthglm$beta0[i] <- mod2$fixef[1]
    esthglm$beta4[i] <- mod2$fixef[5]
    esthglm$se_beta0[i] <- mod2$SeFe[1]
    esthglm$se_beta4[i] <- mod2$SeFe[5]
    esthglm$var_id[i] <- mod2$varRanef
    esthglm$se_var_id[i] <- 0.5*sqrt(mod2$varRanef)*mod2$SummVC2[[1]][2]

    n <- length(mod2$fv)
    k <- (n-(mod2$dfReFe))
    t05 <- qt(0.975, n-k)
    esthglm$coverBet0hglm[i] <-
      mod2$fixef[1]-t05*mod2$SeFe[1]<= b0 &
      mod2$fixef[1]+t05*mod2$SeFe[1]>= b0
    esthglm$coverBet4hglm[i] <-
      mod2$fixef[5]-t05*mod2$SeFe[5]<= b4 &
      mod2$fixef[5]+t05*mod2$SeFe[5]>= b4
    rm(n, k, t05)
  } el  se{
      esthglm$time[i] <- NA
  }
} else{
    esthglm$time[i] <- NA
}
rm(mod2)

# Fit of an INLA model
######################
formula <- Y~offset(log(ncombat))+ category+fallswinner+f(nwrestler, model="iid",
           hyper=list(theta=list(prior="logtgaussian", param=c(0, 0.0001))))
estinla$time[i] <- system.time(mod3 <- try(inla(formula, data=dd,
   family="poisson", control.compute=list(dic=T, cpo=TRUE),
   control.inla=list(strategy="laplace", int.strategy="grid"))))[3]
if (is.list(mod3)& !is.null(mod3$mode$mode.status)){
  estinla$conv[i] <- mod3$mode$mode.status==0
  if (mod3$mode$mode.status==0){
    estinla$beta0[i] <- mod3$summary.fixed[1, 1]
    estinla$beta4[i] <- mod3$summary.fixed[5, 1]
    estinla$se_beta0[i] <- mod3$summary.fixed[1, 2]
    estinla$se_beta4[i] <- mod3$summary.fixed[5, 2]

    prec.marg <- (mod3$marginals.hyperpar$'Precision for nwrestler')
    marg.variance <- inla.tmarginal(
```

```
              function(x) 1/x,
              prec.marg
      )
      m <- inla.emarginal(function(x) x, marg.variance)
      mm  <- inla.emarginal(function(x) x^2, marg.variance)

      estinla$var_id[i] <- m
      estinla$se_var_id[i] <- sqrt(mm-m^2)

      estinla$coverBet0inla[i] <- mod3$summary.fixed[1, 3] <= b0 &
                                  mod3$summary.fixed[1, 5] >= b0
      estinla$coverBet4inla[i] <- mod3$summary.fixed[5, 3] <= b4 &
                                  mod3$summary.fixed[5, 5] >= b4
      rm(prec.marg, marg.variance, m, mm)
    } else{
        estinla$time[i] <- NA
    }
  } else{
      estinla$time[i] <- NA
  }
  rm(mod3, formula, dd)
}
rm(i)

### Save lists in list SimDetails1to8
#################################
SimDetails1to8[[s]][[1]] <- estglm
SimDetails1to8[[s]][[2]] <- estglmer
SimDetails1to8[[s]][[3]] <- esthglm
SimDetails1to8[[s]][[4]] <- estinla

### Computation of comparison criteria
######################################

### Convergence rates
conv.rate.glm   <- sum(estglm$conv)/nrep*100
conv.rate.glmer <- sum(estglmer$conv)/nrep*100
conv.rate.hglm  <- sum(esthglm$conv)/nrep*100
conv.rate.inla  <- sum(estinla$conv)/nrep*100


### Bias of beta0 and beta4
bias.glm.b0  <- mean(estglm$beta0, na.rm=T)-b0
bias.glm.b4  <- mean(estglm$beta4, na.rm=T)-b4
bias.lme4.b0 <- mean(estglmer$beta0, na.rm=T)-b0
bias.lme4.b4 <- mean(estglmer$beta4, na.rm=T)-b4
bias.hglm.b0 <- mean(esthglm$beta0, na.rm=T)-b0
```

```
bias.hglm.b4 <- mean(esthglm$beta4, na.rm=T)-b4
bias.inla.b0 <- mean(estinla$beta0, na.rm=T)-b0
bias.inla.b4 <- mean(estinla$beta4, na.rm=T)-b4

### Variances and standard errors
var1.glm.b0 <- mean(estglm$se_beta0^2, na.rm=T)
var1.glm.b4 <- mean(estglm$se_beta4^2, na.rm=T)
se.glm.b0   <- sqrt(var1.glm.b0)
se.glm.b4   <- sqrt(var1.glm.b4)
var2.glm.b0 <- var(estglm$beta0, na.rm=T)
var2.glm.b4 <- var(estglm$beta4, na.rm=T)

var1.lme4.b0 <- mean(estglmer$se_beta0^2, na.rm=T)
var1.lme4.b4 <- mean(estglmer$se_beta4^2, na.rm=T)
se.lme4.b0   <- sqrt(var1.lme4.b0)
se.lme4.b4   <- sqrt(var1.lme4.b4)
var2.lme4.b0 <- var(estglmer$beta0, na.rm=T)
var2.lme4.b4 <- var(estglmer$beta4, na.rm=T)

var1.hglm.b0 <- mean(esthglm$se_beta0^2, na.rm=T)
var1.hglm.b4 <- mean(esthglm$se_beta4^2, na.rm=T)
se.hglm.b0   <- sqrt(var1.hglm.b0)
se.hglm.b4   <- sqrt(var1.hglm.b4)
var2.hglm.b0 <- var(esthglm$beta0, na.rm=T)
var2.hglm.b4 <- var(esthglm$beta4, na.rm=T)

var1.inla.b0 <- mean(estinla$se_beta0^2, na.rm=T)
var1.inla.b4 <- mean(estinla$se_beta4^2, na.rm=T)
se.inla.b0   <- sqrt(var1.inla.b0)
se.inla.b4   <- sqrt(var1.inla.b4)
var2.inla.b0 <- var(estinla$beta0, na.rm=T)
var2.inla.b4 <- var(estinla$beta4, na.rm=T)

### Confidence intervals for standard errors
CI.se.glm.b0  <- quantile(estglm$se_beta0, c(0.025, 0.975))
CI.se.glm.b4  <- quantile(estglm$se_beta4, c(0.025, 0.975), na.rm=T)
CI.se.lme4.b0 <- quantile(estglmer$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.lme4.b4 <- quantile(estglmer$se_beta4, c(0.025, 0.975), na.rm=T)
CI.se.hglm.b0 <- quantile(esthglm$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.hglm.b4 <- quantile(esthglm$se_beta4, c(0.025, 0.975), na.rm=T)
CI.se.inla.b0 <- quantile(estinla$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.inla.b4 <- quantile(estinla$se_beta4, c(0.025, 0.975), na.rm=T)

### MSE for both parameters
MSE.glm.b0  <- bias.glm.b0^2 + var2.glm.b0
MSE.glm.b4  <- bias.glm.b4^2 + var2.glm.b4
MSE.lme4.b0 <- bias.lme4.b0^2+ var2.lme4.b0
```

```
MSE.lme4.b4 <- bias.lme4.b4^2+ var2.lme4.b4
MSE.hglm.b0 <- bias.hglm.b0^2+ var2.hglm.b0
MSE.hglm.b4 <- bias.hglm.b4^2+ var2.hglm.b4
MSE.inla.b0 <- bias.inla.b0^2+ var2.inla.b0
MSE.inla.b4 <- bias.inla.b4^2+ var2.inla.b4

### Coverage
coverBet0glm  <- sum(estglm$coverBet0glm, na.rm=T)/sum(estglm$conv)*100
coverBet4glm  <- sum(estglm$coverBet4glm, na.rm=T)/sum(estglm$conv)*100
coverBet0lmer <- sum(estglmer$coverBet0lmer, na.rm=T)/sum(estglmer$conv)*100
coverBet4lmer <- sum(estglmer$coverBet4lmer, na.rm=T)/sum(estglmer$conv)*100
coverBet0hglm <- sum(esthglm$coverBet0hglm, na.rm=T)/sum(esthglm$conv)*100
coverBet4hglm <- sum(esthglm$coverBet4hglm, na.rm=T)/sum(esthglm$conv)*100
coverBet0inla <- sum(estinla$coverBet0inla, na.rm=T)/sum(estinla$conv)*100
coverBet4inla <- sum(estinla$coverBet4inla, na.rm=T)/sum(estinla$conv)*100

### Variances of random effects
bias.lme4.var_id <- mean(estglmer$var_id, na.rm=T)-sigZu^2
se.lme4.var_id   <- sd(estglmer$var_id, na.rm=T)
MSE.lme4.var_id  <- bias.lme4.var_id^2 + var(estglmer$var_id, na.rm=T)
CI.lme4.var_id   <- quantile(estglmer$var_id, c(0.025, 0.975), na.rm=T)

bias.hglm.var_id <- mean(esthglm$var_id, na.rm=T)-sigZu^2
se.hglm.var_id   <- sd(esthglm$var_id, na.rm=T)
MSE.hglm.var_id  <- bias.hglm.var_id^2 + var(esthglm$var_id, na.rm=T)
CI.hglm.var_id   <- quantile(esthglm$se_var_id, c(0.025, 0.975), na.rm=T)
se.bias.hglm.var_id <- mean(esthglm$se_var_id, na.rm=T)

bias.inla.var_id <- mean(estinla$var_id, na.rm=T)-sigZu^2
MSE.inla.var_id  <- bias.inla.var_id^2 + var(estinla$var_id, na.rm=T)
CI.inla.var_id   <- quantile(estinla$se_var_id, c(0.025, 0.975), na.rm=T)
se.bias.inla.var_id <- mean(estinla$se_var_id, na.rm=T)


### Mean computation times
mean.time.glm   <- mean(estglm$time, na.rm=T)
mean.time.glmer <- mean(estglmer$time, na.rm=T)
mean.time.hglm  <- mean(esthglm$time, na.rm=T)
mean.time.inla  <- mean(estinla$time, na.rm=T)

### Save comparison criteria in list SimComparison1to8
#########################################################
package <- c('glm', 'glmer', 'hglm', 'INLA')

### Data frame with convergence rates
conv.rt <- c(conv.rate.glm, conv.rate.glmer, conv.rate.hglm, conv.rate.inla)
conv.df <- data.frame("Conv.Rates"=round(conv.rt, 1))
```

```
rownames(conv.df) <- package
SimComparison1to8[[s]][[1]] <- conv.df

# Cleaning up
rm(conv.rt, conv.df)
rm(conv.rate.glm, conv.rate.glmer, conv.rate.hglm, conv.rate.inla)

### Data frame with main results
method <- paste(rep(c('Intercept', 'Winner', 'SigmaWrest'), each=4),
                  c('glm', 'glmer', 'hglm', 'INLA'), sep='.')
bias   <- c(bias.glm.b0, bias.lme4.b0, bias.hglm.b0, bias.inla.b0,
            bias.glm.b4, bias.lme4.b4, bias.hglm.b4, bias.inla.b4,
            NA, bias.lme4.var_id, bias.hglm.var_id, bias.inla.var_id)
MSE    <- c(MSE.glm.b0, MSE.lme4.b0, MSE.hglm.b0, MSE.inla.b0,
            MSE.glm.b4, MSE.lme4.b4, MSE.hglm.b4, MSE.inla.b4,
            NA, MSE.lme4.var_id, MSE.hglm.var_id, MSE.inla.var_id)
Var    <- c(var2.glm.b0, var2.lme4.b0, var2.hglm.b0, var2.inla.b0,
            var2.glm.b4, var2.lme4.b4, var2.hglm.b4, var2.inla.b4,
            NA, var(estglmer$var_id, na.rm=T), var(esthglm$var_id, na.rm=T),
            var(estinla$var_id, na.rm=T))
sder   <- c(se.glm.b0, se.lme4.b0, se.hglm.b0, se.inla.b0,
            se.glm.b4, se.lme4.b4, se.hglm.b4, se.inla.b4,
            rep(NA, 2), se.bias.hglm.var_id, se.bias.inla.var_id)
ic95l  <- c(CI.se.glm.b0[1], CI.se.lme4.b0[1], CI.se.hglm.b0[1], CI.se.inla.b0[1],
            CI.se.glm.b4[1], CI.se.lme4.b4[1], CI.se.hglm.b4[1], CI.se.inla.b4[1],
            rep(NA, 2), CI.hglm.var_id[1], CI.inla.var_id[1])
ic95r  <- c(CI.se.glm.b0[2], CI.se.lme4.b0[2], CI.se.hglm.b0[2], CI.se.inla.b0[2],
            CI.se.glm.b4[2], CI.se.lme4.b4[2], CI.se.hglm.b4[2], CI.se.inla.b4[2],
            rep(NA, 2), CI.hglm.var_id[2], CI.inla.var_id[2])
cove <- c(coverBet0glm, coverBet0lmer, coverBet0hglm, coverBet0inla,
          coverBet4glm, coverBet4lmer, coverBet4hglm, coverBet4inla, rep(NA, 4))
results <- data.frame(Bias=round(bias, 3), MSE=round(MSE, 3), Var=round(Var, 3),
                      Std.Error=round(sder, 3), "SE.CI95%Lower"=round(ic95l, 3),
                      "SE.CI95%Upper"=round(ic95r, 3), Coverage=round(cove, 2))
rownames(results) <- method
SimComparison1to8[[s]][[2]] <- results

# Cleaning up
rm(results, method, bias, MSE, Var, sder, ic95l, ic95r, cove)
rm(bias.glm.b0, bias.lme4.b0, bias.hglm.b0, bias.inla.b0, bias.glm.b4,
   bias.lme4.b4, bias.hglm.b4, bias.inla.b4, bias.lme4.var_id,
   bias.hglm.var_id, bias.inla.var_id)
rm(MSE.glm.b0, MSE.lme4.b0, MSE.hglm.b0, MSE.inla.b0, MSE.glm.b4, MSE.lme4.b4,
   MSE.hglm.b4, MSE.inla.b4, MSE.lme4.var_id, MSE.hglm.var_id, MSE.inla.var_id)
rm(var1.glm.b0, var1.lme4.b0, var1.hglm.b0, var1.inla.b0, var1.glm.b4,
   var1.lme4.b4, var1.hglm.b4, var1.inla.b4)
rm(var2.glm.b0, var2.lme4.b0, var2.hglm.b0, var2.inla.b0, var2.glm.b4,
```

```
    var2.lme4.b4, var2.hglm.b4, var2.inla.b4)
  rm(se.glm.b0, se.lme4.b0, se.hglm.b0, se.inla.b0, se.glm.b4, se.lme4.b4,
    se.hglm.b4, se.inla.b4, se.bias.hglm.var_id, se.bias.inla.var_id,
    se.hglm.var_id, se.lme4.var_id)
  rm(CI.se.glm.b0, CI.se.lme4.b0, CI.se.hglm.b0, CI.se.inla.b0, CI.se.glm.b4,
    CI.se.lme4.b4, CI.se.hglm.b4, CI.se.inla.b4, CI.inla.var_id, CI.hglm.var_id,
    CI.lme4.var_id)
  rm(coverBet0glm, coverBet0lmer, coverBet0hglm, coverBet0inla, coverBet4glm,
    coverBet4lmer, coverBet4hglm, coverBet4inla)


  ### Data frame with computation times
  comp.time  <- c(mean.time.glm, mean.time.glmer, mean.time.hglm, mean.time.inla)
  ctime.df <- data.frame(ElapsedTime=round(comp.time, 2))
  rownames(ctime.df) <- package
  SimComparison1to8[[s]][[3]] <- ctime.df

  # Cleaning up
  rm(comp.time, ctime.df)
  rm(mean.time.glm, mean.time.glmer, mean.time.hglm, mean.time.inla)
  rm(alldata, b0, sigZu, estglm, estglmer, esthglm, estinla, package)

  list.comparison <- SimComparison1to8[[s]]
  list.details    <- SimDetails1to8[[s]]
  save(list.comparison, list.details, file=paste0("SimulationResults", s, ".RData"))
  rm(list.comparison, list.details)
}

rm(s, X, Z, b1, b2, b3, b4, dd.reg, nrep, simfun)
comment(SimDetails1to8) <-
  'List for all 8 simulations with detailed results of all packages'
comment(SimComparison1to8) <-
  'List for all 8 simulations with comparison criteria'
save(SimSettings1to8, SimDetails1to8, SimComparison1to8,
  file='SimulationResults1to8.RData')
```

C.2 Simulation study 2

```
##################################################
## Simulation 2: Simulated Data (Poisson response)
##################################################

rm(list=ls(all=TRUE))
setwd("C:\\Users\\klangohr\\Documents\\A_UPC\\Marti\\PhD")
```

```
### Global settings and needed packages
########################################
library(lme4)
library(hglm)
library(INLA)
library(R.utils)


### Function to generate simulation data
########################################
## Info: b0: intercept;
## b1--b4 correspond to 'category': b1, heavy; b2, light; b3, semi-heavy; b4, Winner

simfun2 <- function(nw=100, ns=6, nco=100, sigmaZu=2.1, b0=-4.4, b1=0.4, b2=0.25,
                    b3=0.1, b4=-0.5)
{
  each <- sample(1:6, nw, replace=T)  # number of seasons of each wrestler
  nwrestler <- rep(1:nw, each)
  ntot <- length(nwrestler)
  category <- factor(rep(sample(c('Light', 'Medium', 'Semi heavy', 'Heavy'),
           nw, replace=T), each), levels=c('Light', 'Medium', 'Semi heavy', 'Heavy'))
  fallswinner <- factor(rep(sample(c('No', 'Yes'), nw, replace=T), each))
  ncombat <- rpois(ntot, nco)
  dd <- data.frame(nwrestler, ncombat, category, fallswinner)
  obs <- 1:length(nwrestler)
  b <- c(b0, b1, b2, b3, b4)
  eta0 <- model.matrix(~category+fallswinner, data=dd)%*%b
  alfa <- rep(rnorm(nw, sd=sigmaZu), each)
  eta <- with(dd, log(ncombat)+eta0+alfa)
  mu <- exp(eta)
  dd$Y <- with(dd, rpois(ntot, lambda=mu))
  dd
}


### Simulation settings
#######################
# Fixed values
# b1 <- 0.4; b2 <- 0.25; b3 <- 0.1;
b4 <- -0.5

# Varying parameters
nw <- rep(c(30, 100), each=16)
nco <- rep(rep(c(60, 100), each=8), 2)
bet0 <- c(-4.1, -1.7, -4.5, -1.75, -4.35, -1.85, -4.8, -2.09, -4.6, -2.21, -4.7,
          -2.27, -4.87, -2.34, -5.35, -2.55, -4.1, -1.7, -4.5, -1.75, -4.35,
          -1.85, -4.8, -2.09, -4.6, -2.21, -4.7, -2.27, -4.87, -2.34, -5.35, -2.55)
```

```
sigmaZu <- c(0, 0, 0.5, 0.05, 1, 0.2, 2.1, 0.67, 0, 0, 0.4, 0.05, 1, 0.2, 2.1,
             0.65, 0, 0, 0.5, 0.05, 1, 0.2, 2.1, 0.67, 0, 0, 0.4, 0.05, 1, 0.2,
             2.1, 0.65)
# Save in data frame
SimSettings940 <- data.frame(Combination=9:40, NumbWrestlers=nw, NumbCombats=nco,
                             Beta0=bet0, Sigma=sigmaZu)
comment(SimSettings940) <- 'Simulation settings for Scenario 2'
rm(bet0, sigmaZu, nco, nw)


# Number of data sets to be generated
nrep <- 1000


### Preparation of lists with results from simulation settings 1 to 8
#########################################################################
SimDetails940 <- vector('list', 32)
SimComparison940 <- vector('list', 32)

### SimDetails940 contains 8 lists, each with all results of each method
#########################################################################
names(SimDetails940) <- paste("SimSetting", 9:40, sep='')
for (i in 1:32){
  SimDetails940[[i]] <- vector('list', 4)
  names(SimDetails940[[i]]) <- c('GLM', 'GLMM', 'HGLM', 'INLA')
}
rm(i)

### SimComparison940 contains 8 lists, each with the comparison criteria
#########################################################################
names(SimComparison940) <- paste("SimSetting", 9:40, sep='')
for (i in 1:32){
  SimComparison940[[i]] <- vector('list', 3)
  names(SimComparison940[[i]]) <- c('ConvergenceRates', 'EstimationResults',
                                    'ComputationTimes')
}
rm(i)


### La simulació
################
for (s in 1:32){
  set.seed(111114)
  nw    <- SimSettings940$NumbWrestlers[s]
  nco   <- SimSettings940$NumbCombats[s]
  b0    <- SimSettings940$Beta0[s]
  sigZu <- SimSettings940$Sigma[s]
```

```
# Generate nrep data sets for each setting
alldata <- vector('list', nrep)
for (i in 1:nrep){
  alldata[[i]] <- simfun2(nw=nw, ns=6, nco=nco, b0=b0, sigmaZu=sigZu)
}
rm(i)


### Application of all packages to nrep data sets:
### * Temporary lists for each package are created.
### * 3 data frames will contain results for each setting
############################################################
# List for function glm
estglm <- vector('list', 8)
names(estglm) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4', 'time',
                   'coverBet0glm', 'coverBet4glm')

# List for package lm4
estglmer <- vector('list', 9)
names(estglmer) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4',
                     'var_id', 'time', 'coverBet0lmer', 'coverBet4lmer')

# List for package hglm
esthglm <- vector('list', 10)
names(esthglm) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4',
                    'var_id', 'se_var_id', 'time', 'coverBet0hglm', 'coverBet4hglm')

# List for package INLA
estinla <- vector('list', 10)
names(estinla) <- c('conv', 'beta0', 'beta4', 'se_beta0', 'se_beta4',
                     'var_id', 'se_var_id', 'time', 'coverBet0inla', 'coverBet4inla')

# Default values (FALSE and NA)
estglm$conv   <- rep(FALSE, nrep)
estglmer$conv <- rep(FALSE, nrep)
esthglm$conv  <- rep(FALSE, nrep)
estinla$conv  <- rep(FALSE, nrep)

for (l in 2:8){
  estglm[[l]]   <- rep(NA, nrep)
  estglmer[[l]] <- rep(NA, nrep)
  esthglm[[l]]  <- rep(NA, nrep)
  estinla[[l]]  <- rep(NA, nrep)
}
estglmer[[9]] <- rep(NA, nrep)
for (l in 9:10){
```

```
  esthglm[[l]] <- rep(NA, nrep)
  estinla[[l]] <- rep(NA, nrep)
}
rm(l)

for (i in 1:nrep){
  # Data set to be used
  dd <- alldata[[i]]

  # Fit of an GLM
  ################
  estglm$time[i] <- system.time(mod0 <- try(glm(Y~offset(log(ncombat))+category
    +fallswinner, family='poisson', data=dd)))[3]
  if (is.list(mod0)){
    estglm$conv[i] <- mod0$converged
    if (mod0$converged){
      estglm$beta0[i] <- coef(summary(mod0))[1, 1]
      estglm$beta4[i] <- coef(summary(mod0))["fallswinnerYes", 1]
      estglm$se_beta0[i] <- summary(mod0)$coef[1, 2]
      estglm$se_beta4[i] <- summary(mod0)$coef["fallswinnerYes", 2]
      n.glm <- mod0$df.residual
      t05 <- qt(0.975, n.glm)
      estglm$coverBet0glm[i] <-
        coef(summary(mod0))[1, 1]-t05*summary(mod0)$coef[1, 2] <= b0 &
        coef(summary(mod0))[1, 1]+t05*summary(mod0)$coef[1, 2] >= b0
      estglm$coverBet4glm[i] <-
        coef(summary(mod0))["fallswinnerYes", 1]-
          t05*summary(mod0)$coef["fallswinnerYes", 2] <= b4 &
        coef(summary(mod0))["fallswinnerYes", 1]+
          t05*summary(mod0)$coef["fallswinnerYes", 2] >= b4
      rm(n.glm, t05)
    } else{
        estglm$time[i]
    }
  } else{
      estglm$time[i] <- NA
  }
        rm(mod0)

  # Fit of an GLMM
  #################
  estglmer$time[i] <- system.time(mod1 <- try(glmer(Y~category+fallswinner
    +(1|nwrestler)+offset(log(ncombat)), family="poisson", data=dd)))[3]
  mod1 <- update(mod1, nAGQ=5)
  if (class(mod1)=="glmerMod"){
    estglmer$conv[i] <- is.null(mod1@optinfo$conv$lme4$messages)
    if (is.null(mod1@optinfo$conv$lme4$messages)){
```

```
    row.fw <- which(rownames(summary(mod1)$coefficient) == 'fallswinnerYes')
    estglmer$beta0[i] <- mod1@beta[1]
    estglmer$beta4[i] <- mod1@beta[row.fw]
    estglmer$se_beta0[i] <-coef(summary(mod1))[1, 2]
    estglmer$se_beta4[i] <- coef(summary(mod1))[row.fw, 2]
    estglmer$var_id[i] <- attr(lme4::VarCorr(mod1)$nwrestler, "stddev")^2

    n <- length(fitted(mod1))
    k <- attr(logLik(mod1), "df")
    t05 <- qt(0.975, n-k)
    estglmer$coverBet0lmer[i] <-
      mod1@beta[1]-t05*coef(summary(mod1))[1, 2] <= b0 &
      mod1@beta[1]+t05*coef(summary(mod1))[1, 2] >= b0
    estglmer$coverBet4lmer[i] <-
      mod1@beta[5]-t05*coef(summary(mod1))[row.fw, 2] <= b4 &
      mod1@beta[5]+t05*coef(summary(mod1))[row.fw, 2] >= b4
    rm(n, k, t05, row.fw)
  } else{
      estglmer$time[i] <- NA
  }
} else{
    estglmer$time[i] <- NA
}
rm(mod1)

# Fit of an HGLM
#################
esthglm$time[i] <- system.time(mod2 <- try(hglm(fixed=Y~category+fallswinner,
  random=~1|nwrestler, offset=(log(ncombat)), fix.disp=1, family=poisson(),
  data=dd, maxit=200, conv=1e-8)))[3]
if (is.list(mod2)){
  esthglm$conv[i] <- mod2$Converge=='converged'
  if (mod2$Converge=='converged'){
    esthglm$beta0[i] <- mod2$fixef[1]
    esthglm$beta4[i] <- mod2$fixef["fallswinnerYes"]
    esthglm$se_beta0[i] <- mod2$SeFe[1]
    ene <- length(mod2$SeFe)
    esthglm$se_beta4[i] <- mod2$SeFe[ene]
    esthglm$var_id[i] <- mod2$varRanef
    esthglm$se_var_id[i] <- 0.5*sqrt(mod2$varRanef)*mod2$SummVC2[[1]][2]

    n <- length(mod2$fv)
    k <- (n-(mod2$dfReFe))
    t05 <- qt(0.975, n-k)
    esthglm$coverBet0hglm[i] <-
      mod2$fixef[1]-t05*mod2$SeFe[1]<= b0 &
      mod2$fixef[1]+t05*mod2$SeFe[1]>= b0
```

```
    esthglm$coverBet4hglm[i] <-
      mod2$fixef["fallswinnerYes"]-t05*mod2$SeFe[ene]<= b4 &
      mod2$fixef["fallswinnerYes"]+t05*mod2$SeFe[ene]>= b4
    rm(n, k, t05, ene)
  } else{
      esthglm$time[i] <- NA
  }
} else{
    esthglm$time[i] <- NA
}
rm(mod2)

# Fit of an INLA model
######################
formula <- Y~offset(log(ncombat))+ category+fallswinner+f(nwrestler, model="iid",
  hyper=list(theta=list(prior="logtgaussian", param=c(0, 0.0001))))
estinla$time[i] <- system.time(mod3 <- try(inla(formula, data=dd, family="poissor
  control.compute=list(dic=T, cpo=TRUE), control.inla=list(diff.logdens = 10,
  strategy="laplace", int.strategy="grid"))))[3]
if (is.list(mod3)& !is.null(mod3$mode$mode.status)){
  estinla$conv[i] <- mod3$mode$mode.status==0
  if (mod3$mode$mode.status==0){
    estinla$beta0[i] <- mod3$summary.fixed[1, 1]
    estinla$beta4[i] <- mod3$summary.fixed["fallswinnerYes", 1]
    estinla$se_beta0[i] <- mod3$summary.fixed[1, 2]
    estinla$se_beta4[i] <- mod3$summary.fixed["fallswinnerYes", 2]

    prec.marg <- (mod3$marginals.hyperpar$'Precision for nwrestler')
    marg.variance <- inla.tmarginal(
      function(x) 1/x,                        prec.marg
    )
    m <- inla.emarginal(function(x) x, marg.variance)
    mm  <- inla.emarginal(function(x) x^2, marg.variance)

    estinla$var_id[i] <- m
    estinla$se_var_id[i] <- sqrt(mm-m^2)

    estinla$coverBet0inla[i] <- mod3$summary.fixed[1, 3] <= b0 &
                                mod3$summary.fixed[1, 5] >= b0
    estinla$coverBet4inla[i] <- mod3$summary.fixed["fallswinnerYes", 3] <= b4 &
                                mod3$summary.fixed["fallswinnerYes", 5] >= b4
    rm(prec.marg, marg.variance, m, mm)
  } else{
      estinla$time[i] <- NA
  }
} else{
    estinla$time[i] <- NA
```

```
  }
  rm(mod3, formula, dd)
}
rm(i)

### Save lists in list SimDetails940
#################################
SimDetails940[[s]][[1]] <- estglm
SimDetails940[[s]][[2]] <- estglmer
SimDetails940[[s]][[3]] <- esthglm
SimDetails940[[s]][[4]] <- estinla

### Computation of comparison criteria
######################################

### Convergence rates
conv.rate.glm   <- sum(estglm$conv)/nrep*100
conv.rate.glmer <- sum(estglmer$conv)/nrep*100
conv.rate.hglm  <- sum(esthglm$conv)/nrep*100
conv.rate.inla  <- sum(estinla$conv)/nrep*100


### Bias of beta0 and beta 4
bias.glm.b0  <- mean(estglm$beta0, na.rm=T)-b0
bias.glm.b4  <- mean(estglm$beta4, na.rm=T)-b4
bias.lme4.b0 <- mean(estglmer$beta0, na.rm=T)-b0
bias.lme4.b4 <- mean(estglmer$beta4, na.rm=T)-b4
bias.hglm.b0 <- mean(esthglm$beta0, na.rm=T)-b0
bias.hglm.b4 <- mean(esthglm$beta4, na.rm=T)-b4
bias.inla.b0 <- mean(estinla$beta0, na.rm=T)-b0
bias.inla.b4 <- mean(estinla$beta4, na.rm=T)-b4

### Variances and standard errors
var1.glm.b0 <- mean(estglm$se_beta0^2, na.rm=T)
var1.glm.b4 <- mean(estglm$se_beta4^2, na.rm=T)
se.glm.b0   <- sqrt(var1.glm.b0)
se.glm.b4   <- sqrt(var1.glm.b4)
var2.glm.b0 <- var(estglm$beta0, na.rm=T)
var2.glm.b4 <- var(estglm$beta4, na.rm=T)

var1.lme4.b0 <- mean(estglmer$se_beta0^2, na.rm=T)
var1.lme4.b4 <- mean(estglmer$se_beta4^2, na.rm=T)
se.lme4.b0   <- sqrt(var1.lme4.b0)
se.lme4.b4   <- sqrt(var1.lme4.b4)
var2.lme4.b0 <- var(estglmer$beta0, na.rm=T)
var2.lme4.b4 <- var(estglmer$beta4, na.rm=T)
```

```
var1.hglm.b0 <- mean(esthglm$se_beta0^2, na.rm=T)
var1.hglm.b4 <- mean(esthglm$se_beta4^2, na.rm=T)
se.hglm.b0   <- sqrt(var1.hglm.b0)
se.hglm.b4   <- sqrt(var1.hglm.b4)
var2.hglm.b0 <- var(esthglm$beta0, na.rm=T)
var2.hglm.b4 <- var(esthglm$beta4, na.rm=T)

var1.inla.b0 <- mean(estinla$se_beta0^2, na.rm=T)
var1.inla.b4 <- mean(estinla$se_beta4^2, na.rm=T)
se.inla.b0 <- sqrt(var1.inla.b0)
se.inla.b4 <- sqrt(var1.inla.b4)
var2.inla.b0 <- var(estinla$beta0, na.rm=T)
var2.inla.b4 <- var(estinla$beta4, na.rm=T)

### Confidence intervals for standard errors
CI.se.glm.b0  <- quantile(estglm$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.glm.b4  <- quantile(estglm$se_beta4, c(0.025, 0.975), na.rm=T)
CI.se.lme4.b0 <- quantile(estglmer$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.lme4.b4 <- quantile(estglmer$se_beta4, c(0.025, 0.975), na.rm=T)
CI.se.hglm.b0 <- quantile(esthglm$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.hglm.b4 <- quantile(esthglm$se_beta4, c(0.025, 0.975), na.rm=T)
CI.se.inla.b0 <- quantile(estinla$se_beta0, c(0.025, 0.975), na.rm=T)
CI.se.inla.b4 <- quantile(estinla$se_beta4, c(0.025, 0.975), na.rm=T)

### MSE for both parameters
MSE.glm.b0  <- bias.glm.b0^2 + var2.glm.b0
MSE.glm.b4  <- bias.glm.b4^2 + var2.glm.b4
MSE.lme4.b0 <- bias.lme4.b0^2+ var2.lme4.b0
MSE.lme4.b4 <- bias.lme4.b4^2+ var2.lme4.b4
MSE.hglm.b0 <- bias.hglm.b0^2+ var2.hglm.b0
MSE.hglm.b4 <- bias.hglm.b4^2+ var2.hglm.b4
MSE.inla.b0 <- bias.inla.b0^2+ var2.inla.b0
MSE.inla.b4 <- bias.inla.b4^2+ var2.inla.b4

### Coverage
coverBet0glm  <- sum(estglm$coverBet0glm, na.rm=T)/sum(estglm$conv)*100
coverBet4glm  <- sum(estglm$coverBet4glm, na.rm=T)/sum(estglm$conv)*100
coverBet0lmer <- sum(estglmer$coverBet0lmer, na.rm=T)/sum(estglmer$conv)*100
coverBet4lmer <- sum(estglmer$coverBet4lmer, na.rm=T)/sum(estglmer$conv)*100
coverBet0hglm <- sum(esthglm$coverBet0hglm, na.rm=T)/sum(esthglm$conv)*100
coverBet4hglm <- sum(esthglm$coverBet4hglm, na.rm=T)/sum(esthglm$conv)*100
coverBet0inla <- sum(estinla$coverBet0inla, na.rm=T)/sum(estinla$conv)*100
coverBet4inla <- sum(estinla$coverBet4inla, na.rm=T)/sum(estinla$conv)*100

### Variances of random effects
bias.lme4.var_id <- mean(estglmer$var_id, na.rm=T)-sigZu^2
se.lme4.var_id   <- sd(estglmer$var_id, na.rm=T)
```

```
MSE.lme4.var_id  <- bias.lme4.var_id^2+ var(estglmer$var_id, na.rm=T)
CI.lme4.var_id   <- quantile(estglmer$var_id, c(0.025, 0.975), na.rm=T)

bias.hglm.var_id <- mean(esthglm$var_id, na.rm=T)-sigZu^2
se.hglm.var_id   <- sd(esthglm$var_id, na.rm=T)
MSE.hglm.var_id  <- bias.hglm.var_id^2+var(esthglm$var_id, na.rm=T)
CI.hglm.var_id   <- quantile(esthglm$se_var_id, c(0.025, 0.975), na.rm=T)
se.bias.hglm.var_id <- mean(esthglm$se_var_id, na.rm=T)

bias.inla.var_id <- mean(estinla$var_id, na.rm=T)-sigZu^2
MSE.inla.var_id  <- bias.inla.var_id^2+var(estinla$var_id, na.rm=T)
CI.inla.var_id   <- quantile(estinla$se_var_id, c(0.025, 0.975), na.rm=T)
se.bias.inla.var_id <- mean(estinla$se_var_id, na.rm=T)


### Mean computation times
mean.time.glm   <- mean(estglm$time, na.rm=T)
mean.time.glmer <- mean(estglmer$time, na.rm=T)
mean.time.hglm  <- mean(esthglm$time, na.rm=T)
mean.time.inla  <- mean(estinla$time, na.rm=T)

### Save comparison criteria in list SimComparison940
#####################################################
package <- c('glm', 'glmer', 'hglm', 'INLA')

### Data frame with convergence rates
conv.rt <- c(conv.rate.glm, conv.rate.glmer, conv.rate.hglm, conv.rate.inla)
conv.df <- data.frame("Conv.Rates"=round(conv.rt, 1))
rownames(conv.df) <- package
SimComparison940[[s]][[1]] <- conv.df

# Cleaning up
rm(conv.rt, conv.df)
rm(conv.rate.glm, conv.rate.glmer, conv.rate.hglm, conv.rate.inla)

### Data frame with main results
method <- paste(rep(c('Intercept', 'Winner', 'SigmaWrest'), each=4),
                   c('glm', 'glmer', 'hglm', 'INLA'), sep='.')
bias   <- c(bias.glm.b0, bias.lme4.b0, bias.hglm.b0, bias.inla.b0,
            bias.glm.b4, bias.lme4.b4, bias.hglm.b4, bias.inla.b4,
            NA, bias.lme4.var_id, bias.hglm.var_id, bias.inla.var_id)
MSE    <- c(MSE.glm.b0, MSE.lme4.b0, MSE.hglm.b0, MSE.inla.b0,
            MSE.glm.b4, MSE.lme4.b4, MSE.hglm.b4, MSE.inla.b4,
            NA, MSE.lme4.var_id, MSE.hglm.var_id, MSE.inla.var_id)
Var    <- c(var2.glm.b0, var2.lme4.b0, var2.hglm.b0, var2.inla.b0,
            var2.glm.b4, var2.lme4.b4, var2.hglm.b4, var2.inla.b4,
            NA, var(estglmer$var_id, na.rm=T), var(esthglm$var_id, na.rm=T),
```

```
            var(estinla$var_id, na.rm=T))
sder   <- c(se.glm.b0, se.lme4.b0, se.hglm.b0, se.inla.b0,
            se.glm.b4, se.lme4.b4, se.hglm.b4, se.inla.b4,
            rep(NA, 2), se.bias.hglm.var_id, se.bias.inla.var_id)
ic95l  <- c(CI.se.glm.b0[1], CI.se.lme4.b0[1], CI.se.hglm.b0[1], CI.se.inla.b0[1],
            CI.se.glm.b4[1], CI.se.lme4.b4[1], CI.se.hglm.b4[1], CI.se.inla.b4[1],
            rep(NA, 2), CI.hglm.var_id[1], CI.inla.var_id[1])
ic95r  <- c(CI.se.glm.b0[2], CI.se.lme4.b0[2], CI.se.hglm.b0[2], CI.se.inla.b0[2],
            CI.se.glm.b4[2], CI.se.lme4.b4[2], CI.se.hglm.b4[2], CI.se.inla.b4[2],
            rep(NA, 2), CI.hglm.var_id[2], CI.inla.var_id[2])
cove <- c(coverBet0glm, coverBet0lmer, coverBet0hglm, coverBet0inla,
          coverBet4glm, coverBet4lmer, coverBet4hglm, coverBet4inla, rep(NA, 4))
results <- data.frame(Bias=round(bias, 3), MSE=round(MSE, 3), Var=round(Var, 3),
                      Std.Error=round(sder, 3), "SE.CI95%Lower"=round(ic95l, 3),
                      "SE.CI95%Upper"=round(ic95r, 3), Coverage=round(cove, 2))
rownames(results) <- method
SimComparison1to8[[s]][[2]] <- results

# Cleaning up
rm(results, method, bias, MSE, Var, sder, ic95l, ic95r, cove)
rm(bias.glm.b0, bias.lme4.b0, bias.hglm.b0, bias.inla.b0, bias.glm.b4,
   bias.lme4.b4, bias.hglm.b4, bias.inla.b4, bias.lme4.var_id,
   bias.hglm.var_id, bias.inla.var_id)
rm(MSE.glm.b0, MSE.lme4.b0, MSE.hglm.b0, MSE.inla.b0, MSE.glm.b4, MSE.lme4.b4,
   MSE.hglm.b4, MSE.inla.b4, MSE.lme4.var_id, MSE.hglm.var_id, MSE.inla.var_id)
rm(var1.glm.b0, var1.lme4.b0, var1.hglm.b0, var1.inla.b0, var1.glm.b4,
   var1.lme4.b4, var1.hglm.b4, var1.inla.b4)
rm(var2.glm.b0, var2.lme4.b0, var2.hglm.b0, var2.inla.b0, var2.glm.b4,
   var2.lme4.b4, var2.hglm.b4, var2.inla.b4)
rm(se.glm.b0, se.lme4.b0, se.hglm.b0, se.inla.b0, se.glm.b4, se.lme4.b4,
   se.hglm.b4, se.inla.b4, se.bias.hglm.var_id, se.bias.inla.var_id,
   se.hglm.var_id, se.lme4.var_id)
rm(CI.se.glm.b0, CI.se.lme4.b0, CI.se.hglm.b0, CI.se.inla.b0, CI.se.glm.b4,
   CI.se.lme4.b4, CI.se.hglm.b4, CI.se.inla.b4, CI.inla.var_id, CI.hglm.var_id,
   CI.lme4.var_id)
rm(coverBet0glm, coverBet0lmer, coverBet0hglm, coverBet0inla, coverBet4glm,
   coverBet4lmer, coverBet4hglm, coverBet4inla)


### Data frame with computation times
comp.time  <- c(mean.time.glm, mean.time.glmer, mean.time.hglm, mean.time.inla)
ctime.df <- data.frame(ElapsedTime=round(comp.time, 2))
rownames(ctime.df) <- package
SimComparison940[[s]][[3]] <- ctime.df

# Cleaning up
rm(comp.time, ctime.df)
```

```
  rm(mean.time.glm, mean.time.glmer, mean.time.hglm, mean.time.inla)
  rm(alldata, nco, nw, b0, sigZu, estglm, estglmer, esthglm, estinla, package)

  list.comparison <- SimComparison940[[s]]
  list.details    <- SimDetails940[[s]]
  save(list.comparison, list.details, file=paste0("SimulationResults", s+8, ".RData")
  rm(list.comparison, list.details)
}

rm(s, b4, nrep, simfun2)
comment(SimDetails940) <-
  'List for all 32 simulations with detailed results of all packages'
comment(SimComparison940) <-
  'List for all 32 simulations with comparison criteria'
save(SimSettings940, SimDetails940, SimComparison940,
  file='SimulationResults9to40.RData')
```

3.5 Article 3: Incidence of Injuries in Traditional Wrestling and Associated Factors. (Sota revisió a la revista *American Journal of Sports Medicine*)

En aquest treball s'han analitzat les dades de lluita però amb més temporades (2005–2013), i les següents pagines mostren el treball enviat (està sota revisió actualment) a la revista *American Journal of Sports Medicine*, revista situada al primer quartil i amb factor d'impacte 4.362 (JCR, 2014).

L'objectiu de l'article ha estat avaluar els factors associats a la incidència de lesions tenint en compte covariables de risc com la categoria del pes del lluitador, el fet de ser guanyador o no, l'edat, i entre d'altres. Existeixen pocs estudis en la literatura científica internacional sobre la incidència de lesions en esports de lluita i els seus factors associats. El fet de ser un esport de contacte, el risc de lesió està sempre present, i des del punt de vista epidemiològic, l'objectiu ha estat conèixer la incidència de lesions i els seus possibles factors potencials, com a pas previ per a dur a terme programes de prevenció i control de lesions en aquest tipus d'esport.

Title: Incidence of Injuries in Traditional Wrestling and Associated Factors

ABSTRACT

Background: Traditional wrestling is considered to be a cultural heritage of humanity and it should be protected. The study of injuries and their associated factors can be useful to protect this heritage.

Purpose: The aim of this study was to analyse the incidence of injuries in traditional wrestling and its risk factors in order to carry out prevention and control programs in these sport.

Study Design: Prospective cohort study which collected injuries during the summer seasons from 2005 through 2013.

Methods: The incidence rates of injuries were calculated by 1000 athletic exposures (AEs) and as a function of age at the start of the season, age at initiation in wrestling, regularity in combats, winner type, and weight category. At the multivariate level, a generalized linear mixed model (GLMM) was used assuming the frequency of the injuries followed a Poisson distribution.

Results: A total of 277 wrestlers and 366 injuries were reported in 26944 AEs. The incidence was 12.5 injuries per 1000 AEs. Higher incidence among those: who did not usually participate in the league (IRR=1.47; 95% CI:1.18-1.83), receiving more falls than leading to falls (Winner type) (IRR=1.51; 95% CI:1.22- 1.87), who started as teenager (IRR=1.51; 95% CI: 1.18-1.54). Non-winner type wrestlers are at much higher risk of injuries than winner type wrestlers in the semi heavy (IRR:1.89) and heavy weight categories (IRR: 1.80).

Conclusion: The incidence of injuries in this traditional wrestling is consistent with that incidence expected in combat sports. The non-regularity and late start in their practice

are risk factors for the incidence of injuries. The technical quality is particularly relevant at heavy weights category.

Key terms: Traditional wrestling; Incidence, Injuries, Associated Factors.

What is known about the subject: Combat sports have a high incidence of injuries. In traditional wrestling, the number of combats, the number of competitions, and the internal or external wrestlers factors may contribute to these injuries, many of them preventable.

What this study adds to existing knowledge: Our results show that in traditional wrestling, a late start in the sport or not regularity in the competitions can be associated with a greater number of injuries. Moreover, in the higher weight categories, it seems that the technical quality can be an important aspect.

INTRODUCTION

Traditional wrestling is considered an intangible cultural heritage that should be recognized and protected.[46] One of the forms of wrestling that has the longest history despite not having the status of an Olympic sport is belt wrestling.[27,40] Among the varieties of this type of sport, Leonese wrestling or *aluche* holds a prominent place, being officially recognized by United World Wrestling (UWW), the European Traditional Wrestling Association (AELT) and the International Belt Wrestling Association (IBWA).[17] As a combat sport, it is naturally not free from injuries, many of which might be preventable.[2] A study of the incidence of such lesions and the factors behind them in this form of fighting sport might be of considerable use for a number of reasons like develop effective preventive measures,[28] avoid premature retirement of

wrestlers increasing the number of participants and encouraging the practice of a physical activity good for health, especially among the young.[26,31,42,48] Besides the various forms of belt wrestling share an internal logic and a set of technical and tactical actions that are quite similar from one to another,[40] hence this may serve to develop preventive strategies for other similar types of wrestling including judo. Finally, the present study has as it aim in compliance with the mandate from the International Olympic Committee (IOC) to prevent injuries so as to encourage participating in safe sport,[15,29] would make a significant contribution to protecting and perpetuating a cultural heritage such as is constituted by engaging in traditional wrestling. With this in view, the present study has as its aim an analysis of the incidence of injuries and the factors associated with them in one of the traditional forms of belt wrestling, Leonese wrestling.

MATERIAL AND METHODS

Study Design: Observational, prospective cohorts.

Period: Official competitions (known as *corros*) for seniors (16 years of age or older) in the male summer league between 2005 - 2013.

Leonese Wrestling Rules: Leonese wrestling is a combat sport in which two participants, with a set hold on their opponent's leather belt, attempt to throw their opponent over by means of a series of Leonese wrestling skills and techniques.[31,37]

The winner is the person who after a fixed period of combat has gained the higher score or the person who first achieves two falls or four points. The way of scoring depends on the type of fall.[44]

Injury Criteria: Injuries were defined as any action arising in a fight that through harm done to a wrestler that prevented the combat from continuing or required it to be halted

for medical attention to be provided, which for precautionary reasons excluded making any efforts to train or compete in other bouts or similar activities for at least the following twenty-four hours.[23]

Injury Classification: Slight lesions were those which required less than a week for recovery. Moderate injuries involved over one week but not more than four to get over, while serious injuries took four or more weeks for recovery.[18]

Athlete-Exposure: An athlete-exposure (AE) was defined by the Injury Surveillance System (ISS) as one athlete participating in one competition in which there was exposure to the possibility of athletic injury.[13]

Data Collection and Injury Report Form: Information was obtained directly from competitions on the basis of statements by the medical services covering them, of reports and assistance from friendly societies providing accident insurance, and of personal interviews at the end of each season.

The data presented an unbalanced study design with repeated measures, given that not all wrestlers were observed for the same number of seasons, and that the number of matches per season varied from one wrestler to another. The possible risk factors for injuries considered were: age of wrestler, regularity, winner type, and weight category (light, medium, semi-heavy, and heavy). The binary variable "regularity" referred to regular participation in *corros*. Wrestlers were considered regular when they participated in at least two-thirds of the *corros* in each season which were open for their participation. While wrestlers were incapacitated by injuries, no account was taken of the *corros* in which they did not participate in calculating this two-thirds ratio. The variable "winner type" was defined as a function of the falls during a match and set to "yes" if the wrestler had more falls in favour than against; otherwise, the value of winner type was set to "no".

Statistical Analysis:

A descriptive study of all variables of interest was carried out: in the case of categorical variables, absolute and relative frequencies are presented, and in the case of numeric variables, measures of central tendency (mean and median) and of statistical dispersion (standard deviation, inter-quartile range, and range) were calculated. The incidence rates for injuries were calculated in terms of cases per 1000AEs both for the entire population and as a function of age at the start of the season, age at initiation into Leonese wrestling, regularity, winner type, and weight category, respectively. To study the possible risk factors for the incidence of injuries, at the univariate level, incidence rate ratios (IRR) were calculated together with the corresponding 95% confidence intervals. At the multivariate level, a generalized linear mixed model (GLMM) was used assuming the frequency of the injuries followed a Poisson distribution. On the basis of the works of Thiele [45], Bolker [5] and Casals [10], a list of relevant information and basic characteristics of the GLMM analysis was reported. The model expression for wrestler i in his jth season is the following: $\log(yij) = \log(\lambda ij) + \beta' Xij + ui$, where y is the number of injuries, λ is the number of AEs, which is the offset of this model, and X includes all independent variables of interest. The vector β contains the fixed effects, whereas ui is the random effect corresponding to wrestler i. The random effects are assumed to be independent and normally distributed: $ui \sim N(0;\sigma 2)$, where σ_u^2 is the variance of random effect. The model accounted for repeated measures and the fact that the values of X could change from one year to the next.

A model simplification was performed by backward selection of variables from the full model, and models were compared using the likelihood ratio test (LRT) until a minimal adequate model was obtained. Model selection was based on the Akaike Information Criterion (AIC). [12] There are several ways to approximate the likelihood function used

to estimate the model parameters. Here, the Gauss-Hermite quadrature (GHQ) was used, which presents advantages over other methods, since there was only one random effect. [5,6] The statistical significance of the fixed effects associated with the covariates included in the model was assessed using the Wald test. In addition to the variables considered (age, weight category, regularity, winner type), potential second-order interactions were also considered and kept in the model if they were statistically significant.

A possible over-dispersion in the model was studied using Pearson's dispersion parameter, using values above 1.5 as the criterion for over-dispersion. All statistical analyses were performed with the statistical package R (The R Foundation for Statistical Computing, Vienna, Austria), version 3.1.1. In particular, the R package lme4 (Bates & Maechler, 2014) was used to fit the GLMMs. [3] Statistical significance was set at $p < 0.05$.

Confidentiality and Ethics Approval: Informed consent was requested and obtained from all the wrestlers to gain access to information about their injuries.

RESULTS

Wrestlers: The total number of wrestlers taking part in the official summer Leonese wrestling competitions during the seasons studied was 277. The number of wrestlers participating varied from 97 in the season with the smallest contingent to 107 in the season with the largest.

As may be seen from Table 1, wrestlers were mainly young, most being aged under thirty. They had generally started wrestling before reaching adolescence. Participation in the sport was not a regular habit for most of the wrestlers, who did not take part in

even half of the seasons considered and did not attend even half the official *corros* in each season. Three-quarters of the wrestlers did not complete as many as two bouts in each *corro* and an even higher proportion had a negative falls average.

Incidence of Injuries: A total of 336 lesions were recorded. Of these, 80 were serious, that is approximately one out of every five injuries, while 105 were moderate, one out of every three, and 151 slight, one out of two. During the same period a total of 26,944 AEs were noted (13,472 bouts). Hence, the incidence of lesions per thousand AEs was 3.9 serious, 6.9 moderate and serious taken together, and 12.5 was the figure for all types of injury.

As can be observed from Table 2, the incidences of injuries were higher in the 21 to 25 age group (13.6), in those who started *aluche* wrestling as adolescents (15.8), in those with a negative average fall ratio (15.2), those who fought in under 66% of the *corros* each season (16.0) and those in the heavy weight category (14.9). The highest incidence corresponded to wrestlers in the heavy weight category with a negative average fall differential (21.8).

Incidence of Injuries, Multivariate Analysis: Table 3 shows the generalized linear mixed model for all types of injuries, which includes all variables of interest. According to this model, the variables associated with the incidence of injuries are weight category, regularity and winner type. In addition, the model includes the interaction between weight category and winner type.

In order to facilitate interpretation, Table 4 gives an account of the model parameters and their interaction in terms of incidence rate ratios (IRR). Adjusted for the remaining variables in the model, there are higher incidence rates for injuries (even though not statistically significant at a 0.05 level) among non-regular wrestlers (compared with regular wrestlers; IRR: 1.231; 95% CI: [0.95, 1.594]) and among wrestlers who started

their involvement with Leonese wrestling as teenagers (IRR: 1.215; 95% CI: [0.926, 1.592]). The statistical interaction between winner type and weight category implies that non-winner type wrestlers are at much higher risk of injuries than winner type wrestlers in the semi-heavy (IRR: 1.887; 95% CI: [1.085, 3.281]) and heavy weight categories (IRR: 1.798; 95% CI: [1.024, 3.158]), whereas the differences are not statistically significant in the light (IRR: 0.809; 95% CI: [0.505, 1.296]) and medium weight categories (IRR: 1.26; 95% CI: [0.819, 1.937]), as is shown in Figure 1.

Incidence of Moderate and Severe Injuries: In respect of the incidence of serious and moderate lesions, it may be seen from Table 5 that they were more frequent in the 21 to 25 age group (7.8), in those who started *aluche* wresting as adolescents (9.0), in those with negative average fall differentials (8.4), in those who did not fight in at least 66% of the *corros* each season (9.4) and in the light weight category (7.9). The highest incidence was found among wrestlers in the heavy category with negative average fall ratios (10.2).

Considering only moderate and severe injuries, the model obtained was as shown in Table 6. According to this, the incidence of severe and moderate injuries was associated with winner type and age at the start of the season.

It can be seen that, when adjustment is made for the remaining variables, non-winner type wrestlers were at greater risk than winner type wrestlers (IRR: 1.549; 95% CI: [1.063, 2.258]). In addition, wrestlers older than 20 years at the start of the season were at greater risk of moderate and severe injuries than those aged 20 or under.

DISCUSSION

There are few articles on the incidence of lesions in traditional wrestling, especially in belt or arm wrestling. Analytic studies assessing risk factors by means of a standard

methodology are even less common, and it would appear that there are none that have run over as long a period as nine years. The present paper covers the incidence of injuries in one traditional form of wrestling, using a standard method based on the number of lesions per 1,000 AEs. This renders it easier to make comparisons with other combat sports and even with other sports disciplines.[13]

Multilevel data structures arise in longitudinal studies where measurements are clustered within individuals (wrestlers in our study). The statistical methods commonly used in medical literature focus on generalized linear models (GLMs) such as counts or proportions. In sports sciences and epidemiology, the data are often shown repeatedly. Even though sports injuries are often recurrent, only a small number of studies consider the correlation structure of recurrent events. Future studies should consider recurrent injuries and apply the appropriate model. Ignoring the correlation of observations between wrestlers may, for example, lead to an underestimation of the standard error. For this reason, sports researchers have started to use statistical models such as GLMM that take into account the heterogeneity between teams.[8,11,39]

The incidence of injuries in this traditional form of wrestling is consistent with the incidence expected in combat sports[1,48] and hence could be reduced. Lack of regularity and starting late in the practice of this sport are risk factors for the incidence of injuries. Technical quality is particularly relevant for the heavy weight category.

The practice of Leonese wrestling, just like other combat sports or indeed sports in general, has obvious positive effects on health, helps participants gain physical and mental strength, teaches self-discipline, moulds the character and increases self-esteem.[7,20] However, sports also have negative effects upon the health of participants to which attention should be paid, particularly in combat sports, which show higher injury rates as compared with other sports disciplines.[9,18,32] Among Olympic sports and during

Olympic Games, combat sports such as judo and wrestling lie in the intermediate to high zone in respect of the proportion of participants injured, whilst taekwondo comes either first or second.[16,41]

The technical and competitive characteristics of Leonese wrestling mean that it is most like judo and freestyle wrestling, in which incidences per exposure are very variable. In American universities and high schools the injury rate in wrestling runs from 2.3 to 9.6 lesions per 1,000 AEs, with a clustering in the range 7.3 to 9.6 per 1,000 AEs, that is, an incidence that is slightly lower than the rate observed in the present study.[23] However, as compared with judo championships, incidences in *aluche* are lower, since for judo the rates range between 25.2 and 72.1 lesions per 1,000 AEs.[19,22,33,34] Leonese wrestling thus had incidences a little higher than in scholastic wrestling but lower than in judo, and hence within the expected range. The rate found was lower than that reported for this same sport in the league for teams, where over the course of seven seasons an incidence of 18.1 per 1,000 AEs was recorded. Nevertheless, it may be considered a high rate, especially in a context of amateur competition in which the frequency of injuries ought to be lower than in sports that are more professional or that require greater dedication, where competition is very strong.[21] Thus, the incidences found can be seen as high and it is likely that an appreciable number of injuries might be avoidable, either through modifications in the rules, or the technical, physical and psychological preparation of wrestlers, or both.[31,35,41]

Practising Leonese wrestling, as is true for any other sport, but particularly combat sports, demands an appropriate physical, technical and psychological state, especially during competitions.[4,24,30,36,43] The results of this study support the hypothesis that sporadic participation and techniques used are factors to be kept in mind when strategies are being designed for preventing lesions in this sport, especially in the heavier weight

categories. The result is that these wrestlers suffer between 25% and 30% more injuries than those who compete throughout the season and are more accustomed to competition. Likewise, participants who have fewer technical skills and are less physically fit, that is, those with a negative average falls differential, show a higher incidence of lesions, especially in heavier weight categories, with an injury rate varying from 50% to 80% higher. Other authors have not observed a higher incidence of lesions in the heavier weight categories of other combat sports, for instance Green et al. for judo[19] or Jarret et al. for wrestling.[23] Nonetheless, it would appear that the higher the weight, the greater the accumulated energy exerted by wrestlers in colliding or twisting, so that the likelihood of an injury should be greater. It may be that the use of interaction techniques is what has allowed this association to be detected.

Another factor associated with injuries is age. In this study the incidence of severe and moderate lesions showed a climb as age rose, the risk of injury being 80% greater in wrestlers aged over 25 as compared with those under 20. It is a constant that the greater one's age the higher the chances of suffering an injury,[21] although other authors have discovered that in combat sports in particular, it is the younger participants that are more often hurt, so that special attention should be paid to the prevention of lesions in the youngest groups.[38,48] As traditional forms of wrestling sometimes have a playful or leisure nature, this would be consistent with a higher injury rate among older wrestlers.

There is also a large group of wrestlers who enter *aluche* wrestling at ages above 14 years, when it is harder to learn and assimilate technical movements. Hence, it is not surprising that they run a greater risk of injury, in this instance a likelihood of suffering a lesion that is around 20% higher. Whilst results found in the literature do speak of greater risks of injury when sports are begun at a young age, the items in question refer

to elite sportspeople who make great efforts and have heavy training loads, which is not the case for Leonese *aluche* or indeed other traditional wrestling styles.[25]

Nevertheless, it is not just the risk factors noted above that are relevant to the prevention of injuries, as the large numbers of *corros* and bouts also to some extent explain the results observed. It is obvious that the greater the exposure, the greater the danger of getting hurt. Unlike what happens in combat sports that are not traditional, where participants commonly give over a considerable portion of their time to training, but competitions are few and short, in Leonese wrestling this is not the situation for the majority of those taking part.

The greatest number of injuries occurs among wrestlers who fight regularly and have good techniques. This is also true of winners, as one-fourth of the wrestlers have a neutral or positive falls differential and yet a total of the order of 40% of injuries. Hence, when attempts are made to reduce the injury rate, account must also be taken of the number of bouts, and thus of exposures. The total of competitions is very large and concentrated in time, so that rationalizing the quantity and frequency of contests would reduce the number of injuries, especially if complemented by an increase in the amount of training intended to improve wrestler's technical, physical and psychological status.[14,23,47]

The injury rate in Leonese wrestling lies within the range expected in combat sports. However, it is still high and could be lessened. There should be a reduction in the number of wrestlers participating only sporadically, and improvements in the technical, physical and psychological preparation of wrestlers, so that only wrestlers demonstrating clear mastery of *aluche* techniques and a high level of physical fitness would be allowed to compete. The number of competition should be cut back, whilst training and preparation sessions should be increased.

299

300 REFERENCES

301 1. Armed Forces Health Surveillance Center (AFHSC). Injuries associated with combat
302 sports, active component, U.S. Armed Forces, 2010-2013. *MSMR.* 2014;21(5):16-8.
303 2. Ayan C, Molina A, Garcia H, et al. Rules modification effect's in incidence of
304 injuries in lucha leonesa (leonesa wrestling). *Apunts: Medicina de l'esport*.
305 2010;45(165):17-22.
306 3. Bates D, Maechler M, Bolker B, Walker S. Ime4: Linear mixed-effects models using
307 Eigen and S4. R package version 1.1-7. 2014.
308 4. Blasco M, Lopez C. *Incidencia de lesiones en Lucha Leonesa y factores asociados*.
309 León: Universidad de León; 2013.
310 5. Bolker B, Brooks M, Clark C, et al. Generalized linear mixed models: a practical
311 guide for ecology and evolution. *Trends Ecol Evol*. 2009;24(9):127-35.
312 6. Bolker B. *Ecological models and data in R*. Princeton: Princeton University Press;
313 2008.
314 7. Bu B, Haijum H, Yong L, Chaohui Z, Xiaoyuan Y, Singh M. Effects of martial arts
315 on health status: a systematic review. *J Evid Based Med*. 2010;3(4):205-19.
316 8. Bullock N, Hopkins W. Methods for tracking athletes' competitive performance in
317 skeleton. *J Sports Sci.* 2009;27(9):937-40.
318 9. Caine D, Young K, Howe W. Wrestling. In: Caine D, Harmer P, Schiff M, ed.
319 *Epidemiology of Injury in Olympic Sports*. Oxford; 2009.
320 10. Casals M, Girabent-Farre's M, Carrasco J. Methodological quality and reporting of
321 generalized linear mixed models in clinical medicine (2000-2012): a systematic review.
322 *PLoS One*. 2014;9(11).

11. Casals M, Martinez J. Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sport.* 2013;13:64-82.

12. Crawley M. *The R book.* Chichester: J.Wiley; 2007.

13. Dick R, Agel J, Marchall S. National Collegiate Athletic Association Injury Surveillance System Commentaries: Introduction and Methods. *J Athl Train.* 2007;42(2):173-82.

14. Ekstrand J, Gillquist J, Möller M, Oberg B, Liljedahl S. Incidence of soccer injuries and their relation to training and team success. *Am J Sports Med.* 1983;11:63-7.

15. Engebretsen L, Bahr R, Cook J, et al. The IOC Centres of Excellence bring prevention to Sports Medicine. *Br J Sports Med.* 2014;48(17):1270-5.

16. Engebretsen L, Soligard T, Steffen K, et al. Sports injuries and illnesses during the London Summer Olympic Games 2012. *Br J Sports Med.* 2013;47(7):407-14.

17. Federación Territorial de Castilla y León. Estatutos de la Federación de Lucha de Castilla y León. http://www.luchaleonesa.es/fcllftp/estatutos_federacion_lucha_cyl_a.pdf Published October 31, 2007. Accessed June 6, 2015.

18. Fuller C, Ekstrand J, Junge A, et al. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Scand J Med Sci Sports*. 2006;16(2):83-92.

19. Green C, Petrou M, Fogarty-Hover M, Rolf C. Injuries among judokas during competition. *Scand J Med Sci Sport*. 2007;17:205-10.

20. Hefferon K, Mutrie N. Physical Activity as a 'Stellar' Positive Psychology Intervention. In: Acevedo E, ed. *The Oxford Handbook of Exercise Pychology.* Oxford University Press, Inc; 2012.

21. Herrero H, Salinero J, Del Coso J. Injuries among Spanish male amateur soccer players: a retrospective population study. *Am J Sports Med*. 2014;42:78-85.

22. James G, Pieter W. Injury rates in elite judokas. *Biol Sport*. 2003;20:25-32.

23. Jarret G, Orwin J, Dick R. Injuries in collegiate wrestling. *Am J Sports Med*. 1998;26(5):674-80.

24. Johnson M, Yesalis C. Wrestling: Strength training and conditioning for wrestling: the Iowa approach. *Strength and Conditioning Journal*. 1986;8:56-61.

25. Kox L, Kuijer P, Kerkhoffs G, Maas M, Frings-Dresen M. Prevalence, incidence and risk factors for overuse of the wrist in young athletes: a systematic review. *Br J Sports Med*. 2015.

26. Lam K, Snyder Valier A, Valovich McLeod T. Injury and treatment characteristics of sport-specific injuries sustained in interscholastic athletics: a report from the athletic training practice-based research network. *Sport Health*. 2015;7(1):67-74.

27. Levine E. The Wrestling-Belt Legacy in the New Testament. *New Testament Study*. 1982;28(4):556-4.

28. Lin Z, Chen Y, Chia F, Wu H, Lan L, Lin J. Episodes of Injuries and Frequent Usage of Traditional Chinese Medicine for Taiwanese Elite Wrestling Athletes. *Am J Chin Med*. 2011;39(2):233-41.

29. Ljungqvist A. Sports injury prevention: a key mandate for the IOC. *Br J Sports Med*. 2008;42(6):391.

30. López C. *El entrenamiento en los deportes de lucha*. Federación Territorial de Lucha, ed. León;2000.

31. Martin V, Fernandez V, Ayan C, et al. A success story: New rules and fewer injuries in traditional Leonese Wrestling (2006-2012). *Apunts: Medicina de l'esport*. 2013;48(178):55-61.

32. Parkkari J, Kannus P, Natri A, et al. Active living and injuy risk. *Int J Sports Med*. 2004;25(3):209-16.

33. Pieter W, De Crée C. *Competition injuries in young and adult judo athletes*. In: Second Annual Congress of the European College of Sport Science. Copenhagen, Denmark: Springer; 1997.

34. Pieter W, Talbot C, Pinlac V, Bercades L. Injuries at the Konica Asian Judo Championship. *Acta Kines Universit Tartuensis.* 2001;6:102-11.

35. Pocecco E, Ruedl G, Stankovic N, et al. Injuries in judo: a systematic literature review including suggestions for prevention. *Br J Sports Med*. 2013;47(18):1139-1143.

36. Powell K, Paluch A, Blair S. Physical activity for health; What kind? How much? How intense? On top of what?. *Annu Rev Public Health*. 2011;32:349-65.

37. Robles J. *Análisis y enseñanza de los gestos técnicos de la Lucha Leonesa: las mañas de los Aluches*. León: Celarayn; 2007.

38. Salanne S, Zelmat B, Rekhroukh H, Claudet I. Judo injuries in children. *Arch Pediatr*. 2010;17:211-8.

39. Sampaio J, Drinkwater E, Leite N. Effects of season period, team quality, and playing time on basketball players' game-related statistics. *Eur J Sport Sci*. 2010;10:141-9.

40. Sayenga D. The Problem of Wrestling 'Styles' in the Modern Olympic Games - a Failure of Olympic Philosophy. *Journal of Olympic History*. 1995;3(3):19-30.

41. Shadgan B, Feldman B, Jafari S. Wrestling injuries during the 2008 Beijing Olympic Games. *Am J Sports Med.* 2010;38(9):1870-6.

42. Sokolowski M, Kaiser A, Czerniak U, Tomczak M, Breczewski G. Wrestlers' health – biological, behavioural and axiological aspects. *Arch Budo.* 2012;8(1):37-43.

43. Starosta W. Kinesthetic Sense and Awareness in Wrestling: The Structure, Conditions and Development of an “Opponent's Feeling”. *International Journal of Wrestling Science.* 2013;3(2):29-50.

44. Territorial Federation of Castilla and Leon of Wrestling. *Sport regulations of Leonese Wrestling*. http://luchaleonesa.es/fcllftp/reglamento_lucha%20_leonesa.pdf Published 2005. Accessed June 1, 2015.

45. Thiele J, Markussen B. Potential of GLMM in modelling invasive spread. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources. 2012;7(16).

46. United Nation Educational, Scientific and Cultural Organization (UNESCO). http://unesdoc.unesco.org/images/0014/001464/146400E.pdf Published 2006. Accessed June 1, 2015.

47. van Mechelen W, Hlobil H, C.G. Kemper H. Incidence, Severity, Aetiology and Prevention of Sports Injuries. *Sports Medicine*. 1992;14(2):82-99.

48. Yard E, Collins C, Dick R, Comstock R. An Epidemiologic Comparison of High School and College Wrestling Injuries. *Am J Sports Med*. 2008;36(1):57-64.

421 **LEGENDS**

422 **Figure 1**. Incidence Rates of Injuries (in injuries per 1,000 AEs) According to Winner

423 Type and Weight Category.

Table 1. Characteristics of Wrestlers Relative to Participation in Competitions.

	Minimum	P25	Median	P75	Maximum	Mean	SD
Seasons	1	1	2	6	9	3.4	2.7
Corros	1	5	19	70	277	49	64.3
Corros/season	1	3	10	17	31	10.9	8.4
Bouts	1	7	27	109	929	97.3	161.9
Bouts/season	1	4	14	26	103	19.0	19.1
Bouts/*corros*	1	1.11	1.41	2.0	4	1.63	0.66
Falls in favour	0	2.5	14.5	86.5	1920	108	248
Falls against	0	11	40	154	710	108	142
Falls difference	-376.5	-43	-13	-2	1700.5	0	177.5
Falls diff. /bout	-2.25	-1.53	-0.88	-0.31	1.83	-0.82	0.86
Starting Age	4	9	12	16	54	13.6	6.8
Mid-season age	16	20	23	28	54	24.7	6.6

Table 2. Incidence rates of injuries (per 1,000 AEs) According to Several Variables.

Variables		Injuries	AEs	Rate	95% Confidence Interval
Age	≤ 20	82	6851	12.0	[9.5-14.9]
	21 to 25	125	9180	13.6	[11.3-16.2]
	26 to 30	78	6434	12.1	[9.6-1.5]
	> 30	51	4373	11.7	[8.7-15.3]
Starting Age	Teenager	128	8107	15.8	[13.2-18.8]
	Child	208	18731	11.1	[9.7-12.7]
Fall Winner	Yes	144	14286	10.1	[8.5-11.9]
	No	192	12658	15.2	[13.1-17.5]
≥ 66% *corros*	Yes	203	18636	11.0	[9.5-12.5]
	No	133	8308	16.0	[13.4-19.0]
Category	Light	90	7093	12.7	[10.2-15.6]
	Medium	101	8766	11.5	[9.4-14.0]
	Semi-Heavy	76	6454	11.8	[9.3-14.7]
	Heavy	69	4631	14.9	[11.6-18.9]
Category Fall Winner (Yes/no)	Light Yes	49	3701	13.2	[9.8-17.5]
	Light No	41	3392	12.1	[8.7-16.4]
	Medium Yes	46	4763	9.7	[7.1-12.9]
	Medium No	55	4003	13.7	[10.3-17.9]
	Semi-Heavy Yes	23	3161	7.3	[4.6-10.9]
	Semi-Heavy No	53	3293	16.1	[12.1-21.1]
	Heavy Yes	26	2661	9.8	[6.4-14.3]
	Heavy No	43	1970	21.8	[15.8-29.4]

Table 3. Estimates for the Parameters in the Generalized Linear Mixed Model for All Types of Injuries. This includes winner type, weight category, and the interaction of these two, as well as regularity.

Coefficients	Estimate	SE	p-value
Intercept	-4.922	0.198	< 0.001
Regularity (Ref.: Yes)	0.214	0.132	0.106
Initiation (Ref.: Teenager)	-0.192	0.137	0.163
Weight category (Ref.: Light)			
Medium	-0.289	0.239	0.228
Semi-Heavy	-0.515	0.295	0.081
Heavy	-0.281	0.285	0.325
Winner type (Ref.: Yes)	-0.209	0.239	0.382
Interaction terms (Winner type × Weight)			
No × Medium	0.439	0.317	0.166
No × Semi heavy	0.846	0.363	0.020
No × Heavy	0.798	0.366	0.029
Variance of random effect	0.093		
Dispersion parameter	1.390		

Table 4. Incidence Rate Ratio (IRR) for the Variables Associated with the Incidence of Injuries.

Coefficients	**IRR**	**95% - CI**
Regularity: No vs. Yes	1.231	[0.950, 1.594]
Initiation: Teenager vs. Child	1.215	[0.926, 1.592]
Winner type: No vs. Yes per weight category		
Light	0.809	[0.505, 1.296]
Medium	1.260	[0.819, 1.937]
Semi-Heavy	1.887	[1.085, 3.281]
Heavy	1.798	[1.024, 3.158]

Table 5. Incidence rates of Severe and Moderate Injuries (per 1,000 AEs) According to Several Variables.

Variables		Injuries	AEs	I. rate	95% Confidence Interval
Age	<=20	39	6851	5.7	[4.0-7.8]
	21-25	72	9180	7.8	[6.1-9.9]
	26-30	45	6434	7.0	[5.1-9.4]
	>30	29	4373	6.6	[4.4-9.5]
Starting Age	Teenager	73	8107	9.0	[7.1-11.3]
	Child	112	18731	6.0	[4.9-7.2]
Fall Winner	Yes	79	14286	5.5	[4.4-6.9]
	No	106	12658	8.4	[6.9-10.1]
≥ 66% *corros*	Yes	107	18636	5.7	[4.7-6.9]
	No	78	8308	9.4	[7.4-11.7]
Category	Light	56	7093	7.9	[6.0-10.3]
	Medium	60	8766	6.8	[5.2-8.8]
	Semi-Heavy	39	6454	6.0	[4.3-8.3]
	Heavy	30	4631	6.5	[4.4-9.3]
Category Fall Winner (Yes/no)	Light Yes	29	3701	7.8	[5.3-11.3]
	Light No	27	3392	8.0	[5.3-11.6]
	Medium Yes	27	4763	5.7	[3.7-8.3]
	Medium No	33	4003	8.2	[5.7-11.6]
	Semi-Heavy Yes	13	3161	4.1	[2.2-7.0]
	Semi-Heavy No	26	3293	7.9	[5.2-11.6]
	Heavy Yes	10	2661	3.8	[1.8-6.9]
	Heavy No	20	1970	10.2	[6.2-15.7]

450

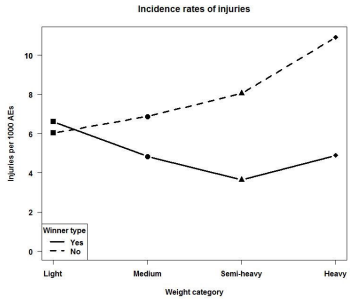
451 **Table 6:** Estimates for the Parameters in the Generalized Linear Mixed Model for

452 Moderate and Severe Injuries and Model-Based Estimation of Incidence Rate Ratios.

Coefficients	Estimate	IRR	95% CI	p-value
Intercept	-5.673			< 0.001
Regularity (Ref.: Yes)	0.266	1.305	[0.913, 1.865]	0.143
Winner type (Ref.: Yes)	0.438	1.549	[1.063, 2.258]	0.023
Weight category (Ref.: Light)				
Medium	-0.134	0.874	[0.564, 1.355]	0.548
Semi-Heavy	-0.213	0.808	[0.488, 1.336]	0.406
Heavy	-0.193	0.824	[0.475, 1.432]	0.493
Age at start of season (Ref.: $\leq$ 20)				
21 to 25	0.574	1.775	[1.162, 2.713]	0.008
26 to 30	0.616	1.852	[1.123, 3.054]	0.015
> 30	0.617	1.853	[1.025, 3.351]	0.041
Variance of random effect	0.334			
Dispersion parameter	1.430			

453

454


Incidence rates of injuries
Injuries per 1000 AEs
0
2
4
6
8
10
Winner type
Yes
No
Light
Medium
Semi-heavy
Heavy
Weight category

4

Anàlisis de la Supervivència. Frailty models

4.1 Introducció

L'anàlisi de supervivència ha esdevingut un paper important als darrers anys en els diferents àmbits d'investigació, en especial, el de la recerca mèdica. Com ja s'ha comentat al Capítol 1, un tret característic d'aquestes anàlisis és el fet de considerar la presència de la censura. La situació més comú que ens podem trobar és el de la censura per la dreta. El nom d'anàlisi de supervivència va lligat indirectament al del model de Cox o riscos proporcionals, el model de supervivència més àmpliament utilitzat. Tal com es va mencionar en el primer capítol, el model de Cox estableix la següent relació entre la funció de risc $\lambda(t|X)$ en el moment t d'un individu amb perfil $X = (X_1, \ldots, X_n)$ i la funció de risc basal en el mateix moment t:

$$\lambda(t|X) = \lambda_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n\right).$$

En aquest model se suposa que la raó entre les funcions de risc es manté constant al llarg del temps, ja que es verifica:

$$\frac{\lambda(t|X)}{\lambda_0(t)} = \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n\right),$$

on el terme de la dreta solament depèn dels valors de les covariants i no del temps t. Aquest terme correspon al ***hazard rati*** en el moment t d'un individu amb perfil X respecte d'un individu amb $X = 0$.

En l'epidemiologia de malalties infeccioses s'han descrit diferents mètodes d'anàlisis per a estudis de cohorts on s'ha tingut en compte el període d'observació com a temps de seguiment o temps calendari, especialment en el tractament del VIH [74–76]. Normalment, en els mètodes emprats en l'anàlisi de supervivència es considera la transició d'un estat inicial (viu) fins a produir-se l'esdeveniment d'interès (mort), i l'efecte de les covariables de l'evolució de la malaltia és generalment modelada utilitzant extensions del model de Cox [18, 77]. A nivell multivariant s'han utilitzat extensions com per exemple els models de cura, models de riscos competitius, models amb variables canviants en el temps, models d'esdeveniments recurrents, models de múltiples estats (*multi-state*) i els *frailty models* [78–83].

El concepte de *frailty* és una solució convenient per tractar la heterogeneïtat no observada i les associacions dins d'un model de Cox en dissenys complexos on les dades presenten una estructura jeràrquica. Una de les limitacions del model de Cox és la restricció d'independència en els temps de supervivència entre les diferents observacions. El model suposa que tots els subjectes tenen la mateixa mateixa distribució dels temps de supervivència. Això voldria dir que tots els subjectes amb valors similars en les covariables tindrien el mateix risc d'experimentar l'esdeveniment.

Per tenir en compte una possible dependència dels diferents grups/*cluster* dels temps de supervivència s'utilitzen els anomenats *frailty models*. Els *frailty models* són una extensió del model de riscos proporcionals o model de Cox, discutit per Cox [18], on s'incorporen efectes aleatoris dins del risc basal per modelar la correlació intragrup. Als darrers anys, l'ús dels

frailty models ha crescut, donada l'estructura de les dades amb mesures repetides que hi figuren per exemple en estudis longitudinals o assajos clínics. Els *frailty models* s'utilitzen quan l'esdeveniment d'interès pot afectar reiteradament al mateix individu, com en el cas d'esdeveniments recurrents, o quan el *frailty* és comú en individus que pertanyen a un grup específic. Aquests models també poden presentar-se de forma aniuada, coneguts com a *nested models*, com per exemple quan els individus d'una família tenen un *frailty* compartit, i les famílies dins d'una comunitat tenen en comú un altre *frailty*.

En general, el terme *frailty* representa per exemple a diferents grups/estrats (famílies, lliteres, districtes, diagnòstic d'unitat, centres hospitalaris, escola, les regions geogràfiques com comarca, província, municipi, districte o secció censal) i aquests actuen multiplicant sobre la funció del risc basal.

Els primers avenços en la modelització de les dades de temps fins a l'esdeveniment (anàlisi de supervivència) davant dades amb una estructura jeràrquica es varen produïr a la dècada del 1980 i a principis del 1990 [84, 85], i les implementacions del *software* estadístic que permet l'anàlisi d'aquest tipus de dades amb dos o més nivells jeràrquics van aparèixer al voltant de l'any 2000 [86]. Tot i així, els *frailty models* com passava amb els models mixtes encara no són tan fàcilment accessibles com els models GLM [87, 88].

L'anàlisi de supervivència és un camp més divers que el dels GLM, amb diferents models i mètodes d'estimació desenvolupats per a diferents situacions. El model de Cox és potser l'analogia més propera d'un GLM, però la seva funció de versemblança (parcial) no dóna analíticament les mateixes formes d'integració d'efectes aleatoris i, a més a més, els efectes aleatoris gaussians no tenen els mateixos avantatges computacionals com en els GLM. La presència de variables en el temps o predictors poden crear patrons de predicció més complexes que en els GLM, fins i tot per al model de Cox [89].

Els *frailty models* són també anomenats models de supervivència amb efectes aleatoris [90–92]. En aquest sentit un efecte aleatori és una vari-

able continua que descriu l'excés de risc o fragilitat (*frailty*) per diferents categories, com individus o grups d'individus. Els *frailty models* són doncs una extensió dels models de Cox introduint els termes *frailties* (covariables aleatòries no observables) en el model. En aquest cas, el *hazard rati* no és només una funció de les covariables, sinó també una funció dels *frailties*. El *frailty model* es pot definir en termes de risc condicional:

$$\lambda(t|Z,X) = Z\lambda_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n\right)$$

on s'afegeix al clàssic model de Cox el terme Z que correspon al *frailty*. Un punt important és que aquest terme Z és una variable aleatoria no observable que varia sobre la mostra que incrementa el risc individual si $Z > 1$ o decreix si $Z < 1$. El *frailty* actua multiplicant a la funció de risc, és a dir, sota la hipòtesi de risc proporcional. La idea intuïtiva d'aquest efecte aleatori seria que l'associació entre els temps de supervivència prové d'un factor desconegut anomenat *frailty* i no degut a que la incidencia del primer temps T_1 influeixi directament en el valor del segon temps T_2. Per exemple, que un pare es mori d'infart no influeix d'una forma directa en que el fill ho vagi a fer però sí que existeix un factor que tenen els dos, per exemple, un factor genètic que lliga el pare amb el fill.

Pel que fa a la interpretació, el hazard rati de la covariable ($\exp(\beta)$) no es modifica i té la mateixa interpretació que un model de Cox sense *frailties*.

El concepte de *frailty* va ser introduït per Vaupel [90] a la bioestadística i per Lancaster [93] a l'econometria. Clayton, va utilitzar ja el terme *frailty* des d'un altre context, per modelar associacions entre els temps de supervivència, més coneguts com *event times* [94].

El model de Cox amb efectes aleatoris segons Hua Zhao [95] és semblant al model lineal mixt, però evitant la complexitat en la notació dels *frailty models*. Això representa una extensió del model general de Cox per incloure-hi efectes aleatoris i per modelitzar el parentesc entre els individus d'un estudi. Entre els models proposats amb aquestes característiques [96–98] trobem el marc de Therneau com un dels més complerts. La funció de versemblança és similar a la funció de versemblança parcial

del model de Cox general [77,97]. Hi ha dues classes de models *frailty*, a nivell univariat amb censura per la dreta i un esdeveniment, i els models que descriuen un anàlisi de supervivència multivariat amb diferents esdeveniments i considerant diferents censures o truncaments, més coneguts com els *shared frailty models*. Els *frailty models* univariats són un cas especial dels *shared frailty* models amb mida de cluster 1 i que pot ser utilitzat per modelar l'heterogeneïtat entre els subjectes.

4.2 *Model Shared frailty*

Els *shared frailty* models s'utilitzen per modelitzar la dependència entre els temps de supervivència. Això també és conegut com un *mixture model* perquè el risc comú en cada grup o cluster (*frailty*) és assumit com aleatori. El model assumeix que tots els temps d'interès en un grup són independents donat un *frailty*. En altres paraules, és un model condicional on el *frailty* és comú per a tots els subjectes en un grup (de forma independent) i per tant és responsable de crear dependència entre els temps d'interès. Els *shared frailty models* van ser introduïts per Clayton (1978) sense utilitzar la noció *frailty* i va ser més tard extensament estudiat [77,87,99,100].

Des del punt de vista d'especificació del model, els *shared frailty models* són modelats com una extensió del clàssic model de Cox utilitzant un efecte *frailty* Z_i, $i = 1, \ldots, n$, idèntic per a tots els subjectes j, $j = 1, \ldots, n_i$ en un cluster. Els temps de supervivència en el cluster i, $1 \leq i \leq n$ són assumits independents i les seves funcions de risc es mostren com:

$$\boldsymbol{\lambda}(t|X_{ij}, Z_i) = Z_i \lambda_0(t) \exp(\beta' X_{ij})$$

on X_{ij} és el vector de les covariables del subjecte j en el ièssim *cluster*, β és el vector relacionat amb els paràmetres dels efectes fixes, i Z_i és la variable no observada (el terme *frailty*), no negativa, amb funció de

densitat $f(z)$ i s'assumeix amb mitjana 1 i variància σ^2 . El terme *frailty* Z_i representa la variabilitat addicional dels factors de risc observats.

Els *shared frailty models* poden ser models de supervivència amb un grup de variació (*frailty*) i una variació individual descrita per la funció de risc. Els models mixtes, a diferència dels models *shared frailty*, mostren un maneig més simètric d'aquestes dues fonts de variació. El fet que en els *shared frailty models* hi hagi observacions censurades fa que no pertanyin o formin part dels models mixtes. Tot i així, tenen un gran vincle, que és el fet que són models condicionals (no marginals) on diferents assumpcions sobre els efectes aleatoris (o *frailties*) poden crear diferent dependència en l'estructura dels *clusters*/grups en els *shared frailty models*. Els *shared frailty models* són doncs models que assumeixen que tots els temps de supervivència en un grup són condicionalment independents, donats els *frailties*.

Els *shared frailty models* són útils en l'anàlisi multivariant de dades de supervivència que sorgeixen quan els individus experimenten events recurrents (ex: tumors, reinfeccions, etc) o quan hi ha una agrupació d'individus, com en el cas de l'estudi de famílies. El clàssic estudi de bessons seria un exemple d'esdeveniments amb dades bivariades [88, 91, 101]. En aquest cas, per a l'estudi de famílies, el component genètic comporta un factor clau per veure l'associació d'una malaltia que hi pot haver dins d'una família. En els assajos clínics multicèntrics el fenomen de *clustering* mitjançant els *clustered survival data* són també freqüents. En aquests estudis es descriuen *frailties* com la variabilitat entre centres hospitalaris que no és explicada per covariables observades. Moltes vegades, ens trobem que s'estudien els temps de supervivència de cadascun dels pacients en els diferents centres hospitalaris. És important destacar que aquests centres sovint presenten diferències pel que fa al tractament als pacients, cures o diagnòstic dels pacients, i per tant s'ha tenir en compte aquesta correlació.

Els *shared frailty models* són una extensió dels *frailty models* univariats, en què es permet que els subjectes d'un mateix *cluster* comparteixin un mateix valor *frailty*. D'aquesta manera, el *frailty* és definit com una

mesura de risc relatiu que els subjectes comparteixen en un grup. Quan un *frailty* és *shared* o compartit, vol dir que es genera dependència entre els individus que comparteixen *frailties*. Tot i així, quan es condiciona sobre el *frailty*, els individus són independents entre sí.

El terme *frailty* pot seguir diferents distribucions com la Gamma (la més comuna), distribució Positiva Estable o més coneguda com *Positive Stable*, Log-Normal (la coneguda com Normal), la distribució dels 3 paràmetres *Compound Poisson* i la Inversa Gaussiana. Segons la distribució del *frailty*, la seva variància determinarà en principi el grau d'associació del *cluster*.

Al centrar-nos en el model semiparamètric frailty model, la funció de risc basal λ_0 s'assumeix com a desconeguda (model semi paramètric), però també pot seguir una distribució amb un paràmetre vectorial (per exemple distribucions com Weibull, Gompertz).

La funció de supervivència conjunta condicionada al frailty Z_i en què es comparteix per tots els individus en el *cluster* i, es pot especificar com:

$$\begin{aligned} S(t_{i1},\ldots,t_{in_i}|Z_i) &= S(t_{i1}|X_{i1},Z_i)\cdots S(t_{in}|X_{in},Z_i) \\ &= \exp(-Z_i\sum_{i=1}^{n_i} M_0(t_{ij})\exp(\beta' X_{ij})) \end{aligned}$$

on $M_0(t)$ és la funció de risc basal acumulada, i la X_i és la matriu de covariables dels individus en el *cluster* i.

A l'any 1984 Hougaard [102] va demostrar la importància de la transformació de Laplace per als càlculs. Fent una mitjana de l'expressió respecte el *frailty* Z_i la funció marginal de supervivència es descriu:

$$\begin{aligned} S(t_{i1},\ldots,t_{in_i}|Z_i) &= \mathrm{E}\big(S(t_{i1},\ldots,t_{in_i}|X_i,Z_i)\big) \\ &= \mathrm{E}\Big(\exp(-Z_i\sum_{j=1}^{n_i} M_0(t_{ij})\exp(\beta' X_{ij})\Big) \\ &= \mathbf{L}\Big(\sum_{j=1}^{n_i} M_0(t_{ij})\exp(\beta' X_{ij})\Big), \end{aligned}$$

on **L** representa la transformació de Laplace de la variable del *frailty*.

4.2.1 Limitacions dels *shared frailty models*

Una de les primeres limitacions seria que la dependència del paràmetre i l'heterogeneïtat de la població es determina al mateix temps, és a dir, la dependència entre els temps de supervivència dins del *cluster* es basa en distribucions marginals dels temps de supervivència. Així doncs, quan les covariables són presents en per exemple un *shared frailty model*, la dependència de la població i l'heterogeneïtat poblacional són confoses, significant que la distribució conjunta dels temps de l'esdeveniment es pot identificar des d'una distribució marginal.

Una de les principals característiques dels *shared frailty models* és la simetria. En assajos clínics multicentrics s'assumeix una relació simètrica entre tots els possibles parells de pacients en un estudi perquè els pacients d'un centre són intercanviables. No és així, en estudis famíliars, on aplicar un *shared frailty model*, implica la mateixa relació (correlació) entre els integrants de la família. Això, contradiu la relació de famílies, que es basa en estudis genètics. Per solucionar aquests problemes metodològics hi ha els *correlated frailty models*. Així doncs, els *shared frailty models* són els que tenen un terme *frailty* a nivell familiar (on es considera un mateix *frailty* per a tots els individus d'una mateixa família). En canvi, els *correlated frailty models* tenen en compte la correlació dels individus dins de cada familia, és a dir, el terme *frailty* a nivell individual (on es consideren diferents termes *frailties* per a tots els individus d'una mateixa família). Per tant, per als *correlated frailty models*, es pot explorar el component genètic.

4.2.2 *Model semiparamètric shared Gamma frailty*

La distribució Gamma és de les més conegudes i utilitzades per descriure el terme *frailty*. Aquesta distribució presenta la simplicitat de la transformació de Laplace per obtenir expressions de supervivència no

condicionals com la funció de densitat acumulada i funció de risc. La distribució *frailty* apareix en la versemblança condicional i pot ser integrada fora d'ella i per tant, dóna simples expressions de la versemblança marginal. De fet, és fàcil obtenir l'estimació de paràmetres via la maximització de la versemblança marginal, tal com s'especifica més en detall en l'apartat posterior.

La funció de densitat de la *shared gamma frailty model* ve definida per:

$$f(z) = \frac{b^{\rho} z^{\rho-1} \exp^{-bz}}{\Gamma(\rho)},$$

on $Z \sim \Gamma(\rho, b)$.

La distribució Gamma és una distribució flexible que disposa d'una varietat de formes quan p varia. Per exemple, quan $p = 1$ coincidiria amb la coneguda distribució Exponencial, o bé quan la p és gran tindria una forma de campana semblant a la distribució Normal. Segons Abbring i van den Berg [103], en alguna classe de models *frailty* la distribució del *frailty* entre els supervivents convergeix segons una distribució Gamma quan el temps tendeix a infinit.

Amb l'assumpció de que $p = b$, on dos paràmetres es transformen en un paràmetre de la distribució, aleshores:

$$E(Z) = 1, V(Z) = \sigma^2 = \frac{1}{b}, Z \sim \Gamma(b, b)$$

Tot i els avantatges que presenta la distribució gamma, no hi ha raons biològiques que la distribució Gamma sigui preferible respecte a altres distribucions [99].

4.2.3 *Model semiparamètric shared Log-Normal frailty*

Una altra distribució important en els frailty models és la Log-Normal. La seva popularitat és deguda a la flexibilitat que presenta en models multivariats amb estructura de correlació. De fet, la distribució Log-normal del *frailty* deriva principalment de la relació amb els models lineals mixtes, on s'assumeix que l'efecte aleatori (o ara en el nostre cas amb el terme

frailty) segueix una distribució Normal. El model semi paramètric Log-Normal *frailty* és una extensió del model de Cox. La forma de la funció de versemblança no té una forma explicita i s'han d'utilitzar mètodes numèrics.

L'especificació del *shared Log-Normal frailty model* és expressada:

$$\lambda_{ij}(t) = Z_i \lambda_0(t) \exp(\beta' X_{ij} + W_j)$$

on

- λ_0 és la funció de risc basal.
- $\beta' = (\beta_1, \ldots, \beta_p)$ és el vector dels coeficient regressors.
- $X_{ij} = (X_{ij1}, \ldots, X_{ijp})$
- $W_j \sim \mathcal{N}(0, s^2)$
- Els temps de l'event són condicionalment independents donat el frailty.

4.2.4 Mètodes d'estimació dels *shared frailty models*

És conegut que si λ_0 és assumit com a una funció no paramètrica, aleshores es considera el model com a model semiparamètric de riscos proporcionals, i les estimacions són sovint obtingudes mitjançant la versemblança parcial.

En el cas dels models semiparamètrics *shared frailty models* el mètode clàssic d'estimació de màxima versemblança no serà apropiat per estimar els paràmetres. Els mètodes més coneguts per estimar els paràmetres en els *shared semi paramètric frailty models* són mitjançant l'algoritme EM, el *penalized partial likelihood*(PPL) i el *Markov chain Monte Carlo* (MCMC).

L'algoritme EM té algunes limitacions, ja que, normalment la seva convergència i temps de computació és més lent que procediments de

Newton-Raphson. L'algoritme EM sovint requereix un gran nombre d'iteracions i per estimar les variàncies es requereix també més computació. Una alternativa a l'algoritme EM és el PPL on el terme *frailty* es tracta com un coeficient de regressió addicional que és construït a partir d'una funció de penalització afegida al logaritme de la versemblança [87]. Tenen doncs similituds a altres mètodes per penalitzacions de regressió com els *ridge regression*, *lasso* i *smoothing splines*. Aquest mètodes normalment convergeixen i optimitzen més ràpidament, tot i que molts cops és difícil l'obtenció vàlida d'estimacions de l'error estàndard. Un altre procediment com és el cas del mètode *Penalized Likelihood* és més fàcil d'obtenir l'estimació dels errors estàndard. Els algoritmes EM i PPL donen els mateixos resultats en el cas del Gamma *frailty model*. Per això, el procediment PPL pot utilitzar-se també per la funció de densitat Gamma [104]. En canvi, en el cas del Log-Normal *frailty model* la transformació de Laplace de la distribució és intractable.

Els diferents mètodes d'estimació són discutits amb més detall en els llibres de Therneau i Grambsch, Hougaard, Duchateau, Wienke, i Hanagal [77, 87, 88, 99, 104].

4.3 Revisió dels softwares per ajustar *frailty models*

Tal com vàrem veure en els GLMM, en els *frailty models* hi ha una gran varietat de *softwares* com SAS, MATLAB, aML, GAUSS, Stata, R, S-Plus que inclouen l'opció d'ajustar *frailty models* amb les distribucions Gamma i Log-Normal [77, 99, 105–107]. Des de la filosofia bayesiana el WinBugs permet analitzar els *shared frailty models* amb diferents distribucions pel *frailty* utilitzant el procediment de MCMC i l'INLA és un altre dels paquets que permet ajustar aquests models [108, 109].

Fins ara, el més semblant a un article de revisió general sobre *softwares* estadístics per a *frailty models* és el treball de Kelly [92] a l'any 2004, i recentment el de Hirsch i Wienke [100] a l'any 2012.

En el treball de Kelly es revisen sis *softwares* (SAS, Stata, S-Plus, R, MLwiN i WinBUGS) per als models de Cox i per a *parametric accelerated*

failure time(AFT) models. Kelly i col.laboradors conclouen que els *frailty models* que podem trobar als paquets d'R o SAS per exemple tenen implementats majoritàriament per defecte les distribucions Gamma o Normal per al terme *frailty*. En cas de voler ajustar altres distribucions és necessari implementar-ho amb grans coneixements de programació pel *software* estadístic, amb l'ajuda per exemple d'algun treball de Therneau [107] o macros en diferents *softwares* que s'han anat implementant. Per exemple, en el SAS s'ha implementat una macro per a la distribució *positive stable frailty model* [110], i en Stata o MLwin per ajustar models multinivell, fent ús de distribucions no lineals (Logistic, Poisson, Log-Normal, Weibull, Gamma i Log-Logistic). A través d'aquesta revisió de Kelly s'han realitzat estudis de simulació per a models de vida accelerada [111] on es troben resultats molt similars tan en els paquets R com en SAS, considerant això sí que l'estimació amb la versemblança per grups de mostres petites poden ocasionar estimacions distorsionades. Tot i així, en alguns estudis de simulació fent ús de distribucions Gaussian (Normal) *frailty,Gamma frailty* [112] i *Log-Normal frailty* [113] conclouen que la precisió dels *frailties* es pobre i menys flexible que en els models marginals i que encara s'ha de treballar i explorar altres possibles distribucions més adequades.

L'altre treball de revisió de *softwares* per a *frailty* models és el treball de Hirsch i Wienke que es centren concretament en el cas dels models semiparamètric *shared frailty* models per a distribucions Gamma i Log-Normal. En aquesta revisió es fa una explicació detallada de la disponibilitat dels principals llibreries d'R ($Survival$, $Kinship$, $Phmm$, $Frailtypack$) i macros de SAS ($SPGAM$, $SPLN3$) que hi ha implementats en la literatura. Hirsch i Wienke descriuen en la propera figura 4.1 les principals característiques dels procediments en l'estudi de simulació dut a terme, centrant-se en els *softwares* R i SAS [100].

En aquest estudi de simulació els autors acaben concloent que la funció $Coxph$ del paquet $Survival$ i la funció $Coxme$ del paquet $Kinship$ d'R proporcionen menys biaix en les estimacions en els *shared frailty model* amb distribució Gamma, i amb distribució Log-normal, respectivament.

Table 1 – Main features of the procedures used in the simulation study.

Function/macro	Coxph	Coxme	Phmm	FrailtyPenal	SPGAM	SPLN3
Software	R	R	R	R	SAS	SAS
Library	Survival	Kinship	Phmm	Frailtypack	–	–
Author	Therneau	Therneau	Donohue, Xu	Rondeau, Gonzalez	Vu	Vu
Version	2.35–8	1.1.0–23	0.6.3	2.2–16	–	–
Model[a]	Semiparametric shared frailty	Semiparametric shared frailty	Semiparametric shared frailty	Semiparametric[b] shared frailty, joint frailty, nested frailty	Semiparametric shared frailty	Semiparametric shared frailty
Distribution	Log-normal, t, gamma	Log-normal	Log-normal	Gamma	Gamma	Log-normal
Algorithm[c]	PPL	PPL	MCEM	ML	EM	EM
Censoring	Right, interval, Left	Right, interval, Left	Right, interval, Left	Right	Right	Right
Truncation	no	no	no	yes	no	no
Regulatory parameters	Method, eps, iter.max, outer.max	Ties, eps, inner.iter, iter.max	Maxtime, maxstep, varstart	n.knots, kappa, maxit	Maxiter, epsilon	n.epsilon, maxiter
s.e. variance	no	no	yes	yes	yes	yes

[a] Frailty models, which can be handled by library.
[b] Quasi-semiparametric model, see details about frailtypack.
[c] PPL: penalized partial likelihood; EM: expectation-maximization, MCEM: EM with MCMC, ML: maximum likelihood.

Figura 4.1: Principals característiques i procediments dels frailty models presentats en l'estudi de Hirsch i Wienke a l'any 2012

Tot i així, les dues funcions no proporcionen errors estàndard del terme *frailty* o efecte aleatori, tal com passava també en el paquet *lme*4 per als GLMM. Altres bones alternatives serien les macros de SAS (*SPGAM* i *SPLN*3) que permeten calcular l'error estàndard, o les llibreries de R (*Frailtypack* i *Phmm*) més adients per ajustar models més complexes.

4.4 Article 4: Incidence of infectious diseases and survival among the Roma population: a longitudinal cohort study. Eur J Public Health. 2011

En aquest capítol es mostra un exemple il·lustratiu on a partir d'unes dades reals comprovem que no es pot assumir algunes de els consideracions bàsiques del model de Cox i sembla més adient utilitzar un *frailty model*.

En aquest exemple es va estudiar la incidència de malalties infeccioses d'una cohort d'ètnia gitana i es va analitzar la supervivència ajustada per edat i sexe d'aquest barri marginal durant 23 anys. Es va ajustar un model semiparametric clàssic de Cox amb l'edat com a covariable (censura per la dreta) i un model de Cox utilitzant l'edat com a resposta el que implica el truncament per l'esquerra. A més a més, es va ajustar el

model de Cox amb efectes aleatoris (semiparametric Log-Normal frailty model), tenint en compte la variable família com efecte aleatori.

En aquest estudi es va concloure que tot i que el model de Cox amb l'edat actuant com a covariable produïa estimacions adequades, el model on l'edat era la resposta, l'estimació i l'interpretació era més apropiada per aquest estudi. L'ajust del *frailty model* ens va mostrar que la família a la que pertanyien els subjectes de l'estudi era un component rellevant per mesurar la supervivència i que els homes tenien pitjor supervivència. Les estimacions en el *frailty model* canvien i potser es pot parlar d'una infraestimació dels errors estandard en la covariable sexe. D'aquest estudi es va publicar un article a la revista European Journal of Public Health que a l'any de la seva publicació (2011) estava situada al segon cuartil amb factor d'impacte 2.59 (JCR, 2014).

En les properes pàgines de la memòria es presenta el treball explicat prèviament i que és citat a continuació:

Casals M, Pila P, Langohr K, Millet JP, Caylà JA and the Roma Population Working Group. Incidence of infectious diseases and survival among the Roma population: a longitudinal cohort study. Eur J Public Health. 2011.

interdisciplinary research]. In: Longo G, Morrone A, editors. *Cultura Salute Immigrazione [Culture Health Immigration]*. Rome: Armando Editore, 1995: 62–74.

15 Stronks K. Public health research among immigrant populations: still a long way to go. *Eur J Epidemiol* 2003;18:841–2.

16 Pennazza F, Boldrini R. Il ricovero ospedaliero degli stranieri in Italia nell'anno 2000 [Hospital admission of foreign citizens in Italy in 2000]. Rome (IT): Ministry of Health, 2002.

17 Geraci S. La sfida della medicina delle migrazioni [The challenge of migration medicine]In: Caritas-Migrantes, editor. *Dossier Statistico Immigrazione 2005* [*Immigration Statistical Dossier 2005*]. Rome (IT): Idos Centro Studi e Ricerche; 2005 Report No. XV: 179–88.

18 Ziglio E, Barbosa R, Charpak Y, Turner S. Health systems confront poverty 2003. Copenhagen (DK): WHO Regional Office for Europe 2003 *Public Health Case Studies* No.: 1.

19 Morrone A, Pugliese E, Sgritta GB. *Gli immigrati nella provincia di Roma - Rapporto 2005* [The immigrants in the Province of Rome - Report 2005]. Milan: Franco Angeli Editore, 2005.

20 Yip R. Iron nutritional status defined. In: Filer IJ, editor. *Dietary iron: birth to two years*. New York: Raven Press, 1989: 19–36.

European Journal of Public Health, Vol. 22, No. 2, 262–266

doi:10.1093/eurpub/ckq204 Advance Access published on 7 January 2011

Incidence of infectious diseases and survival among the Roma population: a longitudinal cohort study

Martí Casals[1,2,3,4], Pilar Pila[1], Klaus Langohr[3], Juan-Pablo Millet[1,2], Joan A. Caylà[1] and the Roma Population Working Group*

1 Epidemiology Service, Public Health Agency of Barcelona, Barcelona, Spain
2 CIBER de Epidemiología y Salud Pública (CIBERESP), Spain
3 Universitat Politècnica de Catalunya, Barcelona, Spain
4 Departament de Salut Pública, Universitat de Barcelona, Barcelona, Spain

Correspondence: Marti Casals, Pl. Lesseps, 1. 08023 Barcelona, Spain, tel: 34 932384545 (ext 376), fax: 34 932182275, e-mail: mcasals@aspb.cat

*Teresa Sorde (Sociology Department of the Universitat Autònoma de Barcelona), Angels Orcau, Patricia Garcia de Olalla, Jeanne Nelson (Epidemiology Service, Public Health Agency of Barcelona), Joan Batalla (Department of Health, Generalitat of Catalonia) and Rafa Guerrero (Department of Justice, Generalitat of Catalonia).

Background: Roma ethnicity is greatly affected by tuberculosis (TB), AIDS, injecting drugs use (IDU) and imprisonment. **Methods:** We assessed the incidence of several health problems by means of a retrospective cohort study performed in Camp de la Bota, Barcelona (Spain). The 380 individuals included in the 1985 TB outbreak investigation were followed-up until 31 December 2008. One hundred ninety-two subjects (50.5%) were men and 188 (49.5%) women. Information sources included questionnaires taken at the time of this outbreak, a population census and other registries from Barcelona and Catalonia. Cox proportional hazards mixed models were employed in the multivariate survival analysis. **Results:** By the end of the follow-up, the survival rate was 79.4%; 50 persons (13.1%) had deceased and 28 (7.3%) had emigrated. The incidence of AIDS was 104 cases per 100 000 person-years of follow-up (pyf), IDU was 240 cases pyf, imprisonment was 642 cases pyf and that of TB was 91 cases pyf. Male survival was lower [hazard ratio (HR) 4.22], when the effect of family was taken into account, than when it was not taken into account (HR 3.67). **Conclusions:** High incidences of AIDS, TB, IDU, imprisonment and poor survival rates have been observed among Roma. Family was found to be an important factor influencing the survival rates: when not considered, the risk of death among men was underestimated.

Introduction

The Roma population represent an ethnic group of 12 million individuals living in almost all European and American countries, as well as in some areas of Asia and Oceania. The precise number of Roma in the European Union is difficult to establish with accuracy. Estimates range from 3 to 7 million cited in the 2004 European Commission report to 10 million, reported in a 2008 European Parliament Resolution.[1] In Spain, the Roma community consists of people between 600 000 and 800 000.[2] Despite their poverty[3] and nomadism, the Roma have conserved part of their original culture, although they also have a history of marginalization and persecution. This is demonstrated by the slavery and murder they have suffered, notably as victims of the Holocaust.[4]

The few health studies that exist about the Roma show that their life expectancy is lower and infant mortality is higher than the general population.[5] The mean life expectancy of the Roma population in Spain was only 58 years.[3] One study reported that Roma women have a life expectancy of 6 years lower than men.[6] It was also observed that 47% of the scientific articles about the Roma population dealt with congenital malformations, paediatric diseases and transmissible diseases.[7]

In England, the health status of nomadic Roma was correlated to age, education and smoking, with higher rates of anxiety, respiratory problems, chest pain, involuntary abortions and mortality.[8,9] They are known to have poorer reproductive health than the rest of the population and a higher prevalence of infectious diseases, injuries and poisoning due to environmental causes. The premature death in this population is three times higher than in the general population.[5,10]

Socially disadvantaged groups are always greatly affected by transmissible infectious diseases.[11] In a study carried out in the Young Offenders Penitentiary in Barcelona, the highest prevalence of HIV infection was found among Roma drug addicts (70%).[12] In some settings, outbreaks of hepatitis have also been seen in families of Roma ethnicity with tattooing being a possible risk factor.[11,12] Given the poor living conditions of many ethnic minorities, tuberculosis (TB) is a concern.[13] In recent years, it has become clear that this ethnic group is not free of TB or HIV/AIDS, due to

shortcomings in prevention and control, and the presence of risk factors such as alcoholism, smoking and injecting drug use (IDU).[14–16]

Poor and deteriorated neighbourhoods can contribute to unhealthy living conditions. A TB outbreak was detected among Roma in Camp de la Bota, a marginalized neighbourhood of Barcelona in 1985, which gave rise to a contact study. A total of 20.5% had a positive tuberculin skin test and 14 TB cases (prevalence rate of 3.63%) were identified. This TB incidence was 115 times higher than the TB incidence rate of Barcelona in the same year. A total of five patients (a smear-positive father and his four sons) were living in the same shack.[17]

The objective of this study was to follow the Roma inhabitants of Camp de la Bota since the 1985 TB outbreak until 2008, analysing the incidence of infectious diseases, IDU, imprisonment and the survival in this population.

Methods

Design

A retrospective cohort study was performed.

Population

The cohort was composed of individuals identified during a TB outbreak investigation in 1985 in a Roma community called Camp de la Bota, Barcelona, Spain, as well as the new births until the settlement was removed in 1990. A tuberculin skin test was performed by public health nurses from the Barcelona TB Programme, in collaboration with nuns who attended the nursery, the community clinic and social centres in the neighbourhood. A thoracic X-ray was also performed for tuberculin skin test positive individuals. A total of 14 cases of active TB were detected (prevalence of 3.63%). The initial cohort included 427 subjects. Of these, data of birth was unknown for 27 (6.3%) and 20 (4.7%) were not known to be alive or dead at the end of the study (figure S3 of the online Supplementary Digital Content, SDC). These 47 individuals were excluded from the analyses. The cohort was followed-up from 31 December 1985 until 31 December 2008.

Information sources

The following registries were consulted: TB registry of Barcelona, communicable diseases and drug user registries of Barcelona and Catalonia, the Central Insurance Registry, the drug user register of the Catalan Autonomous Government Department of Health, the Penitentiary Services registries and the mortality registry of the Department of Health (to identify individuals who had died and their date of death).

Variables

The following variables were collected from the 1985 TB outbreak investigation: sex, date of birth, shack number and presence of TB. During follow-up, any additional names, place of residence, subsequent TB episode, HIV or AIDS diagnosis, IDU, type of drug, imprisonment, censored date and living status (date of death, if deceased) were also collected.

For data collection, we collaborated with a nun who worked in Camp de la Bota between 1974 and 1989 performing community work in the clinic as a nurse and in the nursery as a teacher. An epidemiologic questionnaire was administered for each TB case.

Statistical analyses

The mean, median and SD were calculated for quantitative variables. Frequency tables were used for qualitative variables. The incidences of TB, AIDS, IDU and imprisonment in cases per 100 000 person-years of follow-up (pyf) were calculated for the entire population as a function of gender and of city of residence. In each case, the denominator consisted of the sum of follow-up times since 1st December 1985 of the subjects under study. In case of newborns after that date, follow-up started at their birth date. Survival probabilities were calculated using the Kaplan–Meier method. For the multivariate analyses, the Cox proportional hazards model was applied to compare survival between men and women, using both the general model and the mixed effects model with family as a random effect. These survival analyses were applied to two distinct response variables. First, survival time was calculated as elapsed time since the beginning of the study, 1st December 1985. These times were potentially right-censored. Second, the life expectancy of this population was also of interest. For this purpose, age at death was modelled as the response variable; hence survival times were potentially right-censored and left-truncated (also called 'late entries').[18] In both cases, analyses were carried out taking into account these censoring/truncation schemes. We computed 95% confidence intervals (CIs) for finite populations. The proportionality of risks in the Cox models was verified graphically using a Schoenfeld residuals plot.

The statistical analyses were performed using the statistical package R (The R Foundation for Statistical Computing, Vienna, Austria), version 2.8.1. Incidence rates were calculated using the epitools library[19] and the survival analysis was performed using the survival library.[20]

Ethical considerations

All of the data are part of normal public health practise and, therefore, ethical approval was not required. The data were handled in a strictly confidential manner according to the principles of the Declaration of Helsinki, 1964, reviewed and updated by the World Medical Organisation (Edinburgh, 2000). The Spanish statute 15/1999 on data protection was followed at all times.

Results

A total of 380 individuals were included in the study, of which 192 (50.5%) were male and 188 (49.5%) were female. On December 1st 1985, the median age was 14.6 years (13.9 years among males and 15.0 years among females) and 25 cases (6.6%) were born after this date. The cohort was composed of a total of 65 families, living in 65 shacks, with a mean number of 5.7 inhabitants per shack at the beginning of the study (figure S5 of the SDC). A number was assigned to each shack to identify the residing family. During follow-up, 37 (19.3%) males and 12 (6.4%) females died. A total of 62 (16.3%) individuals were imprisoned (59.7% were males and 40.3% were females) and the median age at the time of imprisonment was 29 years.

In 2008, 19 years after the elimination of this marginal neighbourhood, the residential redistribution of inhabitants was as follows: 14.5% lived in Barcelona, 50.6% in the Barcelona metropolitan area, 25.0% had emigrated to other areas of Catalonia and the residence of 9.8% was unknown.

The follow-up showed that the AIDS incidence was 104 cases per 100 000 pyf, the incidence of drug addiction was 240 cases pyf, the incidence of imprisonment was 642 cases pyf and TB incidence was 91 cases pyf (table 1). TB was the AIDS-defining disease for five cases (62.5%). During the study period, two individuals presented with only TB, four presented with only AIDS and five had TB and AIDS. Twenty-six cases were IDU.

The survival analysis indicated that the probability of surviving at 20 years was 0.87. Survival was greater among women and worse among individuals >30 years of age (figure 1). The probability of death after 15 years of follow-up was 0.15 among males and 0.05 among females.

Figure 2 shows the global survival stratified by sex, with age at death as the response variable. The median age of death during the period of study was 68 years (figure 2) and higher in females than in males (86.7 vs. 56.8, respectively).

In the multivariate analysis, males' and females' survival was compared and adjusted for age (table 2). Model I analysed the time passed since December 1st, 1985 as a function of sex and age at the beginning of the study, including the latter as a continuous variable. According to this model, the hazard ratio (HR) for males vs. females was 3.59. The remaining models in table 2 show the age at death as the response variable, and treated data as right-censored and left-truncated. In this case, the only predictive variable was gender. Using this model, the HR

estimation that compared males and females was similar to that of the previous model (3.67).

Finally, the use of the Cox proportional hazards model with family as a random effect and age at death as the response variable, revealed that the family effect increased the HR to 4.22, a higher risk than previously calculated. The family-level variance was 0.279. The Schoenfeld residual plots (not shown) demonstrated that the distribution of these residuals was random, confirming the validity of the assumption of proportional risks in these models.

Discussion

This study has allowed us to establish that, after many years of follow-up, a marginalized population of Roma ethnicity presents poor survival and elevated incidences of AIDS, TB, IDU and imprisonment. These incidences (table 1) were higher than in Barcelona, where the rates of TB and AIDS during the same period (1987–2008) were 45 and 18 cases pyf, respectively.[21,22] Prison data also indicated that the rate of imprisonment in this population is higher than in the general population. Because 7.3% of the population below the age of 20 years had died after 17 years of follow-up, we conclude that this represents a high rate of mortality in our cohort. Barcelona's health report states that the mortality rate among young adults (15–44 years) in 1985 in our city was 126.9 in 100 000 men and 61.9 in women.[23] In our study, the life expectancy during the study period was 68 years (figure 2) and higher in females than in males (86.7 vs. 56.8, respectively). This number is below the life expectancy of being born in 1999 in Spain (78.9 years) or in Catalonia[3] (79.08 years for the whole population, 82.5 in females and 75.5 in males).

Table 1 Incidence of TB, AIDS, imprisonment and injecting drug use according to sex and last living place in a Roma population. Barcelona (1985–2008)

	Cases	Follow-up (person-years)	Rate/100 000pyf[a] (95% CI)
Global			
TB	7	7702.7	90.9 (68.5–113.2)
AIDS	8	7686.3	104.1 (80.1–128.0)
Imprisonment	45	7009.2	642.0 (579.7–704.3)
IDU	18	7506.3	239.8 (203.0–276.6)
Men			
TB	3	3791.4	79.1 (49.4–108.9)
AIDS	5	3782.3	132.2 (93.7–170.7)
Imprisonment	24	3359.4	714.4 (619.5–809.4)
IDU	17	3359.4	506.0 (426.1–585.9)
Women			
TB	4	3911.3	102.3 (69.0–135.6)
AIDS	3	3904.0	76.8 (48.0–105.7)
Imprisonment	21	3649.8	575.4 (493.6–657.1)
IDU	1	3908.0	25.6 (8.9–42.2)
Last living place			
Barcelona			
TB	5	1097.3	455.7 (323.0–588.3)
AIDS	5	1070.9	466.9 (331.0–602.8)
Imprisonment	10	921.7	1085.0 (861.6–1308.3)
IDU	4	1044.0	383.1 (258.4–507.9)
Elsewhere[b]			
TB	2	6464.8	30.9 (16.7–45.2)
AIDS	3	6474.8	46.3 (28.9–63.7)
Imprisonment	35	5946.9	588.5 (523.8–653.3)
IDU	14	6321.7	221.5 (182.9–260.0)

a: Person-years
b: Place of residence in the follow-up

The median age was 14 years at the beginning of the study and the population pyramid of Camp de la Bota in 1985 by age and sex showed a young population structure, different from that of Barcelona (figure S4 of the SDC). This is characterized by the wide base and narrow peak of the population pyramid similar to that of low-income countries but different from Barcelona which has an older population structure.[17] Regarding other indicators in this cohort, only 17.1% of men and 5.1% of women were employed. A total of 38.6% subjects were illiterate and higher among women than men (P. Pilar, Personal Communication).

In all developed countries, some groups or populations, such as ethnic minorities and individuals of low income, have greater burdens of health problems.[24] Poor health is a general reality for the Roma population. In fact, the Roma population is one of the most stigmatized and has high levels of social exclusion and poverty which affects education, employment, health, social and political participation, housing and prevalence of infectious diseases.[3] The high levels of poverty, lack of education, overpopulation and unemployment are probably principal causes, but little is known about the specific disease patterns and how these differ compared to other populations.[5,25]

The living conditions of this population have resulted in a general increase in drug use and crime, which translates in turn to imprisonment[16] and has repercussions on education and lifestyle.[5,26] Unemployment hinders social integration and explains many of the causes of death by accident among minority groups.[27] These realities have given rise to the definition of 'ethnification of poverty' and 'postmodern racism'.[28] Because of the marginalization of this group, many of the individuals in our study had problems with drug addiction (mainly heroin), such that they committed crimes just to survive, and leading to imprisonment. Syringe sharing also increases the risk of diseases transmission and poor survival.

The young Roma community suffers from these inequalities, with lack of access to education and health services, and poor use of these services

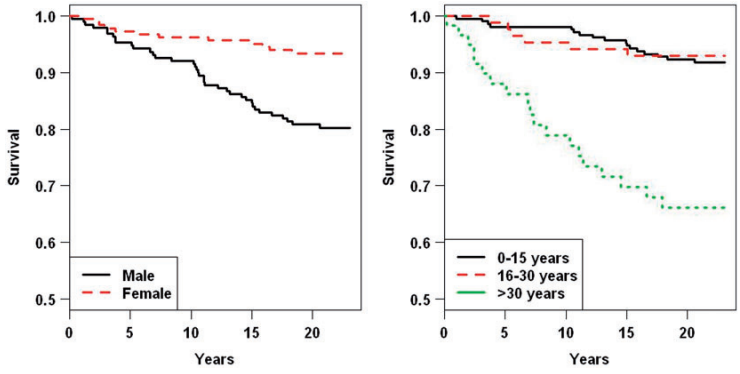


Figure 1 Survival according to various covariates in a population of Roma ethnicity. Camp de la Bota, Barcelona (1985–2008)

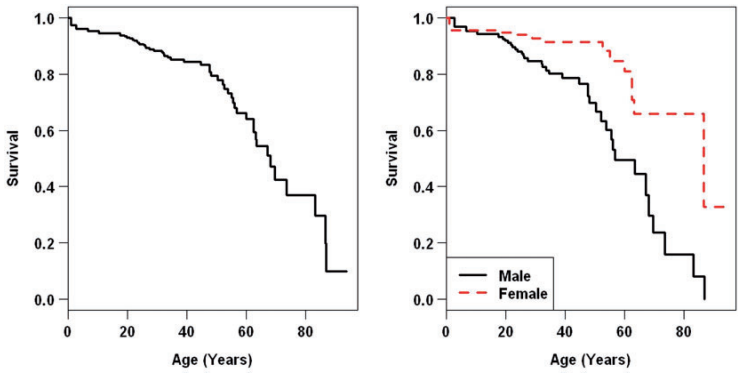


Figure 2 Survival using age at death as the response variable in a marginal population of Roma ethnicity. Camp de la Bota, Barcelona (1985–2008)

Table 2 Predictors of death according to a series of Cox models in a population of Roma ethnicity. Camp de la Bota, Barcelona (1985–2008)

	β	Se(β)	HR (95% CI)
Model I			
Men	1.277	0.333	3.59 (2.89–4.45)
Age	0.051	0.006	1.05 (1.05–1.06)
Model II			
Men	1.30	0.335	3.67 (2.95–4.56)
Model III			
Men	1.44	0.44	4.22 (3.17–5.62)
Model IV			
Men	1.41	0.43	4.10 (3.10–5.42)

Model I: Cox model with gender and age as covariates
Model II: Cox model with gender as covariate considering left truncation
Model III: Mixed effects Cox model with gender as covariate considering left truncation
Model IV: Mixed effects Cox model with gender as covariate considering left truncation, eliminating individuals without information on family

when available.[29] The result is higher infant and youth mortality, illiteracy rates of up to 60% and high rates of IDU. One of the most common causes of morbidity among this population in Spain is transmissible diseases caused by poor living conditions and lack of use of preventive programmes.[16,26,30–33]

Although the analysis used for time elapsed since December 1st, 1985 produces valid estimates, using age at death as the response variable is more appropriate, since it permits interpretation of the results in terms of life expectancy. It is important to note that the family variable was not verified for all study subjects. Thus, in order to compare the random effects model with the previous one (without the random family effect), individuals with no family were discarded. In this case, the HR was 4.10. Given that Model V with random effects is valid, the missing family effect would result in underestimation of the differences between male and female survival. Therefore, we can confirm that family was an important component for estimating survival and that the males had poorer survival. It would also be interesting to explore whether a survival model that incorporates degree of relation between individuals within each family provides a better fit of the data. However, this was beyond the scope of the study.

A limitation of the study, which exists in all studies about the Roma population, is that follow-up is hindered by the fact that subject information is found under different names, shared names and many that do not appear in the registries. We also observed different incidence figures for different registries which indicate that information sources outside of Barcelona could have been misused during the study period. We also think that these results can be extrapolated only to other marginal Roma populations. Nonetheless, high incidence rates of IDU, imprisonment, AIDS and TB were present in our cohort, compared with the general population. To intervene in these marginal populations, public health policies should focus on prevention and prioritize marginal populations for public health services.

In recent years, a number of political initiatives were proposed, such as the recognition of the Roma culture as integral and enriching. In 2003, a study of the Roma population was initiated with the aim of creating an integrated plan for assisting them in the Catalan Autonomous community. The approximate number of Roma in Catalonia in 1999 was 52 937 individuals, but the lack of a census data for this population and their nomadic tendency make it difficult to provide a reliable figure.[6]

Some Europe-wide decisions also have special importance. The European Commission has insisted on the need to eliminate social exclusion of all minority groups, to recognize the Roma population as a minority and promise actions and programmes confronting the exclusion of these populations. The National Reform Programmes (previously the National Action Plans) in many European countries also reflect this need, which now clearly address the underprivileged social and economic situation of the Roma population. In several countries of central and Eastern Europe, the Roma ethnic minority is officially recognized.[34]

Reducing health inequalities is a primary objective for the public health issues. This article as well as other information about Roma should favour specific public health interventions. Future research should take into consideration the use of a Cox model with random effects to accounts for family relationship. Ignoring the correlation of observations within the families may lead to an underestimation of the standard error.

In conclusion, the struggle to eliminate this population's exclusion from society and to avoid social marginalization is the first step to avoid these poverty conditions and thus prevent or decrease as much as possible infectious disease incidence. We trust that the year 2010, declared by the European Union as 'Year against poverty and social exclusion' will contribute to make these dreams a reality.[35]

Supplementary data

Supplementary data are available at *EURPUB* online.

Acknowledgements

CIBER de Epidemiologia y Salud Pública (CIBERESP), Spain.

To the managers of: the registries of Diseases of Obligatory Declaration, the penitentiary centres of the Department of Justice of

Catalonia, the deaths registry and the central social insurance registry of the Department of Health.

To Pilar Estrada (Casc Antic Primary Health Center), Anna Rodes, Gloria Ribas (Department of Health, Generalitat of Catalonia), Guadalupe Gómez (Universitat Politècnica de Catalunya, Barcelona, Spain), Rosa Doménech (Institut de Treball Social i Serveis Socials), Josep M[a] Jansa (Public Health Agency of Barcelona), Ronald Geskus (Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, the Netherlands) and the team of the Epidemiology Service of Barcelona Public Health Agency.

The authors are grateful to the GRASS group (http://www-eio.upc.es/grass/) for their discussions.

Funding

The Ministerio de Ciencia y Tecnologia Grant (MTM 2008-06747-C02-00, partial).

Conflicts of interest: None declared.

Key points

- Some Roma populations continue to be disadvantaged groups.
- We observed high incidence of infectious diseases, such as AIDS and TB, prison history and IDU in this cohort.
- All of these problems, in addition to social exclusion and poverty, explain the poor survival of this community.
- Public health policies should focus on prevention interventions directed at these ethnic minorities and the Roma population should always be one of the top priorities.

References

1 *The situation of Roma EU citizens moving to and settling in other EU Member States*. Vienna: European Union Agency for Fundamental Rights, November 2009. Available at: http://fra.europa.eu/fraWebsite/attachments/ROMA-Movement-Comparative-report_en.pdf (10 September 2010, date last accessed).

2 OSCE. Office for Democratic Institutions and Human Rights, Status Report. 2008. *Implementation of the Action Plan on Improving the Situation of Roma and Sinti Within the OSCE Area*. Available at: http://www.osce.org/publications/odihr/2008/09/33130_1186_en.pdf (10 September 2010, date last accessed).

3 Departament de Governació i Administracions Publiques [on line]. Guia de bones practiques per a la inclusió social del poble gitano a europa. S.l: S.n., 2005–2007. Available at: http://www.romainclusion.org/rcs_gene/085-160_Guia_catala_1.pdf (8 December 2008, date last accessed).

4 Gómez J, Vargas J. Why Roma do not like mainstream schools: voices of a people without territory. *Harvard Educ Rev* 2003;73:559–90.

5 Ginter E, Krajcovicova-Kudlackova M, Kacala O, et al. Health status of romanies (gypsies) in the Slovak republic and in the neighbouring countries. *Bratisl Lek Listy* 2001;102:479–84.

6 Tarrés F. Estudi sobre la població gitana de Catalunya (Generalitat de Catalunya ed.) [Study on the Roma population in Catalonia (Government of Catalonia ed.)]. Direcció General de Serveis Comunitaris [General Direction of Community Services]. Departament de Benestar i Família [Department of Family Welfare]. (October 2005).

7 Hajioff S, McKee M. The health of the Roma people: a review of the published literature. *J Epidemiol Community Health* 2000;54:864–9.

8 Van Cleemput P, Parry G. Health status of gypsy travellers. *J Public Health Med* 2001;23:129–34.

9 Van Cleemput P, Parry G, Thomas K, et al. Health-related beliefs and experience of gypsies and travellers: a qualitative study. *J Epidemiol Community Health* 2007;61:205–10.

10 Estrada A. Epidemiology of HIV/AIDS, hepatitis B, hepatitis C, and tuberculosis among minority injection drug users. *Public Health Rep* 2002;117:126–34.

11 García de Olalla P, Cayla J, Mila C, et al. Tuberculosis screening among immigrants holding a hunger strike in churches. *Int J Tuberc Lung Dis* 2003;7:412–6.

12 Martin V, Bayas J, Laliga A, et al. Seroepidemiology of HIV-1 infection in a catalonian penitentiary. *AIDS* 1990;4:1023–6.

13 Marinac J, Willsie S, McBride D, et al. Knowledge of tuberculosis in high-risk populations: survey of inner city minorities. *Int J Tuberc Lung Dis* 1998;2:804–10.

14 Steele C, Richmond-Reese V, Lomax S. Racial and ethnic disparities in HIV/AIDS, sexually transmitted diseases, and tuberculosis among women. *J Womens Health* 2006;15:116–22.

15 Kelly J, Amirkhanian Y, Kabakchieva E, et al. Gender roles and HIV sexual risk vulnerability of Roma (gypsies) men and women in Bulgaria and Hungary: an ethnographic study. *AIDS Care* 2004;16:231–45.

16 Iraurgi I, Jimenez-Lerma J, Landabaso M, et al. Gypsies and drug addictions. Study of the adherence to treatment. *Eur Addict Res* 2000;6:34–41.

17 Batalla J, Cayla JA. Prospeccion tuberculínica en un barrio de poblacion gitana [Tuberculosis prospect in a neighbourhood of gipsy population]. *Gaceta Sanitaria* 1987: 53–7.

18 Lamarca R, Alonso J, Gómez G, et al. Left-truncated data with age as time scale: an alternative for survival analysis in the elderly. *J Gerontol A Biol Med Sci* 1988;53:337–43.

19 Tomas Aragon. *Epitools: epidemiology tools*. R package version 0.5-3. Available at: http://CRAN.R-project.org/package=epitools (8 December 2008, date last accessed).

20 Therneau T, Lumley T. & original R port by Thomas Lumley. *Survival: Survival analysis, including penalised likelihood*. R package version 2.35-7. 2009. Available at: http://CRAN.R-project.org/package=survival (8 December 2008, date last accessed).

21 Orcau A, García de Olalla P, Caylà JA. La Tuberculosis en Barcelona. Agencia de Salud Pública de BarcelonaInforme. 2007. Available at: http://www.aspb.es/quefem/docs/Tuberculosi_2007.pdf (8 December 2008, date last accessed).

22 García de Olalla P, Clos R, Orcau A, et al. *La SIDA en Barcelona. Agencia de Salud Pública de Barcelona. Boletín SIDA 83*. Available at: http://www.aspb.es/quefem/docs/sida83.pdf (8 December 2008, date last accessed).

23 Agència de Salut Pública de Barcelona. *La salut a Barcelona 2008*. Barcelona: Agència de Salut Pública de Barcelona, 2009, Available at: http://www.aspb.es/quefem/docs/Salut_bcn_2008.pdf (10 September 2010, date last accessed).

24 Krieger N, Chen J, Waterman P, et al. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures the public health disparities geocoding project. *Am J Public Health* 2003;93:1655–71.

25 McKee M. The health of gypsies. Lack of understanding exemplifies wider disregard of the health of minorities in Europe. *BMJ* 1997;315:1172–73.

26 Teira R, Lizarralde E, Muñoz P, et al. A cross-sectional study on the epidemiological and clinical characteristics of HIV-1 infection in gypsies and in other minorities in Bilbao, Northern Spain. *Med Clin* 2002;119:653–6.

27 Koupilova I, Epstein H, Holcik J, et al. Health needs of the Roma population in the Czech and Slovak republics. *Soc Sci Med* 2001;102:479–84.

28 Flecha R. Modern and postmodern racism in Europe: dialogic approach and anti-racist pedagogies. In: *Cultural studies and education: perspectives on theory, methodology, and practice*. Harvard Educational Review Reprint Series No. 38. 2004: 61–80.

29 Kraigher A, Vidovic M, Kustec T, et al. Vaccination coverage in hard to reach Roma children in Slovenia. *Coll Antropol* 2006;30:789–94.

30 Ruiz C, Rodríguez E. Brote por shigella sonnei en el area 1 de salud de la comunidad de Madrid durante 1998. *Gac Sanit* 1999;13:25–5.

31 Perdiguero E, Bernabeu J, Huertas, et al. History of health, a valuable tool in public health. *J Epidemiol Community Health* 2001;55:667–73.

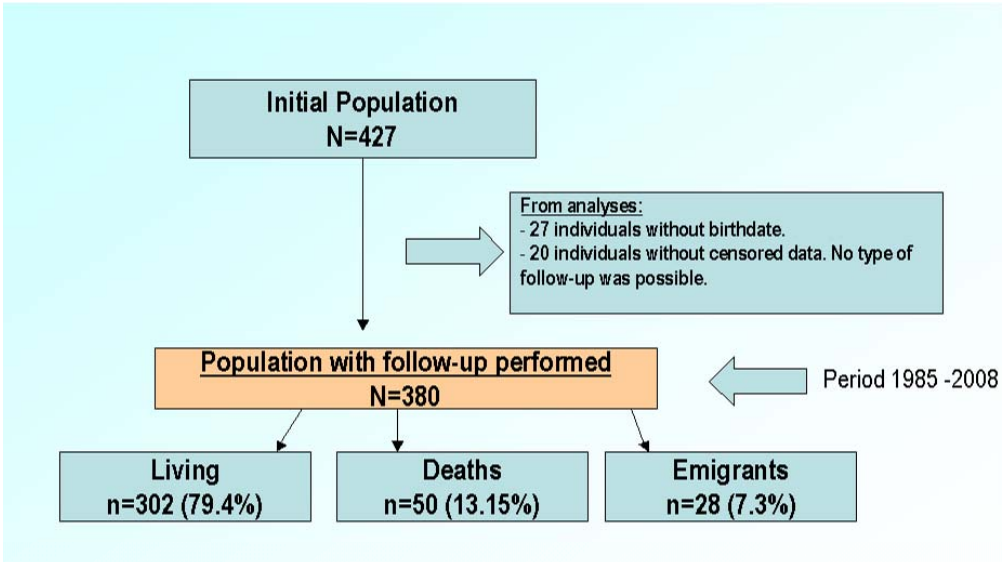
32 Cabedo V, Ortells E, Baquero L, et al. Como son y de que padecen los gitanos. *Aten Primaria* 2000;26:21–5.

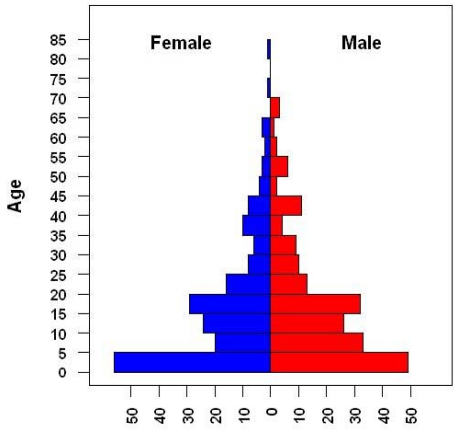
33 Agudelo-Suárez A, Gil-Gonzalez D, Ronda-Perez E, et al. Discrimination, work and health in immigrant populations in Spain. *Social Sci Med* 2009;68:1866–74.

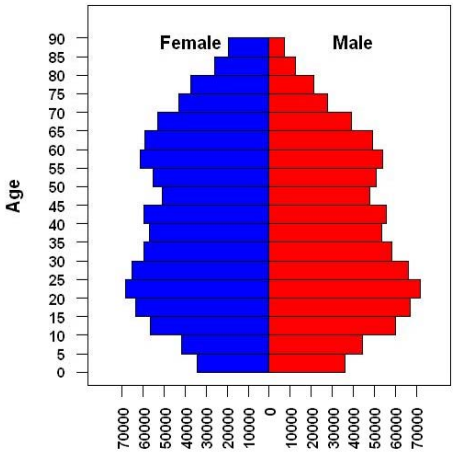
34 Gresham D, Morar B, Underhill P, et al. Origins and divergence of the Roma (gypsies). *Am J Hum Genet* 2001;69:1314–31.

35 European Comisión. *Employment, social affairs and equal opportunities. 2010 European year for combating poverty and social exclusion*. Available at: http://www.2010againstpoverty.eu/?langid=en (10 September 2010, date last accessed).

El material complementari de l'article és presentat a l'apartat de l'annex.


Initial Population
N=427
From analyses:
- 27 individuals without birthdate.
- 20 individuals without censored data. No type of follow-up was possible.
Population with follow-up performed
N=380
Period 1985 -2008
Living
n=302 (79.4%)
Deaths
n=50 (13.15%)
Emigrants
n=28 (7.3%)


Female
Male
Age
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
50
40
30
20
10
0
10
20
30
40
50


Female
Male
Age
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
70000
60000
50000
40000
30000
20000
10000
0
10000
20000
30000
40000
50000
60000
70000

4.5 Línies de recerca futura

Aprofitant les dades de l'exemple il.lustratiu presentat anteriorment es va ajustar de nou el model semiparametric de Cox amb efecte aleatori (*frailty model*) considerant diferents distribucions per als *frailties* (Gamma o Log-Normal) i el model de Cox on la variable edat actua com a covariable (censura per la dreta) i com a variable resposta (truncament per l'esquerra). Per fer l'ajust del model es va decidir explorar diferents filosofies estadístiques implementades a nivell de software d'R (INLA [72], frailtyHL [114], coxme [115], survival [116] amb la funció Coxph) centrant-nos bàsicament amb la covariable sexe.

Fins ara, molts dels articles de *frailty models* que s'han fet són més a nivell d'exploració o aplicació com el del Camp de la Bota [117–119]. En altres, s'ha fet per exemple la comparació d'un paquet d'R (kinship) a nivell Bayesià versus freqüentista [117], però fins ara no s'han explorat els 3 paquets que hem emprat posteriorment, i que impliquen indirectament diferents filosofies estadístiques en aquest camp de l'anàlisi de supervivència. Recentment però, s'ha fet algun estudi semblant d'interès, on es compara els *frailty models* amb sobredispersió amb una *Gamma frailty* versus *Normal random effects* [120].

Alhora d'utilitzar la filosofia més emprada que correspondria a la més clàssica (freqüentista) en el camp de la supervivència, ens hem centrat en els dos paquets més populars, *Kinship* i *Survival*. Un altre dels paquets emprats, el frailtyHL, correspondria a la filosofia jeràrquica basant-se en la versemblança *H-likelihood*. Un dels treballs dels seus autors en forma de tutorial fan una exploració aplicada d'aquest paquet [114]. Per últim, hem utilitzat la filosofia bayesiana mitjançant el paquet INLA. Destacar que una de les principals dificultats inicials ha estat poder ajustar el model amb totes les seves característiques proposades a nivell de distribució del *frailty* i censura en cadascun dels paquets. Per exemple en el paquet *Kinship* amb la funció *Coxme* no és possible utilitzar una distribució Gamma en el *frailty*. En el paquet frailtyHL no era possible fins ara ajustar el model quan hi havia truncament per a l'esquerra. Després de

contactar amb el creador del seu paquet s'ha pogut implementar i és ara possible aquest ajust. Una de les altres dificultats que ens hem trobat, ha estat que amb INLA fins fa poc no era possible ajustar un *Gamma frailty model*, però un article recent explica i aplica aquesta possibilitat mitjançant una transformació [109].

Respecte als resultats, com es pot veure en les següents taules, en principi les estimacions dels coeficients del predictor lineal entre els diferents paquets és força similar. Només veiem un interval de confiança més estret al model INLA, probablement influït per la 'prior' definida. En aquestes dades i tal com suggereixen alguns autors no s'ha utilitzat la 'prior' per defecte sinó la Normal truncada per la desviació estàndard [73, 121]. Els resultats però respecte les estimacions de la variància, ja són més diferents tal com es mostra en la taula 4.1 i figura 4.2.

Taula 4.1: Resultats de les estimacions de la covariable sexe en els diferents *frailty models*

	Parameter				
Paquets/Mètodes	β_{Sexe}	***SE***	β_{LI}	β_{LS}	***V(Z)***
coxph Gamma (no left truncation)	1,446	0,440	0,583	2,308	0,014
coxph Log-normal (no left truncation)	1,464	0,443	0,596	2,333	0,130
coxph Gamma (left truncation)	1,360	0,438	0,502	2,219	0,007
coxph Log-normal (left truncation)	1,390	0,442	0,524	2,256	0,145
coxme Log-normal (no left truncation)	1,457	0,442	0,591	2,323	0,017
coxme Log-normal (left truncation)	1,389	0,441	0,525	2,253	0,140
INLA Gamma (no left truncation)	1,330	0,339	0,693	2,025	0,041
INLA Log-normal (no left truncation)	1,330	0,339	0,693	2,025	0,041
INLA Gamma (left truncation)	1,343	0,341	0,703	2,042	0,034
INLA Log-normal (left truncation)	1,303	0,335	0,673	1,992	0,030
FrailtyHL Gamma (no left truncation)	1,440	0,439	0,580	2,300	NA
FrailtyHL Log-normal (no left truncation)	1,460	0,442	0,594	2,326	0,117
FrailtyHL Gamma (left truncation)	1,390	0,442	0,524	2,256	0,171
FrailtyHL Log-normal (left truncation)	1,390	0,441	0,526	2,254	0,142

Per a la comparació d'aquests resultats amb les diferents metodologies, una primera opció seria mitjançant mesures de bondat d'ajust com el AIC, BIC o DIC. Tot i que en alguns estudis realitzats de Hanagal

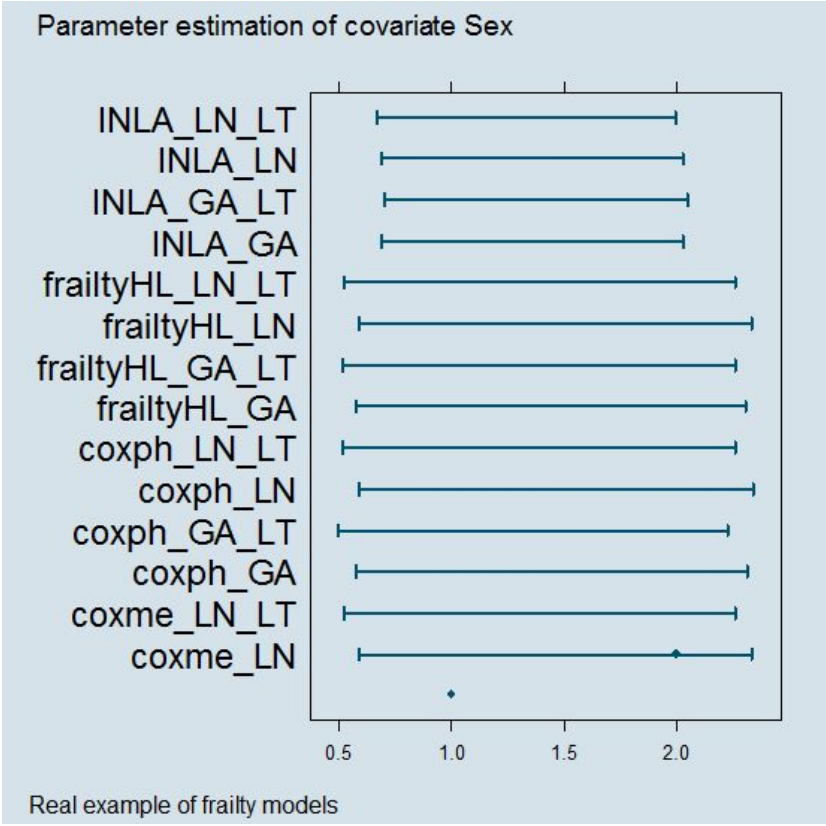


Figura 4.2: Resultats de les estimacions de la covariable sexe amb els seus intervals de confiança en els diferents *frailty models*. LT:'Left truncation'; GA: Distribució Gamma; LN: Distribució Log-Normal

aplica aquestes mesures [122, 123], per a la comparació en aquest exemple real, on s'apliquen diferents filosofies estadístiques, es fa difícil fer una comparació d'aquests models. Una possible recerca futura per tal de comparar aquestes filosofies estadístiques i d'aquesta manera, escollir el model més apropiat dels ajustats, seria via un estudi de simulació.

5 Resum i conclusions

L'objectiu d'aquesta tesi ha estat analitzar i comparar els procediments de models amb efectes aleatoris a partir de dos exemples reals, un relacionat amb les lesions d'un esport de contacte, i l'altre amb la supervivència d'un estudi longitudinal en una població gitana.

Un aspecte a tenir en compte en els models estadístics és que per poder contrastar adequadament les hipòtesis d'investigació d'interès eliminant el biaix de les estimacions i tenint en compte les fonts de variabilitat en les dades, normalment ens trobem amb dissenys cada vegada més complexes, on les dades es presenten amb una estructura jeràrquica, tipus *cluster* o multinivell. Tot i que en el primer capítol s'han presentat els models clàssics més utilitzats, s'ha volgut també fer èmfasi a certes limitacions que ens podem trobar i d'aquesta manera presentar models més sofisticats davant dissenys més complexes.

En el segon capítol s'introdueixen els models amb efectes aleatoris. L'especificació i l'elecció dels tipus de models mixtes depèn com hem dit abans de l'estructura de les dades (balancejades o no), del tipus de variable resposta (continua, ordinal, nominal) i també de la relació en la forma funcional entre la variable resposta i les covariables en el model (lineal, lineal generalitzat, no lineal). En l'àmbit de la medicina les variables resposta són habitualment binàries o recomptes, i els models no lineals assumeixen una major importància. Això ha comportat l'augment de l'ús de GLMM en l'àmbit de la medicina. És per això que en aquest mateix

capítol s'ha revisat i avaluat també la qualitat de la informació aportada en relació a l'anàlisi amb GLMM. L'objectiu és que el professional de la medicina vegi l'ús d'aquests models i la importància d'informar les seves característiques de forma adient. Tal com s'ha comprovat, durant els darrers anys, l'ús de GLMM en la literatura mèdica presenta una clara tendència a l'alça. La revisió va incloure articles de revistes mèdiques indexades amb factor d'impacte, que va consistir principalment en estudis longitudinals amb mesures repetides. La majoria dels articles seleccionats pertanyien a àrees de salut pública ambiental i ocupacional, neurologia clínica, malalties infeccioses, oncologia, i pediatria. La distribució de la variable resposta es va informar en un $89,6\%$ dels articles, predominant la distribució Binomial o de Poisson. No obstant això, la majoria dels articles no informaven adequadament de les característiques de les anàlisis, mètode d'estimació, la validació i la selecció del model d'acord amb les recomanacions actuals. Després d'analitzar i revisar la qualitat de la informació aportada en relació a l'anàlisi amb GLMM, creiem que és important tenir en compte l'ús d'unes guies estandaritzades com a directrius per a la presentació de resultats quan es fa ús de models GLMM en revistes mèdiques.

En el Capítol 3 s'ha estudiat la principal dificultat dels GLMM que és l'estimació dels paràmetres mitjançant tres filosofies, ja que, moltes vegades no hi ha una solució analítica viable que permeti maximitzar la versemblança de les dades. L'objectiu va ser comparar tres filosofies estadístiques (*marginal likelihood, hierarchical likelihood, Bayesian*) tenint en compte els seus mètodes d'estimació i algoritmes corresponents. L'exemple real fa referència a dades sobre lesions d'un esport de contacte popular a la provincia de Lleó ('Lucha Leonesa') durant les temporades $2005 - 2010$. Pel que fa als resultats inicials es va comprovar que les estimacions dels coeficients del predictor lineal en les tres filosofies estadístiques no diferien gaire, només pel que fa al component de la variància i al paràmetre de dispersió. És per això que per tal de seleccionar el model més adient es va prendre l'estratègia de fer un estudi de simulació per comparar les tres filosofies de models GLMM i a més a més el model

GLM sense efecte aleatori, tenint en compte la sobredispersió produïda i diferents escenaris. Els resultats de la simulació semblen indicar que les tres filosofies de models GLMM comparades mostren un bon rendiment excepte en combinacions amb tamany mostral i mitjanes marginals petites i amb gran sobredispersió. En aquest darrer cas, amb gran sobredispersió, probablement caldria ajustar el model amb altres alternatives, ja que, no estaríem davant de la distribució plantejada ni especificant correctament el model. Les tres filosofies comparades demostren que l'amplia disponibilitat computacional, la millora a nivell d'optimització i convergència poden ajudar a investigadors que volen aplicar aquestes eines a les seves dades.

En aquest mateix capítol en l'estudi de lesions de ('Lucha Leonesa') però durant les temporades $2005 - 2013$ s'ha ajustat un model de Poisson generalitzat mixt per estudiar l'associació de les lesions per combat que hi havia entre els diferents lluitadors. L'ús d'aquest model ha permès conèixer aquesta associació i la possibilitat d'iniciar programes de prevenció i control de lesions en aquest esport.

En el Capítol 4 s'introdueixen els models de supervivència amb efectes aleatoris, els anomenats *frailty models*. Concretament ens hem centrat en els *shared semi paramètrics frailty models*. A partir d'un exemple real i ajustant aquests models, s'ha estudiat la incidència i supervivència de malalties infeccioses d'una població gitana que vivia a la ciutat de Barcelona durant 23 anys. En aquest exemple il.lustratiu es va utilitzar primer el model clàssic de Cox amb l'edat com a covariable, i després el model de Cox amb l'edat com a variable resposta que implicava el truncament per l'esquerra. A més a més, es va ajustar el model de Cox amb un efecte aleatori, que representava la família de la població marginal estudiada. Pel que fa als resultats, tot i que l'anàlisi amb el model de Cox quan l'edat és una covariable produeix estimacions adequades, utilitzar l'edat com a resposta resulta una estimació i interpretació més apropiada per aquest estudi. El fet de que les estimacions entre els diferents models utilitzats canvien podria produir una infraestimació dels errors estàndard.

En resum, les principals conclusions que es poden extreure d'aquesta

tesi són:

1. Els models clàssics més utilitzats poden presentar certes limitacions i per tant, s'han de tenir en compte models més sofisticats davant dissenys més complexes on es presenten dades amb una estructura jeràrquica, tipus *cluster* o multinivell.

2. S'ha observat un increment de l'ús dels models GLMM als darrers anys en l'àmbit de la medicina clínica.

3. Després d'avaluar i revisar la qualitat de la informació aportada en relació a l'anàlisi dels GLMM en revistes de mèdiques indexades s'ha observat que la majoria pertanyien a àrees de salut pública ambiental i ocupacional, neurologia clínica, malalties infeccioses, oncologia, i pediatria. Una de les principals troballes és que la majoria dels articles no informaven adequadament les característiques de les anàlisis, mètode d'estimació, la validació i la selecció del model d'acord amb les recomanacions actuals.

4. Un cop s'ha analitzat i revisat la qualitat de la informació aportada en relació a l'anàlisi amb GLMM, creiem que és important tenir en compte l'ús d'unes guies estandaritzades com a directrius per a la presentació de resultats quan es fa ús de models GLMM en revistes mèdiques.

5. Després de comparar el rendiment de tres filosofies estadístiques en un model Poisson generalitzat lineal mixt via simulació es conclou que en general les estimacions dels paràmetres entre les tres filosofies són similars. Tot i així, les estimacions del component de la variància és on es presenta més variabilitat principalment en les situacions més extremes: on el grau de sobredispersió és alt, i la mitjana marginal i el tamany mostral són petits.

6. Els tres paquets (lme4, hglm, INLA) utilitzats que representaven les tres filosofies estadístiques (*marginal likelihood, hierarchical likelihood, Bayesian*) no han mostrat problemes greus de convergència. El paquet INLA és més lent que lme4 i hglm, ambdós similars.

7. En l'exemple real del Camp de la Bota, l'ús del model *shared semi parametric frailty model* amb l'edat com a variable resposta mostra una interpretació més apropiada que el model de Cox quan l'edat és una covariable.

Bibliografia

[1] Clayton D, Hills M. Statistical Models in Epidemiology. Oxford University Press; 1993.

[2] Adèr HJ, Mellenbergh GJ. Advising on Research Methods: A consultant's companion. Johannes van Kessel Publishing; 2008.

[3] Konishi S, Kitagawa G. Information Criteria and Statistical Modeling. Springer; 2008.

[4] Porta M, Greenland S, Hernán M, Silva IDS, Last JM. A Dictionary of Epidemiology. Oxford University Press; 2014.

[5] Snow J, Richardson B. Snow on Cholera: Being a Reprint of Two Papers by John Snow, MD, Together with a Biographical Memoir by BW Richardson, and an Introduction by Wade Hampton Frost, MD. Hafner; 1965.

[6] Fisher RA, et al. Cigarettes, cancer and statistics. Centennial Review. 1958;2:151–166.

[7] Fisher RA. Lung Cancer and Cigarettes? Nature. 1958;182:108.

[8] Fisher RA. Cancer and Smoking. Nature. 1958;182:596.

[9] Louçã F. Should The Widest Cleft in Statistics - How and Why Fisher opposed Neyman and Pearson. ISEG - School of Eco-

nomics and Management, Department of Economics, University of Lisbon; 2008. Working paper 2008/02. Available from: http://ideas.repec.org/p/ise/isegwp/wp22008.html.

[10] McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman and Hall/CRC; 1989.

[11] Nelder JA, Wedderburn RW. Generalized Linear Models. Journal of the Royal Statistical Society, A. 1972;135:370–384.

[12] Cox PR. Life Tables. John Wiley & Sons; 1972.

[13] Altman D, De Stavola B, Love S, Stepniewska K. Review of survival analyses published in cancer journals. British Journal of Cancer. 1995;72(2):511.

[14] Miller Jr RG. Survival Analysis. 2nd ed. John Wiley & Sons; 2011.

[15] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457–481.

[16] Oller R, Gómez G, Calle ML. Interval censoring: model characterizations for the validity of the simplified likelihood. Canadian Journal of Statistics. 2004;32(3):315–326.

[17] Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 1994;12(3):601–607.

[18] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society, B. 1972;34(2):187–220.

[19] Cox DR. Regression Models and Life-Tables. In: Breakthroughs in Statistics. Springer; 1992. p. 527–541.

[20] Lamarca R, Alonso J, Gómez G, Muñoz A. Left-truncated Data With Age as Time Scale: An Alternative for Survival Analysis in the Elderly Population. The Journals of Gerontology. 1998;53(5):337–343.

[21] Agresti A. Categorical Data Analysis. 2nd ed. Wiley Series in Probability and Statistics. New Jersey: John Wiley & Sons; 2002.

[22] Campbell MJ. Statistics at Square Two: Understanding Modern Statistical Applications in Medicine. 2nd ed. John Wiley & Sons; 2006.

[23] Brown H, Prescott R. Applied Mixed Models in Medicine. 2nd ed. John Wiley & Sons; 2006.

[24] Crowder MMJ, Hand DJ. Analysis of Repeated Measures. Chapman & Hall/CRC; 1990.

[25] Diggle P, Heagerty P, Liang KY, Zeger S. Analysis of longitudinal data. Oxford University Press; 2002.

[26] Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society, A. 1996;159:505–513.

[27] McCulloch CE, Searle SR. Generalized, Linear and Mixed Models. New York: John Wiley & Sons; 2001.

[28] O'Hara RB; BioOne. How to make models add up — a primer on GLMMs. Annales Zoologici Fennici. 2009;46(2):124–137.

[29] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Springer; 2009.

[30] Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association. 1988;83(404):1014–1022.

[31] Iavarone L, Gomeni R. An application of nonlinear mixed-effects modeling to pharmacokinetic data exhibiting nonlinear and time-dependent behavior. Journal of Pharmaceutical Sciences. 2003;92(1):27–34.

[32] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal Data Analysis. Chapman & Hall/CRC; 2008.

[33] Davidian M, Giltinan DM. Nonlinear models for repeated measurement data: an overview and update. Journal of Agricultural, Biological, and Environmental Statistics. 2003;8(4):387–419.

[34] Pinheiro JC, Bates DM. Model Building in Nonlinear Mixed Effects Models. University of Wisconsin; 1995. Technical Report #91.

[35] Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models; 2015. R package version 3.1-122. Available from: http://CRAN.R-project.org/package=nlme.

[36] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993;88(421):9–25.

[37] McGilchrist C. Estimation in generalized mixed models. Journal of the Royal Statistical Society, B. 1994;56(1):61–69.

[38] Schall R. Estimation in generalized linear models with random effects. Biometrika. 1991;78(4):719–727.

[39] Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution. 2009;24(3):127–135.

[40] Aitkin M. A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing. 1996;6(3):251–262.

[41] Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. Biometrics. 1999;55(1):117–128.

[42] Lee Y, Nelder JA, Pawitan Y. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. Boca Raton: Chapman & Hall/CRC; 2006.

[43] Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society, B. 1999;61(1):265–285.

[44] Rue H, Held L. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC; 2005.

[45] Nakagawa S, Schielzeth H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution. 2012;4(2):133–142.

[46] Li B, Lingsma HF, Steyerberg EW, Lesaffre E. Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. BMC Medical Research Methodology. 2011;11(1):77.

[47] Austin PC. Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. The International Journal of Biostatistics. 2010;6(1):Article 16.

[48] Birnbaum A. On the foundations of statistical inference. Journal of the American Statistical Association. 1962;57(298):269–306.

[49] Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press; 2001.

[50] Bjørnstad JF. On the generalization of the likelihood function and the likelihood principle. Journal of the American Statistical Association. 1996;91(434):791–806.

[51] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society, B. 2009;71(2):319–392.

[52] Lee Y, Nelder JA. Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. Biometrika. 2001;88(4):987–1006.

[53] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society, B. 2009;71(2):319–392.

[54] Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York: Springer; 2009.

[55] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1):1–14.

[56] Jang W, Lim J. PQL estimation biases in generalized linear mixed models. Working paper 05-21. Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA; 2006. Available from: ftp://isds.duke.edu/pub/WorkingPapers/05-21.pdf.

[57] Thiele J, Markussen B. Potential of GLMM in modelling invasive spread. CAB Reviews. 2012;7(16):1–10.

[58] Bates DM, DebRoy S. Converting a Large R Package to S4 Classes and Methods. In: Hornik K, Leisch F, Zeileis A, editors. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003); 2003. Available from: https://www.r-project.org/conferences/DSC-2003/Proceedings/BatesDebRoy.pdf.

[59] Gilmour A, Gogel B, Cullis B, Welham S, Thompson R. ASReml User Guide Release 3.0. Hemel Hempstead, HP1 1ES, UK www.vsni.co.uk; 2009.

[60] Rowe K. Practical multilevel analysis with MLwiN & LISREL: An integrated course. Melbourne, Australia; 2007.

[61] Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008;59(4):390–412.

[62] Dean CB, Nielsen JD. Generalized linear mixed models: A review and some extensions. Lifetime Data Analysis. 2007;13(4):497–512.

[63] Casals M, Martínez JA. Modelling player performance in basketball through mixed models. International Journal of Performance Analysis in Sports. 2013;13:64–82.

[64] Broström G, Holmberg H. glmmML: Generalized linear models with clustering; 2013. R package version 1.0. Available from: http://CRAN.R-project.org/package=glmmML.

[65] Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4; 2015. R package version 1.1-8. Available from: http://CRAN.R-project.org/package=lme4.

[66] Rönnegård L, Shen X, Alam M. hglm: A Package for Fitting Hierarchical Generalized Linear Models. The R Journal. 2010;2(2). Available from: http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Roennegaard$\sim$et$\sim$al.pdf.

[67] Molas M, Lesaffre E. Hierarchical Generalized Linear Models: The R Package HGLMMM. Journal of Statistical Software. 2011;39(13):1–20. Available from: http://www.jstatsoft.org/v39/i13/.

[68] Brown P. glmmBUGS: Generalised Linear Mixed Models and Spatial Models with WinBUGS, BRugs, or OpenBUGS; 2014. R package version 2.3. Available from: http://CRAN.R-project.org/package=glmmBUGS.

[69] Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software. 2005;12(3):1–16. Available from: http://www.jstatsoft.org.

[70] Komárek A, Lesaffre E. Generalized linear mixed model with a penalized Gaussian mixture as a random-effects distribution. Computational Statistics and Data Analysis. 2008;52(7):3441–3458.

[71] Hadfield JD. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software. 2010;33(2):1–22. Available from: http://www.jstatsoft.org/v33/i02/.

[72] Lindgren F, Rue H. Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software. 2015;63(19):1–25. Available from: http://www.jstatsoft.org/v63/i19/.

[73] Martins TG, Simpson DP, Riebler A, Rue H, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors; 2014. Submitted. arXiv:1403.4630.

[74] Samet JM, Muñoz A. Evolution of the cohort study. Epidemiologic Reviews. 1998;20(1):1–14.

[75] Geskus RB. Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. Statistics in Medicine. 2001;20(5):795–812.

[76] Muñoz A, Sabin CA, Phillips AN. The incubation period of AIDS. Aids. 1997;11(Suppl A):S69–S76.

[77] Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer; 2000.

[78] Borges R. Análisis de supervivencia aplicado a un caso de diálisis renal: diálisis peritoneal en el Hospital Clínico Universitario de Caracas y Hemodiálisis en el Hospital de Clínicas Caracas, 1980–2000 [Ph.D. thesis]. Instituto de Estadística Aplicada y Computación, UCLA, Mérida; 2002.

[79] Gómez G. Análisis de Supervivencia. Universitat Politècnica de Catalunya; 2004.

[80] Brenner H, Gefeller O, Hakulinen T. Period analysis for up-to-date cancer survival data: theory, empirical evaluation, computational realisation and applications. European Journal of Cancer. 2004;40(3):326–335.

[81] Cox C, Chu H, Schneider MF, Muñoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine. 2007;26(23):4352–4374.

[82] Muga R, Langohr K, Tor J, Sanvisens A, Serra I, Rey-Joly C, et al. Survival of HIV-infected injection drug users (IDUs) in the highly active antiretroviral therapy era, relative to sex-and age-specific survival of HIV-uninfected IDUs. Clinical Infectious Diseases. 2007;45(3):370–376.

[83] Cadarso-Suárez C, Meira-Machado L, Kneib T, Gude F. Flexible hazard ratio curves for continuous predictors in multi-state models an application to breast cancer data. Statistical Modelling. 2010;10(3):291–314.

[84] Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. Biometrics. 1991;47(2):467–485.

[85] McGilchrist C, Aisbett C. Regression with Frailty in Survival Analysis. Biometrics. 1991;47(2):461–466.

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

[87] Duchateau L, Janssen P. The Frailty Model (Statistics for Biology and Health). Springer, New York; 2008.

[88] Wienke A. Frailty Models in Survival Analysis. Chapman & Hall/CRC; 2010.

[89] Stryhn H, Christensen J. The analysis—Hierarchical models: Past, present and future. Preventive Veterinary Medicine. 2014;113(3):304–312.

[90] Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography. 1979;16(3):439–454.

[91] Wienke A. Frailty Models. In: Wiley Encyclopedia of Clinical Trials. John Wiley & Sons; 2007. Available from: http://dx.doi.org/10.1002/9780471462422.eoct022.

[92] Kelly PJ. A review of software packages for analyzing correlated survival data. The American Statistician. 2004;58(4):337–342.

[93] Lancaster T. Econometric methods for the duration of unemployment. Econometrica: Journal of the Econometric Society. 1979;47(4):939–956.

[94] Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978;65(1):141–151.

[95] Zhao JH. Mixed-effects Cox models of alcohol dependence in extended families. BMC Genetics. 2005;6 (Suppl 1):S127.

[96] Li H, Zhong X. Multivariate survival models induced by genetic frailties, with application to linkage analysis. Biostatistics. 2002;3:57–75.

[97] Therneau T. On mixed-effect Cox models, sparse matrices, and modeling data from large pedigrees; 2003.

[98] Zhong X, Li H. Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model. Biostatistics. 2004;5:307–327.

[99] Hougaard P. Analysis of Multivariate Survival Data. New York: Springer; 2000.

[100] Hirsch K, Wienke A. Software for semiparametric shared gamma and log-normal frailty models: An overview. Computer Methods and Programs in Biomedicine. 2012;107(3):582–597.

[101] Wienke A, Holm NV, Skytthe A, Yashin AI. The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. Twin Research. 2001;4(4):266–274.

[102] Hougaard P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. Biometrika. 1984;71(1):75–83.

[103] Abbring JH, Van Den Berg GJ. The unobserved heterogeneity distribution in duration analysis. Biometrika. 2007;94(1):87–99.

[104] Hanagal DD. Modeling Survival Data Using Frailty Models. Chapman & Hall/CRC; 2011.

[105] Karim ME. Comparisons of Different Frailty Models: A Simulation Study [Ph.D. thesis]. Institute of Statistical Research and Training, University of Dhaka. Dhaka, Bangladesh; 2007.

[106] Gutierrez R. On Frailty Models in Stata. In: 7th UK meeting, The Royal Statistical Society; 2001. p. 14–15.

[107] Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. Journal of Computational and Graphical Statistics. 2003;12(1):156–175.

[108] Martino S, Akerkar R, Rue H. Approximate Bayesian inference for survival models. Scandinavian Journal of Statistics. 2011;38(3):514–528.

[109] Martins TG, Rue H. Extending INLA to a class of near-Gaussian latent models; 2012. Submitted. arXiv:1210.1434.

[110] Shu Y, Klein JP. A SAS Macro for the positive stable frailty model. Milwaukee, Wisconsin: Medical College of Wisconsin; 1999. Technical Report #33. Available from: http://www.mcw.edu/FileLibrary/Groups/Biostatistics/TechReports/TechReports2550/tr033.pdf.

[111] Valença DM, Santos PB. Estimação em Modelos Paramétricos para Dados de Sobrevivência Correlacionados no Software R: Simulação e Aplicação. In: Gonçalves B, editor. XLII SBPO 2010; 2010. Available from: http://w3.ufsm.br/42sbpo/programa_xlii_sbpo.pdf.

[112] Nielsen GG, Gill RD, Andersen PK, Sørensen TI. A counting process approach to maximum likelihood estimation in frailty models. Scandinavian Journal of Statistics. 1992;19(1):25–43.

[113] McGilchrist CA. REML Estimation for Survival Models with Frailty. Biometrics. 1993;49(1):221–225.

[114] Do Ha I, Noh M, Lee Y. frailtyHL: A Package for Fitting Frailty Models with H-likelihood. R Journal. 2012;4(2):28–37.

[115] Therneau T. coxme: Mixed Effects Cox Models; 2015. R package version 2.2-5. Available from: http://CRAN.R-project.org/package=coxme.

[116] Therneau T. A Package for Survival Analysis in S; 2015. R package version 2.38. Available from: http://CRAN.R-project.org/package=survival.

[117] Garibotti G, Smith KR, Kerber RA, Boucher KM. Longevity and correlated frailty in multigenerational families. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2006;61(12):1253–1261.

[118] David I, Lorino T, Sanaa M. Bayesian Versus Frequentist Approach of the Frailty Cox Model, Application to Calf Gastroenteritis. Communications in Statistics – Simulation and Computation. 2007;36(6):1309–1320.

[119] Tundo J. Frailty models for the between center variation in survival following rectum cancer diagnosis; 2010. Master's Thesis.

[120] Molenberghs G, Verbeke G, Efendi A, Braekers R, Demétrio CG. A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. Statistical Methods in Medical Research. 2015;24(4):434–452.

[121] Gelman A, et al. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006;1(3):515–534.

[122] Hanagal DD, Sharma R. Comparison of Shared Gamma Frailty Models for Acute Leukemia Data. Pune, India: Department of Statistics, University of Pune; 2012. Technical Report No. 2012/1. Available from: http://stats.unipune.ernet.in/P7.pdf.

[123] Hanagal DD, Dabade AD. A Comparative Study of Shared Frailty Models for Kidney Infection Data with Generalized Exponential Baseline Distribution. Journal of Data Science. 2013;11(1):109–142.