UNIVERSITEIT GENT

Academiejaar 2008 - 2009

Het adequate gebruik van Multivariabele Logistische Regressie Analyse in de Intensieve Zorg literatuur anno 2006.

Pieter Lambrecht
Pieter Verslype

Promotor: Prof. Dr. Dominique Benoit
Co-promotor: Prof. Dr. Johan Decruyenaere

Scriptie voorgedragen in de 2de Proef in het kader van de opleiding tot
ARTS

UNIVERSITEIT GENT

Academiejaar 2008 - 2009

Het adequate gebruik van Multivariabele Logistische Regressie Analyse in de Intensieve Zorg literatuur anno 2006.

Pieter Lambrecht
Pieter Verslype

Promotor: Prof. Dr. Dominique Benoit
Co-promotor: Prof. Dr. Johan Decruyenaere

Scriptie voorgedragen in de 2de Proef in het kader van de opleiding tot
ARTS

Datum: 11 mei 2009

Pieter Lambrecht Pieter Verslype Prof. Dr. Dominique Benoit

VOORWOORD

Deze scriptie kon onmogelijk tot stand gebracht worden zonder de gewaardeerde medewerking van:

Prof. Dr. Dominique Benoit, onze promotor die ons de mooie kans heeft gegeven om deze thesis te maken en die deze thesis tot een hoger niveau heeft getild.

Prof. Dr. Johan Decruyenaere, onze copromotor die onze thesis kritisch heeft nagelezen.

Maarten Bekaert, voor de hulp met het coderen van de artikels en het specificeren van de tekortkomingen.

Prof. Dr. Georges Van Maele en Dr. Ellen Deschepper, voor de kritische nalezing en hun hulp bij het uitleggen van statistische methodes.

Waarvoor onze hartelijke dank.

Pieter LAMBRECHT
2de Master Geneeskunde, academiejaar 2008-2009

Pieter VERSLYPE
2de Master Geneeskunde, academiejaar 2008-2009

Inhoudstafel

1. Abstract

DOEL: Multivariabele logistische regressie analyses (MLRA) worden met toenemende frequentie gebruikt in de medische literatuur. Onduidelijke weergave van deze modellen kan de interpretatie van de resultaten bemoeilijken, misleidend of onjuist maken. Het doel van deze studie is om de adequaatheid van rapportage van MLRA in twee grote tijdschriften binnen het domein van de Intensieve Zorg (IZ) geneeskunde anno 2006 te evalueren.

METHODOLOGIE: Alle artikels in Critical Care Medicine (CCM) en Intensive Care Medicine (ICM) gepubliceerd in 2006 werden in extenso onderzocht op het correct gebruik van logistische regressie. Dit gebeurde aan de hand van een lijst met 13 potentiële tekortkomingen die onder andere gebaseerd zijn op eerdere studies. Ook werd onderzocht indien enkele kwalitatieve verschillen tussen de artikels aanleiding gaven tot een betere rapportage van MLRA.

RESULTATEN: In CCM en ICM tezamen melden 106 artikels het gebruik van MLRA en in 89 (84%) artikels werden de resultaten ook getoond. In deze 89 artikels werden 265 modellen gebruikt. De initiële variabelen in het model werden in 7% niet gespecificeerd en in 15% was er een incorrecte codering van de eindpunten. In 14% werd de selectieprocedure niet gespecificeerd en in 53% werd bij een verklarend model een automatische selectieprocedure gebruikt. Er werd geen ROC-curve voorzien in 71% van de modellen en in 56% van de predictieve modellen. In 60% van de modellen die gebruik maken van continue variabelen werden de eenheden niet vermeld. In 51% van de modellen was er overfitting en in 60% werd de conformiteit naar de lineaire gradiënt werd niet nagekeken. Slechts in 16% van de modellen werd getest op interactietermen, 26% en 31% meldde respectievelijk het testen op multicollineariteit en het nagaan van de calibratiestatistiek. In slechts 5 (2%) modellen werd crossvalidatie uitgevoerd. Een model komende uit een “feature” artikel in CCM had een randsignificant (p= 0,051) hogere kans op ernstige tekortkomingen terwijl een model komende uit een “Continuous Medical Education” artikel minder kans had op ernstige

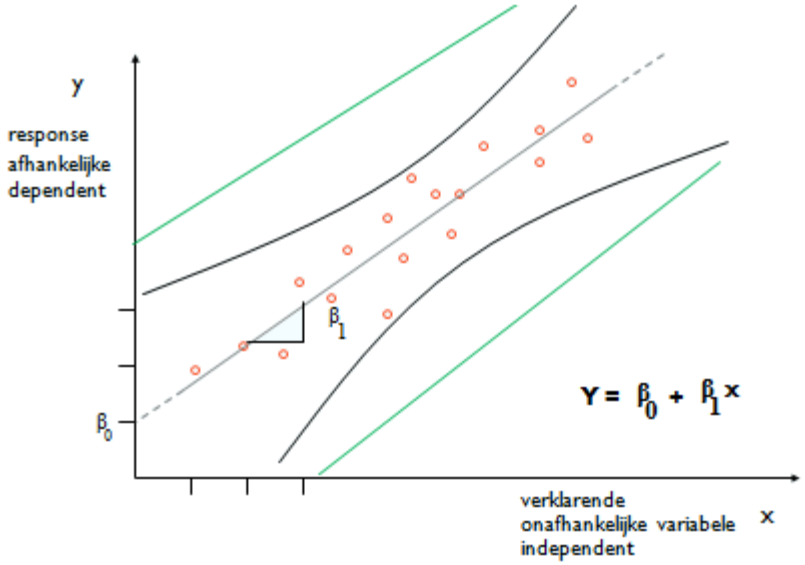
tekortkomingen (p= 0,039). Ernstige en middelmatige tekortkomingen (met beide p= 0,001) kwamen minder voor wanneer een biostatisticus betrokken was bij het onderzoek. Een begeleidende editoriaal was zowel in CCM, ICM als in CCM en ICM samen niet protectief voor ernstige, middelmatige of kleine terkortkomingen (p > 0.05). In ICM was een biostatisticus protectief voor middelmatige tekortkomingen, net zoals in CCM en ICM tezamen (p= 0,001).

CONCLUSIE: Onze resultaten tonen aan dat de rapportage van MLRA in de IZ literatuur vaak onvolledig is. Dit maakt het moeilijk voor de kritische lezer om het artikel nauwkeurig te interpreteren. Tijdschriften, en meer specifiek CCM en ICM zouden duidelijke richtlijnen over het gebruik en rapporteren van MLRA moeten uitbrengen waaraan auteurs zich zouden moeten houden voor het aanbieden van een artikel voor “peer-review” en publicatie.

2. Inleiding

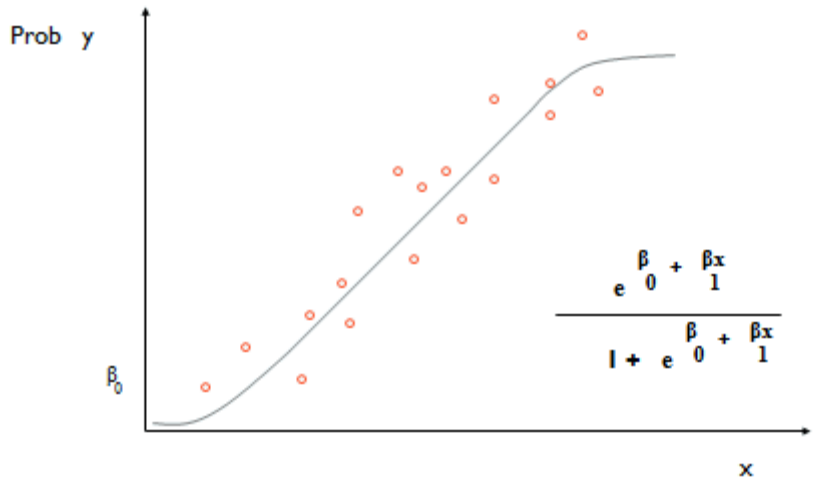
Multivariabele Logistische Regressie Analyse (MLRA) is een statistische methode die gebruikt wordt om aan de hand van zo weinig mogelijk (=“parsimonie”) variabelen, die doorgaans verklarende of onafhankelijke variabelen worden genoemd, de probabiliteit uit te rekenen dat een bepaalde gebeurtenis, zijnde de afhankelijke variabele, zich voordoet (=“predictieve modellen”). MLRA wordt echter in de geneeskunde veel vaker gebruikt om de impact van een variabele op een gebeurtenis te onderzoeken, na het in acht nemen van andere belangrijke variabelen (“confounders”), zonder dat men primair de intentie heeft om aan predictie te doen (=“verklarende modellen”) (Palmas et al., 1993).

Logistische regressie is een extensie van lineaire regressie. Daarom zullen we eerst een korte uitleg geven over lineaire regressie. Bij lineaire regressie wordt naar de lineaire relatie gekeken tussen één of meer onafhankelijke variabelen en het gemiddelde van een continue variabele. Hierbij wordt op basis van een at random genomen steekproef een rechte geconstrueerd aan de hand van de “ordinary least squares” (OLS) methode, waarbij de afstand tot alle punten zo klein mogelijk wordt gehouden. De formule van de “ideale” rechte die door deze methode wordt bekomen, kan weergegeven worden door $y = \beta_0 + \beta_1 x$ (zie figuur 1). Bij het nemen van een steekproef is er echter altijd een fout op de schatting (stochastisch) van de positie van deze rechte. De foutmarge van de ligging van de rechte wordt weergegeven door de 95% confidentie intervallen (CI) in figuur 1. Dit interval wordt groter naarmate men verder van het middelste punt gaat (inwendige hyperbolen). Naast een CI voor een bepaalde waarde van de afhankelijke variabele, kunnen ook predictieintervallen voor een nieuwe meting opgesteld worden (uitwendige hyperbolen).



Figuur 1: Lineaire regressie

Bij logistische regressie is de afhankelijke variabele niet meer continu, maar dichotoom (= een ja of nee fenomeen zoals bv. levend of dood) en berekent men de probabiliteit dat een dichotoom fenomeen zich voordoet. Gezien de probabiliteit om een dichotome uitkomst te voorspellen in de realiteit nooit 0% of 100% zal zijn, zal men i.t.t. lineaire regressie een sigmoidale curve construeren waarbij de rechte naar oneindig afbuigt naarmate men dichter bij de 0% en 100% probabiliteit komt. De constructie van deze sigmoïdale curve bij logistische regressie gebeurt aan de hand van de maximum likelihood estimation (MLE) methode, nadat de afhankelijke variabele getransformeerd werd in een logit variabele (de natuurlijk logaritme van de odds dat de afhankelijke variabele zich voordoet of niet, zie formule figuur 2). Deze methode probeert de log likelihood (LL), die weergeeft hoe waarschijnlijk het is dat de geobserveerde waarden van de afhankelijke variabele voorspeld worden aan de hand van de geobserveerde waarden van de onafhankelijke variabelen, te maximaliseren. Hierbij wordt een ideale sigmoidale curve opgebouwd tussen de geobserveerde probabiliteiten (Kleinbaum, 1994). Logistische regressie is dus sterk vergelijkbaar met lineaire regressie. Logistische regressie veronderstelt echter geen lineaire relatie tussen de onafhankelijke variabelen en de afhankelijke variabele, vereist geen normaal verdeelde variabelen en stelt in het algemeen minder strikte vereisten.

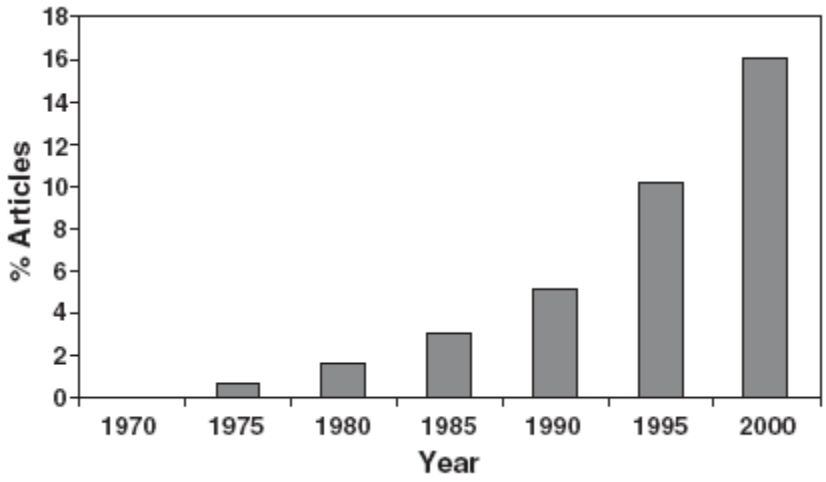


Figuur 2: Logistische regressie

Bij multivariabele logistische regressie wordt de invloed van multipele onafhankelijke variabelen op de afhankelijke variabele samen beoordeeld door gebruik te maken van de correlaties tussen deze variabelen en de uitkomst. Deze variabelen kunnen zowel continu, discreet, dichotoom of een mengeling hiervan zijn. Het is belangrijk om in te zien dat het “onafhankelijk zijn” van een variabele in deze omstandigheden een relatief begrip is. We kunnen namelijk drie grote categorieën zogenaamde “onafhankelijke” predictoren onderscheiden die in werkelijkheid een continuüm vormen. De eerste soort zijn de “echte” of volledig onafhankelijke predictoren. Deze variabelen bezitten informatie over de uitkomst die niet gedeeld wordt door andere predictoren. Deze maken de meeste kans om als onafhankelijke variabelen in het model weerhouden te worden, althans indien ze voldoende vertegenwoordigd zijn in de steekproef (voldoende “power”). Aangezien dit volledig onafhankelijke variabelen zijn, zal hun odds ratio niet beïnvloed worden door andere variabelen die in het model worden opgenomen. Het feit dat deze variabelen volledig onafhankelijk zijn betekent niet noodzakelijk dat er een causaal verband is tussen deze variabelen en de outcome. Dit komt omdat residuele confounding nooit kan uitgesloten worden in observationele en slecht gerandomiseerde studies en omdat er geen rekening wordt gehouden met de tijdsvolgorde (oorzaak en gevolg) bij MLRA. De tweede soort zijn de partieel uitwisselbare predictoren. Deze bezitten informatie over de uitkomst die gedeeld wordt door andere variabelen. Deze zullen aldus in competitie met elkaar treden

indien ze allemaal in het model blijven alsook tijdens de selectieprocedure. De aan- of afwezigheid in het model van deze variabelen zal de odds ratio's van andere variabelen beïnvloeden met uitzondering van de volledig onafhankelijke variabelen. De derde soort zijn de volledig uitwisselbare predictoren. Dit zijn predictoren die, theoretisch gezien, identieke informatie bezitten over de uitkomst en samen niet in het model mogen opgenomen worden omdat MLRA niet zal kunnen "kiezen" tussen deze variabelen waardoor er een lineaire combinatie zal achterblijven in het model (zie verder begrip multicollineariteit). Enkele voorbeelden van hoog gecorreleerde variabelen zijn: ernst-van-ziekte scores, oorzaak (antihypertensieve medicatie) en gevolg (bloeddruk) en verschillende waarden van eenzelfde variabele doorheen de tijd. Meteen vloeit hieruit voort dat het resultaat van MLRA een relatief en geen absoluut resultaat is, gezien dit resultaat sterk afhankelijk is van de variabelen die in het model worden opgenomen. Het verwijderen en/ of toevoegen van een variabele aan het model zal een impact hebben op alle variabelen, behalve op de echt onafhankelijke variabelen. Hierdoor is het uitermate van belang om alle variabelen die opgenomen werden in het model duidelijk op te sommen (zie verder).

In de geneeskunde wordt MLRA in toenemende mate gebruikt bij observationele studies. Dit werd onder andere door Chin (2001) beschreven.



Figuur 3:

Percentage van de artikels gedurende 30 jaar in the American Journal of Epidemiology, American Journal of Public Health, International Journal of Epidemiology, en Journal of Epidemiology and Community Health met MeSH heading 'logistic regression' of met sleutelwoord 'logistic regression' of 'adjusted odds ratio.' (Chin, 2001)

Nieuwe statistische software heeft er namelijk voor gezorgd dat relatief moeilijke statistische berekeningen nu gemakkelijk beschikbaar zijn voor niet-statistici die m.a.w. vaak weinig of geen basiskennis hebben van de gebruikte methode. Hierdoor is het resultaat dat bekomen wordt niet altijd even betrouwbaar. Een correct gebruik en rapportering van MLRA is belangrijk aangezien de berekende resultaten (odds ratio's, 95%CI en p-waarde) moeilijk interpreteerbaar of zelfs onbetrouwbaar kunnen zijn en kunnen leiden tot foutieve conclusies bij de lezer (Campillo, 1993 en Concato et al., 1993). Verschillende studies hebben reeds nagegaan in welke mate publicaties al dan niet voldoen aan de gangbare richtlijnen voor MLRA. Zo onderzochten Concato et al.(1993) 44 studies, gepubliceerd in The New England Journal of Medicine en The Lancet tussen 1985-1989, en vonden ze dat logistische regressie slechts in een 20-30% van de gevallen goed werd uitgevoerd (Concato et al., 1993). Recentere studies stellen nog altijd hetzelfde probleem vast (Ottenbacher et al., 2001). Moss et al. (2003) onderzochten in vijf

tijdschriften (American Journal of Respiratory and Critical Care Medicine, Chest, Critical Care Medicine, The European Respiratory Journal en Thorax) gedurende een periode van 6 maanden in het jaar 2000 alle publicaties (964) en vonden 81 publicaties die MLRA gebruikten. Zo was er bijvoorbeeld in 39% van deze publicaties “overfitting” vast te stellen. Ook had slechts 12% vermeld of interactietermen onderzocht waren en had slechts 1 artikel getest op collineariteit. Ons vermoeden was dat in de huidige Intensieve Zorg literatuur nog steeds in een minderheid van de publicaties MLRA goed wordt uitgevoerd en/of gerapporteerd.

Onze doelstelling is na te gaan in welke mate publicaties in de Intensieve Zorg literatuur in het jaar 2006 voldeden aan de gangbare en een aantal bijkomende, ons inziens, belangrijke voorwaarden voor het uitvoeren en rapporteren van MLRA. Dit doen we door in twee tijdschriften alle publicaties waarin MLRA werd gebruikt na te lezen en te beoordelen volgens vooraf vastgelegde criteria. Deze tijdschriften zijn Critical Care Medicine en Intensive Care Medicine en dit zijn de tijdschriften met de hoogste impactfactor in dit vakgebied (respectievelijk 5.07 en 4.10). De bedoeling is om deze resultaten te publiceren in één van deze bovenstaande tijdschriften en uit te leggen, o.a. aan de hand van een aantal concrete voorbeelden, waarom het zo belangrijk is dat voldaan wordt aan de gangbare criteria. Verder zouden we ook een aantal richtlijnen willen voorstellen waaraan iedere studie op zijn minst zou moeten voldoen om in aanmerking te komen voor publicatie in deze tijdschriften. Voortijdige resultaten hiervan werden reeds voorgesteld op de stafvergadering Intensieve Zorg van het UZ Gent op 19/12/2008.

In de zomer (2007-2008) hebben we een cursus, gegeven door Prof. Dr. D. Benoit, over MLRA gevolgd. Hierin werd het begrip MLRA en andere statistische methodes uitgelegd en hebben we ook de eerste potentiële problemen gezien die gepaard gaan met het gebruik van MLRA in de medische literatuur. Om het gebruik van MLRA en andere beginselen van de statistiek beter te begrijpen zijn we dan nog meerdere malen samengekomen om dit verder uit te diepen. Tijdens deze samenkomsten hebben we verder geleerd over het correcte gebruik van MLRA en hebben we zelf logistische regressie analyses uitgevoerd op een reële database. Hierdoor hebben we een beter begrip gekregen van de mogelijke problemen door 'trial and error'.

Voor ons onderzoek hebben we samengewerkt met Prof. Dr. Dominique Benoit (Vakgroep Inwendige Ziekten, Departement Intensieve Zorg 12K12IB, promotor), Maarten Bekaert (Master in Statistics sinds 2007 en nu Doctoraat student aan het Department of Applied Mathematics and Computer Sciences, U Gent), Prof. Stijn Vansteelandt (Department of Applied Mathematics and Computer Sciences, U Gent), Professor Georges Van Maele en Dr. Ellen Deschepper (Biostatistics Unit, Ghent University Hospital) en Prof. Dr. Johan Decruyenaere (Vakgroep Inwendige Ziekten, Departement Intensieve Zorg, copromotor). Prof. Dr. D. Benoit, M. Bekaert en S. Vansteelandt hielpen bij het opstellen van de criteria voor de MLRA en Prof. Decruyenaere, Prof. Van Maele en Dr. Deschepper hielpen ons door onze thesis aan een kritische lezing te onderwerpen. Voor het lezen en analyseren van de publicaties en voor de analyse van de resultaten werkten we samen met prof Dr. D. Benoit en M. Bekaert.

3. Methodologie

Critical Care Medicine en Intensive Care Medicine zijn de twee tijdschriften die we onderzochten naar het correct gebruik en /of rapportering van MLRA. We zijn begonnen met het nakijken van Critical Care Medicine omdat dit online kan nagelezen worden op Pubmed en omdat er geschreven exemplaren van aanwezig zijn in de Biomedische bibliotheek. We hebben het aantal artikels onderling verdeeld. Alle artikels in Critical Care Medicine anno 2006 werden manueel nagekeken in de “Abstract”, “Materials and Methods” en in de “Results” sectie op het gebruik van MLRA.

Om dit te doen hadden we in samenwerking met onze promotor een aantal criteria opgesteld waaraan een artikel zeker moest voldoen, o.a. gebaseerd op de studies van Concato et al.(1993) en Moss et al.(2003). Daarna hebben we een aantal artikels apart gelezen en deze vervolgens in groep besproken. De problemen die we ondervonden werden dan in deze sessies verduidelijkt. Tijdens deze sessies werd het ook snel duidelijk dat we onze criteria moesten uitbreiden omdat niet alle problemen in onze vragenlijst vervat waren (validatie van de vragenlijst). Daarna werden de gangbare criteria die opgenomen werden in onze lijst omgezet voor verwerking in SPSS.

A. Potentiële tekortkomingen:

Bij ons onderzoek hebben we de volgende mogelijke tekortkomingen nagekeken:

1. Het gebruik van een correcte selectieprocedure naargelang het doel van de MLRA (predictie of impact)

MLRA kan gebruikt worden op een predictieve of een verklarende manier. Beide vereisen enigszins een andere aanpak, althans zeker in de selectie van de onafhankelijke variabelen die in het model worden opgenomen.

Bij predictieve modellen is het hoofddoel om een binaire outcome, zoals bijvoorbeeld mortaliteit (aan- of afwezig), te voorspellen aan de hand van een minimum aantal variabelen. De bedoeling is hier om een “zo eenvoudig mogelijk”

model te bekomen zonder hierbij te veel predictieve performantie te verliezen. Om dit “parsimoneus” model te bekomen wordt vaak gebruik gemaakt van automatische selectieprocedures, zijnde procedures die de software gebruikt om variabelen te selecteren op basis van statistische criteria.

Bij verklarende modellen is het daarentegen de bedoeling om een zo accuraat mogelijke impact van een specifieke variabele (bv. nosocomiale infectie) op de outcome (mortaliteit) te berekenen, na het in acht nemen (=”adjusteren”) van potentiële confounders (bv. onderliggende ziekte die aanleiding kan geven tot nosocomiale infectie maar ook rechtstreeks invloed kan en zal hebben op de mortaliteit). Hierbij zal het van belang zijn om alle (gekende) potentiële confounders in het model op te nemen en te houden, zelfs indien ze op het eerste zicht statistisch niet belangrijk lijken en zal men aldus voorzichtig moeten omspringen met automatische statistische selectieprocedures. (zie verder)

In de praktijk is het niet altijd even gemakkelijk om predictieve van verklarende modellen te onderscheiden en vaak worden ze door elkaar gebruikt. Wanneer onderzoekers verschillende onafhankelijke variabelen samen wensen te beoordelen zonder aandacht te schenken aan één specifieke variabele hebben ze vooral een predictieve intentie, van zodra ze aandacht schenken aan één specifieke variabele hebben ze vooral een verklarende intentie.

Men kan kiezen om alle variabelen gezamenlijk in een model te houden (“entermethode” in SPSS) of via een stapsgewijze selectieprocedure (“stepwise logistic regression”) een selectie te maken van de, althans statistisch, meest belangrijke variabelen.

Bij verklarende modellen blijven alle potentiële confounders best in het model om een zo accuraat mogelijk effect van een fenomeen of therapie in te schatten. Men maakt hier aldus best gebruik van de entermethode. Bij de entermethode worden alle variabelen die geselecteerd zijn door de onderzoekers in één keer in het model gebracht en wordt er geen selectie gemaakt van de variabelen aan de hand van statistische criteria. De onderzoekers beslissen welke variabelen in het model blijven (of niet) aan de hand van a priori kennis uit de literatuur en de resultaten van de univariate analyse in de cohorte. Zo kan men enkele variabelen in het model houden die dit niveau niet halen, maar die men toch klinisch belangrijk acht. Zelfs wanneer geen uni- of multivariate significantie wordt behaald, worden deze

variabelen best in het model “geforceerd” omdat zij de odds ratio’s van de onderzochte variabele kunnen beïnvloeden. Weliswaar moet men een expert zijn in een vakgebied om te weten welke variabelen in een model moeten geforceerd worden en aldus om het model ten volle te begrijpen. Een verklaring voor het feit dat deze variabelen als klinisch belangrijk worden gezien (in vorige studies of uit eigen ervaring), maar toch niet significant blijken, is dat de studiecohorte die onderzocht wordt niet vergelijkbaar is met vorige cohorten en aldus niet representatief is. Een andere reden hiervoor kan zijn dat het model te weinig power heeft, dit wil zeggen dat er te weinig observaties zijn om een variabele als belangrijk te kunnen onderscheiden. Het niet opnemen van klinisch belangrijke variabelen in het model alleen op basis van statistische criteria kan aldus een verkeerd beeld scheppen van de relatie tussen de belangrijkste variabelen en de outcome. Indien men in dergelijke omstandigheden een automatische selectieprocedure gebruikt zullen de niet-significante, doch mogelijk klinisch relevante variabelen, automatisch verwijderd worden uit het model waardoor de berekende odds ratio van de variabele waarin we geïnteresseerd zijn onderschat of overschat zal worden, tenzij het om een volledig onafhankelijke variabele gaat. Deze procedures worden dan ook best niet gebruikt in verklarende modellen. (Hosmer and Lemeshow, 2000)

Bij predictieve modellen kan gebruik worden gemaakt van stapsgewijze automatische selectie procedures (stepwise logistische regressie). Stepwise logistische regressie is een automatische selectieprocedure waarbij variabelen in het model gehouden of verwijderd worden op basis van statistische criteria. Tijdens deze procedure zal de software nakijken in welke mate de performantie van het model om de outcome te voorspellen verandert in functie van het toevoegen of verwijderen van variabelen. De stepwise methode wordt gebruikt in de verkennende fase van het onderzoek. Er worden geen a priori veronderstellingen gemaakt naar de onderlinge relaties tussen de variabelen en het is de bedoeling om met zo weinig mogelijk variabelen een zo juist mogelijke voorspelling te maken. Dit kan echter een probleem geven, want als er geen onderlinge verbanden worden nagekeken, zou het kunnen dat bepaalde variabelen aan het model worden toegevoegd omdat ze elkaar versterken, of dat bepaalde variabelen uit het model worden weggelaten doordat ze elkaar verzwakken (zie multicollineariteit).

Door de automatische selectieprocedure kunnen relevante gegevens worden weggelaten die in deze subpopulatie niet significant zijn, maar die dit wel zijn in de algemene populatie. In kleine datasets is de selectie van de variabelen typisch onstabiel. Men verkrijgt vaak verschillende modellen afhankelijk van de procedure die gebruikt werd. Het toevoegen of verwijderen van een klein aantal gegevens kan het model aanzienlijk veranderen (Steyerberg et al., 2000). Bij grote datasets worden automatische selectieprocedures cruciaal en daarom is het noodzakelijk dat statistici het model valideren en indien nodig het model corrigeren.

Bij forward logistic regression wordt de meest significante variabele toegevoegd aan het model en daarna wordt de fitness van het model berekend. Deze stappen herhalen zich totdat er na het toevoegen van een variabele geen significante verbetering meer optreedt van het model.

Bij backward logistic regression wordt vertrokken van een verzadigd model waarbij telkens de minst significante variabele uit het model wordt weggelaten totdat er bij het weglaten van een variabele een significante verslechtering ontstaat van het model. Telkens nadat er een variabele wordt verwijderd uit het model wordt de fitness van het model getest om na te gaan indien het model nog steeds overeenkomt met de data.

2. <u>Duidelijke opsomming van de variabelen</u>

Het is belangrijk te onderstrepen dat resultaten van een MLRA een relatieve bevinding zijn en afhankelijk zijn van de variabelen die opgenomen werden in het model. Daarom is het van groot belang aan te geven welke factoren wel en welke niet in het model opgenomen werden. Onderzoekers zouden steeds in hun manuscript alle variabelen moeten opsommen die initieel in het model werden opgenomen, of nog beter deze vermelden of opsommen in de tabel die het uiteindelijk resultaat weergeeft. Ook is het belangrijk een duidelijke beschrijving te geven hoe men tot deze selectie van variabelen is gekomen. Variabelen worden best gekozen volgens de literatuur en de klinische praktijk. Er kan dan beslist worden om eerst een univariate analyse uit te voeren en om enkel variabelen te includeren die voldoen aan het vereiste significantieniveau. Er wordt aangeraden om een significantieniveau van 0.25 te hanteren als screeningscriterium. Bij gebruik van een strenger niveau (bv. 0.05) slaagt men er vaak niet in om

variabelen te identificeren die potentieel belangrijk zijn of gekend zijn als klinisch belangrijk. Het gebruik van een hoger niveau heeft het nadeel van variabelen te includeren die van bedenkelijk belang zijn. Hierdoor is het belangrijk om alle variabelen kritisch te evalueren vooraleer ze definitief in het model op te nemen. Een ander probleem bij de univariate aanpak is dat er geen rekening wordt gehouden met het feit dat een verzameling van variabelen, welke allen zwak geassocieerd zijn met de outcome, tezamen een belangrijke predictor van de outcome kunnen zijn (Hosmer and Lemeshow, 1989). Zo kan men enkele variabelen in het model houden die dit niveau niet halen, maar die men toch klinisch belangrijk acht. Ook is het nuttig om te vermelden welke variabelen men in het model forceert en waarom men dit doet. Zoals voorheen al aangehaald is het bij verklarende modellen de bedoeling om de impact van een potentieel belangrijke variabele op de outcome, na adjusteren voor potentiële confounders, na te gaan. Daarom is het van belang dat alle potentieel belangrijke confounding variabelen in het model blijven.

3. Correcte weergave van de eenheid van continue variabelen

Odds ratio's zijn een manier om aan te geven hoe waarschijnlijk een verschijnsel is vergeleken tussen twee groepen. De odds ratio is het quotiënt van de waarschijnlijkheid van een verschijnsel in een eerste groep en de waarschijnlijkheid van dit verschijnsel in de andere groep. Wanneer men resultaten onder de vorm van odds ratio's tracht te interpreteren is het van belang dat duidelijk gesteld wordt wat de eenheid is van deze odds ratio's. Wanneer men deze niet weergeeft bij continue variabelen is het moeilijk om deze ratio's te interpreteren omdat men niet kan weten indien het bijvoorbeeld gaat om een risicostijging per eenheid of een veelvoud hiervan (bv. één jaar versus één decade, één kilogram versus 10 kilogram,…) (Concato et al., 1995).

4. Correcte weergave van de referentie categorie bij categorische variabelen met meer dan twee groepen.

Als bij categorische variabelen de odds ratio's bekeken worden en wanneer er slechts twee categorieën worden gebruikt, dan wordt de ene categorie altijd vergeleken ten opzichte van de andere categorie. Als er echter meer dan twee categorieën gebruikt worden is het belangrijk om te weten welke de

referentiecategorie is. Indien de referentiecategorie niet gespecificeerd wordt weet met men niet met welke groep er vergeleken wordt. De odds ratio van de referentiecategorie wordt gelijk gesteld aan 1.

5. Correcte codering van het eindpunt

De meeste gedichotomiseerde outcomes zullen een dichotomie vertonen tussen een gunstige en een ongunstige observatie. Wanneer we het risico beschrijven kan dit verwijzen naar het risico voor een gunstige of een ongunstige observatie. Zo kan een daling van het risico dus goed of slecht zijn. Het is belangrijk duidelijk te omschrijven of een observatie de gunstige of de ongunstige outcome voorstelt, want de resultaten veranderen als we de gunstige en de ongunstige outcome omdraaien. Er zijn dus twee risico ratio's: de odds ratio voor een goede outcome en de odds ratio voor een slechte outcome, die elkaars reciproque zijn. De algemene regel (volgens de Cochrane Collaboration) is dat voor outcomes die we proberen te verhinderen (bv. mortaliteit, terugkeren of verergeren van symptomen, …), het best is om de observatie te omschrijven als de ongunstige outcome, wat normaal gezien de intuïtieve keuze is. Voor outcomes waar men probeert de gezondheid te verbeteren (bv. genezing, verdwijnen van symptomen, …) is er nog geen consensus. Sowieso moet duidelijk gesteld worden welke outcome men gebruikt. Een andere onduidelijkheid die kan ontstaan is het aangeven van een ander eindpunt in "Materials and Methods" dan het eindpunt dat in werkelijkheid gebruikt wordt in de resultaten (The Cochrane Collaboration open learning material, 2009).

6. Respecteren van de 10/1 regel

Multivariabele methoden van analyse geven problematische resultaten wanneer te weinig outcome observaties beschikbaar zijn in verhouding tot het aantal onafhankelijke variabelen. Overfitting doet zich voor wanneer te veel variabelen, waarvan sommige "ruisvariabelen" zijn, geselecteerd worden voor het definitieve model. Het praktisch gevolg hiervan is dat het gemaakte model van toepassing is op de steekproef (die afwijkt van de vooropgestelde populatie), maar dat de resultaten hiervan niet meer reproduceerbaar zijn bij andere gelijkaardige steekproeven. Men kan dan dus geen goede voorspelling doen aan de hand van

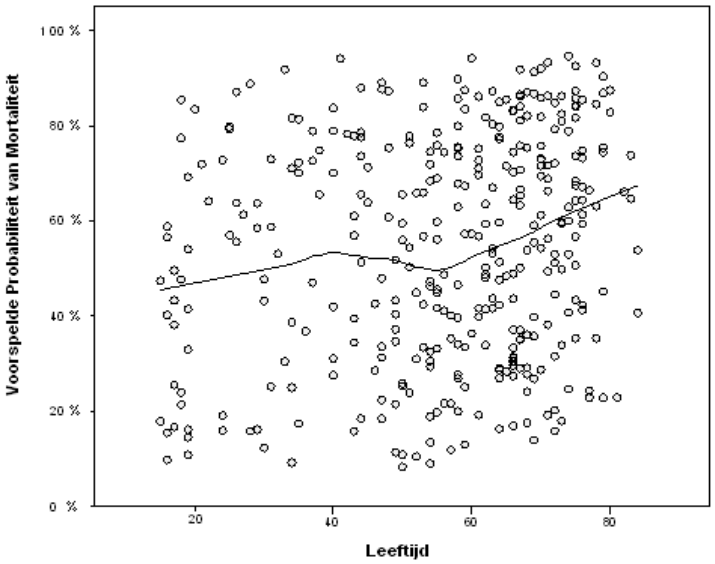
enkele variabelen, of geen correcte impact van een variabele berekenen. De algemene regel bij multivariabele logistische regressiemodellen is dat een minimum van 10 observaties per predictor variabele (EPV, events per predictor variable) wenselijk is om de validiteit van het model te bewaren. Dit is gebaseerd op simulatie studies. De resultaten hiervan toonden aan dat indien niet wordt voldaan aan de 10/1 regel, er een toenemende bias is in twee richtingen: de odds ratio's waren soms verhoogd en soms verlaagd. In ons onderzoek hebben we deze 10/1 regel niet strikt toegepast. Enkel in gevallen van grove schending van de 10/1 regel hebben we dit zo genoteerd. (Concato et al. 1993 en 1995, Peduzzi et al. 1995 en 1996).

7. Onderzoeken lineair verband continue variabelen

Het verband tussen een continue variabele en de outcome is lineair als het veranderen van de probabiliteit (logit) voor de outcome constant is voor alle waarden van de continue variabele. Als een continue variabele in een MLRA model wordt opgenomen, wordt automatisch verondersteld dat deze variabele een lineair verband heeft met de outcome. Dit is echter niet altijd het zo. Daarom is het belangrijk om niet-lineariteit na te gaan, want standaard multipele regressie analyse kan enkel het verband tussen de afhankelijke variabele en onafhankelijke continue variabelen correct schatten als de verbanden lineair zijn. Als het verband tussen de afhankelijke en de onafhankelijke variabelen niet lineair is, zal de regressie analyse het verband van de variabelen onderschatten. Er is dan een verhoogd risico voor een type II-fout (kans op het ten onrechte aanvaarden van de nulhypothese), wat de uiteindelijke conclusie van het model kan veranderen (Osborne and Waters, 2002). Het modelleren van een continue variabele helpt ook om inzicht te krijgen in de reële verbanden met de outcome zoals in de "reële wereld", althans wanneer men te maken heeft met een grote database van goede kwaliteit. Een mooi voorbeeld in de IZ geneeskunde is de verhouding tussen BMI en mortaliteit bij geventileerde patiënten op IZ. Twee studies met een groot aantal patiënten hebben aangetoond dat het verband tussen BMI en mortaliteit niet lineair, maar U-vormig is. Verrassend genoeg bleek uit deze studie dat enkel een zeer lage BMI statistisch significant geassocieerd is met een hogere mortaliteit (Garrouste-Orgeas, 2004; O'Brien, 2006).

Er zijn verschillende methoden om het probleem van niet-lineariteit na te kijken. De meest gebruikte methodes zijn de “methode van kwartielen” en de “Locally Weighted Scatterplot Smoothing” (LOESS curve). In de eerste methode wordt de continue variabele gecategoriseerd in verschillende categorieën volgens kwartielen of percentielen, afhankelijk van het totale aantal observaties. De categorieën worden dan in het model ingebracht en er wordt dan nagegaan indien de bekomen odds ratio’s een lineair verband hebben met het eindpunt. Volgens de LOESS methode wordt een curve gefit naar het model, door kleine modellen te fitten van lokale deeltjes van de dataset. Op deze manier wordt een functie bepaald die het kerngedeelte van de dataset vastlegt (Härdle, 1992). Het nagaan van lineaire verbanden wordt belangrijker naarmate er meer gegevens in de dataset zijn, want dan wordt het verband ook duidelijker (zie n ≥ 500) (Stone-Romero and Rosopa, 2008; McKillup, 2006).

Als voorbeeld kunnen we een database van 310 kritisch zieke patiënten met een hematologische maligniteit gebruiken. Als we in een MLRA model het verband tussen de continue variabele leeftijd en de outcome mortaliteit met een LOESS-curve nakijken na het adjusteren voor belangrijke covariaten, dan merken we dat er geen lineair verband is. In figuur 4 zien we dat de probabiliteit om te sterven lichtjes toeneemt tussen 20 en 50 jaar, stabiliseert tussen 50 en 55 jaar en enkel significant stijgt boven de 55 jaar. Als de leeftijd in het model wordt geïncludeerd als een dichotome variabele (< en > 55 jaar), worden bepaalde variabelen uit het model weggelaten.



Figuur 4:

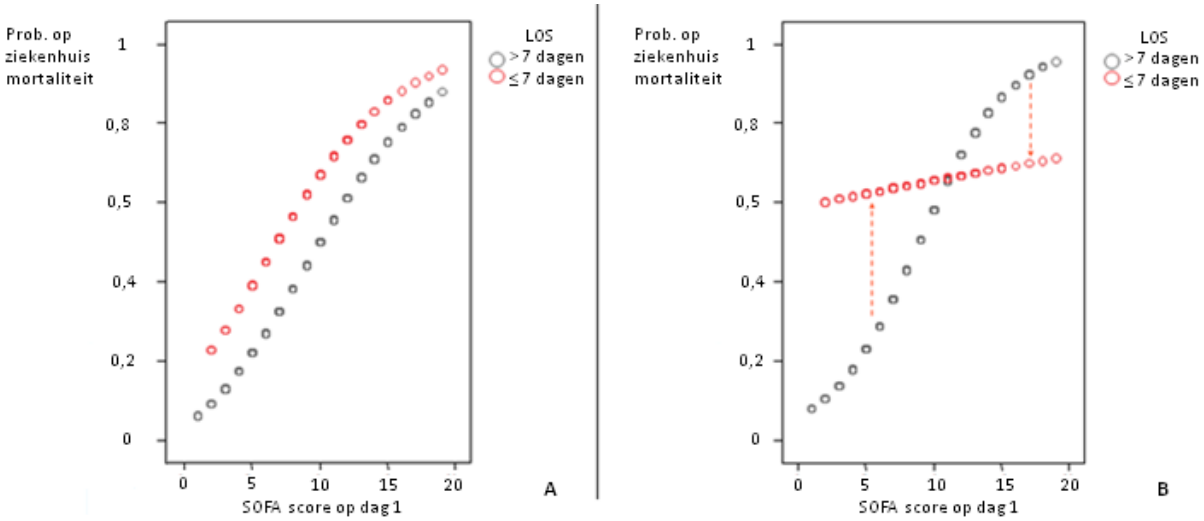
Verband tussen leeftijd en mortaliteit, waarbij de volle lijn de LOESS-curve voorstelt.

8. Onderzoeken van interactietermen

In elk model moet nagegaan worden of er een klinisch significante interactieterm aanwezig is. Dit betekent dat er een interactie is tussen twee variabelen waarvoor het effect van één van de variabelen niet constant is over het verloop van de andere. Het nakijken van een interactieterm komt op hetzelfde neer als de analyse van een effect van een bepaalde variabelen binnen verschillende subgroepen. Een interactieterm wordt mathematisch in het model opgenomen door het product te nemen van twee variabelen en beide variabelen in het model te houden. In figuur 5 wordt een voorbeeld uit een reële database gegeven.

Eerst wordt een lijst van mogelijke paren van variabelen gemaakt die enige wetenschappelijke basis hebben om een interactie te maken met elkaar. De interactievariabelen worden, één voor één, in het model opgenomen en hun significantie wordt beoordeeld aan de hand van de likelihood ratio test. Interactietermen moeten gekozen worden op traditionele niveaus (0.05) van statistische significantie. Inclusie van een interactieterm in het model die niet significant is doet typisch de geschatte standaardfouten stijgen zonder de

puntschattingen te veranderen. In het algemeen moet een interactieterm statistisch significant zijn om zowel punt- en intervalschattingen te veranderen. De finale beslissing of een interactieterm al dan niet opgenomen moet worden in het model moet zowel gebaseerd zijn op statistische, als op praktische overwegingen. Elke interactieterm in het model moet zin hebben vanuit een klinisch perspectief (Hosmer and Lemeshow, 1989). Het niet onderzoeken van interactietermen werd in ons onderzoek pas aanzien als een fout indien meer dan 500 observaties in het model zaten. Hoe meer observaties men namelijk in een model heeft, hoe gemakkelijker het is om een duidelijke interactie tussen twee variabelen waar te nemen (Kleinbaum, 1994).



Figuur 5:

Voorbeeld uit een reële database: Op figuur 5 A is te zien dat wanneer men de SOFA score (orgaanfalen score) op dag één uitzet tegenover de probabiliteit om te sterven in het ziekenhuis, de probabiliteit op sterven stijgt naarmate deze score stijgt. Het lijkt er tevens op dat patiënten die langer dan 7 dagen op IZ verblijven, een 20% exces aan mortaliteit hebben onafhankelijk van de SOFA score op dag 1. Na het introduceren van een interactieterm tussen de SOFA score en het aantal ligdagen > 7 is op figuur 5 B echter te zien dat de impact van > 7 dagen opname op IZ niet dezelfde is bij patiënten met weinig en veel orgaanschade op dag 1. Patiënten met weinig orgaanschade bij opname die toch langer dan 7 dagen op IZ verblijven hebben een exces aan mortaliteit van ongeveer 40% terwijl deze met veel orgaanschade en langer dan 7 dagen verblijven op IZ een daling in mortaliteit hebben van 20%. Dit is logisch: minder zieke patiënten die > 7 dagen op IZ liggen hebben een hoger mortaliteit omdat ze vermoedelijk een complicatie hebben opgelopen, terwijl dat ernstig zieke patiënten die > 7 dagen op IZ liggen een lagere mortaliteit hebben dan verwacht gezien het gewoon betekent dat ze ondertussen niet overleden zijn (bias door drop-out van de zwaarst zieke patiënten).

9. Onderzoeken van collineariteit

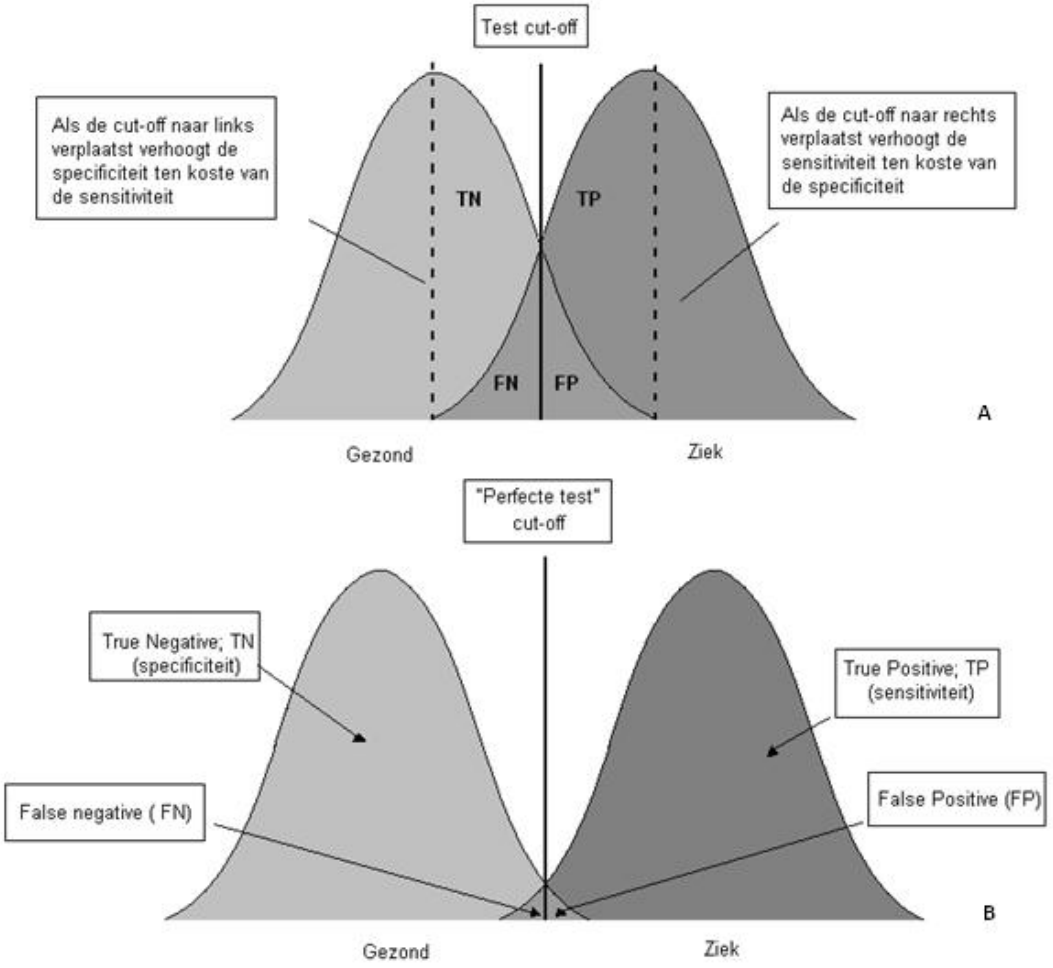
Multicollineariteit in een logistisch regressiemodel refereert naar de collineariteit (lineaire combinatie van twee variabelen) van twee of meerdere onafhankelijke variabelen. Collineariteit ontstaat wanneer variabelen in het model worden opgenomen die sterk met elkaar gecorreleerd zijn. Door de sterke correlatie is er een overbodigheid van informatie van de variabelen met een overschatting van de variantie van de parameters als gevolg. Dit kan, vooral in kleine populaties, ervoor zorgen dat sommige onafhankelijke variabelen op zichzelf statistisch niet significant zijn, maar door de lineaire combinatie van die variabelen toch leiden tot een sterk significant eindmodel. Door de overschatting van de variantieparameter kunnen er verkeerde verbanden gelegd worden tussen de onafhankelijke en afhankelijke variabelen en dus ook verkeerde conclusies getrokken worden.

Multicollineariteit kan opgespoord worden door de correlaties tussen onafhankelijke variabelen na te kijken, maar er kan ook gebruik gemaakt worden van multicollineariteit diagnostische statistiek. Het probleem van multicollineariteit kan op verschillende manieren worden opgelost. Er kan een gemiddelde berekend worden van de variabelen en dan wordt dit gemiddelde gebruikt als een onafhankelijke variabele in het logistische regressie model. Bepaalde variabelen kunnen na overleg uit het model weggelaten worden en sommige variabelen kunnen worden gecombineerd. Bv. gewicht en lengte kunnen in een bepaald model een collineair verband hebben en dit kan opgelost worden door er één variabele van te maken, namelijk BMI. (Dohoo et al., 1997)

10. Weergave van discrimination statistics (ROC-curve)

Een ROC (Receiver Operating Characteristic) curve geeft de probabiliteit van een goede voorspelling (sensitiviteit) en een valse voorspelling (1-specificiteit) weer voor het geheel aan mogelijke afkapwaarden. Een ROC-curve is een diagnostisch middel om de keuze te vergemakkelijken tussen een dichotoom testresultaat. Deze curve biedt de mogelijkheid optimale modellen te behouden en suboptimale modellen achterwege te laten. Naargelang de behoefte van de onderzoeker kan de ideale afkapwaarde gekozen worden. Een lage afkapwaarde wordt voornamelijk gebruikt bij screening, terwijl een hoge afkapwaarde meer wordt gebruikt bij diagnostische tests. Als er een lage afkapwaarde wordt gekozen is de drempel

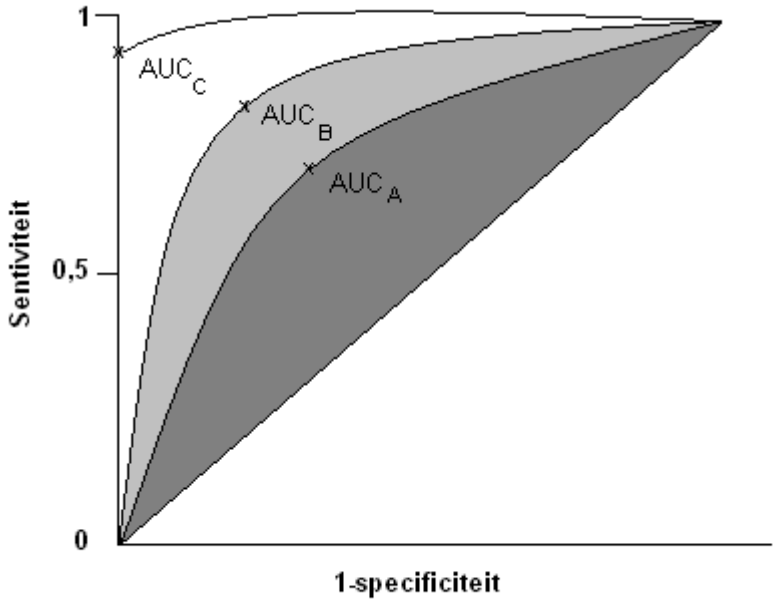
verlaagd om een positief resultaat te hebben. De sensitiviteit van de test is verhoogd, maar door de hoge sensitiviteit is de specificiteit verlaagd, want er zijn nu meer mensen die een positief resultaat hebben zonder dat ze werkelijk een positief resultaat hebben. Als er een hoge afkapwaarde wordt gekozen is de drempel verhoogd om een positief resultaat te hebben. De specificiteit wordt verhoogd, maar door de hoge drempel zullen sommige mensen die een positief resultaat hebben niet gediagnosticeerd worden. Dit betekent dat de sensitiviteit is verlaagd is (Søreide, 2009).



Figuur 6:

(A) Een “levensechte” situatie waarbij de populaties een aanzienlijke overlap vertonen in het testspectrum. Dit zorgt voor een verminderd onderscheidend vermogen van de test. De sensitiviteit kan verbeterd worden ten koste van de specificiteit door de cut-off voor de test te veranderen of omgekeerd. TP, true positive; TN, true negative. (B) Een goede onderscheidende test met een bijna perfect onderscheidend vermogen tussen de zieke en de gezonde populatie. De twee populaties tonen een kleine overlap in het testspectrum wat bijdraagt tot het goed onderscheidend vermogen van de test. (Søreide, 2009)

De AUC (area under the ROC-curve), die kan variëren van 0 tot en met 1, geeft een beschrijving van de mogelijkheid om te discrimineren tussen de gevallen die al dan niet een positieve outcome zullen vertonen. Een ROC-curve met een AUC die gelijk is aan 0.5, is een nonsens test. Als algemene regel wordt gesteld dat een AUC tussen 0.7 en 0.8 als aanvaardbare discriminatie wordt beschouwd en tussen 0.8 en 0.9 beschouwt men dit als excellente discriminatie. Is de AUC echter boven 0.9, dan wordt dit gezien als een uitstekende discriminatie. Bij 1 is er een perfecte discriminatie tussen positieve en negatieve resultaten.



Figuur 7:

Drie ROC-curves en hun respectievelijke oppervlakte onder de curve (AUC) zijn te zien. De diagnostische accuraatheid van model C (witte oppervlakte) is beter dan deze van model B en A, aangezien de AUC van C>B>A.

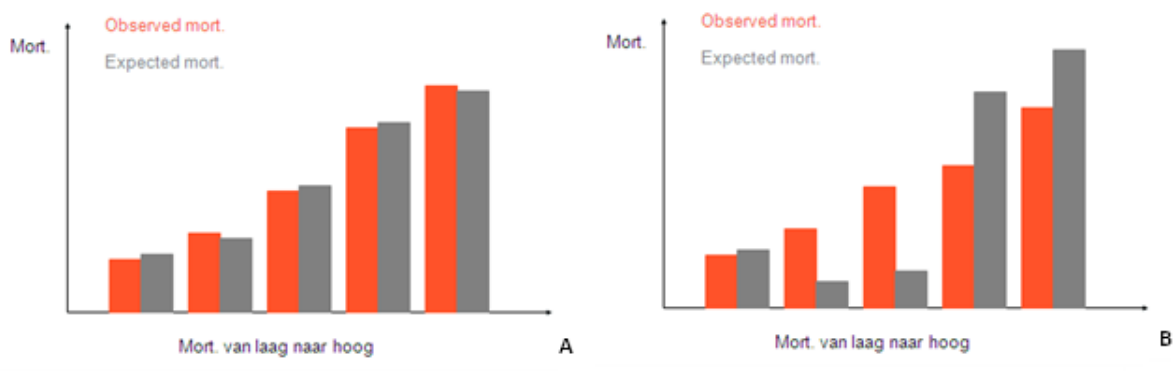
Een ROC-curve moet zeker gegeven worden bij logistische regressie modellen wanneer er een intentie is om een predictie te doen. Aan de hand van de AUC kan men namelijk afleiden of het predictieve model betrouwbare voorspellingen kan doen. Naar ons inziens zou men ook bij modellen, die de intentie hebben om de impact van een variabele op de outcome na te gaan, een ROC-curve mogen weergeven. Dit geeft namelijk een idee over de residuele confounding en de performantie van het model (Hosmer and Lemeshow, 1989).

In ons onderzoek werd ook nagegaan of een Nagelkerke statistic (R^2) weergegeven werd. Een Nagelkerke statistic geeft een idee van het discriminerend vermogen van een model. R^2 waarden zijn bij logistische regressie typisch laag vergeleken met lineaire regressie. Dit geeft een probleem wanneer men deze waarden rapporteert aan een publiek dat gewoon is om waarden te zien zoals deze in lineaire regressiemodellen. Routinematig weergeven van R^2 waarden wordt dus niet aangeraden, maar kan wel behulpzaam zijn bij het bouwen van een model om verschillende modellen met elkaar te vergelijken. (Hosmer and Lemeshow, 1989)

11. Weergave van calibration statistics (Hosmer-Lemeshow test)

De goodness-of-fit van een statistisch model beschrijft hoe goed het logistisch regressie model past of overeenkomt bij een set van observaties. De Hosmer-Lemeshow test is een calibratietest die gebruikt wordt bij logistische regressie analyse. Het is een test die kijkt hoe goed het model overeen komt met de werkelijke populatie.

De Hosmer-Lemeshow statistiek bekijkt de goodness-of-fit door tien geordende groepen te maken van subjecten zoals in tabel 1. De tien geordende groepen worden gemaakt volgens verwachtte probabiliteiten voor de outcome van 0 tot 0,1; 0,1 tot 0,2 en zo verder tot 0,9 tot 1. Elk van deze groepen wordt verder verdeeld in twee categorieën: succes /falen of afwezig /aanwezig. De verwachtte frequenties voor elk van de cellen zijn verkregen uit het model. Indien het model goed is, worden de subjecten met succes in de hogere decielen geplaatst en diegene met falen in de lagere decielen. In elke categorie wordt een onderlinge vergelijking gemaakt tussen de werkelijke populatie in de groep (observed) en tussen het voorspelde aantal door het logistische regressie model (expected). Er wordt dus een chi^2- test uitgevoerd, waar niet-significantie de verhoopte uitkomst is, want dan kan er geen significant verschil worden aangetoond tussen het predictieve model en de populatie (zie figuur 8 en tabel 1) (Hosmer and Lemeshow, 2000).



Figuur 8:

Op figuur 8 A is er geen significant verschil tussen observed en expected. Er is dus een goede goodness-of-fit van het model. Op figuur 8 B is er een significant verschil tussen observed en expected. Dit betekent dat het logistische regressiemodel geen goede weergave geeft van de onderzochte populatie.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10,239	8	,249

Contingency Table for Hosmer and Lemeshow Test

		in hospital mortality = nee		in hospital mortality = ja		
		Observed	Expected	Observed	Expected	Total
Step 1	1	32	31,346	8	8,654	40
	2	24	27,652	16	12,348	40
	3	22	25,204	18	14,796	40
	4	27	21,687	13	18,313	40
	5	21	18,232	19	21,768	40
	6	17	15,298	23	24,702	40
	7	11	12,473	29	27,527	40
	8	7	10,009	33	29,991	40
	9	11	8,150	29	31,850	40
	10	3	4,951	36	34,049	39

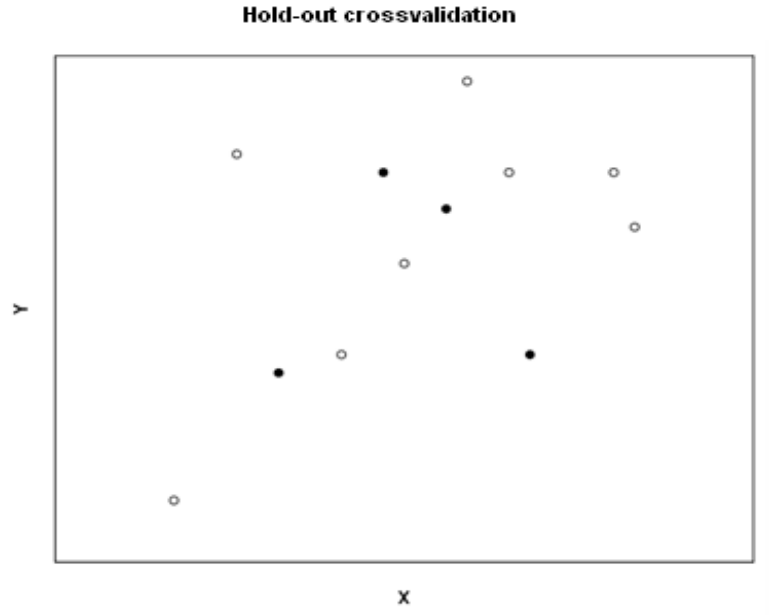
Tabel 1:

Er wordt een chi²- test gedaan om te onderzoeken indien er een significant verschil bestaat tussen het logistische regressiemodel en de onderzochte populatie. In dit voorbeeld is de test niet-significant en kan de nulhypothese (er is geen verschil tussen de waargenomen en de verwachte waarden) niet verworpen worden.

12. Uitvoeren van crossvalidatie

Crossvalidatie is een methode die voornamelijk wordt gebruikt bij predictieve modellen om na te gaan of de resultaten van een gebruikte statistische analyse toepasbaar zijn op een onafhankelijke database. Het doel is dus om het level-of-fit te schatten van een model op een dataset die onafhankelijk is van de data die gebruikt werden om het model te maken. Het belang van crossvalidatie neemt toe bij een groter wordende dataset (zie n ≥ 500). Crossvalidatie is op verschillende manieren mogelijk.

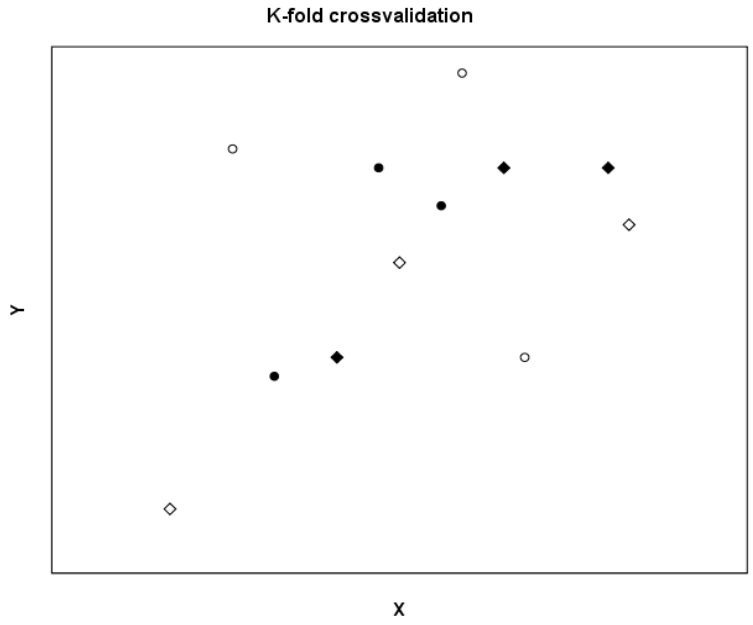
Bij de simpelste vorm van crossvalidatie (Hold-out crossvalidatie) wordt de data at random in een trainingset (70%) en een testset (30%) verdeeld (Trujillano, 2008). De statistische analyse wordt uitgevoerd op de trainingset en de bekomen resultaten worden dan gevalideerd op de testset. Om de variabiliteit te verminderen is het mogelijk om meerdere crossvalidaties uit te voeren met telkens verschillende at random verdelingen van de dataset, terwijl de validatieresultaten uitgemiddeld worden over het aantal ronden (Efron and Tibshirani, 1993).



Figuur 9:

1. Kies at random 30% van de gegevens als testset.
2. De resterende gegevens zijn de trainingset.
3. Voer de regressie uit op de trainingset.
4. Valideer de resultaten op de testset.

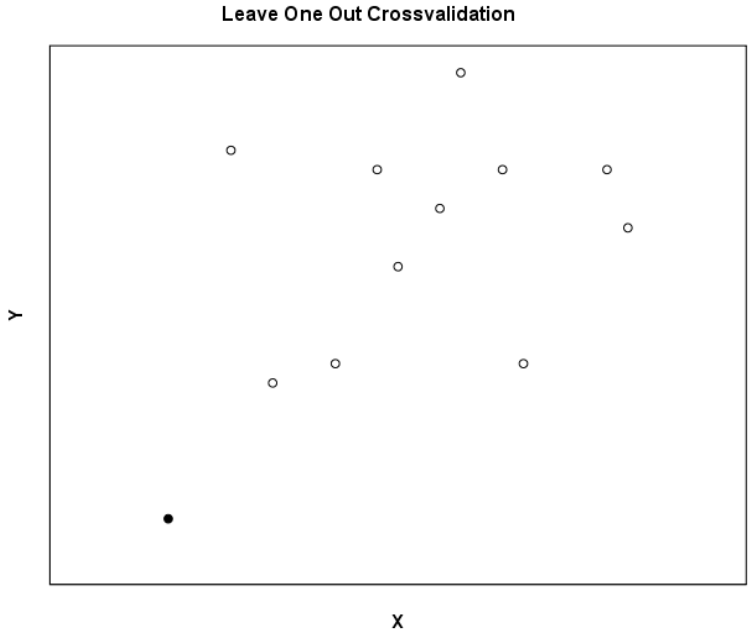
In *K*-fold crossvalidatie wordt de dataset verdeeld in *K*-delen. Van de *K*-delen wordt één enkel deel weerhouden als de testset terwijl de overgebleven *K*-1 delen worden gebruikt als trainingset. Dit proces wordt *K*-keer herhaald waarbij elk *K*-deel één keer wordt gebruikt. De resultaten van de validatieset worden dan uitgemiddeld over het aantal *K*-delen en hiermee wordt er een schatting van het level-of-fit bekomen. Vaak wordt er gekozen voor 10-fold crossvalidatie (Baumann, 2003; Efron and Tibshirani, 1993).



Figuur 10:

1. Verdeel de gegevens at random in *K*-delen. K= 4
2. Gebruik elk deel eenmaal als testset terwijl de rest de trainingsset is.
3. Voer de regressie uit op de trainingsset.
4. Geef de gemiddelde fout weer.

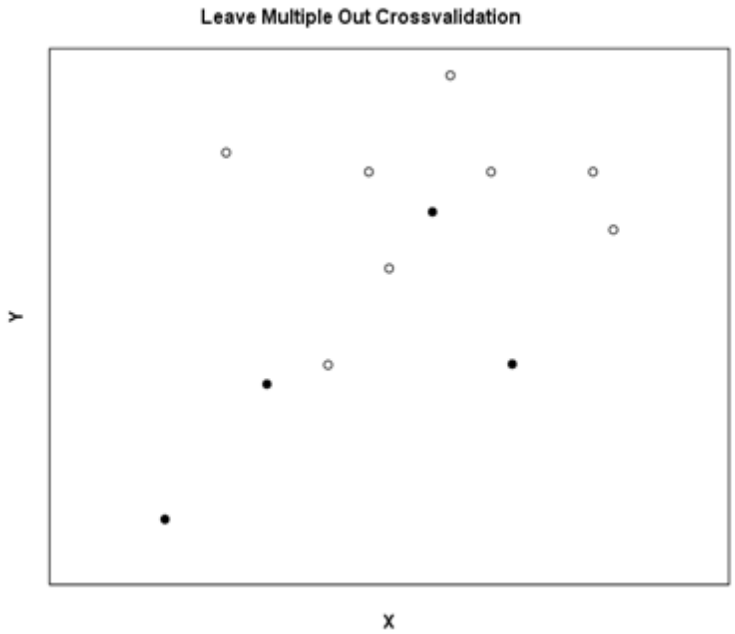
In Leave-One-Out crossvalidatie wordt één observatie van de dataset gebruikt als de testset, terwijl de rest van de gegevens gebruikt wordt als trainingset. Dit wordt herhaald totdat elke observatie 1 keer gebruikt werd als testset (Baumann, 2003; Efron and Tibshirani, 1993).



Figuur 11:

1. Kies één observatie en gebruik deze als testset.
2. De overgebleven observaties zijn de trainingsset.
3. Voer de regressie uit op de trainingsset.
4. Doe dit voor elke variabele.
5. Geef de gemiddelde fout weer.

In Leave-Multiple-Out crossvalidatie worden meerdere gegevens van de dataset gebruikt als testset. Er zijn M observaties die verklaard worden door N variabelen. Het aantal variabelen N dat wordt gebruikt als testset is $1 \leq N < M$ en de trainingsset is M-N. Theoretisch gezien is $N = M - M/ (\ln(M)-1)$ het ideale aantal (Baumann, 2003).



Figuur 12:

1. Kies N variabelen als testset.
2. De overgebleven variabelen zijn de trainingsset.
3. Voer de regressie uit op de trainingsset.
4. Valideer de resultaten op de testset.

Bootstrapping is een valideringstechniek die soms gebruikt wordt in plaats van crossvalidatie. Uit een dataset met N gegevens worden at random n gegevens gekozen waarbij, in tegenstelling tot bij crossvalidatie, elk gegeven meerdere malen mag gebruikt worden. Deze nieuwe dataset wordt gebruikt als trainingsset, terwijl de gegevens die niet werden gekozen uit de originele dataset gebruikt worden als testset. Dit proces wordt een voorafbepaald K-keer herhaald. Hoogstwaarschijnlijk zal het aantal gegevens in de testset variëren per K-keer. (zie figuur 13). De ware fout wordt geschat als het gemiddelde van het aantal fouten van de testsets (Steyerberg et al., 2000).

Figuur 13:

Visuele voorstelling van bootstrapping.

13. Rekening houden met het tijdsaspect

Wanneer men een logistisch regressiemodel maakt, moet men rekening houden met het ogenblik wanneer de variabelen verzameld werden. Aangezien logistische regressie geen rekening houdt met de tijdsequentie, maar enkel met de correlaties, zou men in predictieve modellen waarbij men bv. naar de IZ mortaliteit kijkt best enkel variabelen opnemen die in de eerste 24u of in de eerste dagen van opname voorkomen. Variabelen in het model opnemen die in de tijd zeer dicht staan bij het eindpunt (bv dezelfde dag voorkomen als het eindpunt), is niet echt zinvol aangezien ze geen reële predictieve waarde hebben, maar wel zeer sterk gecorreleerd zijn met het eindpunt. Daarbij komt nog, zeker bij verklarende modellen, dat men het risico loopt om de oorzaak en het gevolg samen in het model op te nemen. Wanneer men observaties in het model opneemt die zich later voordoen (die eigenlijk als gevolg kunnen gezien worden in plaats van als oorzaak), zullen deze vaak in het model behouden worden aangezien het gevolg meestal dichter (sterker gecorreleerd) staat bij het eindpunt. Zo zullen deze observaties als onafhankelijke (potentieel beïnvloedbaar) predictor worden beschouwd en kunnen ze eventueel de echte oorzakelijke predictoren uit het model verdringen. Daarbij kan ook nog het probleem van collineariteit ontstaan.

B. Classificatie van de tekortkomingen

Bij ons onderzoek hebben we bovenstaande tekortkomingen geclassificeerd als ernstig, middelmatig en minder ernstig. Als er ten minste één van de voorwaarden voor een tekortkoming voldaan was, werd dit beschouwd als een tekortkoming. De ernstige tekortkomingen zijn: overfitting volgens de 1/10 regel, geen ROC-curve weergeven bij predictieve modellen en onduidelijke vermelding van de variabelen die in het initiële model werden opgenomen. Middelmatige tekortkomingen zijn: onduidelijke vermelding van de selectieprocedure, het niet-onderzoeken van interactietermen in modellen met meer dan 500 observaties en het niet-onderzoeken of er voldaan is aan een lineaire gradiënt in modellen met meer dan 500 observaties. Als minder ernstige tekortkomingen classificeerden we onduidelijke codering van het eindpunt, het niet-onderzoeken van collineariteit, het niet-weergeven van een calibration statistic, het niet-weergeven van de eenheid van continue variabelen, een onjuiste selectieprocedure in verklarende modellen en het niet-uitvoeren van crossvalidatie in modellen met meer dan 500 observaties. De tijdssequentie tussen variabelen werd niet nagekeken, gezien het voor de niet-medische expert moeilijk was om na te gaan indien oorzaak en gevolg samen werden opgenomen in het model. Ook is het voor de niet-medische expert moeilijk om te weten welke variabelen potentieel eerst of later in de tijd kwamen indien de auteurs dit niet vermeldden in hun artikel.

De resultaten worden telkens weergegeven naargelang er meer of minder dan 500 observaties waren in het model. Aangezien 500 observaties als een aanzienlijke databank kan beschouwd worden, is het van belang om er zo volledig en betrouwbaar mogelijke resultaten uit te verkrijgen.

C. Invloed van enkele kwalitatieve verschillen tussen de artikels

In ons onderzoek hebben we ook onderzocht of een aantal kwalitatieve verschillen tussen de artikels een invloed hadden op de kwaliteit van de rapportage van de MLRA. Zo hebben we ten eerste onderzocht wat de invloed was indien het artikel een “feature” artikel was. Een feature artikel is een speciaal of belangrijk artikel waarin iets nieuws wordt uitgelegd of waarin nieuwe bevindingen worden besproken die belangrijk worden geacht voor de doorsnee lezer. Ten tweede hebben we onderzocht wat de invloed was indien het artikel een “continuing medical education” (CME) artikel was. Een Continuing Medical Education artikel is een artikel waar men na het lezen een vragenlijst kan invullen over de inhoud van het artikel. Die vragenlijst wordt gebruikt voor de accreditering van artsen en heeft als doel om de huidige kennis te onderhouden en de nieuwe kennis efficiënt te gebruiken om zo de kwaliteit van de medische zorg te verbeteren voor de patiënten en de gemeenschap. Aangezien “feature” en “CME” artikels enkel voorkomen in CCM en niet in ICM hebben we enkel bij CCM artikels de invloed ervan kunnen onderzoeken. Ten derde werd ook het effect onderzocht indien er een biostatisticus betrokken was bij de publicatie (biostatisticus of epidemioloog in de lijst met auteurs). Een biostatisticus is een statisticus die zijn vaardigheden en kennis toepast in gezondheidsgeoriënteerde domeinen. Ze werken vaak samen met andere onderzoekers uit andere gezondheidsdisciplines en zijn gespecialiseerd in het opzetten van studies, de datacollectie en het analyseren van de dataset. Ten vierde hebben we ook onderzocht wat het effect was indien er een editoriaal over het artikel geschreven was. Ten laatste onderzochten we in CCM wat het effect was van het percentage PhD’s in de auteurslijst op de tekortkomingen. Dit hebben we enkel in CCM nagekeken omdat in ICM niet vermeld werd als een auteur een doctoraat had.

4. Resultaten uit de Intensieve Zorgen literatuur gepubliceerd in 2006

A. Resultaten Critical Care Medicine (CCM)

In CCM melden 70 artikels van de 346 het gebruik van MLRA en in 61 artikels werden de resultaten ook getoond. In de 61 artikels werden 170 modellen gevonden. Deze modellen werden opgesplitst in twee groepen volgens de grootte van de steekproef (zie classificatie tekortkomingen). In tabel 2 worden de resultaten van CCM getoond volgens de grootte van de steekproef. In de eerste groep zijn er 100 modellen met n < 500 en in de tweede groep zijn er 70 modellen met n ≥ 500. De initiële variabelen in het model zijn niet gespecificeerd in 9,0% van de n < 500 groep en in 7,1% van de n ≥ 500 groep. Er is overfitting volgens de 1/10 regel in 53,0% van de modellen met n < 500 en in 42,9% van de modellen met n ≥ 500. Er wordt geen ROC-curve voorzien in 86,0% van de modellen met n < 500. In 33 van de 100 modellen werd aan predictie gedaan waarvan er in 84,8% van de modellen geen ROC-curve werd berekend in de groep n < 500. In de groep met n ≥ 500 model was dit respectievelijk zo in 87,1% van de modellen en in 80,0% van de 25 predictieve modellen. In acht van de 170 modellen in CCM wordt een Nagelkerke statistic weergegeven in plaats van een ROC-curve met corresponderende AUC. De selectieprocedure wordt niet verduidelijkt in 16,0% van de groep n < 500 en in 20,0% van de groep n ≥ 500 en er wordt een automatische selectieprocedure in de verklarende modellen gebruikt in respectievelijk 68,7% en 46,7%. De conformiteit naar de lineaire gradiënt wordt in 80,6% van de modellen met n < 500 en in 66,7% van de modellen met n ≥ 500 niet nagekeken. Er werd geen controle voor interactietermen gedaan in 92,0% van de modellen in de groep n < 500 en in 68,6% van de modellen van de n ≥ 500 groep. De eenheden van continue variabelen worden niet getoond in 32,4% van de 37 modellen met continue variabelen in de groep n < 500 en in 44,7% van de 38 modellen met continue variabelen in de n ≥ 500 groep. De calibratiestatistiek werd niet weergegeven in 79,0% van de modellen van de groep n < 500 en in 65,7% van de modellen van de n ≥ 500 groep. De collineariteit werd niet nagekeken in

respectievelijk 95,0 % en 85,7% van de modellen. In 16% van de modellen in groep n < 500 was er een incorrecte codering van de eindpunten en in de n ≥ 500 groep was dit 27,1%. Crossvalidatie werd in geen enkel model van de groep n < 500 uitgevoerd en slechts in vier modellen van de n ≥ 500 groep. De resultaten worden weergegeven in tabel 2 met de corresponderende percentages tussen haakjes.

Resultaten CCM	n < 500	n ≥ 500
	n=100	n=70
De initiële variabelen in het model zijn niet gespecificeerd	9 / 100 (9,0)	5 / 70 (7,1)
Overfitting volgens de 1/10 regel	53 / 100 (53,0)	30 / 70 (42,9)
Geen weergave van een ROC-curve	86 / 100 (86,0)	61 / 70 (87,1)
Predictieve modellen waar geen ROC-curve wordt voorzien	28 / 33 (84,8)	20 / 25 (80,0)
De selectieprocedure wordt niet verduidelijkt	16 / 100 (16,0)	14 / 70 (20,0)
Automatische selectieprocedure in verklarende modellen	46 / 67 (68,7)	21 / 45 (46,7)
De conformiteit naar de lineaire gradiënt wordt niet nagekeken	75 / 93 (80,6)	44 / 66 (66,7)
Er wordt geen controle voor interactietermen gedaan	92 / 100 (92,0)	48 /70 (68,6)
De eenheden van continue variabelen worden niet weergegeven	12 / 37 (32,4)	17 / 38 (44,7)
Calibratiestatistiek wordt niet weergegeven	79 / 100 (79,0)	46 /70 (65,7)
Collineariteit wordt niet besproken.	95 / 100 (95,0)	60 / 70 (85,7)
Een incorrecte codering van de eindpunten	16 / 100 (16,0)	19 / 70 (27,1)
Crossvalidatie wordt niet uitgevoerd	100 / 100 (100)	66 / 70 (94,3)

Tabel 2:

De resultaten van CCM volgens n < 500 en n ≥ 500.

In tabel 3 hebben we de ernst van de tekortkomingen nagekeken bij Feature artikels en Continuing Medical Education artikels. In tabel 4 hebben we hetzelfde gedaan bij artikels waarover een editoriaal geschreven is en artikels die gebruik maken van een biostatisticus. Vijftien van de 170 multivariabele logistische

regressiemodellen zijn afkomstig uit feature artikels. Dertien (87%) modellen uit feature artikels en 93 (60%) modellen uit niet-feature artikels hebben ernstige tekortkomingen. Een chi²-test werd uitgevoerd met als uitkomst een p-waarde van 0,051. Dit is een randsignificante p-waarde die aantoont dat modellen uit feature artikels meer kans hebben op ernstige tekortkomingen dan modellen uit niet-feature artikels. Er zijn 4 (27%) modellen uit feature artikels en 66 (43%) modellen uit niet-feature artikels die middelmatige tekortkomingen hebben. De berekende p-waarde van de chi²-test (0,28) is niet significant. Alle modellen uit de feature artikels en uit de niet-feature artikels hebben minder ernstige tekortkomingen. De p-waarde van de chi²-test is 0,999. Er zijn 24 van de 170 modellen afkomstig uit CME artikels. Er zijn ernstige tekortkomingen gevonden bij 10 (42%) van de CME modellen en bij 96 (66%) van de niet-CME modellen. Er is een significante p-waarde van 0,039 berekend met de chi²-test. Dit wil zeggen dat de modellen uit CME artikels significant minder kans hebben om een ernstige tekortkoming te hebben dan modellen uit niet-CME artikels. Er zijn middelmatige tekortkomingen gevonden bij respectievelijk 7 (29%) en 63 (43%) modellen uit CME en niet-CME artikels. De p-waarde van de chi²-test is 0,26. Alle modellen uit de CME artikels en alle modellen uit de niet-CME artikels hebben minder ernstige tekortkomingen. De p-waarde uit de chi²-test is 0,999.

Resultaten CCM	Feature artikel			CME artikel		
	Ja	Nee		Ja	Nee	
	15	155	p-waarde	24	146	p-waarde
<u>Ernstige tekortkomingen</u>	13 (87)	93 (60)	0,051	10 (42)	96 (66)	0,039
De initiële variabelen in het model zijn niet gespecificeerd	3	80		9	74	
Overfitting volgens de 1/10 regel	8	6		1	13	
Predictieve modellen waar geen ROC-curve wordt voorzien	6	44		9	39	
<u>Middelmatig tekortkomingen</u>	4 (27)	66 (43)	0,28	7 (29)	63 (43)	0,26
De selectieprocedure wordt niet verduidelijkt	2	28		1	29	
De conformiteit naar de lineaire gradiënt wordt niet nagekeken*	2	42		7	37	
Er wordt geen controle voor interactietermen gedaan*	1	47		7	41	
<u>Kleine tekortkomingen</u>	15 (100)	155 (100)	0,999	24 (100)	146 (100)	0,999
Automatische selectieprocedure in verklarende modellen	2	65		13	54	
De eenheden van continue variabelen worden niet weergegeven	0	29		10	19	
Calibratiestatistiek wordt niet weergegeven	10	115		14	111	
Collineariteit wordt niet besproken.	13	142		23	132	
Een incorrecte codering van de eindpunten	5	30		1	34	
Crossvalidatie wordt niet uitgevoerd*	2	64		15	51	

Tabel 3:

De resultaten van CCM volgens de ernst van de tekortkomingen bij een feature artikel en een CME artikel

* werden beoordeeld als middelmatige en kleine tekortkomingen in een database met n ≥ 500.

In tabel 4 kan men zien dat er een editoriaal werd geschreven bij 149 van de 170 modellen. Ernstige tekortkomingen werden gevonden bij 92 (62%) modellen met een editoriaal en bij 14 (67%) modellen zonder editoriaal. De p-waarde van de chi²-test is 0,811. Middelmatige tekortkomingen werden gevonden bij 60 (40%) modellen met een editoriaal en bij 10 (48%) modellen zonder editoriaal. De p-waarde van de chi²-test is 0,642. Alle modellen uit artikels die een editoriaal hebben en alle modellen uit artikels die geen editoriaal hebben, hebben ten minste één minder ernstige tekortkoming. De p-waarde uit de chi²-test is 0,999. Een biostatisticus heeft meegewerkt aan de berekening van 98 van de 170 modellen. Hiervan hebben er 47 (48%) ernstige tekortkomingen en bij de 72 modellen waaraan er geen biostatisticus heeft meegewerkt is dit 59 (82%). De p-waarde uit de chi²-test is 0,001 en is sterk significant. Bij middelmatige tekortkomingen is dit respectievelijk 32 (33%) en 38 (53%). De berekende p-waarde 0,01 uit de chi²-test is hier ook sterk significant. Modellen waaraan een biostatisticus heeft meegewerkt hebben een significant lagere kans op ernstige en middelmatige tekortkomingen dan modellen waaraan geen biostatisticus heeft meegewerkt. Alle modellen waaraan een biostatisticus heeft meegewerkt hebben minder ernstige tekortkomingen net zoals de modellen zonder een biostatisticus. De p-waarde uit de chi²-test is 0,999.

Resultaten CCM	Editoriaal			Biostatisticus		
	Ja	Nee		Ja	Nee	
	149	21	p-waarde	98	72	p-waarde
<u>Ernstige tekortkomingen</u>	92 (62)	14 (67)	0,811	47 (48)	59 (82)	0,001
De initiële variabelen in het model zijn niet gespecificeerd	11	3		0	14	
Overfitting volgens de 1/10 regel	71	12		44	39	
Predictieve modellen waar geen ROC-curve wordt voorzien	39	9		24	21	
<u>Moderate tekortkomingen</u>	60 (40)	10 (48)	0,642	32 (33)	38 (53)	0,001
De selectieprocedure wordt niet verduidelijkt	27	3		18	12	
De conformiteit naar de lineaire gradiënt wordt niet nagekeken*	37	7		13	31	
Er wordt geen controle voor interactietermen gedaan*	39	9		18	30	
<u>Kleine tekortkomingen</u>	149 (100)	21 (100)	0,999	98 (100)	72 (100)	0,999
Automatische selectieprocedure in verklarende modellen	54	13		40	27	
De eenheden van continue variabelen worden niet weergegeven	18	11		18	111	
Calibratiestatistiek wordt niet weergegeven	114	11		68	57	
Collineariteit wordt niet besproken.	138	17		90	65	
Een incorrecte codering van de eindpunten	34	1		9	26	
Crossvalidatie wordt niet uitgevoerd*	55	11		31	35	

Tabel 4:

De resultaten van CCM volgens de ernst van de tekortkomingen bij een artikels met een editoriaal en artikels waaraan een biostatisticus heeft meegewerkt

* werden beoordeeld als middelmatige en kleine tekortkomingen in een database met n ≥ 500.

In tabel 5 zien we het verband tussen het percentage PhD's van de auteurs en de tekortkomingen in CCM. Als er ernstige en middelmatige tekortkomingen zijn is het percentage PhD's respectievelijk 16,8 en 10,0. Als er geen ernstige tekortkomingen zijn is het percentage PhD's respectievelijk 35,6 en 33,0. Als men op deze gegevens een Mann-Whitney U test uitvoert, bekomt men bij alle twee een sterk significante p-waarde van 0,001. Hieruit blijkt dat een hoog percentage PhD's in de auteurslijst protectief is voor ernstige en middelmatige tekortkomingen. Er kon geen p-waarde berekend worden bij de kleine tekortkomingen omdat alle artikels in CCM minstens één kleine tekortkoming hadden.

Verband percentage PhD's en tekortkomingen			
	Ja	Nee	p-waarde
Ernstige tekortkomingen	16,8% (0-50)	35,6% (9-71)	0,001
Middelmatige tekortkomingen	10,0% (0-33)	33,0% (0-71)	0,001

Tabel 5:

Verband tussen het percentage PhD's in de auteurslijst en de ernst van de tekortkomingen

B. Resultaten Intensive Care Medicine (ICM)

In ICM melden 36 (23,5%) van de 153 artikels het gebruik van MLRA en in 28 (77,8%) artikels werden de resultaten ook getoond. In tabel 6 worden de 95 ICM modellen getoond en verdeeld volgens de grootte van de steekproef. Er zijn 59 modellen met n < 500 en 36 modellen met n ≥ 500. De initiële variabelen in het model zijn niet gespecificeerd in 8,5% van de n < 500 groep en in geen enkel model van de n ≥ 500 groep. Er is overfitting volgens de 1/10 regel in 83,1% van de modellen met n < 500 en in 5,6% van de modellen met n ≥ 500. In 47,5% van de modellen met een steekproef n < 500 wordt er geen ROC-curve voorzien en in 35 van de 59 modellen werd aan predictie gedaan waarvan er in 22,9% van de modellen een ROC-curve werd getoond. In de groep met n ≥ 500 was dit respectievelijk zo in 38,9% van de modellen en in 33,3% van de 18 predictieve modellen. In twee van de 95 modellen wordt een Nagelkerke statistic weergegeven in plaats van een ROC-curve met een corresponderende AUC. De selectieprocedure wordt niet verduidelijkt in 11,9% van de ene groep en geen enkele keer in de andere groep. De selectieprocedure was automatisch in de verklarende modellen in 8,3% in de n < 500 groep en in 66,7% van de modellen in de n ≥ 500 groep. De conformiteit naar de lineaire gradiënt wordt in 32,2% van de modellen met n < 500 en in 22,2% van de modellen met n ≥ 500 niet nagekeken. Er werd geen controle voor interactietermen gedaan in 88,1% van de modellen in de eerste groep en in 86,1% van de modellen van de tweede groep. In de 46 modellen met continue variabelen van de n < 500 groep worden de eenheden in 76,1% niet getoond en in de 16 modellen met continue variabelen in de n ≥ 500 groep is dit 50,0%. De calibratiestatistiek wordt niet weergegeven in 64,4% van de modellen van n < 500 groep en in 63,9% van de modellen van de n ≥ 500 groep. De collineariteit werd niet nagekeken in respectievelijk 52,5% en 27,8% van de modellen. Er was een incorrecte codering van de eindpunten bij 3,4% van de modellen in de eerste groep en 8,3% in de tweede groep. Crossvalidatie werd in geen enkel model van de eerste groep uitgevoerd en slechts in één model van de tweede groep.

Resultaten ICM	n < 500	n ≥ 500
	n= 59	n= 36
De initiële variabelen in het model zijn niet gespecificeerd	5 / 59 (8,5)	0 / 36 (0)
Overfitting volgens de 1/10 regel	49 / 59 (83,1)	2 / 36 (5,6)
Geen weergave van een ROC-curve	28 / 59 (47,5)	14 / 36 (38,9)
Predictieve modellen waar geen ROC-curve wordt voorzien	8 / 35 (22,9)	6 / 18 (33,3)
De selectieprocedure wordt niet verduidelijkt	7 / 59 (11,9)	0 / 36 (0)
Automatische selectieprocedure in verklarende modellen	2 / 24 (8,3)	12 / 18 (66,7)
De conformiteit naar de lineaire gradiënt wordt niet nagekeken	19 / 59 (32,2)	8 / 36 (22,2)
Er wordt geen controle voor interactietermen gedaan	52 / 59 (88,1)	31 / 36 (86,1)
De eenheden van continue variabelen worden niet weergegeven	35 / 46 (76,1)	8 / 16 (50,0)
Calibratiestatistiek wordt niet weergegeven	38 / 59 (64,4)	23 / 36 (63,9)
Collineariteit wordt niet besproken.	31 / 59 (52,5)	10 / 36 (27,8)
Een incorrecte codering van de eindpunten	2 / 59 (3,4)	3 / 36 (8,3)
Crossvalidatie wordt niet uitgevoerd	59 / 59 (100)	35 / 36 (97,2)

Tabel 6:

De resultaten van ICM volgens n < 500 en n ≥ 500.

Bij het bekijken van de gegevens van ICM zijn er 2 van de 95 modellen die een editoriaal hebben en 43 modellen die gebruik hebben gemaakt van een biostatisticus. In tabel 7 zien we dat 1 (50%) model met een editoriaal en 63% van de modellen zonder editoriaal ernstige tekortkomingen hebben. Een chi^2-test werd uitgevoerd waarvan de p-waarde 0,999 is. Geen enkel model met een editoriaal en 46% van de modellen zonder editoriaal hebben een middelmatige tekortkoming. De p-waarde van de chi^2-test is 0,499. Alle modellen met een editoriaal en 99% van de modellen zonder een editoriaal hebben minder ernstige tekortkomingen. De p-waarde van de chi^2-test is 0,999. Van de 43 modellen waaraan er een biostatisticus heeft meegewerkt zijn er bij 88% ernstige tekortkomingen gevonden. Bij de modellen zonder biostatisticus was dit 42%. De p-waarde van de chi^2-test is 0,001 en is sterk significant. Dit wil zeggen dat als er een biostatisticus heeft mee gewerkt er meer kans is op ernstige tekortkomingen. Een middelmatige

tekortkoming werd in 23% van de modellen met een biostatisticus en in 64% van de modellen zonder biostatisticus gevonden. De p-waarde van de chi²-test is 0,001 en is sterk significant maar deze keer verlaagt het gebruik van een biostatisticus de kans op een middelmatige tekortkoming. Alle modellen met een biostatisticus en 98% van de modellen zonder biostatistus hebben kleine tekortkomingen. De p-waarde van de chi²-test is 0,999.

Resultaten ICM	Editoriaal			Biostatisticus		
	Ja	Nee		Ja	Nee	
	2	93	p-waarde	43	52	p-waarde
<u>Ernstige tekortkomingen</u>	1 (50)	59 (63)	0,999	38 (88)	22 (42)	0,001
De initiële variabelen in het model zijn niet gespecificeerd	0	5		1	4	
Overfitting volgens de 1/10 regel	1	50		35	16	
Predictieve modellen waar geen ROC-curve wordt voorzien	1	13		4	10	
<u>Moderate tekortkomingen</u>	0 (0)	43 (46)	0,499	10 (23)	33 (64)	0,001
De selectieprocedure wordt niet verduidelijkt	0	1		3	4	
De conformiteit naar de lineaire gradiënt wordt niet nagekeken*	/	8		0	8	
Er wordt geen controle voor interactietermen gedaan*	/	31		7	24	
<u>Kleine tekortkomingen</u>	2 (100)	92 (99)	0,999	43 (100)	51 (98)	0,999
Automatische selectieprocedure in verklarende modellen	1	13		13	1	
De eenheden van continue variabelen worden niet weergegeven	1	42		30	13	
Calibratiestatistiek wordt niet weergegeven	2	59		30	31	
Collineariteit wordt niet besproken.	2	39		26	15	
Een incorrecte codering van de eindpunten	0	5		1	4	
Crossvalidatie wordt niet uitgevoerd*	/	35		7	28	

Tabel 7:

De resultaten van ICM volgens de ernst van de tekortkomingen bij artikels met een editoriaal en artikels waaraan een biostatisticus heeft meegewerkt.

* werden beoordeeld als middelmatige en kleine tekortkomingen in een database met n ≥ 500.

C. Resultaten CCM en ICM tezamen

In CCM en ICM tezamen melden 106 artikels het gebruik van MLRA en in 89 (84%) artikels werden de resultaten ook getoond. In de 89 artikels werden 265 modellen gevonden. De gegevens van CCM en ICM tezamen zijn in tabel 8 te zien. In totaal zijn er 159 modellen met n < 500 en 106 modellen met n ≥ 500. De initiële variabelen in het model zijn niet gespecificeerd in 8,8% van de modellen van de n < 500 groep en in 4,7% van de modellen van de n ≥ 500 groep. Er is overfitting volgens de 1/10 regel in 64,2% van de modellen met n < 500 en in 30,2% van de modellen met n ≥ 500. In 71,7% van de modellen met een steekproef n < 500 wordt er geen ROC-curve voorzien. In 36 van de 68 modellen werd aan predictie gedaan waarvan er in 22,9% van de modellen een ROC-curve werd berekend. In de groep met n ≥ 500 wordt er in 69,8% van de modellen en in 60,5% van de 43 predictieve modellen geen ROC-curve voorzien. In tien van de 265 modellen wordt een Nagelkerke statistic weergegeven in plaats van een ROC-curve met een corresponderende AUC.De selectieprocedure wordt niet verduidelijkt in 14,5% van de n < 500 groep en in 13,2% van de n ≥ 500. Respectievelijk wordt er een automatische selectieprocedure gebruikt in 52,7% en 52,4% van de verklarende modellen. De conformiteit naar de lineaire gradiënt wordt in 65,7% van de modellen met n < 500 en in 51,5% van de modellen met n ≥ 500 niet nagekeken. De interactietermen werden niet nagekeken in 90,6% van de modellen in de n < 500 groep en in 74,5% van de modellen van de n ≥ 500groep. In de 83 modellen met continue variabelen van de n < 500 groep worden de eenheden in 56,6% niet getoond en in de 54 modellen met continue variabelen van de n ≥ 500 groep is dit 46,3%. De calibratiestatistiek wordt niet weergegeven in respectievelijk 73,6% en 65,1% van de modellen. De collineariteit wordt niet nagekeken in 79,2% van de modellen in de n < 500 groep en 66,0% van de n ≥ 500groep. Een incorrecte

codering van de eindpunten werd opgemerkt bij 11,3% van de modellen in de n < 500 groep en 20,8% van de n ≥ 500groep. Crossvalidatie werd in de n < 500 groep niet uitgevoerd en slechts vijfmaal in de n ≥ 500 groep.

Resultaten CCM + ICM	n < 500	n ≥ 500
	n= 159	n= 106
De initiële variabelen in het model zijn niet gespecificeerd	14 / 159 (8,8)	5 / 106 (4,7)
Overfitting volgens de 1/10 regel	102 / 159 (64,2)	32 / 106 (30,2)
Geen weergave van een ROC-curve	114 / 159 (71,7)	74 / 106 (69,8)
Predictieve modellen waar geen ROC-curve wordt voorzien	36 / 68 (52,9)	26 / 43 (60,5)
De selectieprocedure wordt niet verduidelijkt	23 / 159 (14,5)	14 / 106 (13,2)
Automatische selectieprocedure in verklarende modellen	48 / 91 (52,7)	33 / 63 (52,4)
De conformiteit naar de lineaire gradiënt wordt niet nagekeken	94 / 143 (65,7)	52 / 101 (51,5)
Er wordt geen controle voor interactietermen gedaan	144 / 159 (90,6)	79 / 106 (74,5)
De eenheden van continue variabelen worden niet weergegeven	47 / 83 (56,6)	25 / 54 (46,3)
Calibratiestatistiek wordt niet weergegeven	114 / 159 (73,6)	69 / 106 (65,1)
Collineariteit wordt niet besproken.	126 /159 (79,2)	70 / 106 (66,0)
Een incorrecte codering van de eindpunten	18 / 159 (11,3)	22 / 106 (20,8)
Crossvalidatie wordt niet uitgevoerd	159 / 159 (100)	101 / 106 (95,3)

Tabel 8:

De resultaten van CCM en ICM tezamen volgens n < 500 en n ≥ 500.

In tabel 9 zijn de gegevens van CCM en ICM onderzocht naargelang er een editoriaal over het artikel geschreven was en naargelang er een biostatisticus meewerkte aan de publicatie. Bij 151 van de 265 modellen werd er een editoriaal geschreven. Bij 93 (62%) van de modellen met een editoriaal en bij 73 (64%) van de modellen zonder editoriaal werden ernstige tekortkomingen gevonden. De p-waarde van de chi^2-test is 0,684. Middelmatige tekortkomingen werden gevonden bij 60 (40%) modellen met een editoriaal en bij 53 (47%) modellen zonder een

editoriaal. De p-waarde van de chi²-test is 0,271. Alle modellen die een editoriaal hebben en alle modellen die geen editoriaal hebben, behalve één, hebben ten minste één minder ernstige tekortkoming. De p-waarde uit de chi²-test is 0,999. Een biostatisticus heeft meegewerkt bij 141 van de 265 modellen. Hiervan hebben er 85 (60%) ernstige tekortkomingen en bij de 124 modellen waaraan er geen biostatisticus heeft meegewerkt is dit 81 (65%). De p-waarde uit de chi²-test is 0,398. Bij middelmatige tekortkomingen is dit respectievelijk 42 (30%) en 71 (57%).De chi²-test heeft hier een sterk significante p-waarde van 0,001. Modellen waaraan een biostatisticus heeft meegewerkt hebben een significant lagere kans op middelmatige tekortkomingen dan modellen waaraan geen biostatisticus heeft meegewerkt. Minder ernstige tekortkomingen komen voor bij alle modellen waaraan een biostatisticus heeft meegewerkt en bij 123 (99%) modellen waaraan geen biostatisticus heeft meegewerkt. De p-waarde uit de chi²-test is 0,281.

Resultaten CCM + ICM	Editoriaal			Biostatisticus		
	Ja	Nee		Ja	Nee	
	151	114	p-waarde	141	124	p-waarde
<u>Ernstige tekortkomingen</u>	93 (62)	73 (64)	0,684	85 (60)	81 (65)	0,398
De initiële variabelen in het model zijn niet gespecificeerd	11	5		1	18	
Overfitting volgens de 1/10 regel	72	62		79	55	
Predictieve modellen waar geen ROC-curve wordt voorzien	40	22		28	34	
<u>Moderate tekortkomingen</u>	60 (40)	53 (47)	0,271	42 (30)	71 (57)	0,001
De selectieprocedure wordt niet verduidelijkt	27	10		21	16	
De conformiteit naar de lineaire gradiënt wordt niet nagekeken*	37	15		13	39	
Er wordt geen controle voor interactietermen gedaan*	39	40		25	54	
<u>Kleine tekortkomingen</u>	151 (100)	113 (99)	0,249	141 (100)	123 (99)	0,281
Automatische selectieprocedure in verklarende modellen	55	26		53	28	
De eenheden van continue variabelen worden niet weergegeven	9	16		14	11	
Calibratiestatistiek wordt niet weergegeven	43	26		24	45	
Collineariteit wordt niet besproken.	140	56		116	80	
Een incorrecte codering van de eindpunten	34	6		10	30	
Crossvalidatie wordt niet uitgevoerd*	55	46		38	63	

Tabel 9:

De resultaten van CCM en ICM tezamen volgens de ernst van de tekortkomingen bij artikels met een editoriaal en artikels waaraan een biostatisticus heeft meegewerkt

* werden beoordeeld als middelmatige en kleine tekortkomingen in een database met n ≥ 500.

5. Discussie

De resultaten van deze studie tonen aan dat de kwaliteit van het weergeven van multivariabele logistische regressie in de Intensieve Zorg literatuur niet optimaal is. Mogelijks is er zelfs een probleem bij het uitvoeren van MLRA, met onbetrouwbare resultaten en conclusie als gevolg. In vergelijkbare studies werden dezelfde vaststellingen gedaan in de cardiopulmonaire literatuur (Moss et al., 2003) en in de gynaecologische en obstetrische literatuur (Khan et al., 1999).

Aangezien de meeste artsen en onderzoekers vertrouwen op het peer review proces voor een correct gebruik en interpretatie van statistische methoden, werd de kwaliteit hiervan onderzocht door Goodman et al.(1998). Ongeveer één derde van de 114 tijdschriften die antwoordden op het onderzoek vereisen statistisch nazicht voor alle opgestuurde manuscripten. Het statistisch nazichtbeleid verschilde tussen de tijdschriften naargelang hun oplage. Van de tijdschriften met een oplage van meer dan 25.000 had 82% een statistische consultant als werknemer, tegenover slechts 31% bij tijdschriften met een oplage van minder dan 4.100. Hieruit valt te besluiten dat er meer problemen te verwachten zijn bij tijdschriften met een lage oplage. Bij navraag aan de redacteurs van de onderzochte tijdschriften schatten zij in dat een grondig statistisch nazicht resulteerde in een belangrijke verandering in ongeveer 50% van de manuscripten. Goodman et al. (1994) onderzochten het peer review proces overigens door de kwaliteit van de manuscripten voor en na revisie te onderzoeken in de Annals of Internal Medicine. De kwaliteit van multivariabele rapportage werd getaxeerd als één van de meest deficiënte factoren op het tijdstip van inzending. Wanneer het manuscript echter werd beoordeeld en herzien, was de kwaliteit van rapporteren van multivariabele analyses aanzienlijk beter. Uit het bovenstaande kan besloten worden dat statistisch nazicht zeker belangrijk is. Gezien het grote aantal manuscripten dat ingezonden wordt met de bedoeling gepubliceerd te worden en de beperkte publicatieruimte, zijn sommige redacteurs echter geneigd om een uitgebreide beschrijving van de methoden te beperken. Zo heeft het tijdschrift AJRCCM *(*American Journal of Respiratory and Critical Care Medicine) zijn

“Instructions for contributors” aangepast met de bedoeling de lengte van elk artikel te beperken (Tobin, 2000). Auteurs zullen nu hun “Methods” deel moeten beperken tot 500 woorden. Een uitgebreid methodeverslag zal in dit tijdschrift echter online beschikbaar worden gesteld.

Wanneer je als lezer een artikel met MLRA probeert te interpreteren is het van belang om in gedachten te houden dat onderzoekers altijd verschillende multivariabele logistische regressie analyses uitvoeren, waaruit dan één of meer modellen worden weergegeven. Bijgevolg moeten lezers zichzelf altijd afvragen waarom de onderzoekers beslisten om precies dit model te tonen. Het feit dat verschillende mogelijkheden bestaan is, in tegenstelling tot wat veel artsen denken, geen probleem zolang de onderzoekers voldoende informatie verschaffen over het hoofddoel van hun studie; de strategie bij het bouwen van een model; hoe goed de data het uiteindelijke model fitten en of het model klinisch of biologisch zinvol is. In ons onderzoek kwam echter duidelijk naar voor dat onderzoekers hierbij vaak tekort schieten. Alhoewel slechts 7% van de modellen niet specificeerden welke variabelen initieel in de analyse werden opgenomen, waren deze variabelen vaak moeilijk in een korte tijdspanne te vinden. We vonden ook vaak tekortkomingen bij de strategie om het MLRA model op te bouwen. Zo bleek uit ons onderzoek dat in 14 % van de modellen geen uitleg gegeven werd over de gebruikte selectieprocedure en dat er in 51% van de modellen een overfitting was volgens de 1/10 regel. Verder bleek dat de conformiteit naar de lineaire gradiënt niet nagekeken werd in 52% van de modellen met meer dan 500 observaties en dat in 69 % van de modellen geen calibratiestatistiek gegeven werd. Een ROC-curve werd ook niet weergeven bij predictieve modellen met meer dan 500 observaties in 60% van de gevallen. In 95 % van de modellen met meer dan 500 observaties werd geen crossvalidatie uitgevoerd, wat de kans op externe validatie van de modellen vermindert. Ook was er in 15% van de modellen een incorrecte codering van de eindpunten, wat verwarring kan geven bij de lezer over het al dan niet protectief of infauste karakter van een variabele. Enerzijds is er in de praktijk aldus vaak onvoldoende informatie voorhanden opdat de kritische lezer met een basiskennis in de MLRA inzicht zou verkrijgen in het model en aldus hoe de resultaten van het onderzoek tot stand zijn gekomen. Anderzijds is het voor lezers

met klinische ervaring in dat specifieke vakgebied doch zonder kennis in de statistiek bijna onmogelijk om de resultaten op een correcte wijze te interpreteren gezien ze geen inzicht hebben in de statische tekortkomingen. Tenslotte zal een statisticus zonder klinische of geneeskundige achtergrond onmogelijk kunnen nagaan indien het MLRA model klinisch relevant is en indien alle gekende en aldus belangrijke confounders werden opgenomen in een model. Onderzoekers zouden dan ook beter een duidelijke beschrijving geven van de strategie die ze gebruikt hebben bij het opbouwen van hun model en eventueel zelfs verschillende modellen tonen, althans wanneer ze een verklarende intentie hebben. Zo kan de lezer uitmaken in welke omstandigheden een bepaald resultaat geldig is en in welke mate dit resultaat al dan niet consistent is. Op deze manier krijgt de lezer inzicht in de databankstructuur en kan hij een onderscheid maken tussen partieel en volledig onafhankelijke bevindingen. Medici zouden ook best samenwerken met statistici, zeker bij grote databanken van hoge kwaliteit en waarbij de onderzoeksvraag zeer relevant lijkt voor de praktijk.

Ons onderzoek wees uit dat een biostatisticus protectief is voor ernstige en middelmatige tekortkomingen in CCM en voor middelmatige tekortkomingen in ICM. Dit lijkt een logisch gevolg van de specialiteit van de biostatisticus, maar dit verklaart niet waarom in ICM een model dat nagekeken werd door een biostatisticus een significant verhoogde kans heeft op ernstige tekortkomingen. Als we de gegevens samenvoegen is een biostatisticus enkel maar protectief voor middelmatige tekortkomingen. Een mogelijke verklaring zou kunnen zijn dat een biostatisticus helpt bij het opstellen van de studie, de datacollectie en de verwerking van de gegevens, maar dat hij niet echt betrokken wordt bij de uiteindelijke interpretatie en “vulgarisatie” van de resultaten of dat de biostasticus enkel maar partieel betrokken werd bij de verwerking van de gegevens (Gøtzsche et al., 2007). Verder had het PhD-percentage een beschermend effect op ernstige en middelmatige tekortkomingen. Dit kan verklaard worden op verschillende manieren. De gedoctoreerde auteurs schrijven waarschijnlijk een groter aantal artikels per jaar en hebben vermoedelijk ook een betere kennis van de literatuur en de statistiek. Hoe hoger het aantal PhD’s, hoe vaker het artikel kritisch zal gelezen worden, wat de kans vergroot dat tekortkomingen opgemerkt worden. Een feature

artikel had een randsignificant grotere kans op ernstige tekortkomingen, terwijl men eigenlijk het omgekeerde zou verwachten. Waarschijnlijk is men door de nieuwe opvattingen of inzichten vooral enthousiast over het resultaat, maar let men minder op de manier waarop deze resultaten verkregen werden. Door het ontbreken van vroegere studies waarop men zich kan baseren, is de database waarop deze studies gebaseerd zijn waarschijnlijk kleiner, zodat het gevaar voor overfitting toeneemt. In feite zou bij een feature artikel extra nadruk moeten gelegd worden op de statistiek, zodat men sneller tot inzicht kan komen of een bepaalde nieuwe techniek of behandeling werkelijk efficiënt is. Bij het onderzoeken of er een discrepantie is tussen het aantal tekortkomingen bij artikels met of zonder een editoriaal, werd geen statistisch significant verschil gevonden, wat enigszins opnieuw verrassend is. Normaal gezien krijgen enkel belangrijke artikels een editoriaal en zou men verwachten dat deze artikels extra goed nagelezen worden. Dit wijst er opnieuw op dat men vooral de nadruk legt op het resultaat en niet voldoende nagaat hoe dit resultaat bekomen werd. Verder bleken Continuing Medical Education artikels ook protectief te zijn voor ernstige tekortkomingen in CCM. Dit is positief, want de conclusie van deze artikels wordt overgenomen door de artsen die zich verder scholen via deze artikels. Het is weliswaar jammer dat in deze artikels die gebruikt worden voor accreditering veel middelmatige tekortkomingen voorkomen.

Wanneer we bovenstaande resultaten interpreteren is het van belang rekening te houden met enkele potentiële studielimitaties die de interpretatie van onze resultaten kunnen beïnvloeden. Ten eerste hadden we geen medische kennis genoeg om na te gaan welke specifieke confounders zeker in bepaalde modellen moesten opgenomen worden en welke niet. Sowieso is het onze aanbeveling dat auteurs dit beschrijven zodat de lezer deze gespecialiseerde kennis niet nodig heeft om het model te begrijpen. Ten tweede was het niet mogelijk om te bepalen of auteurs hun statistische analyses onjuist uitvoerden, of indien ze hun methoden niet accuraat weergaven in hun manuscript. Ten derde hebben we onze resultaten bekeken volgens model en niet per artikel. De kans is groot dat wanneer men in een bepaald artikel een bepaalde tekortkoming aantreft in één model, men ook in de andere modellen deze tekortkoming zal terugvinden. Dit heeft als resultaat dat

er een inflatie aan tekortkomingen is in onze studie. Toch hebben wij de keuze gemaakt om onze resultaten te beschrijven volgens de modellen en niet volgens de artikels, omdat het onmogelijk is om de beschreven tekortkomingen te coderen wanneer in bepaalde modellen van het artikel wel aan deze tekortkoming voldaan is en in andere modellen niet. Ten vierde hebben we niet gekeken naar oorzaak-gevolg problemen. Ten laatste hebben we in ons onderzoek enkel artikels uit het jaar 2006 onderzocht. Het is mogelijk dat de kwaliteit van rapporteren verbeterd is gedurende de laatste jaren.

Het is meer dan vijftien jaar geleden sinds Concato et al.(1993) als eersten de problemen omtrent het rapporteren van multivariabele statistische analyses beschreven. Ons onderzoek wijst uit dat ook in het recente verleden de criteria voor het rapporteren nog niet voldoende geïmplementeerd werden. Verschillende auteurs suggereerden reeds aanbevelingen om het probleem van slechte statistische weergave aan te pakken. Zo stelde Campillo (1993) dat duidelijke publicatiecriteria voor regressiemodellen moeten beschikbaar zijn voor onderzoekers die hun resultaten willen publiceren in een medisch tijdschrift. Lang en Secic (1993) hebben aanbevolen dat een artikel met MLRA een tabel moet weergeven met de coëfficiënt (β), SE, Wald-test waarde, p-waarde, odds ratio en het 95% CI voor elke onafhankelijke variabele. Daarenboven raadde hij aan dat de codering van elke onafhankelijke variabele weergegeven zou worden en dat er een statement opgenomen zou moeten worden dat stelt of het model gevalideerd werd of niet. Uiteindelijk bevolen ze ook aan dat uitleg moet gegeven worden omtrent interactietermen en collineariteit in deze manuscripten en dat het totaal aantal observaties moet weergegeven worden. In ons onderzoek werd echter in 74% van de modellen geen uitleg gegeven indien er onderzocht werd naar collineariteit en werd in modellen met meer dan 500 observaties in 75 % van de modellen niet vermeld of er gecontroleerd werd op interactietermen. Een andere mogelijkheid voor de aanpak van het probleem zou kunnen zijn om het aantal statistische consultants te verhogen. Anderzijds zou ook een kritische nalezing door peers met enige statistische basiskennis over MLRA een groot aantal tekortkomingen kunnen voorkomen. Verder zouden vooral tijdschriften zoals CCM en ICM minimale richtlijnen moeten rapporteren waaraan artikels die MLRA gebruiken zouden moeten voldoen om in aanmerking te komen voor peer-review en publicatie. Bij

ons onderzoek was het niet onze intentie om een zo groot mogelijk aantal tekortkomingen te vinden, maar om duidelijk te maken dat er nog altijd een groot aantal tekortkomingen voorkomt in de literatuur en om een positief beeld te schetsen van hoe men best MLRA aanpakt. Hopelijk zal onze thesis en toekomstig artikel dan ook bijdragen tot verbeteringen in het rapporteren van MLRA en zal het de lezers toelaten studieresultaten beter te interpreteren.

6. Referentielijst

BAUMANN K. : Cross-validation as the objective function for variable-selection techniques. Trends in Analytical Chemistry, 2003; 22(6), 395-406.

CAMPILLO C. Standardizing criteria for logistic regression models. Ann Intern Med 1993;119:540-541.

CHIN S. : The rise and fall of logistic regression. Aust Epidemiol 2001; 8:7–10.

CONCATO J ET AL. : The risk of determining risk with multivariable models. Ann Intern Med 1993; 118: 201-210.

CONCATO J, PEDUZZI P, HOLFOLD TR, ET AL. : Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 1995;48:1495–501.

DAVISON A. C., HINKLEY D. V. : Boostrap methods and their application. Second Edition. Cambrigde University Press, 1997.

DOHOO I.R., DUCROT C., FOURICHON C., DONALD A., HURNIK D. : An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med, 1997, 29, 221-239

EFRON B., TIBSHIRANI R. : An introduction to the bootstrap. CRC Press, 1993.

GARROUSTE-ORGEAS M., TROCHÉ G., AZOULAY E., CAUBEL A. et al. : Body mass index : an additional prognostic factor in ICU patients. Intensive Care Med , 2004, 30, 437-443.

GOODMAN SN, ALTMAN DG, GEORGE SL. : Statistical reviewing policies of medical journals: caveat lector? J Gen Intern Med 1998; 13:753–756.

GOODMAN SN, BERLIN J, FLETCHER SW, ET AL. : Manuscript quality before and after peer review and editing at the *Annals of Internal Medicine*. Ann Intern Med 1994; 121:11–21.

GØTZSCHE PC, HRÓBJARTSSON A, JOHANSEN HK, HAAHR MT, ALTMAN DG, ET AL.: Ghost authorship in industry-initiated randomised trials. PLoS Med 2007 4(1): e19.

HÄRDLE W. : Applied Nonparametric Regression. Cambridge University Press, 1992.

HOSMER D.W., LEMESHOW S. : Applied logistic regression Second Edition. John Wiley and Sons, 2000.

HOSMER DW JR, LEMESHOW S. : Applied logistic regression. New York: John Wiley, 1989:25-37.

KHAN KS, CHIEN PFW, DWARAKANATH LS. : Logistic regression models in obstetrics and gynecology literature. Obstet Gynecol 1999; 93:1014–1020

KJETIL SØREIDE. : Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *J. Clin. Pathol.* 2009;62;1-5.

KLEINBAUM DG. : Logistic regression: a self-learning text. New York, NY: Springer-Verlag, 1994

LANG TA, SECIC M. : How to report statistics in medicine. Philadelphia, PA: American College of Physicians, 1997.

MCKILLUP S. : Statistics explained: an introductory guide for life scientists. Cambridge University Press, 2006.

MOSS M ET AL. : An Appraisal of Multivariate Logistic Models in the Pulmonary and Critical Care Literature. Chest 2003;123;923-928.

O'BRIEN J. M., PHILLIPS J. S., NAEEM A. A., LUCARELLI M., MARSH C. B., LEMESHOW S. : Body mass index is independently associated with hospital mortality in mechanically ventilated adults with acute lung injury. Crit Care Med, 2006, 34, 738-744.

OSBORNE J. W., WATERS E. : Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation, 2002, 8(2). Opgehaald op 10 maart 2009, van http://PAREonline.net/getvn.asp?v=8&n=2

OTTENBACHER KJ ET AL. : A review of two journals found that articles using multivariable logistic regression frequently did report commonly recommended assumptions. J Clin Epidemiol 2001;54;1159-65.

PALMAS W, DENTON TA, DIAMOND GA. : Publication criteria for statistical prediction models. Ann Intern Med 1993, 118:231–232

PEDUZZI P, CONCATO J, FEINSTEIN AR, ET AL. : Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10.

PEDUZZI P, CONCATO J, KEMPER E, ET AL. : A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.

STEYERBERG E. W., EIJKEMANS M. J. C., HARELL Jr F. E., HABBEMA J. D. F : Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small datasets. Statistics in Medicine, 2000, 19, 1059 -1079.

STONE-ROMERO E. F., ROSOPA P. J. : The Relative Validity of Inferences About Mediation as a Function of Research. Organizational Research Methods, 2008; 11; 326. Opgehaald op 10 maart 2009, van http://orm.sagepub.com/cgi/rapidpdf/1094428107300342v1

THE COCHRANE COLLABORATION OPEN LEARNING MATERIAL. Opgehaald op 10 april 2009, van http://www.cochrane-net.org/openlearning/html/mod11-4.htm

TOBIN MJ. : Authors, authors, authors: follow instructions or expect delay. Am J Respir Crit Care Med 2000; 162:1193–1194.

TRUJILLANO J. : Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. Gac Sanit, 2008; 22(1), 65-72.