# Statistical modelling strategies for monitoring antibiotic use and resistance

**Robin Bruyndonckx**

Promotors: prof. dr. Niel Hens and prof. dr. Samuel Coenen
Copromotor: prof. dr. Marc Aerts

# Acknowledgements

Met het einde van mijn doctoraat in zicht, wil ik even de tijd nemen om een aantal mensen te bedanken.

Als eerste wil ik mijn promotoren, prof. dr. Niel Hens, prof. dr. Samuel Coenen en prof. dr. Marc Aerts, bedanken voor hun uitstekende begeleiding, die langzaamaan is veranderd in een aangename samenwerking. Jullie stonden me steeds bij, met raad en daad, maar leerden me vooral steeds zelfstandiger werken. Jullie boden me, zowel binnen het Centrum voor Statistiek als binnen het Labo Medische Microbiologie, een veilige omgeving om te leren en te groeien, en daarvoor ben ik bijzonder dankbaar. Niel, bedankt voor je out-of-the-box ideeën en motiverende woorden. Samuel, bedankt voor je enthousiaste aanpak. Marc, bedankt voor je geduld wanneer je me iets drie keer opnieuw moest uitleggen.

I gratefully thank all my co-authors for the pleasant and fruitful collaboration, which greatly improved this work. Additionally, I thank all the members of the jury for taking the time to read this thesis and provide insightful comments.

Verder wil ik mijn collega's bedanken voor de aangename sfeer waarin ik de voorbije jaren heb mogen werken. JOSS-colleagues, thanks for all the fun-filled activities. Koffie-groepers, bedankt voor de luchtige praatjes die me even het werk deden vergeten. Ann, bedankt om me zo hartelijk te verwelkomen op S6. En Eva, bedankt dat ik de voorbije jaren lief en leed met jou kon delen. Je maakte steeds tijd om me te helpen bij problemen, privé of werkgerelateerd, en kon met je relativeringsvermogen elke paniekaanval temmen. Ik ben je ongelofelijk dankbaar.

Tot slot wil ik ook familie en vrienden bedanken. Mama en papa, bedankt dat jullie me doorheen de jaren steeds alle mogelijke kansen hebben gegeven, ook toen ik na mijn opleiding biomedische wetenschappen besloot nog statistiek te gaan bijstuderen. Quinten & Jasmine, Amber & Jasper, maar ook Annie & Guido, Ruud, Filip & Linda, bedankt voor de momenten van ontspanning, en het begrip wanneer ik weer eens geen tijd had. Aan mijn vzw-collega's bij Zwerfkat in nood, bedankt om me met open armen en hart in jullie grote vrienden-familie op te nemen. Zowel het veldwerk als het opvangen zijn een onbetaalbare ervaring die me als persoon sterker maakten.

En last, maar zeker niet least, Rob!
Liefje, bedankt dat je er elke dag opnieuw voor me bent. Je staat altijd voor me klaar, en daar kan ik je niet genoeg voor bedanken.
DANKJEWEL!

Robin Bruyndonckx
3 May 2016
Diepenbeek

# Contents

# List of Publications

This dissertation is based on the following scientific papers:

**Bruyndonckx, R.**, Hens, N., Aerts, M., Goossens, H., Molenberghs, G. and Coenen, S. (2014). Measuring trends of outpatient antibiotic use in Europe: jointly modelling longitudinal data in defined daily doses and packages. *Journal of Antimicrobial Chemotherapy*, **69**: 1981-1986.

Coenen, S., **Bruyndonckx, R.**, Hens, N., Aerts, M. and Goossens, H. (2014). Comment on: Measurement units for antibiotic consumption in outpatients. *Journal of Antimicrobial Chemotherapy*, **69**: 3445-3446.

**Bruyndonckx, R.**, Hens, N., Aerts, M., Goossens, H., Cortiñas Abrahantes, J. and Coenen, S. (2015). Exploring the association between resistance and outpatient antibiotic use expressed in defined daily doses or packages. *Journal of Antimicrobial Chemotherapy*, **70**: 1241-1244.

Lambert, M., **Bruyndonckx, R.**, Goossens, H., Hens, N., Aerts, M., Catry, B., Neely, F., Vogelaers, D. and Hammami, N. (2015). The Belgian national policy of funding and implementing antimicrobial stewardship in hospitals: impact on selected quality indicators for antimicrobial use. *BMJ Open*, **5**: e006916.

**Bruyndonckx, R.**, Aerts, M. and Hens, N. (2015). Simulation-based evaluation of the performance of the F test in a linear multi-level model setting with sparseness at the level of the primary unit. *Accepted for publication in biometrical journal*

## Papers in Revision and Working Papers:

**Bruyndonckx, R.**, Hens, N., Aerts, M., Goossens, H., Latour, K., Catry, B. and Coenen, S. (2016). Persistence of antimicrobial resistance in respiratory streptococci. *working paper*

Abong, G., Jaspers, S., **Bruyndonckx, R.**, Hens, N., Catry, B., Latour, K. and Coenen, S. (2016). Antimicrobial activity of Nitrofurantoin against multidrug-resistant urinary E. coli from Belgian outpatients. *working paper*

Barker, C., **Bruyndonckx, R.**, Hens, N., Aerts, M., Versporten, A., Goossens, H., Bielicki, J. and Sharland, M. (2016). Evaluation of paediatric beta-lactam prescribing in hospitals. *working paper*

Latour, K., Diba, C., **Bruyndonckx, R.**, Geerdens, C., Coenen, S. and Catry, B. (2016). Route of administration substantially influences antimicrobial resistance in urinary tract pathogens from elderly. *working paper*

Maes, B., Bakkus, M., Boeckx, N., Boone, E., Cauwelier, B., Denys, B., Deschouwer, P., Devos, T., El Housni, H., Hillen, F., Jacobs, K., Lambert, F., Louagie, H., Maes, M., Meeus, P., Moreau, E., Nollet, F., Peeters, K., Saussoy, P., Van Lint, P., Vaerman, J., Vaeyens, F., Vandepoele, C., Vannuffel, P., Ver Elst, K., Vermeulen, K. and **Bruyndonckx R.** (2016). A novel approach for standardizing BCR-ABL1 quantification on the International Scale. *working paper*

# List of Abbreviations

| | |
|---|---|
| AC | Change-point for the first antibiotic awareness day |
| AF | Treatment with J01A or J01F |
| AIC | Akaike information criterion |
| AMT | Antimicrobial management team |
| AMTIN | Change-point for the introduction of an AMT |
| APR-DRG | All patient refined diagnosis related groups |
| AR | Argentina |
| AR(1) | First order autoregressive |
| ARPEC | Antibiotic Resistance and Prescribing in European Children |
| AT | Austria |
| ATC | Anatomical therapeutic chemical |
| AU | Australia |
| AUC | Area under the curve |
| BAPCOC | Belgian Antibiotic Policy Coordination Committee |
| BE | Belgium |
| BG | Bulgaria |
| BLI | $\beta$-lactamase inhibitor |
| BR | Baseline resistance |
| BUN | Blood urea nitrogen |
| CAP | Community-acquired pneumonia |

| | |
|---|---|
| CD | Treatment with J01C or J01D |
| CH | Switzerland |
| CI | Confidence interval |
| CRF | Case report form |
| CRP | C-reactive protein |
| CZ | Czech Republic |
| DDD | Defined daily dose |
| DDDhosp | Number of DDDs consumed during the stay |
| DE | Germany |
| DID | Number of DDD per 1000 inhabitants per day |
| DK | Denmark |
| EARSS | European Antimicrobial Resistance Surveillance System |
| EE | Estonia |
| ENSP | Erythromycin-non-susceptible *Streptococcus pneumoniae* |
| ES | Spain |
| ESAC | European Surveillance of Antimicrobial Consumption |
| EU | European Union |
| FI | Finland |
| FIN | Change-point for revision of the financing mechanism for pneumonia |
| FR | France |
| GE | Georgia |
| GEE | Generalized estimating equations |
| GH | Ghana |
| GP | General practitioner |
| GR | Greece |
| GRACE | Genomics to combat Resistance against Antibiotics in Community-acquired LRTI in Europe |

| | |
|---|---|
| HR | Croatia |
| HU | Hungary |
| IE | Ireland |
| IL | Israel |
| IMA | Intermutualistic Agency |
| IN | India |
| IQR | Interquartile range |
| IR | Iran |
| IT | Italy |
| J01 | Antibacterials for systemic use |
| J01A | Tetracyclines |
| J01B | Amphenicols |
| J01BGR | Other antibiotics |
| J01C | Penicillins |
| J01CA | Penicillins with extended spectrum |
| J01CA01 | Ampicillin |
| J01CA04 | Amoxicillin |
| J01CE01 | Benzylpenicillin |
| J01CR | Combinations of penicillins |
| J01CR02 | Amoxicillin with BLI |
| J01CR05 | Piperacillin with BLI |
| J01D | Cephalosporins |
| J01DB04 | Cefazolin |
| J01DC02 | Cefuroxime |
| J01DD01 | Cefotaxime |
| J01DD02 | Ceftazidime |
| J01DD04 | Ceftriaxone |
| J01DH02 | Meropenem |

| | |
|---|---|
| J01E | Sulphonamides |
| J01F | Macrolides |
| J01FA01 | Erythromycin |
| J01G | Aminoglycosides |
| J01M | Quinolones |
| J01R | Combinations of antibacterials |
| J01X | Urinary antiseptics |
| LRTI | Lower respiratory tract infection |
| LT | Lithuania |
| LU | Luxembourg |
| LV | Latvia |
| ML | Maximum likelihood |
| MW | Malawi |
| MX | Mexico |
| NL | Netherlands |
| NO | Norway |
| OP | Ratio of oral versus parenteral antimicrobial use during the stay |
| PID | Number of packages per 1000 inhabitants per day |
| PL | Poland |
| PN | *Streptococcus pneumoniae* |
| PNSP | Penicillin-non-susceptible *Streptococcus pneumoniae* |
| PPS | Point prevalence survey |
| PSI | Pneumonia severity index |
| PT | Portugal |
| PY | *Streptococcus pyogenes* |
| RCT | Randomized controlled trial |
| RDE | Relative difference between the estimated and true standard error |
| RDM | Relative difference between the estimated and true mean |

| | |
|---|---|
| REML | Restricted maximum likelihood |
| RO | Romania |
| ROC | Receiver operator curve |
| RU | Russian Federation |
| SA | Saudi Arabia |
| SE | Sweden |
| SES | Simulation standard error |
| SI | Slovenia |
| SK | Slovakia |
| TMLS | Tetracycline, macrolide, lincosamide and streptogramin |
| TR | Turkey |
| UK | United Kingdom |
| UN | United Nations |
| US | United States |
| WHO | World Health Organisation |
| XK | Kosovo |

# List of Figures

# List of Tables

# Chapter 1

# General Introduction

Antibiotics are drugs that are used to treat bacterial infections. One prominent example is penicillin, which was discovered by Alexander Fleming in 1929. He observed that while some bacteria are sensitive to penicillin, others are resistant (Fleming, 1929). In his Nobel lecture, Fleming noted that the sensitive bacteria could easily develop resistance by exposure to low doses of the antibiotic (Fleming, 1945). His point was proven to be correct as only a few years later over 50% of *Staphylococcus aureus* strains were no longer susceptible to penicillin (Alanis, 2005). Meanwhile, both ecological studies and randomized controlled trials (RCTs) in individual patients have demonstrated a link between antibiotic use and resistance (Costelloe *et al.*, 2010; Goossens *et al.*, 2005; Malhotra-Kumar *et al.*, 2007).

Over time, the use and misuse of antibiotics has led to resistance of bacteria to several antibiotics (Ilić *et al.*, 2012; Willemsen *et al.*, 2009). This is a major public health problem as resistance in the infecting organisms is related to treatment failure, prolonged hospitalization, increased costs of care and increased mortality (French, 2005). In order to fight this problem an effective national and international approach is needed urgently. One part of the solution, to which this thesis contributes, is to gather trustworthy information on antibiotic use and its relationship with resistance in order to develop targeted interventions (Smith, 1998).

## 1.1 Outline of the thesis

With an incidence of 30 to 50 cases per 1000 patients per year, acute cough is one of the main reasons for consulting in primary care (Gibson *et al.*, 2013). Although antibiotic treatment for acute cough cases has been shown to have little or no effect, both overall and in patients with co-morbidities, and the majority of acute cough cases are caused by a self-limiting lower respiratory tract infection (LRTI), antibiotics are prescribed to over 50% of patients (Butler *et al.*, 2009; Little *et al.*, 2013; Moore *et al.*, 2014). This inappropriately high level of antibiotic prescribing is explained by the difficulty to accurately identify patients that would benefit from antibiotic treatment (e.g. suffering from a bacterial LRTI or pneumonia) (Teepe *et al.*, 2016; Van Vugt *et al.*, 2013). Dinant *et al.* (2007) suggested that the best way forward is to identify early and manage differently those at high risk of an adverse outcome, while adopting a 'wait and see approach' for the others, hence adjusting treatment according to prognosis rather than diagnosis. Therefore, in Chapter 2, we develop a prognostic prediction rule to predict poor prognosis (i.e. admission to hospital or reconsultation with new or worsened complaints) in patients presenting to primary care with acute cough, aiming to enable general practitioners (GPs) to reassure patients at low risk and provide appropriate treatment for patients at high risk.

Although guidelines for appropriate antibiotic dosing in hospitalized children do exist (e.g. British National Formulary for Children), they are not used properly and prescribed doses deviate substantially from recommended doses. Chapters 3 to 5 focus on data collected within the Antibiotic Resistance and Prescribing in European Children (ARPEC) project which was set up to determine factors that cause variation in doses of antibiotics prescribed to hospitalized children. The information in this dataset has a hierarchical structure, with children nested within departments nested within hospitals nested within countries nested within UN macro-geographical regions. Such a complex multi-level structure automatically gives rise to sparseness issues caused by the low number of subunits at different levels of the hierarchy. Whenever a higher-level unit contains only one subunit, this unit is referred to as a singleton. In Chapter 3, we evaluate the performance of the mixed effects model in the presence of singletons at the lowest level (i.e. the child). In Chapter 4, we evaluate the performance of the F test in the presence of singletons at the highest level (i.e. the UN macro-geographical region). In Chapter 5, we determine which factors are causing variation in doses of ceftriaxone prescribed to hospitalized children and use a meta-analytic approach, pooling all antibiotic-specific analyses, to determine which factors are causing

variation in doses of $\beta$-lactam antibiotics prescribed to hospitalized children.

One factor influencing antibiotic dosing that could not be assessed using the ARPEC data is time. The evolution of outpatient antibiotic use over time is assessed in Chapter 6 using quarterly use data expressed in defined daily doses (DDD) or packages per 1000 inhabitants per day (DID and PID, respectively). When linking antibiotic use with resistance, both DID and PID could be used. In Chapter 7, we investigate which measure best explains this association. In Chapter 8, we focus on resistance profiles in streptococci which reside asymptomatically in the oropharynx but can cause e.g. otitis or pneumonia when they invade the ear and lower respiratory tract, respectively. Upon consumption, an antibiotic is most successful in eliminating the sensitive bacterial phenotypes, leaving a small proportion of highly resistant bacterial phenotypes behind. As a result, the proportion of resistant phenotypes directly after antibiotic treatment is elevated. When no new exposure follows, the advantage of being resistant disappears and the proportion of resistant phenotypes decreases as a result of natural selection against a redundant trait. We compare the persistence of this elevated resistance in oropharyngeal streptococci after exposure to penicillins and cephalosporins or macrolides and tetracyclines.

In Belgium, several attempts have been made to lower resistance rates by optimizing antibiotic consumption. This was done through e.g. the introduction of antimicrobial management teams (AMTs) in hospitals and launch of an annual antibiotic awareness day in 2001. In Chapter 9, we assess the impact of these policies on three selected quality indicators using change-point mixed models.

## 1.2 Motivating data

In this section, we introduce the datasets that are used throughout this work. The Acute Cough Data (Section 1.2.1) are used to develop a prediction rule that will help to identify patients with poor prognosis. The stability of linear multi-level models and of the F test will be assessed using the Ceftriaxone Data (Section 1.2.2). Current variations in antibiotic doses prescribed to hospitalized children are assessed using the Inpatient Antibiotic Use Data (Section 1.2.3) and variations in antibiotic doses prescribed to outpatients are studied using the Outpatient Antibiotic Use Data (Section 1.2.4). The Outpatient Antibiotic Use Data are linked with the Yearly Antimicrobial Resistance Data (Section 1.2.5) to investigate the association between antibiotic use and resistance. Persistence of resistance is evaluated using the Bacterial Susceptibility Data (Section 1.2.6) and the impact of Belgian policies on antimicrobial consumption is assessed using the Hospital Stays Data (Section 1.2.7).

### 1.2.1 Acute Cough Data

Data on the presence of poor prognosis (i.e. admission to hospital or reconsultation with new or worsened complaints) in adult patients presenting to primary care with acute cough were collected in work packages 9 (an observational study of adults with an LRTI) and 10$a$ (a randomized trial to assess the clinical effectiveness of antibiotics for community-acquired LRTI) within the GRACE (Genomics to combat Resistance against Antibiotics in Community-acquired LRTI in Europe; www.grace-lrti.org) Network of Excellence. Information on clinical signs, severity of symptoms, co-morbidities and presence of poor prognosis for 3104 patients was obtained using a case report form (CRF), a diary filled out by the patient in the 28 days following consultation and a notes review. Patients that did not meet the imposed inclusion criteria (listed in Table A1)(0.3%) or had no prognosis reported (4.2%) were removed from the dataset. The remaining patients are distributed over 11 countries according to Table 1.1. To avoid computational issues, we selected countries with more than 15 patients with poor prognosis for further analysis (i.e. Belgium, Germany, the Netherlands, Poland, Spain and the UK). The working data contain information on 100 variables recorded for 2604 patients. Included explanatory variables cover information that is available on the day of consultation and concentrations of C-reactive protein (CRP) and blood urea nitrogen (BUN) (variables are listed in Table A2).

In Chapter 2, the Acute Cough Data will be used to develop a framework for the prediction of poor prognosis in patients presenting to primary care with acute cough.

Table 1.1: Number of patients with poor prognosis in the Acute Cough Data.

| Country | Number of patients | Number of poor prognoses |
|---|---|---|
| Belgium | 388 | 76 |
| France | 30 | 7 |
| Germany | 189 | 52 |
| Italy | 18 | 0 |
| Netherlands | 325 | 75 |
| Poland | 590 | 120 |
| Slovakia | 139 | 5 |
| Slovenia | 73 | 6 |
| Spain | 594 | 86 |
| Sweden | 103 | 8 |
| UK | 518 | 113 |

### 1.2.2 Ceftriaxone Data

Data on the use of ceftriaxone (expressed in mg/kg/day) in hospitalized children were collected in a pilot study for a one-day point prevalence survey (PPS) organised within work package 5 (PPS of paediatric hospital antimicrobial consumption) of the ARPEC project. The pilot study was conducted between September and December 2011. The working data contain information on 329 ceftriaxone prescriptions for children hospitalized in 124 departments in 47 hospitals in 20 countries in 10 UN macro-geographical regions. Explanatory variables include characteristics of the patient, the department, the hospital and the country and are listed in Table 1.2. The hierarchical structure of the Ceftriaxone Data gives rise to sparseness issues mainly caused by the low number of subunits at the lowest and the highest level (i.e. the child and the UN macro-geographical region, respectively). In Chapter 3, we evaluate the stability of the linear multi-level model in the presence of worsening sparseness at the lowest level (i.e. the child). In Chapter 4, we study the performance of the F test in the presence of worsening sparseness at the highest level (i.e. the UN macro-geographical unit).

### 1.2.3 Inpatient Antibiotic Use Data

Data on antimicrobial use in hospitalized children were collected in a one-day PPS organised within work package 5 of the ARPEC project. The PPS was organised worldwide in three waves (March - April 2011, September - December 2011 and October 2012 - January 2013) and is described in great detail elsewhere (Versporten *et al.*, 2013).

In this thesis, we focus on $\beta$-lactam antibiotics that were prescribed frequently (i.e. over 200 prescriptions observed in the PPS). Patients with their gender or prescribed dose missing were removed from the dataset (0.7%). The working data contain information on 5228 prescriptions (expressed in mg/kg/day) for children hospitalized in 1217 paediatric departments in 222 hospitals in 41 countries in 9 UN macro-geographical regions. They are aggregated at the level of the active substance in accordance to the Anatomical Therapeutic Chemical (ATC) Classification System (WHO (2011)) and contain information on doses (in mg/kg/day) of 12 $\beta$-lactam antibiotics. Included substances are oral amoxicillin (J01CA04) and amoxicillin with $\beta$-lactamase inhibitor (BLI) (J01CR02) and parenteral ampicillin (J01CA01), benzylpenicillin (J01CE01), amoxicillin with BLI (J01CR02), piperacillin with BLI (J01CR05), cefazolin (J01DB04), cefuroxime (J01DC02), cefotaxime (J01DD01), ceftazidime (J01DD02), ceftriaxone (J01DD04) and meropenem (J01DH02). Explanatory variables include characteristics of the patient, the department, the hospital and the country and are listed in Table 1.2.

Table 1.2: Overview of explanatory variables in the Inpatient Antibiotic Use Data.

| Variable | Description |
|---|---|
| Id_dep | Identification number for the department |
| Id_ins | Identification number for the hospital |
| Country | Identification code for the country |
| Region | UN macro-geographical region (Northern, Eastern, Southern and Western Europe Southern and Northern America, Africa, Asia, Australia) |
| ATC6 | Antibiotic identification code combined with information on the route of administration (oral or parenteral) |
| Dosage | Total dose (in mg/kg/day) |
| Type_hosp | Type of hospital (primary & secondary, tertiary & specialized & infectious disease hospital) |
| Type_dep | Type of department (general paediatric ward, paediatric intensive care unit) |
| Beds | Number of beds in the department |
| Occup | Percentage of occupied beds in a department |
| Prev | Antibiotic prevalence in a department |

Table 1.2 Continued.

| Variable | Description |
|---|---|
| Age | Age of the patient |
| Weight | Weight of the patient |
| Gender | Gender of the patient (male, female, unknown) |
| Ud1 | Presence of a primary underlying diagnosis |
| Ud2 | Presence of a secondary underlying diagnosis |
| Ud3 | Presence of a tertiary underlying diagnosis |
| Type_treat | Type of treatment (empiric, targeted) |
| Vent | Ventilation status (ventilated, not ventilated) |
| Indic | Indication for treatment (community acquired infection, hospital acquired infection, surgical prophylaxis, medical prophylaxis, unknown) |
| Reason | Reason for treatment (mild, moderate, severe, different) |

The most frequently prescribed antimicrobial in the Inpatient Antibiotic Use Data is parenteral ceftriaxone (18.9%). It is a third-generation cephalosporin with broad-spectrum activity against Gram-positive and Gram-negative bacteria. In Chapter 5, we will study the causes of variation in doses of individual $\beta$-lactam antibiotics prescribed to hospitalized children. Using a meta-model we will assess whether the factors causing variation in individual $\beta$-lactam antibiotics (e.g. ceftriaxone) impact all $\beta$-lactam antibiotics in the same manner and test whether higher doses were given to children treated for a severe infection compared to children treated for a mild or moderate infection, with severity of the reason for treatment classified as shown in Table 1.3.

Table 1.3: Classification of severity of the reason for treatment.

| Severity | Reason for treatment |
|---|---|
| Severe | Sepsis, central nervous system infections, cardiac infections, febrile neutropenia/fever in oncologic patients, catheter related blood stream infections |
| Moderate | Surgical disease, lower respiratory tract infections, urinary tract infections, lymphadenitis, skin/soft tissue infections, joint/bone infections, fever of unknown origin, gastrointestinal tract infections |
| Mild | Upper respiratory tract infections, acute otitis media |
| Other | Prophylaxis, tuberculosis, malaria, unknown |

### 1.2.4 Outpatient Antibiotic Use Data

Data on outpatient antibiotic use, expressed in DID and PID, were collected within the European Surveillance of Antimicrobial Consumption (ESAC) project (currently ESAC-Net). Data were measured yearly and quarterly between 2000 and 2007 and aggregated at the level of the active substance in accordance to the ATC classification system and the DDD measurement unit (WHO (2011)). Information was available for 31 countries, being 26 EU member states (all but Cyprus and Malta), two founding members of the European Free Trade Association (Norway and Switzerland) and three other countries (Turkey, Israel and Russian Federation). For most countries,

information on ambulatory care was provided, although for some only information on total care was available (Denmark, Netherlands, Russian Federation, Sweden and Slovenia). This was not considered to be a problem as ambulatory care represents over 90% of total care.

The data contain information on consumption of antibacterials for systemic use (J01) and its eight pharmacological subgroups (i.e. penicillins (J01C), macrolides (J01F), quinolones (J01M), cephalosporins (J01D), tetracyclines (J01A), sulphonamides (J01E), urinary antiseptics (J01X) and other antibiotics (concatenation of J01B, J01G and J01R)). Information on two chemical subgroups of J01C (i.e. penicillins with extended spectrum (J01CA) and combinations of penicillins (J01CR)) is also provided.
The observed country-specific changes in quarterly antibiotic consumption expressed in DID and PID are shown in Figure 1.1. The individual profiles show that there is considerable within-country and between-country variability, indicating the need for random effects in the model. It can also be seen that there is a clear seasonal fluctuation, which could be approximated well by a sine wave and suggests the need for a non-linear term to model the seasonality.

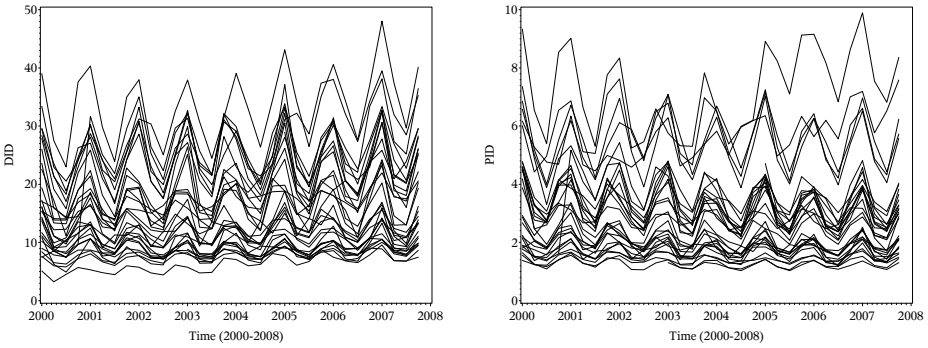



Figure 1.1: Observed country-specific changes in quarterly antibiotics consumption (J01) expressed in DID (left) or PID (right) for 31 European countries.

In Chapter 6, data on quarterly antibiotic consumption will be analysed in order to detect a time-trend in antibiotic dosing. In Chapter 7, data on yearly antibiotic consumption will be coupled with data on antimicrobial resistance (Section 1.2.5) in order to investigate the association.

### 1.2.5 Yearly Antimicrobial Resistance Data

Data on proportions of penicillin-non-susceptible *Streptococcus pneumoniae* (PNSP) and erythromycin-non-susceptible *Streptococcus pneumoniae* (ENSP) isolates were collected within the European Antimicrobial Resistance Surveillance System (EARSS) project (currently EARS-Net). Data were gathered yearly between 2000 and 2009. Information was available for 30 countries, being 27 EU member states (all but Greece), two founding members of the European Free Trade Association (Norway and Switzerland) and one other country (Iceland).

The observed country-specific changes in PNSP and ENSP over time are shown in Figure 1.2. This figure shows that there is a lot of within-country and between-country variability, indicating the need for random effects in the model.

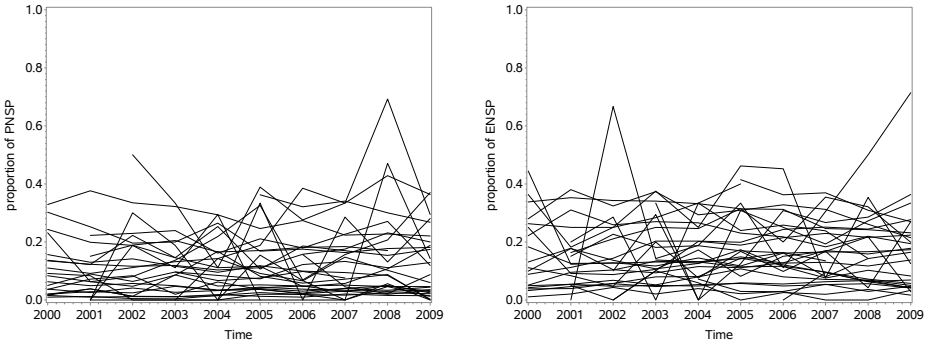



Figure 1.2: Observed country-specific evolution of the proportion of PNSP (left) and ENSP (right) over time for 30 European countries.

In Chapter 7, we will investigate the association between antibiotic use and resistance. Therefore, we will combine data on yearly outpatient antibiotic use (Section 1.2.4) with the data on antimicrobial resistance that were introduced here. Three new datasets will be created by combining use data with resistance data for the same year (time lag = 0), resistance data for one year later (time lag = 1) and resistance data for two years later (time lag = 2). When creating these three datasets, we will only consider countries for which both use and resistance data are available. Each of the combined datasets contains information on 27 European countries, encompassing 25 EU member states (all but Cyprus, Malta and Greece) and two founding members of the European Free Trade Association (Norway and Switzerland).

### 1.2.6 Bacterial Susceptibility Data

Data containing information on individual patients' resistance status linked with information on individual antimicrobial consumption were collected during a multi-centric study conducted within a collaboration of the Intermutualistic Agency (IMA) and the Scientific Institute of Public Health (Catry *et al.*, 2008). The information on resistance status was obtained from 14 voluntary participating laboratories (in 2005). The information on patient prescriptions was obtained from national reimbursement data (collected by IMA for the period July 2004 - December 2005). In this thesis, we focus on prescriptions for an oral dose of J01A or J01F (treatment AF) and J01C or J01D (treatment CD), obtained from the pharmacy, and respiratory tract samples of *Streptococcus pyogenes* (bacteria PY) and *Streptococcus pneumoniae* (bacteria PN) tested for resistance against penicillin (J01CE01) or erythromycin (J01FA01). Samples tested for resistance against penicillin were linked with the most recent CD prescription and samples tested for resistance against erythromycin were linked with the closest AF prescription. Because resistance to penicillin and erythromycin involve different mechanisms, we did not study penicillin resistance after treatment with AF or erythromycin resistance after treatment with CD (Descheemaeker, 2000; Dever and Dermody, 1991). We selected samples for which time between antimicrobial consumption and sampling was at least four days (95% of isolates), to ensure that all patients could have started taking the purchased antimicrobial. The final data used in this thesis contain information on resistance status for 451 test results for 363 patients (Table 1.4). A summary of observed test results is given in Table 1.5.

Table 1.4: Number of tests conducted per patient in the Bacterial Susceptibility Data

| Number of observations per patient | Number of patients |
|---|---|
| 1 | 288 |
| 2 | 67 |
| 3 | 5 |
| 4 | 2 |
| 6 | 1 |

Table 1.5: Resistance status for isolates in the Bacterial Susceptibility Data

| Resistance status | Bacteria | Treatment | Number of isolates |
|---|---|---|---|
| Susceptible | PN | CD | 150 |
| | | AF | 33 |
| | PY | CD | 148 |
| | | AF | 37 |
| Non-susceptible | PN | CD | 38 |
| | | AF | 26 |
| | PY | CD | 6 |
| | | AF | 13 |

AF: treatment with macrolides or tetracyclines

CD: treatment with penicillins or cephalosporins.

PY: *Streptococcus pyogenes*; PN: *Streptococcus pneumoniae*

In Chapter 8, we will assess the difference in persistence of resistance after treatment with penicillins or cephalosporins versus macrolides or tetracyclines. Other studies have suggested that resistance in oral streptococci lasts for more than six months after exposure to macrolides, while it is estimated to be much shorter after exposure to penicillins (Chung *et al.*, 2007; Malhotra-Kumar *et al.*, 2007, 2016). Because designing a new study for every drug-bug combination is expensive and time-consuming, an additional aim of Chapter 8 will be to assess whether routinely collected data on resistance and antibiotic use at the level of the individual patient can confirm the conclusions reached in the studies conducted by Malhotra-Kumar *et al.* (2007), Malhotra-Kumar *et al.* (2016) and Chung *et al.* (2007) and hence could serve as a proxy to study other drug-bug combinations.

### 1.2.7 Hospital Stays Data

Data on hospital stays in acute care hospitals in Belgium were collected yearly between 1999 and 2010 for pathology-based financing purposes by a collaboration of the "National Institute for Health and Disability Insurance" and the "Federal Government Finances". Each entry was classified according to the "all patient refined diagnosis related groups" (APR-DRG) system (3M Health Information Systems, 2003) and includes information on the patient, the hospital, the antimicrobial consumption and the stay itself. Two APR-DRGs with the highest degree of antimicrobial consumption, being APR-DRG 302 (major lower limb surgery without trauma) and APR-DRG 139 (simple pneumonia), were selected for further study. Included explanatory variables are listed in Table 1.6.

Table 1.6: Overview of explanatory variables in the Hospital Stays Data.

| Variable | Description |
|---|---|
| Age | Age of the patient |
| Gender | Gender of the patient (male, female) |
| Los | Length of stay in days |
| Time | Year of stay (with $1999 = 1$) |
| Sev | Severity of the stay (level $1 - 4$) |
| ICU | Stay at intensive care (yes, no) |
| Size | Number of stays in one year |
| Pt_ori | Patient origin (home, long term care, other hospital, unknown) |
| Dis_st | Discharge status (dead, alive) |
| Comp | Compliance to guidelines for limb surgery |
| OP | Ratio of oral versus parenteral antimicrobial use during the stay |
| DDDhosp | Number of defined daily doses consumed during the stay |

For major lower limb surgery, the outcome of interest was compliance to guidelines at patient level. Compliance was defined as use of the correct antimicrobial (i.e. cefazolin) in the correct dose range (i.e. $2-8$g) while no other antimicrobial was given. Hospital stays with secondary infectious diagnoses or with severity of illness levels 3 or 4 were excluded from the analysis to ensure that the antimicrobial was prescribed purely for prophylaxis.

For pneumonia, the data did not allow an assessment of the appropriateness of the antimicrobial use on stay level. Therefore, the information on hospital stays was aggregated at hospital level. Aggregating the explanatory variables was done by using the median for continuous outcomes and using the distribution of the stays (%) according to the levels of the categorical variables. The outcomes of interest at hospital level were the number of total DDD per 100 hospital days (excluding penicillins)(DDDhosp) and the ratio of oral versus parenteral DDD (OP). Figure 1.3 shows that there is a lot of variability between hospitals for both pneumonia outcomes. This suggests that there is a need for subject-specific intercepts and slopes. Outliers were detected using box plots with an observation lying below the lower far fence ($25th$ percentile $-3$IQR) or above the upper far fence ($75th$ percentile $+3$IQR) being labelled outlying and discarded.

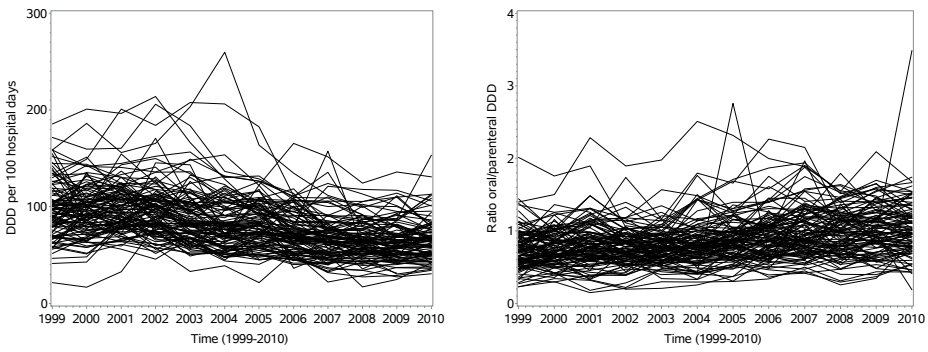



Figure 1.3: Observed hospital-specific changes in DDD per 100 hospital days (left) and ratio oral/parenteral antimicrobial use (right).

In Chapter 9, the Hospital Stays Data will be used to assess the impact of Belgian policies on antimicrobial consumption.

# Chapter 2

# Predicting poor prognosis in patients presenting to primary care with acute cough

Existing prognostic prediction rules have been derived to predict mortality in patients presenting to the emergency department with community-acquired pneumonia (CAP). The Pneumonia severity index (PSI), developed by Fine *et al.* (1997), uses two steps to classify patients in five groups according to their risk of mortality (more information in Figure A1). The final score used to classify patients in risk groups is however not easily computed as it combines 20 different variables. Online tools have been made available to improve usability of PSI (e.g. www.internisten.nl/jniv/calculatoren/longziekten/items/pneumonia-severity-index-of-fine-score). Another way to circumvent this computation is to focus on the first step, separating the patients that are suitable for home management from those at risk of mortality. An alternative rule used to predict mortality in patients presenting to the emergency department with CAP, is the CRB score, which was developed by the British Thoracic Society and later modified to the CURB score by Neill *et al.* (1996) (more information in Figure A2). This prediction rule uses three (for CRB) or four (for CURB) easily measurable clinical features to distinguish between severe and non-severe pneumonia. Extensions of the CRB and CURB score, which combine usability and stratification into several risk groups, are, respectively, the CRB-65 and CURB-65 score, which

were developed by Lim *et al.* (2003) (more information in Figure A2).

Using a meta-analytic approach, Loke *et al.* (2010) and Akram *et al.* (2011) demonstrated that CRB-65, CURB-65 and PSI can be used to predict mortality from CAP in outpatients. However, since both death from CAP and CAP itself are very uncommon in outpatients, several authors have suggested to consider other outcomes (Bont *et al.*, 2008; Francis *et al.*, 2012; Vugt *et al.*, 2012).

In this chapter, we will use information within the Acute Cough Data (Section 1.2.1) that is readily available on the day of consultation to construct a framework for the prediction of poor prognosis in patients presenting to primary care with acute cough. Additionally, we will assess the added value of including information on biomarkers CRP and BUN, and compare the performance of the new prediction rule to the performance of five existing prediction rules (PSI, CRB, CURB, CRB-65 and CURB-65).

## 2.1 Imputation of missing values

Only seven out of the 97 explanatory variables that were available on the day of consultation were complete. For the remaining 90 variables, between 0.03% and 21.71% of records were missing. These missing values were imputed using predictive mean matching, which fits a regression model for each variable with missing observations, conditional on the other variables. The imputed value for a subject that has its record missing is the observed value for the subject with the closest predicted value. Because we expect observations within one country to be more similar than observations between different countries, missing values were imputed per country. To ensure that the uncertainty on the missing values is represented in the imputed observations, we used multiple imputations (available as R package **mice**) (Rubin, 1987; Van Buuren and Groothuis-Oudshoorn, 2011). After the imputation of missing values, the imputed and observed values were combined into five imputed datasets. All analyses were repeated for each imputed dataset, after which results were pooled.
In order to evaluate the added value of the inclusion of biomarkers, missing values for CRP (35.11%) and BUN (37.68%) were imputed, conditional on the imputed country-specific datasets.

## 2.2 Development of new prediction rule

To account for the difference in baseline risk of poor prognosis, countries were grouped according to the observed proportion of patients with poor prognosis (A: $< 15\%$, B: $15-25\%$, C: $> 25\%$) and group-specific prediction rules were constructed. In a second stage, we used these group-specific prediction rules to construct a general prediction rule.

### 2.2.1 Construction of a group-specific prediction rule

We selected the most important variables for each imputed dataset using a forest of conditional inference trees (available as R package **cforest**) (Hothorn *et al.*, 2006). This method was preferred over the random forest approach to avoid bias towards variables with more categories (more elaborately discussed in Section 2.7). The conditional forest approach draws *ntree* bootstrap samples from the original sample and fits a conditional inference tree using *mtry* explanatory variables to each of the bootstrapped samples. At the beginning of each tree, a test statistic is computed for the null hypothesis of independence between each explanatory variable and the response. The explanatory variable rendering the lowest univariate p-value is then used to implement a split. This process repeats itself, until the null hypothesis can no longer be rejected. The importance of the variables in a forest is then represented by the decrease in mean accuracy. This variable importance measure is computed by comparing the prediction accuracy (i.e. the number of observations correctly classified) before and after randomly permuting the predictor of interest. If there is an association between the predictor and the response, the prediction accuracy will decrease substantially by permuting the predictor and the decrease in mean accuracy will therefore be high.

Within each tree, we considered 10 explanatory variables. Within each forest, we used 1000 trees, to eliminate instability of an individual tree. Variance importance measures were then averaged over 100 forests to eliminate instability of an individual forest.

After selection of the important variables, a logistic regression model using these variables was fitted for each imputed dataset. This imputation-specific model can be presented as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki},$$

where $\pi_i$ is the probability of poor prognosis for patient $i$ (with $i = 1, ..., n$), $X_{ki}$ is the $k$th covariate, $\beta_0$ is the intercept, $\beta_k$ are the coefficients for the covariates $X_{ki}$ and $K$ is the number of included covariates.
Insignificant fixed effects ($\alpha = 0.10$) were removed using backwards model building based on p-values obtained by the likelihood ratio test. In a next step, interaction terms between the remaining fixed effects were included whenever the interaction term did not contain sparse levels, and a second round of backwards elimination ($\alpha = 0.05$) was performed.

Variables which were significant in at least two imputation-specific models were retained. The final group-specific model was obtained after a final round of backwards elimination ($\alpha = 0.05$) using p-values obtained by the pooled likelihood ratio test (Meng and Rubin, 1992). This test statistic is defined as:

$$D_L = \frac{\bar{d}_L}{k(1+r_L)},$$

with

$$r_L = \frac{m+1}{k(m-1)}(\bar{d}'_m - \bar{d}_L),$$

where $m$ is the number of imputations, $\bar{d}'_m$ is the likelihood ratio averaged over $m$ imputations, $\bar{d}_L$ is the likelihood ratio averaged over $m$ imputations and evaluated using $\bar{\theta}_F$ and $\bar{\theta}_R$, which are the pooled parameter estimates for the full and reduced model, respectively, and $k$ is the number of parameters of interest. In order to determine significance of test statistic $D_L$, it is compared to an F distribution with numerator degrees of freedom equal to $k$ and denominator degrees of freedom equal to $\nu$, which can be calculated as:

$$\nu = 4 + (km - k - 4)\left[1 + \left(1 - \frac{2}{km-k}\right)\frac{1}{r_L}\right]^2.$$

The final group-specific model was then fitted to each of the imputed datasets and results were pooled over the five imputations.

### 2.2.2 Construction of the general prediction rule

In a second stage, the five imputed datasets were combined into five completed datasets. All variables which were significant in at least one group-specific model were used to construct a general model. This model can be represented as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \sum_{k=1}^{K} \beta_k X_{kij}, \tag{2.1}$$

where $\pi_{ij}$ is the probability of poor prognosis for patient $i$ (with $i = 1, ..., n_j$) in group $j$ (with $j$ = A, B or C), $n_j$ are the number of patients within group $j$, $X_{kij}$ is the $k$th covariate, $\beta_{0j}$ is the group-specific intercept, $\beta_k$ are the coefficients for the covariates $X_{kij}$ and $K$ is the number of included covariates.
The final general model was obtained after a final round of backwards elimination ($\alpha = 0.05$) using p-values obtained by the pooled likelihood ratio test. This model was then fitted to each completed dataset after which the results were pooled over the five datasets.

The model's ability to discriminate between observations at high and low risk was evaluated using a receiver operating characteristic (ROC) curve. This is a plot of sensitivity (i.e. true positive rate) versus 1-specificity (i.e. false positive rate) at different cut-off values. The closer the curve gets to the left top border of the graph, the more accurate the model (i.e. high sensitivity and specificity). The ROC curve can be summarized by the area under the curve (AUC), which reflects the probability that the score for a case exceeds the score for a control in a random case-control pair. It can range from 0.5, corresponding to no discriminative ability, to 1, corresponding to perfect discrimination (Hosmer and Lemeshow, 2000).

The predicted probability of poor prognosis was obtained by filling in the pooled parameter estimates into Equation 2.1 and taking the inverse-logit. If the predicted probability is higher than a cut-off value $c$, the patient is at high risk of poor prognosis. If it is lower than that cut-off value, the patient is at low risk of poor prognosis. Selection of the cut-off value holds an intrinsic trade-off between high sensitivity (for a low cut-off value) and high specificity (for a high cut-off value). We determined the optimal cut-off value $c$ using the Youden index, which maximizes the distance between the ROC curve and the identity line by maximizing the sum of sensitivity and specificity (Figure 2.1)(Youden, 1950).

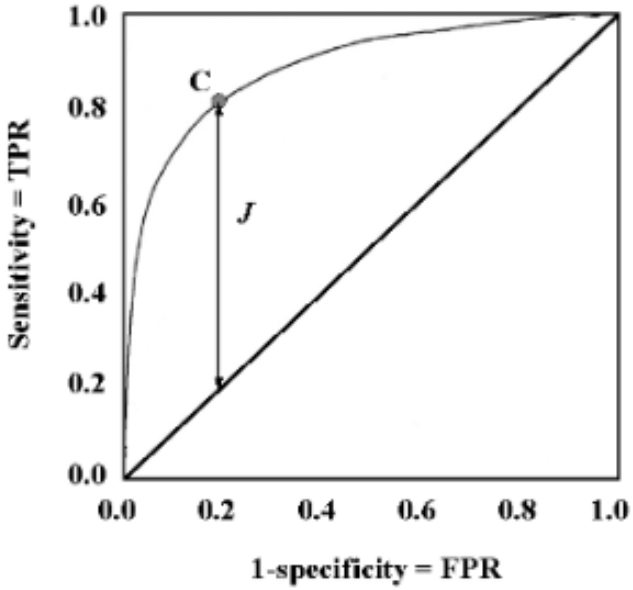



Figure 2.1: Illustration of a ROC curve and its Youden index (J), which maximizes the sum of sensitivity and specificity (C = optimal Youden cut-off point).
*Taken from Zaletel-Kragelj and Božikov (2010)*

## 2.3 Extension of the new prediction rule

The new prediction rule focusses on information that is available at the moment of consultation. However, a GP can order and even perform additional testing during the consultation e.g. determination of CRP. This biomarker is believed to have a high predictive value for pneumonia when combined with signs and symptoms (Van Vugt *et al.*, 2013). Another biomarker that is currently used to distinguish between severe and non-severe pneumonia is BUN.
To assess the relevance of CRP and BUN in predicting poor prognosis, we included them in the final general model (separately) and computed their pooled p-values. The improvement in discriminative ability was assessed using their AUC values.

## 2.4 Cross-validation of the new prediction rule

We used a cross-validation approach to evaluate the stability of the new prediction rule. For this procedure, the completed datasets were split in three sets of equal size. The number of cross-validations was chosen ad hoc, after considering the size of the dataset (2604).

**Cross-validation 1:**
Use set 1 and 2 as learning sample, and set 3 as test sample.

**Cross-validation 2:**
Use set 2 and 3 as learning sample, and set 1 as test sample.

**Cross-validation 3:**
Use set 1 and 3 as learning sample, and set 2 as test sample.

Next, we iterated through the following steps for each Cross-validation:

- Using the learning sample, reduce the full general model through backwards elimination and obtain pooled parameter estimates.
- Using these pooled estimates fit the reduced model on the test sample and determine its AUC.

## 2.5 Comparison to existing prediction rules

Existing prediction rules include PSI, CRB, CURB, CRB-65 and CURB-65. For PSI, we focussed on the first step, and used variables *Age*, *Heart_fail_yn*, *Other_fine_diseases*, *Conf_disor*, *Beats_min*, *Breaths_min*, *Syst_bp* and *Oral_temp* to obtain a 0/1 categorisation (according to Figure A1). Scores for CRB, CURB, CRB-65 and CURB-65 were obtained using variables *Conf_disor* (C), *BUN* (U), *Breaths_min* (R), *Syst_bp* and *Diast_bp* (B) and *Age* (65) (according to Figure A2).
The overall performance of the underlying models for the new and existing prediction rules was evaluated using their AUC values. The performance of the prediction rules themselves, with cut-off values determined by maximizing the Youden index, was compared using this index, which provides an objective means of comparing prognostic tests and is calculated as $sensitivity + specificity - 1$ (Youden, 1950).

## 2.6 Results

### 2.6.1 Development of a group-specific prediction rule

The most important variables for each imputed dataset were obtained using forests of conditional inference trees. Afterwards, logistic regression was used to obtain imputation-specific models. Using variables that were significant in at least two imputation-specific models, the group-specific models were constructed.
The resulting variable importance plots are shown in Figures 2.2 to 2.4. The parameter estimates and standard errors for the final group-specific prediction rules are reported in Tables 2.1 to 2.3. For completeness, we show both imputation-specific and pooled results. Our main interest however are the pooled results, where the standard error accounts for variability both within and between imputations.

## Group A

The variable importance plots for the five imputed datasets for group A are given in Figure 2.2.

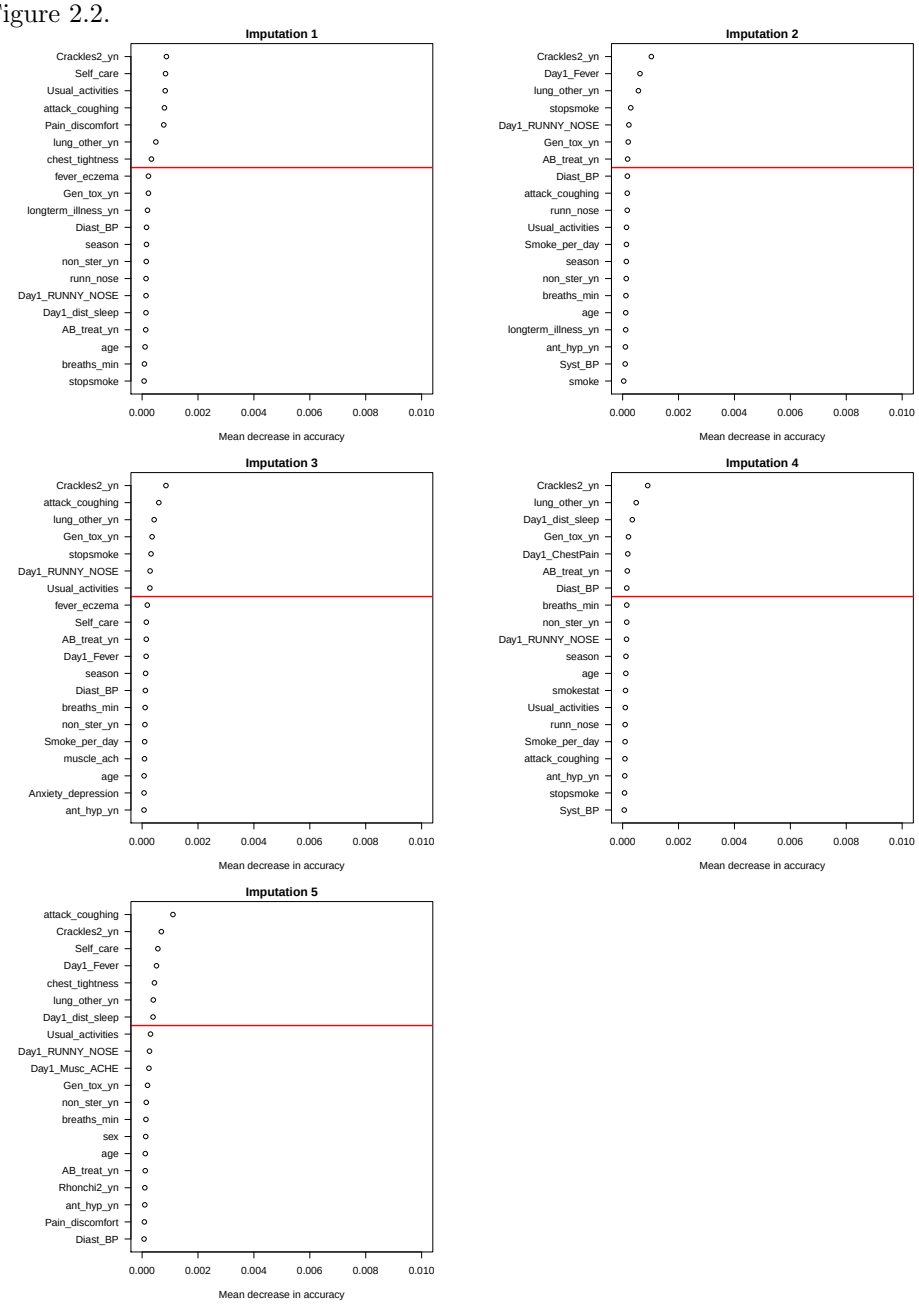



Figure 2.2: Variable importance plots for group A with the top seven predictors.

From the variable importance plots (Figure 2.2), we selected the top seven predictors. This number was chosen ad hoc, after inspection of the plots and considering the number of observations in group A (i.e. 594). Using this set of selected predictors, we constructed five imputation-specific models. Variables that were significant in at least two imputation-specific models were used to construct a group-specific logistic regression model. The final group-specific model was fitted to each imputed dataset, after which the results were pooled. Parameter estimates and standard errors are reported in Table 2.1.

Table 2.1: Parameter estimates (standard errors) for the final group-specific model for group A.

| Parameter | Imputation 1 | Imputation 2 | Imputation 3 | Imputation 4 | Imputation 5 | Pooled results |
|---|---|---|---|---|---|---|
| Intercept | 3.980 (1.056) | 3.423 (1.039) | 3.799 (1.046) | 1.412 (1.046) | 4.300 (1.080) | 3.783 (1.132) |
| Lung_other_yn | -1.297 (0.436) | -1.239 (0.425) | -1.297 (0.433) | -1.296 (0.428) | -1.346 (0.428) | -1.295 (0.432) |
| Attack_coughing | -0.954 (0.269) | -0.640 (0.256) | -0.876 (0.265) | -0.571 (0.255) | -1.141 (0.278) | -0.837 (0.368) |
| Crackles2_yn | -1.054 (0.325) | -1.037 (0.320) | -1.012 (0.323) | -1.022 (0.320) | -1.044 (0.329) | -1.034 (0.324) |

For patients in group A, the odds of poor prognosis is affected by the presence of lung diseases other than asthma or chronic obstructive pulmonary disorder ($p = 0.0031$), the presence of coughing attacks ($p = 0.0308$) and the presence of crackles during the GPs physical examination ($p = 0.0022$).

## Group B

The variable importance plots for the five imputed datasets for group B are given in Figure 2.3.

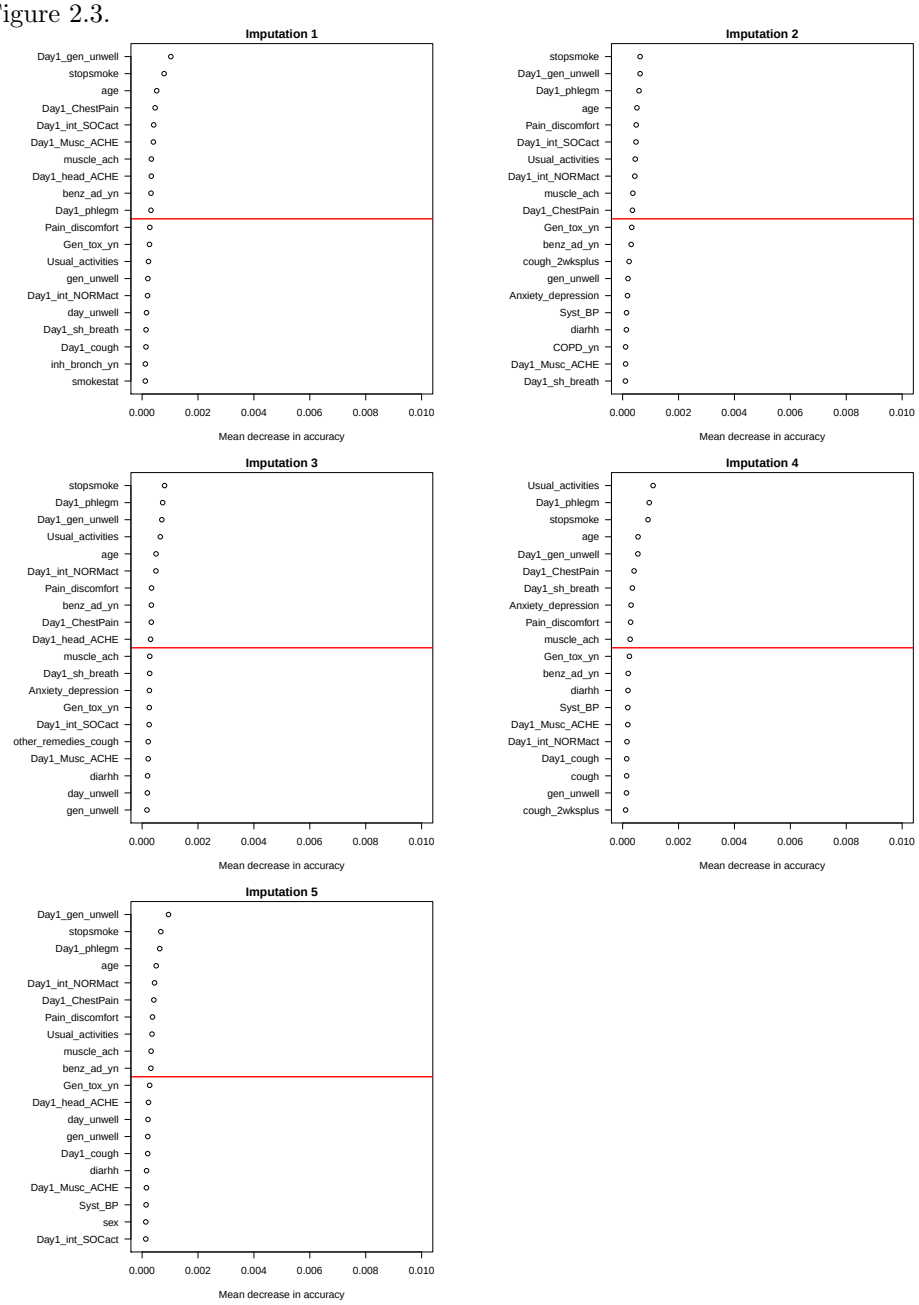



Figure 2.3: Variable importance plots for group B with the top ten predictors.

From the variable importance plots (Figure 2.3), we selected the top ten predictors. This number was chosen ad hoc, after inspection of the plots and considering the number of observations in group B (i.e. 1821). Using this set of selected predictors, we constructed five imputation-specific models. Variables that were significant in at least two imputation-specific models were used to construct a group-specific logistic regression model. The final group-specific model was fitted to each imputed dataset, after which the results were pooled. Parameter estimates and standard errors are reported in Table2.2.

Table 2.2: Parameter estimates (standard errors) for the final group-specific model for group B.

| Parameter | Imputation 1 | Imputation 2 | Imputation 3 | Imputation 4 | Imputation 5 | Pooled results |
|---|---|---|---|---|---|---|
| Intercept | -0.969 (0.379) | -1.013 (0.383) | -1.130 (0.385) | -1.042 (0.382) | -1.042 (0.384) | -1.039 (0.388) |
| Benz_ad_yn | -0.455 (0.181) | -0.439 (0.181) | -0.421 (0.182) | -0.451 (0.182) | -0.451 (0.181) | -0.443 (0.182) |
| Usual_activities (2) | 0.366 (0.123) | 0.407 (0.124) | 0.478 (0.124) | 0.412 (0.124) | 0.412 (0.124) | 0.415 (0.131) |
| Usual_activities (3) | 0.522 (0.216) | 0.719 (0.218) | 0.782 (0.215) | 0.741 (0.211) | 0.741 (0.212) | 0.701 (0.242) |
| Day1_phlegm (1) | -0.381 (0.246) | -0.460 (0.255) | -0.549 (0.264) | -0.334 (0.260) | -0.334 (0.248) | -0.412 (0.274) |
| Day1_phlegm (2) | -0.065 (0.225) | -0.051 (0.225) | -0.021 (0.229) | -0.078 (0.224) | -0.078 (0.228) | -0.059 (0.228) |
| Day1_phlegm (3) | 0.178 (0.192) | 0.206 (0.195) | 0.232 (0.197) | 0.192 (0.192) | 0.192 (0.198) | 0.200 (0.196) |
| Day1_phlegm (4) | 0.509 (0.201) | 0.492 (0.200) | 0.596 (0.202) | 0.583 (0.197) | 0.583 (0.203) | 0.552 (0.207) |
| Day1_phlegm (5) | 0.548 (0.233) | 0.588 (0.238) | 0.472 (0.239) | 0.519 (0.243) | 0.519 (0.241) | 0.529 (0.243) |
| Day1_phlegm (6) | 0.132 (0.302) | -0.038 (0.320) | 0.097 (0.317) | 0.077 (0.312) | 0.077 (0.317) | 0.069 (0.322) |
| Stopsmoke | 0.007 (0.002) | 0.006 (0.002) | 0.007 (0.002) | 0.007 (0.002) | 0.007 (0.002) | 0.007 (0.002) |

For patients in group B, the odds of poor prognosis is impacted by the use of antidepressants ($p = 0.0204$), the time since the patient last smoked ($p = 0.0069$), the severity of interference with daily activities ($p = 0.0016$) and the severity of phlegm as assessed by the patient in its diary ($p = 0.0005$).

### Group C

The variable importance plots for the five imputed datasets for group C are given in Figure 2.4.

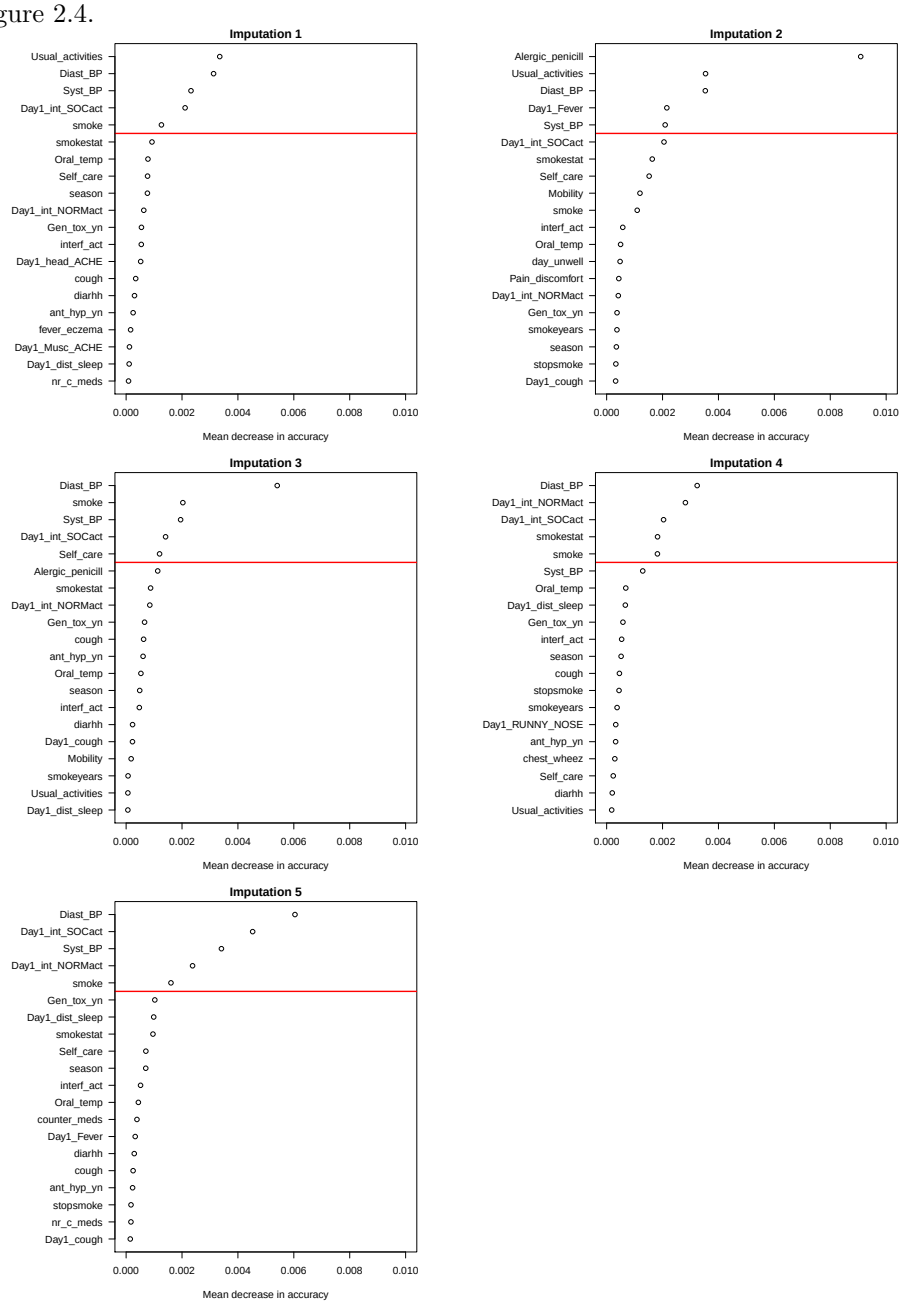



Figure 2.4: Variable importance plots for group C with the top five predictors.

From the variable importance plots (Figure 2.4), we selected the top five predictors. This number was chosen ad hoc, after inspection of the plots and considering the number of observations in group C (i.e. 189). Using this set of selected predictors, we constructed five imputation-specific models. Variables that were significant in at least two imputation-specific models were used to construct a group-specific logistic regression model. The final group-specific model was fitted to each imputed dataset, after which the results were pooled. Parameter estimates and standard errors are reported in Table 2.3.

Table 2.3: Parameter estimates (standard errors) for the final group-specific model for group C.

| Parameter | Imputation 1 | Imputation 2 | Imputation 3 | Imputation 4 | Imputation 5 | Pooled results |
|---|---|---|---|---|---|---|
| Intercept | 2.789 (1.294) | 2.539 (1.278) | 3.029 (1.306) | 2.362 (1.273) | 2.994 (1.299) | 2.743 (1.328) |
| Diast_BP | -0.046 (0.016) | -0.044 (0.016) | -0.049 (0.016) | -0.041 (0.016) | -0.049 (0.016) | -0.046 (0.016) |
| Smoke (2) | -1.057 (0.513) | -1.024 (0.510) | -1.062 (0.514) | -1.062 (0.510) | -1.038 (0.514) | -1.049 (0.513) |
| Smoke (3) | 0.443 (0.376) | 0.479 (0.374) | 0.448 (0.377) | 0.439 (0.375) | 0.441 (0.377) | 0.450 (0.376) |

For patients in group C, the odds of poor prognosis is impacted by the patient's diastolic blood pressure ($p = 0.0038$) and smoking status ($p = 0.0090$).

### 2.6.2 Development of a general prediction rule

Variables that were significant in at least one group-specific model were used to construct the general model on the five completed datasets. A group-specific intercept was included to correct for the baseline risk of poor prognosis. The final general model was fitted to each completed dataset after which the results were pooled. Parameter estimates and standard errors are reported in Table 2.4.

Table 2.4: Parameter estimates (standard errors) for the final model.

| Parameter | Imputation 1 | Imputation 2 | Imputation 3 | Imputation 4 | Imputation 5 | Pooled results |
|---|---|---|---|---|---|---|
| Group A | -0.280 (0.482) | -0.245 (0.482) | -0.366 (0.486) | -0.215 (0.483) | -0.277 (0.483) | -0.277 (0.487) |
| Group B | 0.241 (0.476) | 0.261 (0.476) | 0.150 (0.480) | 0.300 (0.479) | 0.235 (0.477) | 0.237 (0.482) |
| Group C | 0.540 (0.505) | 0.571 (0.502) | 0.474 (0.507) | 0.600 (0.507) | 0.534 (0.503) | 0.544 (0.507) |
| Crackles2_yn | -0.414 (0.153) | -0.413 (0.154) | -0.390 (0.154) | -0.447 (0.153) | -0.395 (0.154) | -0.412 (0.156) |
| Day1_phlegm (1) | -0.530 (0.208) | -0.593 (0.202) | -0.628 (0.218) | -0.789 (0.217) | -0.478 (0.205) | -0.604 (0.249) |
| Day1_phlegm (2) | -0.054 (0.177) | 0.022 (0.176) | 0.030 (0.178) | -0.055 (0.176) | -0.140 (0.178) | -0.039 (0.193) |
| Day1_phlegm (3) | 0.018 (0.157) | 0.095 (0.158) | 0.071 (0.159) | 0.030 (0.156) | -0.029 (0.159) | 0.037 (0.166) |
| Day1_phlegm (4) | 0.194 (0.169) | 0.272 (0.168) | 0.319 (0.168) | 0.275 (0.166) | 0.267 (0.167) | 0.265 (0.175) |
| Day1_phlegm (5) | 0.245 (0.195) | 0.336 (0.198) | 0.226 (0.198) | 0.192 (0.203) | 0.196 (0.201) | 0.239 (0.209) |
| Day1_phlegm (6) | -0.218 (0.242) | -0.310 (0.252) | -0.134 (0.245) | -0.320 (0.248) | -0.343 (0.248) | -0.265 (0.265) |
| Usual_activities (2) | 0.268 (0.106) | 0.291 (0.106) | 0.349 (0.106) | 0.358 (0.106) | 0.303 (0.106) | 0.314 (0.114) |
| Usual_activities (3) | 0.817 (0.177) | 0.875 (0.183) | 0.889 (0.181) | 0.962 (0.179) | 0.861 (0.175) | 0.881 (0.188) |
| Stopsmoke | 0.006 (0.002) | 0.006 (0.002) | 0.006 (0.002) | 0.006 (0.002) | 0.006 (0.002) | 0.006 (0.002) |
| Diast_BP | -0.014 (0.005) | -0.015 (0.005) | -0.014 (0.005) | -0.014 (0.005) | -0.014 (0.005) | -0.014 (0.005) |

In general, the odds of poor prognosis is impacted by the baseline risk for poor prognosis (group, $p < 0.0001$), the presence of crackles during the GPs physical examination ($p = 0.0117$), the severity of phlegm as assessed by the patient ($p = 0.0047$), the severity of interference with daily activities ($p < 0.0001$), the time since the patient last smoked ($p = 0.0045$) and the patient's diastolic blood pressure ($p = 0.0020$).

The AUC values for the final general and group-specific models fitted using their pooled parameter estimates are given in Table 2.5. The ROC curves for the final general model are visualised in Figure 2.5. Both AUC values and ROC curves show that the discriminative power of the new prediction rule is acceptable, although there is still room for improvement.

Table 2.5: Area under the curve for the final group-specific and general models.

| | Group-specific models | | | General |
|---|---|---|---|---|
| Imputation | Group A | Group B | Group C | model |
| 1 | 0.66 | 0.61 | 0.71 | 0.60 |
| 2 | 0.64 | 0.62 | 0.69 | 0.61 |
| 3 | 0.65 | 0.63 | 0.70 | 0.61 |
| 4 | 0.63 | 0.64 | 0.69 | 0.62 |
| 5 | 0.67 | 0.62 | 0.70 | 0.60 |

The new prediction rule was obtained by filling in the final general model's pooled parameter estimates (Table 2.4 last column) into Equation 2.1 and taking the inverse-logit. The optimal cut-off value, averaged over the five completed datasets, was 0.182. The new prediction rule hence classifies a patient to be at high risk for poor prognosis when the predicted probability is over 0.182, and classifies the patient to be at low risk for poor prognosis when the predicted probability is below this threshold. At this threshold, sensitivity and specificity, averaged over the five completed datasets, equal 0.701 and 0.450, respectively.

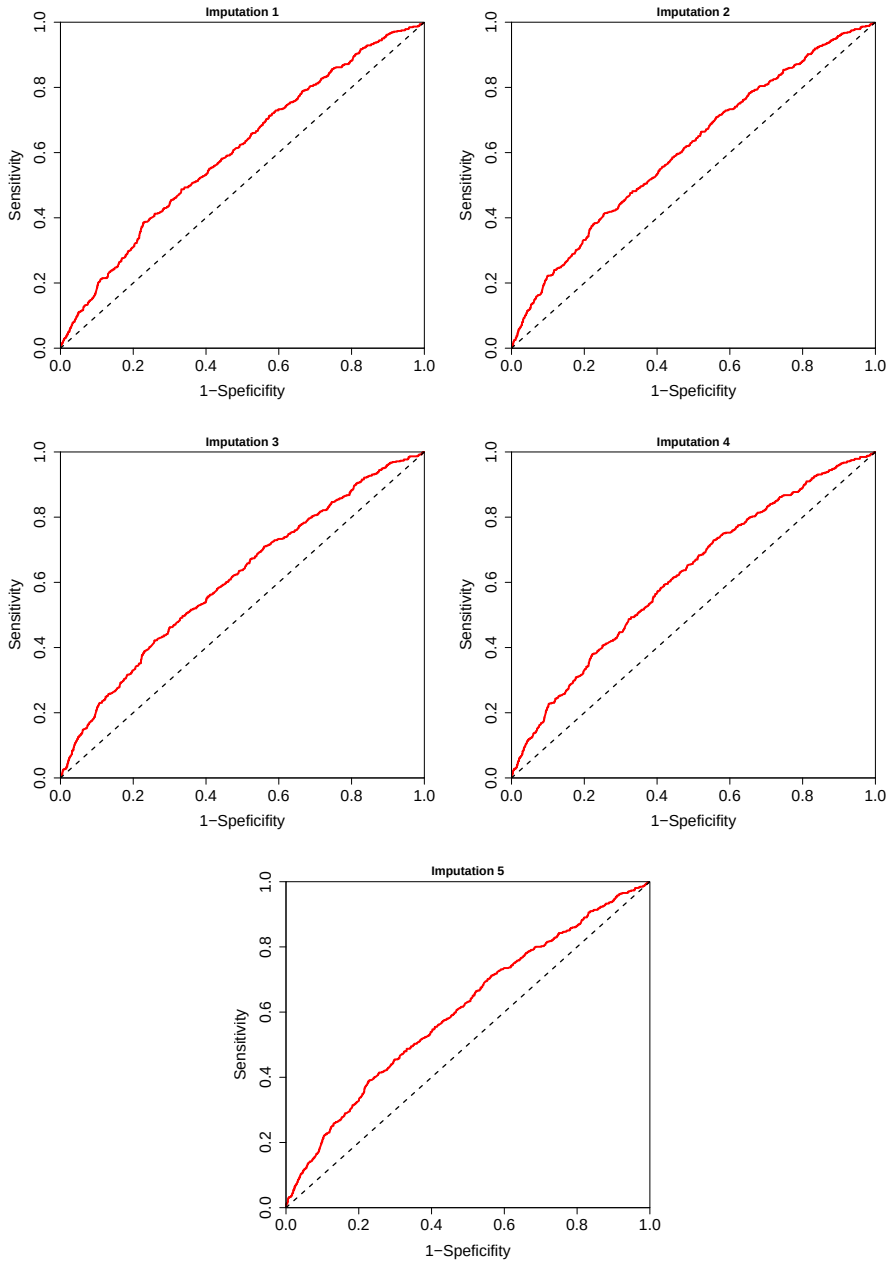



Figure 2.5: ROC curves for the general model fitted to the five completed datasets.

### 2.6.3 Extension of the new prediction rule

CRP and BUN were added to the final general model separately. Pooled p-values for both predictors however showed that they were redundant (p-value= 0.5497 and 0.9964, respectively). AUC values for the final general model before and after inclusion of either CRP or BUN are reported in Table 2.6. The AUC values verify that neither CRP nor BUN considerably improved the final general model used to construct the new prediction rule.

Table 2.6: Area under the curve for the final general model both before and after inclusion of either CRP or BUN.

| | Imputation number | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Final general model | 0.60 | 0.61 | 0.61 | 0.62 | 0.60 |
| Final general model + CRP | 0.63 | 0.63 | 0.64 | 0.65 | 0.63 |
| Final general model + BUN | 0.63 | 0.63 | 0.64 | 0.64 | 0.63 |

### 2.6.4 Cross-validation of the new prediction rule

A three-fold cross-validation approach was used to validate the new prediction rule. For each of the three learning samples, we reduced the full general model using backwards elimination. Pooled parameter estimates and standard errors for the full and the reduced models are reported in Tables A3 and A4. Using the pooled parameter estimates, we computed the AUC of the reduced model fitted to the respective test data (Table 2.7).

Table 2.7: Area under the curve for the three cross-validations.

| | Imputation number | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Cross-validation 1 | 0.59 | 0.60 | 0.60 | 0.60 | 0.59 |
| Cross-validation 2 | 0.60 | 0.59 | 0.60 | 0.60 | 0.59 |
| Cross-validation 3 | 0.59 | 0.60 | 0.60 | 0.61 | 0.61 |

Out of the nine predictors present in the full general model, similar variables were kept in the final general model and all three reduced models. The top three predictors (i.e. predictors with the lowest p-value in the final general model) were present in all three reduced models. This indicates that the new prediction rule is quite stable. AUC values for all reduced models were comparable to the AUC of the final general model (averaged AUC = 0.61), verifying that the model's stability is acceptable.

### 2.6.5 Comparison of available prediction rules

We compared the performance of the underlying models for the new and existing prediction rules (PSI, CRB, CURB, CRB-65 and CURB-65) using their AUC values. The optimal cut-off value for each prediction rule was determined by averaging the optimal cut-off over the five completed datasets. Performance of the prediction rules was compared using the Youden index at this averaged cut-off point. Both comparative measures are reported in Table 2.8. The results show that the new prediction rule outperformed the existing prediction rules in the prediction of poor prognosis in patients presenting to primary care with acute cough.

Table 2.8: Area under the curve (AUC) and Youden index for the new and existing prediction rules.

| | | Imputation number | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| New | AUC | 0.60 | 0.61 | 0.61 | 0.62 | 0.60 |
| | Youden | 0.15 | 0.14 | 0.16 | 0.15 | 0.15 |
| PSI | AUC | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 |
| | Youden | 0.03 | 0.03 | 0.02 | 0.03 | 0.02 |
| CRB | AUC | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 |
| | Youden | 0.06 | 0.06 | 0.06 | 0.05 | 0.06 |
| CURB | AUC | 0.53 | 0.54 | 0.53 | 0.54 | 0.54 |
| | Youden | 0.05 | 0.07 | 0.06 | 0.07 | 0.08 |
| CRB-65 | AUC | 0.53 | 0.53 | 0.53 | 0.53 | 0.54 |
| | Youden | 0.06 | 0.06 | 0.06 | 0.06 | 0.07 |
| CURB-65 | AUC | 0.53 | 0.54 | 0.53 | 0.53 | 0.54 |
| | Youden | 0.05 | 0.06 | 0.05 | 0.05 | 0.07 |

## 2.7 Discussion

Currently there are no prognostic prediction rules to predict poor prognosis in patients presenting to primary care with acute cough. The only alternative available is the use of prognostic prediction rules that were developed to predict mortality in patients presenting to the emergency department with CAP (e.g. PSI and CRB-65) (Fine *et al.*, 1997; Neill *et al.*, 1996). Although the use of these prediction rules has been demonstrated to predict mortality in outpatients (Akram *et al.*, 2011; Loke *et al.*, 2010), we showed that they perform poorly when predicting poor prognosis.

In this chapter, we set out to develop a new prognostic prediction rule to more accurately predict poor prognosis in patients presenting to primary care with acute cough. In order to account for the fact that there is a different baseline probability to experience poor prognosis in different countries, countries were grouped according to this baseline risk (in three groups: $< 15\%$, $15-25\%$ and $> 25\%$). To take into account that different predictors might be important when baseline probability of poor prognosis differs, we started by constructing three group-specific models. Important predictors for poor prognosis in group A are the presence of lung diseases other than asthma or chronic obstructive pulmonary disorder, the presence of coughing attacks and the presence of crackles during the GPs physical examination. Important predictors for group B are the use of antidepressants, the time since the patient last smoked, the severity of interference with daily activities and the severity of phlegm as assessed by the patient. Important predictors for poor prognosis in group C are the patient's diastolic blood pressure and smoking status. These three group-specific models were then combined into a general model, including a group-specific intercept to correct for difference in baseline risk of poor prognosis. Important predictors in the general model are the presence of crackles during the GPs physical examination, the severity of phlegm as assessed by the patient, the severity of interference with daily activities, the time since the patient last smoked and the patient's diastolic blood pressure.

Discriminative power of both the group-specific and the general final models were adequate, although there is still room for improvement. In an attempt to improve the discriminative power of the final general model, measurements of CRP or BUN were included. Both variables however were not significant upon addition to the final model and did not improve the discriminative power.

The final prediction rule was obtained using the pooled parameter estimates for the final general model. The discriminative power and Youden index were compared between the new prediction rule and five existing prediction rules (PSI, CRB, CURB, CRB-65 and CURB-65). In this comparison, we deliberately did not compare sensitivity and specificity as such, because they are highly affected by the choice of the cut-off value and could give a misleading idea. The Youden index was chosen as an alternative because it was maximized in determining the optimal cut-off value, which ensures that we are comparing the optimal Youden values, and was developed to serve as an objective manner to compare between different prognostic tests (Youden, 1950). The comparison between the new and existing prediction rules showed that although there is room for improvement in the new prediction rule, it already outperforms the available prediction rules and hence is the most reliable option to determine the risk for poor prognosis to date.

### 2.7.1 Conditional versus random forest approach

Both the conditional and the random forest approach have the same basis. They both start by taking a bootstrap sample from the original sample and then use this bootstrap sample and a small random selection of predictor variables to fit an unpruned classification tree. To avoid instability of an individual tree caused by small changes in the bootstrap sample, multiple trees are combined into a forest. The difference between both packages lies in the method used to construct an individual tree.

The random forest tree computes a split criterion for each possible cut-point within the range of a predictor. The variable selected for the next split is the one that produces the highest criterion value (e.g. Gini impurity) in its best cut-point. Variables with more potential cut-points (i.e. more categories) are more likely to produce a good criterion value by chance alone. Because of this, random forests show a preference for variables with more categories.
The conditional inference tree computes a test statistic for conditional independence between each predictor and the response. The variable selected for the next split is the one that produced the lowest univariate p-value. Because this test statistic incorporates the number of categories in its degrees of freedom, bias towards variables with many categories is avoided.
For this reason, random forests are not reliable in situations where predictors have different number of categories and conditional forests were used in this chapter (Hothorn *et al.*, 2006; Strobl *et al.*, 2007).

Additionally, Strobl *et al.* (2007) showed that a small preference towards predictors with a higher number of categories still occurs when bootstrapping with replacement while this is not observed when bootstrapping without replacement. We therefore followed their recommendation and used bootstrapping without replacement in the construction of our conditional forests. An explanation for this bias is that, even if sampled under the null hypothesis of complete independence, samples that are bootstrapped with replacement might either exclude or multiply include certain observations by chance, which causes the bootstrap sample distribution to deviate slightly from the null hypothesis. This effect is more pronounced for variables with more categories, because in larger cross-tables the absolute cell counts are smaller than in smaller cross-tables, which enlarges the impact of excluding or doubling observations.

# Chapter 3

# Performance of the linear multi-level model in the presence of sparseness at the lowest level

Data that are collected in e.g. medical sciences often have a hierarchical structure. This means that units at a lower level (secondary units) are nested within units at a higher level (primary units) (Snijders and Bosker, 1999). Some well-known examples of such hierarchies include patients nested within hospitals, workers nested within factories and animals nested within litters. Multi-level hierarchies also occur frequently (e.g. students nested within classes within schools within cities within countries). As subjects that are nested within one unit tend to be more alike than subjects from different units, the observations are typically no longer independent. Ignoring dependency will usually cause a downward bias in the standard errors, resulting in possible misinterpretation of the effect of predictor variables (Garson, 2013; Hox, 1998; Kreft and De Leeuw, 1998; Moulton, 1986). To account for the hierarchical nature of the data, multi-level models, also known as linear mixed models, are often used (Goldstein, 2003).

In this chapter, we will study the impact of an increasing proportion of singletons (i.e. units containing only one subunit) on different aspects of the multi-level model, focussing on a two-level setting and including explanatory variables both at the pri-

mary and at the secondary level. We will assess whether, when high proportions of singletons are present, the model's performance improves by ignoring the dependency within units or by removing or grouping the singletons.

## 3.1 The linear multi-level model

The multi-level model can be used to model hierarchical data with a dependent variable defined at the lowest level (usually the subject) and explanatory variables at all levels. For example, suppose we have gathered data from $J$ hospitals, with $n_j$ patients in the $j$th hospital (with $j = 1, ..., J$). We have a dependent variable $Y_{ij}$ (e.g. antibiotic intake for patient $i$ in hospital $j$), an explanatory variable $X_{ij}$ (e.g. age) at the level of the patient and an explanatory variable $Z_j$ (e.g. hospital size) at the level of the hospital.

At the level of the patient, a regression equation can be set up to predict the outcome from the explanatory variables:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij}, \tag{3.1}$$

where $Y_{ij}$ and $X_{ij}$ are respectively the antibiotic intake and age of patient $i$ in hospital $j$, $\beta_{0j}$ and $\beta_{1j}$ are the intercept and slope and $\epsilon_{ij}$ is the residual error term. The intercept $\beta_{0j}$ and slope $\beta_{1j}$ are hospital-dependent and hence can be split into an overall mean and a hospital-dependent deviation:

$$\beta_{0j} = \beta_0 + \beta_2 Z_j + b_{0j}, \tag{3.2}$$

$$\beta_{1j} = \beta_1 + \beta_3 Z_j + b_{1j}, \tag{3.3}$$

where $\beta_0$ and $\beta_1$ represent the overall means, $\beta_2$ and $\beta_3$ are the deviations from the mean caused by the explanatory variable hospital size $Z_j$ and $b_{0j}$ and $b_{1j}$ represent the hospital-specific deviations from the mean.

We can rewrite the model by substituting Equations (3.2) and (3.3) into Equation (3.1):

$$Y_{ij} = [\beta_0 + \beta_2 Z_j + \beta_1 X_{ij} + \beta_3 Z_j X_{ij}] + [b_{0j} + b_{1j} X_{ij} + \epsilon_{ij}]. \tag{3.4}$$

In this model two parts can be distinguished: a fixed part, which contains the regression coefficients and their associated variables $[\beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + \beta_3 Z_j X_{ij}]$ and a random part, which contains the hospital-specific and residual error terms $[b_{0j} + b_{1j} X_{ij} + \epsilon_{ij}]$. The patient-level errors ($\epsilon_{ij}$) and the hospital-level errors ($b_{0j}$ and

$b_{1j}$) are assumed to be mutually independent and follow a normal distribution.

In this chapter, we will focus on a random intercepts model where the intercept $\beta_{0j}$ is hospital-dependent while the slope $\beta_{1j}$ is not. This implies that Equation (3.4) simplifies to:

$$Y_{ij} = [\beta_0 + \beta_1 X_{ij} + \beta_2 Z_j] + [b_{0j} + \epsilon_{ij}]. \tag{3.5}$$

In general there will be $P$ explanatory variables at the level of the patient and $Q$ variables at the level of the hospital, hence Equation (3.5) generalizes to:

$$Y_{ij} = \left[\beta_0 + \sum_{p=1}^{P} \beta_p X_{pij} + \sum_{q=P+1}^{Q} \beta_q Z_{qj}\right] + \left[b_{0j} + \epsilon_{ij}\right],$$

where $\beta_0$ is the intercept, $\beta_p$ (with $p = 1, ..., P$) represent the fixed effects at the level of the patient, $\beta_q$ (with $q = P+1, ..., Q$) represent the fixed effects at the level of the hospital, $b_{0j}$ is a random effect for hospital and $\epsilon_{ij}$ is the residual error term.

Fitting such models can be done with a statistical software package such as `SAS`. A description on the use of the `SAS PROC MIXED` procedure to fit multi-level models is given by Littell *et al.* (2006) and Singer (1998). For a comprehensive elaboration on multi-level models we refer to the books by Snijders and Bosker (1999), Goldstein (2003), Raudenbush and Bryk (2002), Hox (2010), and Wang *et al.* (2012). For some illustrations of the application of multi-level models to hierarchical data we refer to Goldstein *et al.* (1993), Renard *et al.* (1998) and Lee (2000).

Multi-level settings usually consist of a small number of units that tend to be quite large. However, several specific but frequently studied settings, mainly in longitudinal and family research, involve a large number of units that tend to be quite small. When the unit contains only one element, it is referred to as a singleton. An example of such a setting can be found in the Ceftriaxone Data (Section 1.2.2), where prescribed doses of ceftriaxone (expressed in mg/kg/day) are reported for 329 children, divided over 124 departments as illustrated in Figure 3.1. Here, 47% of the included departments are singletons (i.e. contain only one child). Regardless of sparseness, hierarchical data are generally analysed with a multi-level model.

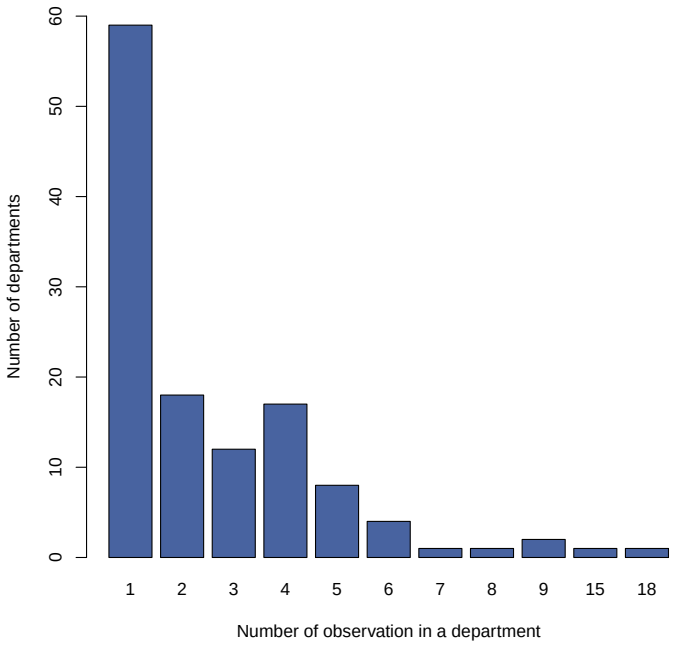



Figure 3.1: Size of the departments included in the Ceftriaxone Data.

Several studies to determine the impact of small sample sizes on different aspects of the multi-level model showed that both residual and random effects variance were biased when the number of subjects within the units is small. The impact on fixed effects appeared to be smaller, as both fixed effects estimates and their standard error were unbiased in the presence of small clusters (Bell *et al.*, 2014; Clarke, 2008; Maas and Hox, 2005).

Although the small sample setting has been extensively studied and renders promising results, we are specifically interested in the setting with different proportions of singletons. To our knowledge, only a few studies assessed the impact of singletons on the multi-level model. Pickering and Weatherall (2007) investigated a setting with 15% of singletons and found that fixed effects estimates and standard errors were unbiased. Sauzet *et al.* (2012) studied a setting with $80 - 99\%$ of singletons and found that parameter estimates for fixed effects were biased when the percentage of singletons became extreme.

While these studies already give an idea about the impact of the presence of singletons, they focus on specific singleton proportions (either very high or fairly low) and use rather simple models only containing explanatory variables at the lowest level of the hierarchy. We will study the impact of an increasing proportion of singletons

$(0-95\%)$ on different aspects of the multi-level model, using a two-level setting with explanatory variables at both levels of the hierarchy.

## 3.2 Simulation study

We set up a simulation study, based on an analysis of the Ceftriaxone Data (Section 1.2.2) where the association between prescribed dose and department size, reason for treatment and age was studied (three covariates model). For this analysis, the variable *Beds* was categorized, based on the distribution of beds in the Ceftriaxone Data, into small ($< 17$ beds), medium ($17-26$ beds) and large ($> 26$ beds).

The three covariates model can be presented as follows:

$$\begin{aligned} Y_{ij} = \beta_0 + b_{0j} + \beta_1 Size_{1j} + \beta_2 Size_{2j} + \beta_3 Age_{ij} + \beta_4 Reason_{1ij} \\ + \beta_5 Reason_{2ij} + \beta_6 Reason_{3ij} + \epsilon_{ij}, \end{aligned} \tag{3.6}$$

where $Y_{ij}$ represents the ceftriaxone dose prescribed to child $i$ ($i = 1, ..., n_j$) in department $j$ ($j = 1, ..., J$), $n_j$ is the number of children in department $j$, $J$ is the number of included departments, $\beta_0$ is the general intercept, $b_{0j}$ is the department-specific intercept, $Size_{1j}$ is 1 if department $j$ is *large*, $Size_{2j}$ is 1 if department $j$ is *medium*, $Age_{ij}$ is the age of child $i$, $Reason_{1ij}$ is 1 if the reason for treatment is $different$, $Reason_{2ij}$ is 1 if the reason for treatment is *mild*, $Reason_{3ij}$ is 1 if the reason for treatment is *moderate*, $\beta_1$ up to $\beta_6$ are the respective coefficients for the listed parameters and $\epsilon_{ij}$ is the residual error term. We assume that the random effect follows a normal distribution with mean zero and variance $\sigma^2_{RE}$ and that the error terms are independent and follow a normal distribution with mean zero and variance $\sigma^2_{Res}$. Parameter estimates and standard errors for the fitted model are reported in Table 3.1.

Because the structure of the Ceftriaxone Data was rather elaborate, we considered a simplified version with 350 children divided over 50 departments. As in the Ceftriaxone Data, all simulated datasets contained 16 large, 18 medium and 16 small departments. The percentage of singleton departments ranged from 0 to 95% (in steps of 5%). The number of singleton departments was rounded upwards (e.g. 5% of singletons implies 2.5 departments containing only one child. Hence, for 5% of singletons, we included 3 departments with one child).

Table 3.1: Parameter estimates and standard errors for the fixed effects in the three covariates model.

| Parameter | Estimate | Std. error |
|---|---|---|
| Intercept | 82.9514 | 4.5132 |
| $\text{Size}_{1j}$ | −3.2239 | 4.7705 |
| $\text{Size}_{2j}$ | 4.4692 | 4.7473 |
| $\text{Age}_{ij}$ | −2.0599 | 0.3383 |
| $\text{Reason}_{1ij}$ | −8.2539 | 4.3193 |
| $\text{Reason}_{2ij}$ | −21.9945 | 7.4318 |
| $\text{Reason}_{3ij}$ | −5.5856 | 3.5735 |
| $\sigma^2_{RE}$ | 180.3700 | 47.1385 |
| $\sigma^2_{Res}$ | 489.5100 | 44.7563 |

For each scenario, 1000 datasets were simulated according to the following procedure:

1. Sample a random intercept from a normal distribution with mean zero and standard deviation $\sigma_{RE}$ for each of the 50 included departments.

2. Group the combination of age and reason for treatment for the 329 children in the Ceftriaxone Data based on the size of the department they are treated in. Then, conditional on the size of the department, sample a combination of age and reason for treatment for 350 children.

3. Sample a residual error term from a normal distribution with mean zero and standard deviation $\sigma_{Res}$ for each of the 350 included children.

4. Simulate the prescribed dose for each child using Equation 3.6 and parameter estimates reported in Table 3.1.

## 3.3 Models fitted

All simulated datasets were analysed with the three covariates model.

For each scenario, we assessed the performance of the fitted model using four performance characteristics. The first is the relative difference between the mean parameter estimate and the true parameter (RDM). The second characteristic is the relative difference between the mean estimated standard error and the empirical standard error (RDE). Here, the estimated standard error reflects the uncertainty within the simulations while the empirical standard error (SES) reflects the uncertainty between simulations. The first is calculated as the mean of the obtained standard errors while the latter is calculated as the standard deviation of obtained parameter estimates. The third performance characteristic is the mean length of the confidence interval. The last performance characteristic is the coverage of the confidence interval, calculated as the percentage of times the true parameter falls within the estimated confidence interval. The stability of the F test was assessed using the number of times the null hypothesis was rejected (rejection rate).

Because some of the simulated scenarios contain a fairly high proportion of singletons, one might doubt the need to correct for clustering. Therefore, we studied the same performance characteristics for models that handle the singletons in three different ways. The first method that comes to mind to handle the singletons is to simply ignore the dependence within departments (i.e. ignoring singletons). This is done by fitting a model containing fixed effects for reason for treatment, age and department size, but no random effect. Other options to deal with a high proportion of singletons are to discard the singletons from the data (i.e. dropping singletons) or to group the singletons into an artificial unit (i.e. regrouping singletons). Both approaches were evaluated by fitting the three covariates model to all simulated datasets either after dropping or after regrouping the included singletons.

## 3.4 Results

### 3.4.1 The three covariates model

All simulated datasets were analysed with the three covariates model (Equation 3.6). We report the mean parameter estimate together with an assessment of its uncertainty (SES) and the four performance characteristics for the multi-level model. These characteristics are reported for one fixed effect at the level of the child and one fixed effect at the level of the department in Tables 3.2 and 3.3, respectively. Performance characteristics for the other fixed effects can be consulted in Tables A5 up to A9. Stability of the F test for parameters at the level of the child (*Age* and *Reason*) and at the level of the department (*Size*) is reported in Table 3.4. Accuracy of the random effects variance and the residual variance is presented in Table 3.5.

Table 3.2: Performance characteristics for the fixed effect $Reason_{1ij}$ in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | -8.047 | 3.720 | -2.5 | 3.7 | 15.171 | 95.8 |
| 5 | -7.981 | 3.729 | -3.3 | 3.2 | 15.148 | 94.5 |
| 10 | -8.281 | 3.790 | 0.3 | 1.4 | 15.116 | 94.7 |
| 15 | -8.354 | 3.784 | 1.2 | 1.7 | 15.143 | 95.2 |
| 20 | -8.130 | 3.826 | -1.5 | -0.4 | 14.999 | 95.0 |
| 25 | -8.093 | 3.860 | -1.9 | -1.1 | 15.023 | 94.5 |
| 30 | -7.929 | 3.807 | -3.9 | 0.5 | 15.050 | 94.8 |
| 35 | -8.331 | 3.862 | 0.9 | -1.2 | 15.018 | 95.5 |
| 40 | -8.204 | 3.901 | -0.6 | -2.2 | 15.021 | 94.3 |
| 45 | -8.421 | 3.848 | 2.0 | -0.6 | 15.042 | 94.7 |
| 50 | -8.189 | 3.746 | -0.8 | 1.3 | 14.936 | 95.1 |
| 55 | -8.139 | 3.842 | -1.4 | -1.4 | 14.907 | 95.0 |
| 60 | -8.355 | 3.827 | 1.2 | -0.6 | 14.961 | 95.2 |
| 65 | -8.144 | 3.813 | -1.3 | -0.7 | 14.891 | 94.5 |
| 70 | -8.151 | 3.756 | -1.2 | 1.1 | 14.942 | 94.6 |
| 75 | -8.149 | 3.798 | -1.3 | 0.2 | 14.972 | 94.3 |
| 80 | -8.122 | 3.718 | -1.6 | 1.8 | 14.892 | 94.8 |
| 85 | -8.199 | 3.860 | -0.7 | -2.5 | 14.804 | 94.0 |
| 90 | -8.183 | 3.903 | -0.9 | -2.4 | 14.990 | 94.7 |
| 95 | -8.230 | 3.969 | -0.3 | -1.2 | 15.426 | 94.7 |

SES: simulation standard error

RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error

The difference between the estimated and the true parameter (RDM) for the fixed effect at the level of the child ($Reason_{1ij}$) was not affected by the percentage of singletons and was consistently small (Table 3.2). This indicates that the parameter is estimated well regardless of the percentage of singletons present in the data. The difference between the estimated and true standard error (RDE) was small throughout the simulation study, indicating that the standard error accurately estimated the true standard error for the three covariates model. Both the true standard error and the length of the confidence interval experienced minor fluctuations. This however did not seem to be related to the increase in the percentage of singleton departments. Coverage of the confidence interval was around 95% throughout the simulation study. Similar findings were reported for other covariates at the level of the child (results shown in Tables A5 to A7).

Table 3.3: Performance characteristics for the fixed effect $Size_{1j}$ in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | -3.323 | 5.605 | 3.1 | -0.4 | 22.438 | 95.2 |
| 5 | -3.496 | 5.732 | 8.4 | -2.8 | 22.428 | 94.4 |
| 10 | -3.466 | 5.736 | 7.5 | -1.2 | 22.812 | 95.9 |
| 15 | -3.226 | 5.821 | 0.1 | -1.0 | 23.230 | 95.3 |
| 20 | -3.484 | 5.859 | 8.1 | -1.9 | 23.176 | 95.2 |
| 25 | -3.452 | 5.850 | 7.1 | 0.5 | 23.721 | 94.6 |
| 30 | -3.252 | 6.022 | 0.9 | -2.0 | 23.852 | 94.5 |
| 35 | -3.063 | 6.053 | -5.0 | -0.9 | 24.254 | 95.1 |
| 40 | -2.905 | 6.045 | -9.9 | 2.5 | 25.058 | 96.1 |
| 45 | -3.313 | 6.346 | 2.8 | -0.7 | 25.513 | 95.8 |
| 50 | -3.031 | 6.512 | -6.0 | -2.7 | 25.716 | 93.5 |
| 55 | -3.243 | 6.888 | 0.6 | -5.6 | 26.419 | 94.1 |
| 60 | -2.796 | 6.547 | -13.3 | -0.1 | 26.632 | 94.9 |
| 65 | -3.321 | 6.501 | 3.0 | 4.4 | 27.686 | 96.3 |
| 70 | -3.373 | 7.186 | 4.6 | -2.5 | 28.563 | 94.9 |
| 75 | -2.780 | 7.753 | -13.8 | -6.5 | 29.633 | 94.3 |
| 80 | -3.137 | 7.798 | -2.7 | -4.3 | 30.611 | 93.9 |
| 85 | -3.083 | 8.242 | -4.4 | -6.2 | 32.048 | 94.6 |
| 90 | -3.255 | 8.296 | 0.9 | -4.0 | 33.245 | 94.9 |
| 95 | -3.591 | 9.047 | 11.4 | -5.5 | 35.352 | 93.6 |

SES: simulation standard error

RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error

The RDM for the fixed effect at the level of the department ($Size_{1j}$) was slightly higher than for a covariate at the level of the child, but fluctuated regardless of the percentage of singletons (Table 3.3). This indicates that the parameter is not optimally estimated. Also the RDE was slightly higher than for a covariate at the level of the child, indicating that the standard error estimated the true standard error less accurately. Both the true standard error and the length of the confidence interval increased with an increasing percentage of singletons. Coverage of the confidence interval remained around 95% throughout the simulation study.
Similar findings were reported for other covariates at the level of the department (results shown in Tables A8 and A9).

Table 3.4: F test rejection rate for the fixed effects in the three covariates model under an increasing percentage of singletons.

| Singletons (%) | Age | Reason | Size |
|---|---|---|---|
| 0 | 1000 | 829 | 216 |
| 5 | 1000 | 841 | 236 |
| 10 | 1000 | 839 | 221 |
| 15 | 1000 | 846 | 197 |
| 20 | 1000 | 843 | 205 |
| 25 | 1000 | 843 | 190 |
| 30 | 1000 | 833 | 199 |
| 35 | 1000 | 843 | 180 |
| 40 | 1000 | 833 | 195 |
| 45 | 1000 | 853 | 183 |

| Singletons (%) | Age | Reason | Size |
|---|---|---|---|
| 50 | 1000 | 861 | 185 |
| 55 | 1000 | 865 | 161 |
| 60 | 1000 | 867 | 163 |
| 65 | 1000 | 853 | 156 |
| 70 | 1000 | 859 | 161 |
| 75 | 1000 | 865 | 159 |
| 80 | 1000 | 835 | 168 |
| 85 | 1000 | 858 | 149 |
| 90 | 1000 | 885 | 140 |
| 95 | 1000 | 902 | 159 |

The F test for the effect at the level of the child is quite stable regardless of the percentage of singletons in the data. The F test for an effect at the level of the department decreased slightly with an increasing proportion of singletons.

The RDM for both the random effects variance and the residual variance was small throughout the simulation study (Table 3.5 and 3.6). This indicates that generally, in the presence of singletons, the estimated variances approach the true variances quite well.

Table 3.5: Performance characteristics for the random effects variance in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | RDM (%) | Singletons (%) | Mean | RDM (%) |
|---|---|---|---|---|---|
| 0 | 178.660 | -0.9 | 50 | 177.770 | -1.4 |
| 5 | 176.120 | -2.4 | 55 | 178.470 | -1.1 |
| 10 | 177.420 | -1.6 | 60 | 174.890 | -3.0 |
| 15 | 178.660 | -0.9 | 65 | 179.550 | -0.5 |
| 20 | 175.120 | -2.9 | 70 | 177.630 | -1.5 |
| 25 | 178.160 | -1.2 | 75 | 175.210 | -2.9 |
| 30 | 176.450 | -2.2 | 80 | 181.920 | 0.9 |
| 35 | 175.750 | -2.6 | 85 | 177.610 | -1.5 |
| 40 | 181.350 | 0.5 | 90 | 180.110 | -0.1 |
| 45 | 178.940 | -0.8 | 95 | 178.590 | -1.0 |

RDM: relative difference between estimated and true mean

Table 3.6: Performance characteristics for the residual variance in the three covariates model with an increasing percentage of singletons.

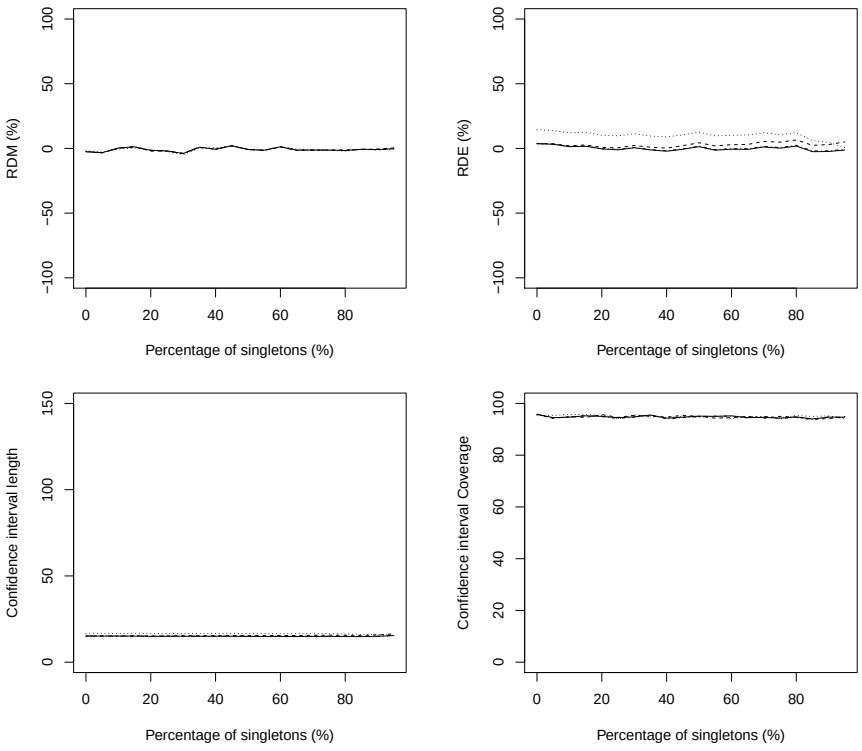
| Singletons (%) | Mean | RDM (%) | Singletons (%) | Mean | RDM (%) |
|---|---|---|---|---|---|
| 0 | 483.980 | -1.1 | 50 | 484.290 | -1.1 |
| 5 | 484.230 | -1.1 | 55 | 482.610 | -1.4 |
| 10 | 483.720 | -1.2 | 60 | 484.990 | -0.9 |
| 15 | 484.570 | -1.0 | 65 | 481.800 | -1.6 |
| 20 | 481.280 | -1.7 | 70 | 482.590 | -1.4 |
| 25 | 482.570 | -1.4 | 75 | 483.300 | -1.3 |
| 30 | 485.050 | -0.9 | 80 | 487.220 | -0.5 |
| 35 | 484.220 | -1.1 | 85 | 482.980 | -1.3 |
| 40 | 483.420 | -1.2 | 90 | 483.470 | -1.2 |
| 45 | 484.200 | -1.1 | 95 | 482.620 | -1.4 |

RDM: relative difference between estimated and true mean

### 3.4.2 Handling the singletons

Next to the analysis with the three covariates model, the simulated datasets were analysed with a model containing fixed effects for age, reason for treatment and department size without correction for clustering (ignoring singletons). Additionally, the three covariates model was fitted to the datasets where singletons were removed (dropping singletons) or grouped into an artificial department (regrouping singletons).

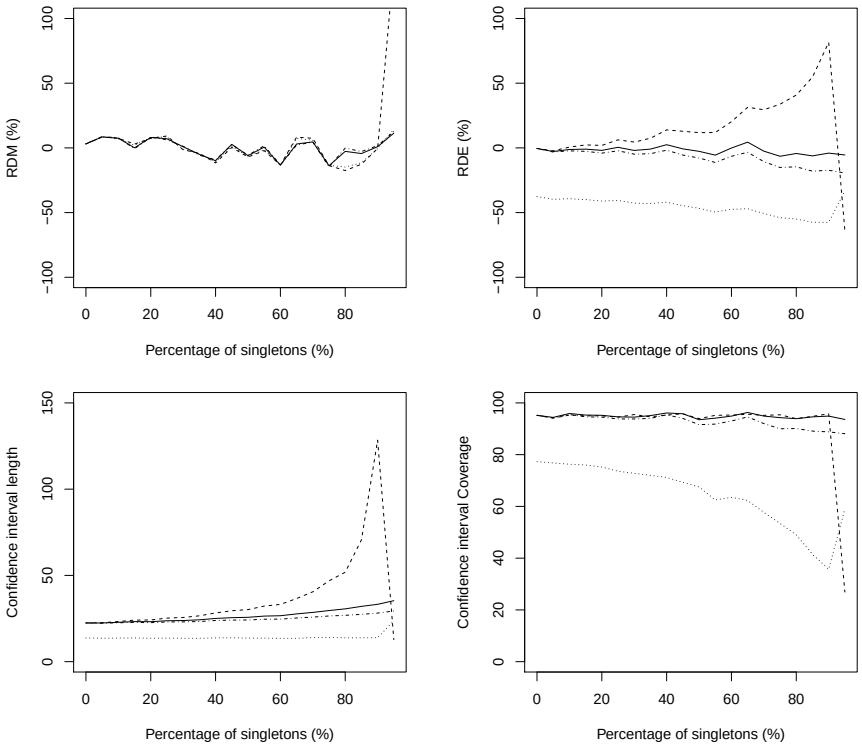
Obtained performance measures for one fixed effect at the level of the child and one fixed effect at the level of the department are visualized in Figures 3.2 and 3.3. Performance characteristics for the other fixed effects can be consulted in Figures A3 up to A7. F test rejection rates for one fixed effect at the level of the child and one fixed effect at the level of the department are shown in Figure 3.4. Rejection rates for the additional fixed effect at the level of the child can be consulted in Figure A8. RDM for both residual and random effects variance are shown in Figure 3.5.




RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure 3.2: Performance measures for the fixed effect $Reason_{1ij}$ when ignoring (dotted lines), dropping (dashed lines) or regrouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model on the original data (full lines).

Figure 3.2 shows that the RDM, confidence interval length and coverage for the fixed effect at the level of the child ($Reason_{1ij}$) were comparable for the three covariates model fitted to the original data and the three options to handle the singletons (ignoring, dropping or regrouping). When clustering was ignored (dotted lines), the RDE was higher compared to the three covariates model fitted to the original data (full lines) or when dropping and regrouping the singletons (dashed and dot-dashed lines, respectively).




RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure 3.3: Performance measures for the fixed effect $Size_{1j}$ when ignoring (dotted lines), dropping (dashed lines) or regrouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model fitted to the original data (full lines).

Figure 3.3 shows that the RDM for the fixed effect at the level of the department ($Size_{1j}$) was comparable for the three covariates model fitted to the original data and the three options to handle the singletons. Ignoring the singletons (dotted lines) resulted in a decreased RDE and confidence interval length. Confidence interval coverage was unacceptably low for all percentages of singletons. When dropping the singletons (dashed lines), RDE and confidence interval length increased with an increasing percentage of singletons. The coverage remained stable throughout the simulation study. For the scenario with 95% of singletons, the plots show a severe drop in RDE, confidence interval length and coverage. When regrouping the singletons into an artificial department (dot-dashed lines), RDE and confidence interval length were slightly lower than for the three covariates model fitted to the original data. The coverage remained acceptable throughout the simulation study.

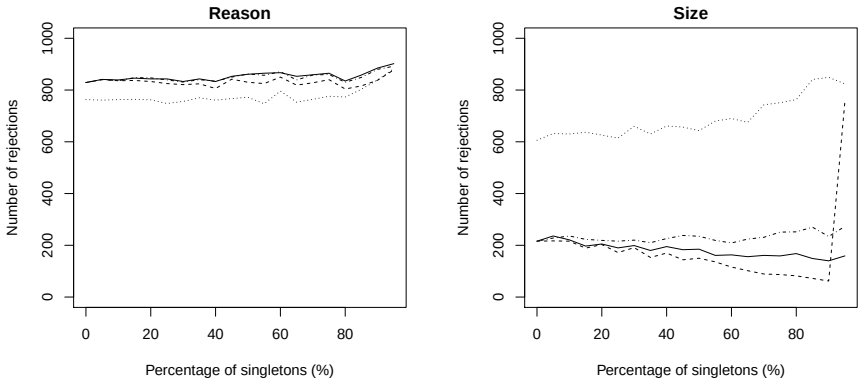



Figure 3.4: F test rejection rates for fixed effects *Reason* and *Size* in the model when ignoring (dotted lines), dropping (dashed lines) or regrouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model fitted to the original data (full lines).

Figure 3.4 shows that dropping or regrouping the singletons (dashed and dot-dashed lines, respectively) does not influence the performance of the F test for an effect at the level of the child. Ignoring the dependency within clusters (dotted lines) causes the rejection rate to be slightly lower compared to the rejection rate for the three covariates model fitted to the original data (full lines).

Dropping and regrouping the singletons (dashed and dot-dashed lines, respectively) cause the rejection rate for the fixed effect at the level of the department to be respectively lower and higher compared to the rejection rate for the three covariates model fitted to the original data (full lines). Ignoring the dependency within clusters (dotted lines) causes the rejection rate to be a lot higher than the rejection rate for the three covariates model fitted to the original data (full lines).

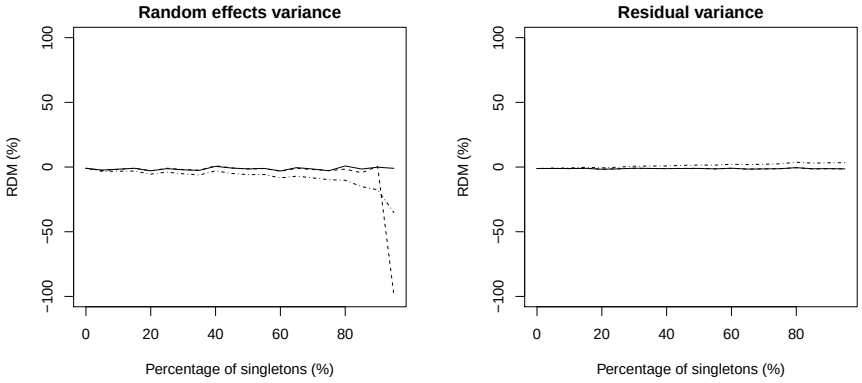



Figure 3.5: Relative difference between estimated and true mean (RDM) for the random effects variance (left) and residual variance (right) when dropping (dashed lines) or regrouping (dot-dashed lines) the singletons, compared to their RDM in the three covariates model fitted to the original data (full lines).

Figure 3.5 shows that when dropping the singletons (dashed lines), the residual variance stayed close to the true residual variance. The random effects variance was close to the true random effects variance throughout the simulation study, but decreased steeply at the end (for the scenario with 95% singletons).
When regrouping the singletons (dot-dashed lines), the residual variance was slightly overestimated while the random effects variance was slightly underestimated, with the difference between estimated and true variance getting bigger with an increasing percentage of singletons.

## 3.5 Discussion

We conducted a simulation study, inspired by the structure of the Ceftriaxone Data, to investigate the impact of an increasing percentage of singletons on different aspects of the linear multi-level model. This impact was assessed using four performance characteristics and revealed that neither the RDM nor the RDE were affected by the percentage of singletons in the data. They were consistently low, with RDM and RDE for an effect at the level of the child being slightly lower than RDM and RDE for an effect at the level of the department. This might be explained by the number of independent observations that are available to estimate both effects, with this number being considerably lower for the effect at the level of the department. Both the SES and width of the confidence interval fluctuated for an effect at the level of the child, while they increased with the percentage of singletons for an effect at the level of the department. This increase goes hand in hand with a decrease in the rejection rate for the F test and can be explained by the more stable estimation of the average dose for a department when the number of children in that department is larger. The coverage approached 95% for explanatory variables at both levels and varying percentages of singletons.

Because some of the simulated datasets contain a fairly high proportion of singletons, one might decide to either ignore the dependency within clusters (ignoring singletons), remove the singletons from the data (dropping singletons) or group them into an artificial department (regrouping singletons). A simulation study was conducted to investigate the consequences of these three options on different aspects of the multi-level model. Impact on the level of the child was minor, while impact on the level of the department was more clear. As mentioned before, this can be explained by the number of independent observations available. When ignoring the singletons, the RDE and confidence interval length were lower while the rejection rate for the F test was higher compared to the three covariates model fitted to the original data. This can be explained by the consistent underestimation of the standard error when ignoring clustering. Confidence interval coverage was unacceptably low for all percentages of singletons, indicating that ignoring the dependency within the clusters is never a good idea. When dropping the singletons, the RDE and confidence interval length were higher while the rejection rate for the F test was lower compared to the three covariates model fitted to the original data, with these differences increasing when the proportion of singletons increases. This can be explained by the increase in standard error due to the decrease in number of remaining departments. For the scenario

with 95% of singletons, there is a severe drop in RDE, confidence interval length and coverage together with a steep increase in F test rejection rate, which is explained by the presence of only one large department in this scenario. The low coverage and narrow confidence intervals resulting from both ignoring and dropping the singletons force us to conclude that both worsen the performance of the multi-level model.
When regrouping the singletons into an artificial department, the RDE and confidence interval length were slightly lower while the rejection rate for the F test was slightly higher compared to the three covariates model fitted to the original data. The residual variance was slightly overestimated while the random effects variance was slightly underestimated, with the difference between estimated and true variance increasing with an increasing percentage of singletons. All these findings can be explained by the grouping of singletons that are not actually related, which decreases the variance between included departments and causes a slight underestimation of the true standard error for the effect at the level of the department. Although regrouping is an option that might be considered when the data at hand contain a high percentage of singletons, the regular multi-level model performs better even when the percentage of singletons increases.
An alternative that could be used in the presence of sparseness at the lowest level of the hierarchy is to select a more convenient clustering level to model the prescribed dose (e.g. hospital in which the department is situated) (Cortiñas Abrahantes *et al.*, 2004). Although this strategy would improve the model's stability, it was not considered here because we wanted to preserve the possibility to obtain department-specific estimates.

# Chapter 4

# Performance of the F test in a linear multi-level model setting with sparseness at the highest level

In Chapter 3, we evaluated the stability of the multi-level model in the presence of singletons at the lowest level. While a typical hierarchical setting will encompass at least a small percentage of singletons at the lowest level, it will often also include singletons at the highest (i.e. primary) level of the hierarchy. This implies that units at the primary level contain only one secondary unit while this unit tends to be quite large. When a primary unit contains a small number of secondary units, this is referred to as primary unit sparseness.

An example of primary unit sparseness can be found in the Ceftriaxone Data (Section 1.2.2), where prescribed doses of ceftriaxone (expressed in mg/kg/day) are reported for 329 children, divided over 20 countries within 10 UN macro-geographical regions according to Table 4.1. Here, some regions only contain one country, while all but one country contain more than ten children. Regardless of such primary unit sparseness, hierarchical data are generally analysed with multi-level models.

Table 4.1: The number of countries and children per region in the Ceftriaxone Data.

| Region number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of countries | 1 | 1 | 1 | 1 | 3 | 1 | 4 | 1 | 2 | 5 |
| Number of children | 4 | 31 | 24 | 11 | 32 | 61 | 43 | 16 | 80 | 27 |

There are several methods to assess significance of fixed effects in a linear multi-level model. A first option is the Wald test with the test statistic for testing the hypothesis $\mathrm{H}_0 : L\beta = 0$ versus $\mathrm{H}_a : L\beta \neq 0$ (for any known matrix $L$) defined as:

$$t = \left(\widehat{\beta} - \beta\right)' L' \left[ L \left( \sum_{i=1}^{N} X_i' V_i^{-1}(\widehat{\alpha}) X_i \right)^{-1} L' \right]^{-1} L \left(\widehat{\beta} - \beta\right), \tag{4.1}$$

where $\beta$ is a vector of fixed effects and $\widehat{\beta}$ a vector of its estimates, $N$ is the number of included observations, $X_i$ is a vector of known covariates and $V_i(\widehat{\alpha})$ is the variance components matrix. This test statistic asymptotically follows a chi-squared distribution with $rank(L)$ degrees of freedom. A disadvantage of the Wald test is the use of standard errors which ignore the variability introduced by estimating the variance components (Dempster *et al.*, 1981). This downward bias can be resolved by using an approximate F test with the test statistic for testing the hypothesis $\mathrm{H}_0 : L\beta = 0$ versus $\mathrm{H}_a : L\beta \neq 0$ (for any known matrix $L$) defined as $t/rank(L)$. This test statistic follows an approximate F-distribution with $rank(L)$ numerator degrees of freedom and denominator degrees of freedom estimated from the data. A third option is the likelihood ratio test which compares the likelihood of the model under $\mathrm{H}_0$ with the likelihood of the full model. This test statistic asymptotically follows a chi-squared distribution with degrees of freedom equal to the difference in the two models' dimensions (Verbeke and Molenberghs, 2009).
As both the Wald and the likelihood ratio test follow an asymptotic distribution while the F test follows an exact distribution, we chose to discuss the Wald and likelihood ratio test briefly and put the main focus of this chapter on the F test. The Satterthwaithe procedure, which uses the matrices of random and fixed effects itself, is used to determine the denominator degrees of freedom for the F test (Kutner *et al.*, 2005; Verbeke and Molenberghs, 1997).

Inference on the fixed effects is usually based on maximum likelihood (ML) estimation, where the likelihood is maximized jointly for the fixed effects and variance components. The ML estimator can however be biased downwards, as it does not take into account the loss in degrees of freedom from estimating the fixed effects (Harville,

1977; Searle *et al.*, 1992). Restricted maximum likelihood (REML) estimation does account for this by maximizing a set of error contrasts rather than the joint maximum likelihood (Patterson and thompson, 1971). For this reason, REML is often preferred over ML (Lee and Kapadia, 1984; McCulloch and Searle, 2001).
The difference between REML and ML estimation usually is rather small. When the number of primary units is small, as often occurs in multi-level modelling, the bias in ML estimation can become substantial (Hox, 1998; Swallow and Monahan, 1984). As a result, the difference between REML and ML estimation will be more pronounced and the preference for REML estimation more outspoken (Kreft and De Leeuw, 1998). This is seen when assessing significance of the fixed effects in the Ceftriaxone Data using a model containing a random effect for country and fixed effects for age, reason, department type and region (four covariates model)(Table 4.2, discussed in more detail in Section 4.1). The substantial difference between REML and ML motivated the study of the performance of the F test under REML and ML in the presence of a decreasing number of secondary units within the primary units (i.e. increasing primary unit sparseness).

Table 4.2: Significance of the fixed effects in the four covariates model obtained using F tests under ML and REML.

| | ML | | REML | |
|---|---|---|---|---|
| Effect | F-value | P-value | F-value | P-value |
| Age | 35.77 | <0.0001 | 32.86 | <0.0001 |
| Reason | 2.14 | 0.0097 | 1.86 | 0.0307 |
| Department type | 3.60 | 0.0005 | 4.18 | <0.0001 |
| Region | 4.86 | **<0.0001** | 0.80 | **0.6385** |

In this chapter, we will focus on a setting consisting of three levels in which the primary units contain a very small number of secondary units. We will point out that it is necessary to be cautious when dealing with such sparse multi-level data and provide guidelines on how to handle similar situations.

## 4.1 Motivating example

The four covariates model was fitted to the Ceftriaxone Data (Section 1.2.2), both under REML and ML. Significance of all fixed effects was assessed using F tests with approximate degrees of freedom obtained through the Sattertwaithe procedure (Kutner *et al.*, 2005; Verbeke and Molenberghs, 1997) (Table 4.2). This resulted in a difference in significance, which is most striking for the fixed effect *Region*. Table 4.3 shows that also the parameter estimates and standard errors for *Region* are affected. This was also observed for the other fixed effects but not shown here.

Table 4.3: Parameter estimates (standard errors) for *Region* in the four covariates model under ML and REML.

| Effect | | ML estimate (s.e.) | REML estimate (s.e.) |
|---|---|---|---|
| Intercept | | 51.475 (17.286) | 53.071 (18.766) |
| Region | 1 | 13.708 (13.034) | 13.122 (18.768) |
| | 2 | 5.645 (6.782) | 4.610 (14.851) |
| | 3 | 39.058 (7.293) | 37.163 (15.161) |
| | 4 | 1.299 (8.637) | 0.786 (15.878) |
| | 5 | -0.196 (6.538) | 3.070 (11.917) |
| | 6 | 8.176 (5.684) | 6.883 (14.378) |
| | 7 | 12.109 (6.491) | 11.254 (11.123) |
| | 8 | 11.628 (8.819) | 10.844 (15.939) |
| | 9 | 8.952 (5.647) | 11.700 (12.199) |
| | 10 | 0 (.) | 0 (.) |

The differences between results obtained under REML and under ML motivated the study of the performance of the F test in the presence of increasing primary unit sparseness. This study started by identifying the most basic yet problematic setting, which is a model with a random effect for country and a fixed effect for region (one covariate model). In this model, the F test for the effect of *Region* was still highly significant under ML (p-value $< 0.0001$) while it was non-significant under REML (p-value $= 0.3649$). Differences in parameter estimates and standard errors remained considerable (values not reported here). This basic setting and the Ceftriaxone Data were used to set up a simulation study which is described in more detail in the next section.

## 4.2 Simulation study

The setting that was identified as most basic, yet problematic and used in setting up the simulation study can be presented using centred parametrisation as follows:

$$Y_{ijk} = \mu_{jk} + \epsilon_{ijk},$$
$$\mu_{jk} = \mu_k + \epsilon_{jk},$$

where $Y_{ijk}$ represents the dose for child $i$ ($i = 1, ..., I_{jk}$) in country $j$ ($j = 1, ..., J_k$) within region $k$ ($k = 1, ..., K$), $K$ is the number of regions, $J_k$ is the number of countries in region $k$ and $I_{jk}$ is the number of children in country $j$ within region $k$, $\mu_{jk}$ represents the average dose for country $j$ in region $k$, $\epsilon_{ijk}$ represents the child-specific deviation from $\mu_{jk}$, $\mu_k$ represents the average dose for region $k$ and $\epsilon_{jk}$ represents the country-specific deviation from $\mu_k$. In this setting, we assume that $\epsilon_{ijk}$ follows a normal distribution with mean zero and variance $v_{jk}$, and that $\epsilon_{jk}$ follows a normal distribution with mean zero and variance $v_k$. This setting is further simplified by assuming that $v_{jk}$ and $v_k$ are both constants.

Data were simulated under the null hypothesis ($H_0$), assuming that *Region* has no effect on dose, as well as under a specific alternative hypothesis ($H_a$). As the structure of the Ceftriaxone Data was rather elaborate, we considered a simplified version with 240 children equally divided over five regions. The number of countries in the regions differed over nine different scenarios as presented in Table 4.4.
Note that Table 4.4 contains some extreme situations which were included merely to illustrate the worsening problems. In practice however, one would generally not fit a model including a random effect for country and a fixed effect for region in such situations.

For each scenario 1000 datasets were simulated, under $H_0$ and under $H_a$, according to the following procedure:

1. Simulate $\mu_{jk}$ from a normal distribution with mean $\mu_k$ and variance $v_k$
2. Simulate $Y_{ijk}$ from a normal distribution with mean $\mu_{jk}$ and variance $v_{jk}$
   As we aim to mimic the Ceftriaxone Data, a simulated value represents a child's dose, which can never be negative. For this reason, the negative values (on average 1.4% of the simulated values per scenario) are simulated again from the same distribution until all are positive. This implies that the simulated datasets are obtained from a truncated normal distribution, but as the percentage of truncation is very small this does not affect the inference.

Table 4.4: Different scenarios with a varying number of countries per region.

| | Number of countries in region | | | | | Total number |
|---|---|---|---|---|---|---|
| Scenario | 1 | 2 | 3 | 4 | 5 | of countries |
| 1 | 1 | 1 | 1 | 1 | 1 | 5 |
| 2 | 2 | 1 | 1 | 1 | 1 | 6 |
| 3 | 2 | 2 | 1 | 1 | 1 | 7 |
| 4 | 2 | 2 | 2 | 1 | 1 | 8 |
| 5 | 2 | 2 | 2 | 2 | 1 | 9 |
| 6 | 2 | 2 | 2 | 2 | 2 | 10 |
| 7 | 3 | 3 | 3 | 3 | 3 | 15 |
| 8 | 4 | 4 | 4 | 4 | 4 | 20 |
| 9 | 8 | 8 | 8 | 8 | 8 | 40 |

The values that were used for $\mu_k$, $v_k$ and $v_{jk}$ were inspired by the Ceftriaxone Data. For $v_k$ we used the variance of the average dose for countries in the largest region (i.e. Western Europe containing five countries), being 225.169 and for $v_{jk}$ we used the variance of the dose for children in the largest country (i.e. Georgia containing 73 children), being 785.984. Under $H_0$ the value for $\mu_k$ was set equal to the overall average dose (70.676) while under $H_a$ average doses for five out of ten associated regions were used (57.081 (lowest average), 62.860, 70.676 (overall average), 84.675 and 98.783 (highest average)).

In a tenth scenario, data that very closely reflect the structure of the Ceftriaxone Data were simulated. These datasets contained 329 children divided unequally over 20 countries within ten regions (Table 4.1). As this scenario covered ten regions rather than five, $\mu_k$ (under $H_a$) consisted of the average doses for all ten associated regions (69.534, 69.173, 98.783, 57.081, 59.966, 71.611, 65.484, 84.675, 70.955 and 62.859).

## 4.3 Models fitted

Based on the Ceftriaxone Data and following a rule of thumb (Snijders and Bosker, 1999), which states that a group-effect should be considered as a fixed effect when the number of levels is small ($< 10$), region was modelled as a fixed effect and country was modelled as a random effect. Therefore all simulated datasets (Scenarios 1-10)

were analysed with a model containing a random effect for country and a fixed effect for region.

The variability across the regions can be represented by:

$$\frac{\sum_{s=1}^{S} \frac{\sum_{k=1}^{K_s} (\hat{\mu}_{ks} - \bar{\hat{\mu}}_s)^2}{K_s - 1}}{S}, \tag{4.2}$$

where $\hat{\mu}_{ks}$ represents the estimate for the effect of region $k$ ($k = 1, ..., K_s$) in simulated dataset $s$ ($s = 1, ..., S$), $K_s$ is the number of regions in simulated dataset $s$, $S$ is the number of simulated datasets and $\bar{\hat{\mu}}_s$ is the average of all $\hat{\mu}_{ks}$ in simulated dataset $s$. The variability across regions that was used in the simulation procedure (true variability) can be calculated from equation (4.2), by replacing $\hat{\mu}_{ks}$ by $\mu_k$ (given in Section 4.2). The true variability under $H_a$ equals 286.563 for Scenarios 1-9 and 152.996 for Scenario 10.

The performance of the F test for the effect of *Region* was studied at the 5% significance level. The first performance characteristic that is presented is the type I error rate, which is computed as the number of times the null hypothesis is rejected in the datasets that were simulated under $H_0$. A second characteristic of interest is the power, which is computed as the number of times the null hypothesis is rejected in the datasets that were simulated under $H_a$. The last characteristic is the corrected power of the test which uses a corrected p-value obtained from comparing the test statistic to the distribution of test statistics simulated under $H_0$. Plots of the observed versus the expected p-values (following a uniform distribution between 0 and 1) were created for a scenario with primary unit sparseness (Scenario 3). Under $H_a$, both observed and corrected p-values were used.

Since the simulated datasets contained information on only five levels, region was modelled as a fixed effect. But as the regions in the simulated datasets are a sample of the regions in the original dataset, which are a sample of all regions worldwide, region could also be considered a random effect. Therefore, a model with a random effect for both country and region was fitted to the simulated datasets (Scenarios 1-10). For these models, presentation of the results is somewhat altered. Variability is represented by the variance of the random intercepts for *Region* averaged over all simulated datasets. The effect of *Region* can no longer be assessed with an F test as the true

parameter value now lies on the boundary of the parameter space (Verbeke and Molenberghs, 1997). Instead, we studied the performance of a likelihood ratio test for the effect of *Region* based on a 50:50 mixture of a $\chi^2_0$ and a $\chi^2_1$ distribution.

As the simulation setting consists of scenarios with a high number of singletons, we investigated different ways to handle the singletons. The first method that comes to mind to handle the singletons is to eliminate them. However, a country that is alone in a region could contain a lot of children (e.g. *Region 6* from the motivating example (Table 4.1)). As this implies that discarding the singletons would result in the elimination of a large percentage of collected data, this option would generally not be accepted. For illustration purposes, the singletons in Scenarios 3 and 10 were dropped. This implied that the dropped datasets (Scenarios 3b and 10b) contained 96 rather than 240 and 182 rather than 329 children, respectively. Due to dropping, Scenario 3b included two instead of five regions while Scenario 10b included four instead of ten regions. An alternative method to handle the singletons is to regroup them into one big region (Scenarios 3c and 10c). Due to regrouping, Scenario 3c included three instead of five regions while Scenario 10c included five instead of ten regions. A third option to deal with singletons is to split each single region into two artificial countries hence decreasing primary unit sparseness (Scenarios 3d and 10d). Due to splitting, Scenario 3d included ten instead of seven countries while Scenario 10d included 26 instead of 20 countries. The total number of children that were used in the analysis did not change by regrouping or splitting.
These additional scenarios (Scenarios 3b-d, 10b-d) were analysed with a linear multilevel model containing a random effect for country and a fixed effect for region. Variability is reported as in Equation 4.2, with true variability under $H_a$ equal to 16.698 for Scenarios 3b and 3d, 21.829 for Scenarios 10b and 10d, 212.385 for Scenario 3c and 40.551 for Scenario 10c. Performance of the F test is presented by the type I error rate, power and corrected power.

Some well-known alternatives to the F test for the effect of *Region* include the Wald test (under REML or ML) and the likelihood ratio test. We assessed the performance of these tests in a scenario with primary unit sparseness (Scenario 3). Plots of the observed versus expected p-values (following a uniform distribution between 0 and 1) were created. Under $H_a$, corrected p-values were added to the plots.
We also studied a non-parametric alternative to the F test for the effect of *Region*, being the permutation test. This procedure is used to determine statistical significance of a parameter by rearranging the data (Lehmann, 2006). For each simulated dataset

in Scenario 3 a permutation test (for the effect of *Region*) was set up according to the following steps (Chihara and Hesterberg, 2011):

1. Reallocate countries to regions by sampling without replacement.
   *Note that we permute countries rather than children to retain the hierarchical structure.*

2. Use a linear multi-level model containing a random effect for country and a fixed effect for region to obtain the F statistic for the effect of *Region* in the permuted dataset.

3. Repeat steps 1 and 2 1000 times.

4. Calculate the permutation p-value as the proportion of times the F statistics obtained from the permuted datasets were at least as extreme as the F statistic obtained from the simulated dataset.

Plots of the permutation p-values versus expected p-values (following a uniform distribution between 0 and 1) were created for a scenario with primary unit sparseness (Scenario 3). Under $H_a$, corrected p-values were added to the plots.

## 4.4 Results

### 4.4.1 Region as a fixed effect

The simulated datasets for Scenarios 1-10 were analysed with a model containing a random effect for country and a fixed effect for region. Variability, type I error rate, power and corrected power under REML and ML are given in Table 4.5.

The variability under REML and ML was similar but the type I error rate was consistently higher under ML than under REML. The same was true for both power and corrected power. When all regions contained at least two countries, the difference in corrected power between REML and ML disappeared. When focussing on Scenarios 1-9, it can be seen that the variability decreased towards the true variability as the number of countries increased. When the number of countries was large, the type I error rate reached the nominal 5% level under REML, but not under ML. The power was high for datasets without singletons. Scenario 1 had an extremely high power and type I error rate.

Table 4.5: Variability, type I error rate, power and corrected power for Scenarios 1-10 and a model with region as fixed effect under REML and ML.

| | Scenario number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Under $H_0$ | Variability REML | 199.95 | 183.70 | 168.54 | 146.76 | 128.40 | 111.52 | 80.85 | 60.35 | 39.33 | 188.38 |
| | Variability ML | 199.95 | 183.70 | 168.54 | 146.76 | 128.40 | 111.52 | 80.85 | 60.35 | 39.33 | 192.78 |
| | Type I error rate REML | 941 | 267 | 112 | 122 | 59 | 57 | 47 | 49 | 51 | 52 |
| | Type I error rate ML | 945 | 747 | 528 | 392 | 341 | 289 | 174 | 127 | 89 | 698 |
| Under $H_a$ | Variability REML | 471.07 | 437.72 | 431.37 | 405.42 | 393.85 | 361.29 | 340.57 | 321.19 | 289.05 | 322.61 |
| | Variability ML | 471.07 | 437.72 | 431.37 | 405.42 | 393.85 | 361.29 | 340.57 | 321.19 | 289.05 | 326.70 |
| | Power REML | 993 | 318 | 166 | 196 | 190 | 337 | 625 | 796 | 979 | 105 |
| | Power ML | 993 | 862 | 820 | 744 | 724 | 785 | 850 | 913 | 994 | 852 |
| | Corrected power REML | 407 | 272 | 238 | 191 | 222 | 316 | 630 | 796 | 978 | 173 |
| | Corrected power ML | 407 | 393 | 391 | 309 | 291 | 371 | 632 | 796 | 978 | 238 |

*Note that type I error rate, power and corrected power are expressed as the number of rejections in* 1000 *simulated datasets.*

Average parameter estimates and standard deviations for the fixed effects are given in Table A10 (for Scenarios 1-9) and A11 (for Scenario 10). These tables show that the estimates under REML and ML were very similar under Scenario 10 and identical under Scenarios 1-9.

Plots of the observed versus the expected p-values (Scenario 3; Figure 4.1) demonstrate that under $H_0$ the p-values under REML are closer to the uniform distribution than under ML, which confirms that REML performs better than ML. Under $H_a$, we can see that ML greatly outperforms REML. After correction for the distribution of test statistics simulated under $H_0$, REML and ML are comparable with ML still doing slightly better.

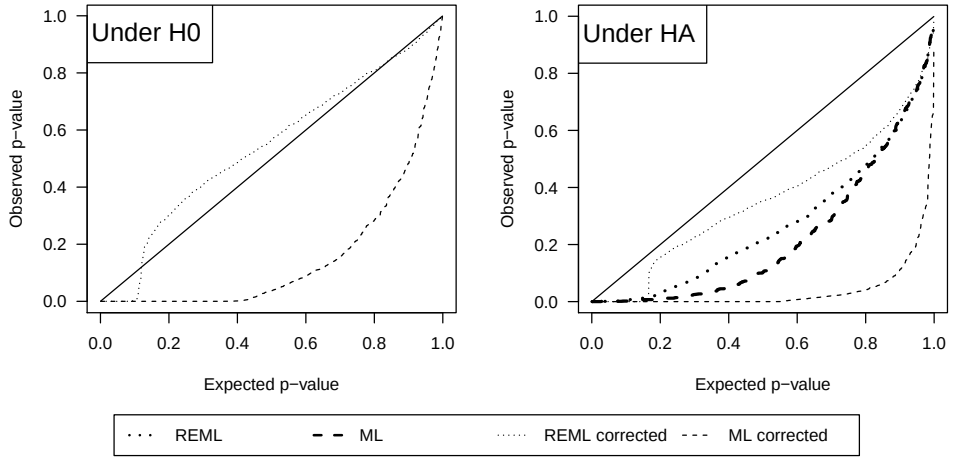



Figure 4.1: Observed versus expected p-values for the F test under $H_0$ (left) and $H_a$ (right) with p-values obtained under REML and ML represented by dotted and dashed lines, respectively. Corrected p-values are represented by bold lines.

### 4.4.2 Region as a random effect

The simulated datasets for Scenarios 1-10 were also analysed with a model containing a random effect for both country and region. Variability, type I error rate, power and corrected power under REML and ML are given in Table 4.6.

Table 4.6: Variability, type I error rate, power and corrected power for Scenarios 1-10 and a model with region as random effect under REML and ML.

| | Scenario number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Under $H_0$ | Variability REML | 184.45 | 89.11 | 68.70 | 55.15 | 48.54 | 41.89 | 26.25 | 18.07 | 11.32 | 37.33 |
| | Variability ML | 144.49 | 52.80 | 44.16 | 35.27 | 31.50 | 26.92 | 15.82 | 10.81 | 6.45 | 27.38 |
| | Type I error rate REML | 0 | 15 | 59 | 67 | 65 | 56 | 38 | 43 | 34 | 56 |
| | Type I error rate ML | 0 | 12 | 38 | 51 | 49 | 44 | 26 | 30 | 17 | 45 |
| Under $H_a$ | Variability REML | 455.65 | 318.74 | 298.20 | 253.90 | 235.90 | 257.26 | 262.19 | 258.87 | 249.53 | 127.12 |
| | Variability ML | 361.44 | 228.66 | 215.22 | 173.49 | 163.40 | 187.72 | 194.82 | 194.80 | 191.72 | 96.53 |
| | Power REML | 0 | 140 | 232 | 182 | 217 | 353 | 598 | 767 | 961 | 169 |
| | Power ML | 0 | 113 | 186 | 140 | 163 | 296 | 514 | 693 | 936 | 121 |
| | Corrected power REML | 0 | 245 | 212 | 139 | 171 | 317 | 630 | 796 | 978 | 152 |
| | Corrected power ML | 6 | 241 | 211 | 139 | 164 | 317 | 630 | 796 | 978 | 139 |

*Note that type I error rate, power and corrected power are expressed as the number of rejections in* 1000 *simulated datasets.*

The variability was systematically larger under REML than under ML. Under $H_0$, the variability decreased and the power increased as the number of countries increased. The combination of the decrease in variability and the increase in power indicates that the null hypothesis was correctly rejected more often in the absence of singletons. The corrected power became identical under REML and ML as soon as the regions contained at least two countries. The type I error rate was rather unstable.

When we compare the model containing region as a random effect (Table 4.6) with the model containing region as a fixed effect (Table 4.5), we see that the variability was substantially lower in the model with region as a random effect. As long as the regions contained two countries or less, the corrected power was larger for the model containing region as a fixed effect. When all regions contained more than two countries, the corrected power was equal under REML and ML, as well as for both models. This indicates that intrinsically the power is equal under REML and ML and for both models.

### 4.4.3 Handling the singletons

Different strategies to handle the singletons gave rise to six additional scenarios (i.e. $3b, 10b, 3c, 10c, 3d$ and $10d$). These scenarios were analysed with a model containing a random effect for country and a fixed effect for region. Variability, type I error rate, power and corrected power under REML and ML are given in Table 4.7. Both dropping the singletons from the analysis (Scenarios 3b and 10b) and regrouping the singletons into one big region (Scenarios 3c and 10c) resulted in a decreased variability, power and corrected power. The type I error rate dropped under REML and ML but reached the nominal 5% level only under REML. Splitting the regions into two artificial countries (Scenarios 3d and 10d) resulted in an increased type I error rate, power and corrected power.

Estimates and standard deviations for the fixed effects in the three additional scenarios are given in Tables A12 to A15.

Table 4.7: Variability, type I error rate, power and corrected power for the three additional scenarios and a model with region as fixed effect under REML and ML.

| | Scenario number | 3 | 3b | 3c | 3d | 10 | 10b | 10c | 10d |
|---|---|---|---|---|---|---|---|---|---|
| Under $H_0$ | Variability REML | 168.54 | 112.09 | 97.39 | 168.54 | 188.38 | 98.87 | 85.60 | 188.92 |
| | Variability ML | 168.54 | 112.09 | 97.39 | 168.54 | 192.78 | 102.05 | 86.83 | 192.09 |
| | Type I error rate REML | 112 | 56 | 62 | 369 | 52 | 57 | 47 | 291 |
| | Type I error rate ML | 528 | 179 | 181 | 713 | 698 | 285 | 167 | 754 |
| Under $H_a$ | Variability REML | 431.37 | 106.79 | 269.03 | 431.77 | 322.61 | 111.05 | 117.16 | 322.97 |
| | Variability ML | 431.37 | 106.79 | 269.03 | 431.77 | 326.70 | 114.70 | 117.92 | 326.12 |
| | Power REML | 166 | 77 | 131 | 761 | 105 | 82 | 66 | 665 |
| | Power ML | 820 | 198 | 388 | 946 | 852 | 366 | 187 | 950 |
| | Corrected power REML | 238 | 54 | 109 | 328 | 173 | 87 | 68 | 234 |
| | Corrected power ML | 391 | 56 | 129 | 410 | 238 | 93 | 44 | 226 |

*Note that type I error rate, power and corrected power are expressed as the number of rejections in* 1000 *simulated datasets.*

### 4.4.4 Alternatives to the F test

As an alternative to th F test, we studied the performance of the Wald test, likelihood ratio test and permutation test for a scenario with a small number of countries per region (Scenario 3). Plots of observed versus expected p-values for the Wald test and the likelihood ratio test (Figure 4.2) indicate that under $H_0$ the p-values were far from the uniform distribution, with the Wald test under REML doing slightly better than the two others. Under $H_a$,we can see that, after correction for the distribution of test statistics simulated under $H_0$, all are comparable with the Wald test under ML doing slightly better. Plots of permutation p-values versus expected p-values (Figure 4.3) show that under $H_0$ the p-values for REML were much closer to the uniform distribution than for ML, which illustrates that also the permutation test performs better under REML. Under $H_a$, the performance of the permutation test under REML and ML is comparable, with ML doing slightly better than REML.

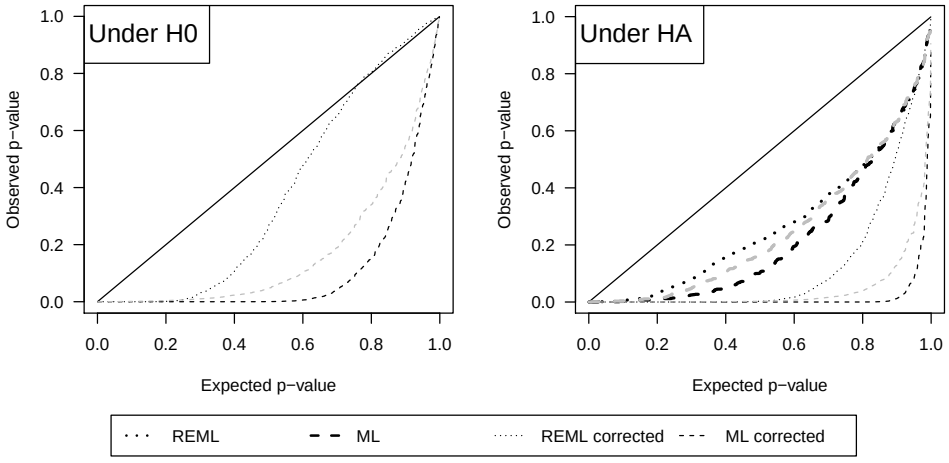



Figure 4.2: Observed versus expected p-values for the Wald test (black lines) and the likelihood ratio test (grey lines) under $H_0$ (left) and $H_a$ (right) with p-values obtained under REML and ML represented by dotted and dashed lines, respectively. Corrected p-values are represented by bold lines.

When comparing the Wald and likelihood ratio tests (Figure 4.2) to the F test (Figure 4.1), we can conclude that under $H_0$ neither the Wald nor the likelihood ratio test are valid alternatives to the F test, while under $H_a$ all are comparable. When comparing the permutation test (Figure 4.3) to the F test (Figure 4.1), it can be seen that, both under REML and under ML, the permutation test outperforms the F test. This is especially clear when comparing the tests under $H_0$, where the p-values from the permutation test matched the expected values quite well while the p-values from the F test did not. Under $H_a$, the permutation p-values were close to the corrected p-values, indicating an equivalent performance of the permutation test and the corrected F test.

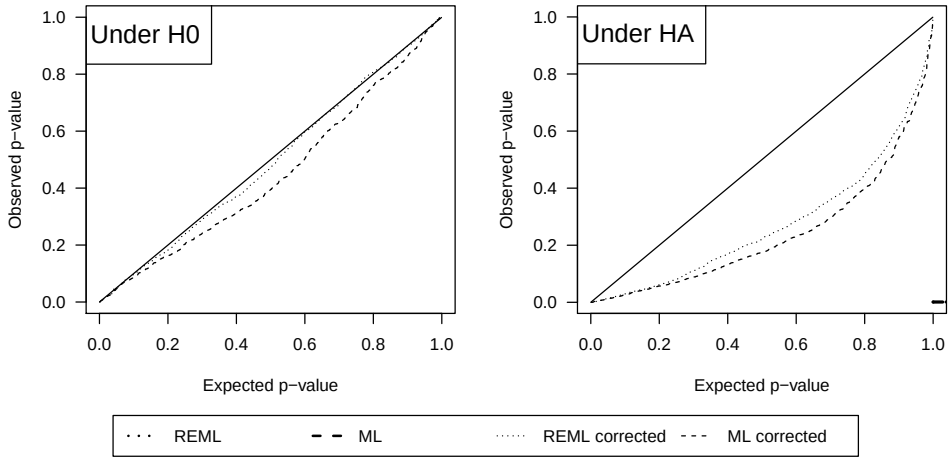



Figure 4.3: Observed versus expected p-values for the permutation test under $H_0$ (left) and $H_a$ (right) with p-values obtained under REML and ML represented by dotted and dashed lines, respectively. Corrected p-values are represented by bold lines.

## 4.5 Revisiting the motivating example

The four covariates model fitted to the Ceftriaxone Data used an F test under both REML and ML to determine significance of the fixed effects (results reported in Table 4.2). The difference in significance of the fixed effect for *Region* under REML and ML was striking.

Here, we assess significance of the fixed effects in the four covariates model using a Wald test, a likelihood ratio test and a permutation test rather than an F test. Significance of the fixed effects is reported in Table 4.8. To illustrate the principle of the permutation test, the simulated null distributions (under REML) are shown in Figure 4.4. Note that we used 1000 permutations and that a permutation test for the effect of region involves permuting countries (as all children in one country come from the same region) while a permutation test for the effect of department type, age and reason all involve permuting children (as each child could have a different age or be treated in a different type of department or for a different reason).

Similar to the p-values obtained using F tests, the p-values obtained using Wald tests under REML and ML were in strong disagreement. However, the p-values obtained using permutation tests under REML and ML were all in agreement. Therefore, using the permutation test allowed for a solid conclusion on the significance of the parameters in the model.

Table 4.8: Significance of the fixed effects in the four covariates model obtained using permutation tests under ML and REML.

| | Permutation test | | Wald test | | Likelihood ratio test |
|---|---|---|---|---|---|
| Effect | ML | REML | ML | REML | ML |
| Age | <.001 | <.001 | <.001 | <.001 | <.001 |
| Reason | 0.020 | 0.029 | 0.008 | 0.026 | 0.011 |
| Department type | <.001 | <.001 | <.001 | <.001 | <.001 |
| Region | 0.402 | 0.675 | <.001 | 0.615 | 0.029 |

*Note that permutation p values are based on* 1000 *permutations.*

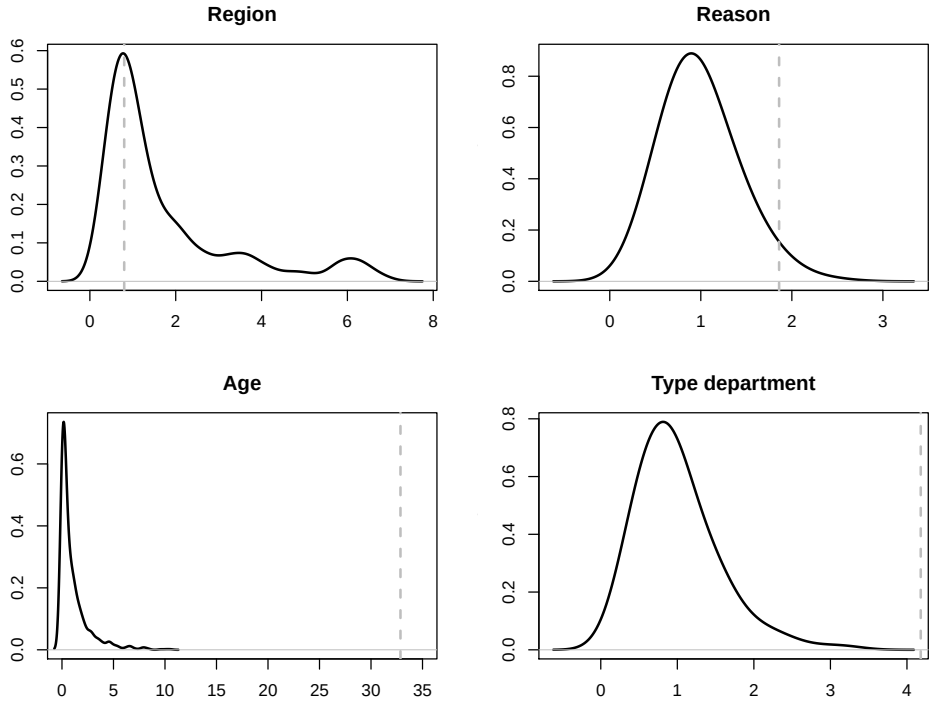



Figure 4.4: Null distributions used in the permutation tests. True F value are highlighted by a dashed line.

## 4.6 Discussion

A data example on ceftriaxone consumption in hospitalized children illustrated that significance of the fixed effects determined using F tests under REML and ML can differ substantially when there is sparseness at the level of the primary unit. We conducted a simulation study, based on the structure of this data example, to investigate the effect of increasing primary unit sparseness on the performance of the F test.

Fitting a model with region as a fixed effect to all scenarios showed that the variability across regions was similar whether the model was fitted under REML or under ML. Type I error rate, power and corrected power were higher under ML than under REML, which can be explained by the downward bias in the estimates for the variance components under ML (Harville, 1977; Searle *et al.*, 1992). When the primary unit sparseness decreased, the variability decreased towards the true variability, which can be explained by the more stable estimation of the average for a region

when the number of countries in that region is larger. For scenarios without primary unit sparseness, the power was high while the type I error rate reached the nominal 5% level under REML, but not under ML. This indicates that without primary unit sparseness, the F test does perform well. In scenarios with primary unit sparseness, type I error rate was high and power low. The high type I error rate implies that the null hypothesis is too often incorrectly rejected. The combination of a low power with high variability indicates that the null hypothesis is not often correctly rejected. These findings suggest that performance of the F test in the presence of primary unit sparseness is inadequate and there is need for an alternative approach.
Scenario 1, which consists of only singletons, has a very high type I error rate and power. This indicates that the null hypothesis got rejected regardless of the underlying truth. An explanation for this behaviour is that the simulated region averages equal the simulated country averages and hence they could by chance be far from the original average dose and far from each other. When a region contains several countries the simulated region averages are the average of the simulated average doses for the countries included in that region and the variability between countries is averaged out. This results in region averages that are closer to the original average dose for that region.

Fitting a model with region as a random effect to all scenarios showed that the variability was systematically larger under REML than under ML which can again be explained by the downward bias in the ML estimates for the variance components (Harville, 1977; Searle *et al.*, 1992). The variability decreased when the number of countries increased which is a result of the average dose for regions getting closer to the original average when more countries are included.
When comparing the model with region as a random effect to the model with region as a random effect, it is clear that the variability was substantially lower in the model with region as a random effect. This can be explained by shrinkage, which causes the variance of the random effects to underestimate the true variability (Verbeke and Molenberghs, 1997, 2009).

To reduce primary unit sparseness and hence improve the performance of the F test, we considered deleting singletons, regrouping them into one big region or splitting them into two artificial countries. Dropping the singletons resulted in a decrease in variability, which can be explained by the removal of small regions, retaining only those regions that contain several countries. Also regrouping the singletons resulted in a decrease in variability, which can be explained by the removal of small regions by

including them into one big artificial region. Both dropping and regrouping singletons resulted in a decreased power, corrected power and type I error rate. Splitting the singletons resulted in an increase in type I error rate, power and corrected power. This can be explained by the resemblance of two artificial countries within one region, which reduces within-region variability considerably and hence increases the overall rejection rate. Therefore, the three options could solve either the problem of a high type I error rate or the problem of low power, whilst worsening the other. This forces us to conclude that neither method acts as a solution to the poor performance of the F test in the presence of singletons.

As an alternative to the F test, we studied the performance of the Wald test, the likelihood ratio test and the permutation test. While performance of the Wald and likelihood ratio test were comparable to performance of the F test, the permutation test outperformed the F test both under REML and ML.

As we have highlighted the problems that arise with sparseness at the level of the primary unit, a small note should be made with regard to study design. If possible, one should strive to avoid inclusion of singletons in the sample when setting up a study by including at least two secondary units within each primary unit. If the presence of singletons in the sample is unavoidable, one should be careful in using the Wald test, F test or likelihood ratio test and the permutation test (under REML) should be considered whenever different test statistics are in disagreement.
Note that we only considered the three-level setting with children nested within countries nested within regions. We do however believe that the results discussed in this chapter can be generalized to other and more complex settings and that if problems arise in a rather simple setting, they will surely surface in more complex settings.

# Chapter 5

# Variation in doses of $\beta$-lactam antibiotics prescribed to hospitalized children

$\beta$-Lactam antibiotics are one of the oldest and most popular classes of antibacterial agents. They contain a $\beta$-lactam ring in their molecular structure and include penicillins, cephalosporins, monobactams and carbapenems. The $\beta$-lactam antibiotics work by inhibiting the last stage of the cell wall synthesis, causing the cell wall to rupture and the bacteria to die (Neu and Gootz, 1996).
Because the different $\beta$-lactam antibiotics have a similar pharmacological profile, we would expect the reasons for deviating from the recommended dose to be similar and the variability in prescribed dose to be rather small. However, large variability in prescribed $\beta$-lactam doses is observed in the Inpatient Antibiotic Use Data (Section 1.2.3).

In this chapter, we will study the causes of variation in the individual $\beta$-lactam antibiotics and use this information in the construction of a meta-model to assess whether factors causing variation in individual $\beta$-lactam antibiotics impact all $\beta$-lactam antibiotics in the same manner. This meta-model will then also be used to test whether higher doses were prescribed to children treated for a severe infection compared to children treated for a mild or moderate infection, with severity of infection classified as shown in Table 1.3.

## 5.1 Assessing the causes of variation in prescribed doses for a single $\beta$-lactam antibiotic

We used the multi-level model to identify the causes of variation in prescribed doses of the 12 individual $\beta$-lactam antibiotics. The full model contained a random intercept for the department (Id_dep), the hospital (Id_ins) and the country. Other explanatory variables (listed in Table 1.2) were included as fixed effects. Due to the high correlation between age and weight, and because dosing guidelines are weight-based rather than age-based, we included weight rather than age in the starting model.

The multi-level model can be presented as follows:

$$Y_{ijkl} = \beta_0 + \beta_R X_{Rl} + b_{0l} + \beta_H X_{Hkl} + b_{0kl} + \sum_{m=3}^{M} \beta_m X_{mjkl} + b_{0jkl} + \sum_{p=M+1}^{P} \beta_p X_{pijkl} + \epsilon_{ijkl},$$

where $Y_{ijkl}$ represents the dosage prescribed to child $i$ ($i = 1, ..., n_j$) in department $j$ ($j = 1, ..., n_k$) in hospital $k$ ($k = 1, ..., n_l$) in country $l$ ($l = 1, ..., N$), $n_j$ are the number of children in department $j$, $n_k$ are the number of departments in hospital $k$, $n_l$ are the number of hospitals in country $l$ and $N$ are the number of participating countries. $X_{Rl}$ is a covariate at country-level (i.e. region), $X_{Hkl}$ is a covariate at hospital-level (i.e. hospital type), $X_{mjkl}$ is covariate $m$ ($m = 1, ..., M$) at department-level (e.g. department type), $M$ is the number of covariates at department-level, $X_{pijkl}$ is covariate $p$ ($p = 1, ..., P$) at child-level (e.g. gender), $P$ is the number of covariates at child-level, $\beta_R$, $\beta_H$, $\beta_m$ and $\beta_p$ are the respective coefficients for the listed parameters, $\beta_0$ is the general intercept and $\epsilon_{ijkl}$ is the residual error term. $b_{0l}$ is the country-specific random intercept, $b_{0kl}$ is the hospital-specific random intercept and $b_{0jkl}$ is the department-specific random intercept.

Box plots were constructed to identify outlying observations at each level of the hierarchy (i.e. for prescribed doses in children, departments, hospitals and countries. Outliers were assumed to be either recording errors or extreme cases and were therefore removed from the analysis (4.4% of included observations). To avoid estimation problems, we removed observations whenever there were less than three observations in a category (0.4% of included observations).

Exploratory plots of total dose (expressed in mg/kg/day) versus weight (expressed in kg) revealed that some doses were prescribed according to and others independent of weight (Figure 5.1). To account for this, weight was included in the model as follows:

$$I_{ijkl} w_{ijkl}^2 + (1 - I_{ijkl}) w_{ijkl},$$

with $w_{ijkl}$ the weight for child $i$ in department $j$ in hospital $k$ in country $l$.

Here,

$$I_{ijkl} = \begin{cases} 0 & \text{if dose occurrence} \leq 1\% \\ \frac{oc}{5} & \text{if } 1\% < \text{dose occurrence} < 5\% \\ 1 & \text{if dose occurrence} \geq 5\% \end{cases}$$

with *oc* representing the percentage of times a specific dose occurs in the Inpatient Antibiotic Use Data.

The covariance structure of the starting model was reduced under REML in a backwards fashion using likelihood ratio tests based on a 50:50 mixture of a $\chi_0^2$ and a $\chi_1^2$ distribution. Afterwards, the mean structure was reduced in a backwards fashion. This was done, as advocated in Chapter 4, by selecting a variable for removal from the model under REML and verifying this action under ML. Whenever there was disagreement between REML and ML, a permutation test under REML was used to decide on the removal of the variable from the model. In a next step, interaction terms between the remaining fixed effects were included whenever the interaction term did not contain sparse levels and a second round of backwards elimination was performed.

## 5.2 Differentiating between two prescribing styles

The most frequently prescribed antimicrobial in the Inpatient Antibiotic Use Data is parenteral ceftriaxone (ATC code J01DD04; 18.9%). The exploratory plot of total ceftriaxone dose (expressed in mg/kg/day) versus weight (expressed in kg) showed that prescriptions were given according to two different styles, i.e. one according to and one independent of the childs weight (Figure 5.1).

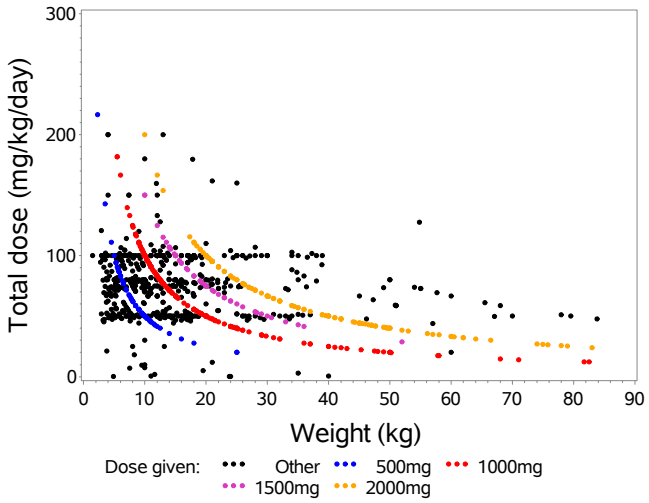



Figure 5.1: Scatter plot of total dose (expressed in mg/kg/day) versus weight (expressed in kg) for ceftriaxone prescriptions in children.

To identify patient characteristics that differentiate between the two styles of prescribing, we constructed a generalized linear mixed model using an indicator, with 1 being a dose prescribed independent of weight, as outcome variable and a logit link. We included a random intercept for country, for hospital (Id_ins) and for department (Id_dep) and fixed effects for the other explanatory variables listed in Table 1.2. Because there was a high correlation between age and weight (0.88), we conducted this analysis once when including age and once when including weight.

To avoid convergence issues while including three random effects in the generalized linear mixed model, we used a Laplace approximation. The covariance structure of the model was reduced in a backwards fashion using likelihood ratio tests based on a 50:50 mixture of a $\chi^2_0$ and a $\chi^2_1$ distribution. Afterwards, the mean structure was reduced in a backwards fashion using F tests.

## 5.3 Assessing causes of variation in prescribed doses for the class of $\beta$-lactam antibiotic

In Section 5.1, we constructed a final antibiotic-specific model for each of the included $\beta$-lactam antibiotics (listed in Section 1.2.3). Using variables (fixed effect, random effect or two-way interaction) that were significant in at least one of these antibiotic-specific models, we build a meta-model. To allow for heterogeneity across antibiotics, we included a fixed effect for the type of antibiotic and an interaction between each included fixed effect and the effect for the type of antibiotic. To avoid estimation problems, we removed observations whenever there were less than five observations in a category (0.8% of included observations). The covariance structure of the meta-model was reduced under REML using likelihood ratio tests based on a 50:50 mixture of a $\chi^2_0$ and a $\chi^2_1$ distribution. The mean structure was reduced by selecting a variable for removal from the model under REML and verifying this action under ML. Whenever there was disagreement between REML and ML, a permutation test under REML was used to decide on the removal of the variable from the model. Goodness of fit for both the final antibiotic-specific models and the final meta-model was represented by $R^2$ and adjusted $R^2$.

Using this meta-model, we set out to answer the question whether higher doses are prescribed to children treated for more severe infections than to children treated for less severe infections, with reason for treatment classified as shown in Table 1.3. Answering this question involves testing the following two hypotheses:

$$\mu_{Severe} > \mu_{Moderate},$$

$$\mu_{Severe} > \mu_{Mild},$$

with $\mu_{Severe}$, $\mu_{Moderate}$ and $\mu_{Mild}$ representing the average dose prescribed to a child for a severe, moderate or mild infection, respectively. Because answering this question requires testing two comparisons using the same data, the significance level was adjusted ($\alpha = 0.025$).

## 5.4 Results

In this section, we will elaborately discuss the results for the $\beta$-lactam antibiotic that was most frequently prescribed to hospitalized children in the Inpatient Antibiotic Use Data, being parenteral ceftriaxone (ATC code J01DD04; 18.9%). This is a third-generation cephalosporin that is administered parenterally and has broad-spectrum activity against Gram-positive and Gram-negative bacteria.
Results for the other included $\beta$-lactam antibiotics (listed in Section 1.2.3) will not be discussed in detail but will be used in the construction of the meta-model.

### 5.4.1 Causes of variation in prescribed doses of ceftriaxone

Likelihood ratio tests indicated that the random effect for hospital could be removed from the model. The random effects for country and department had to be retained. Parameter estimates for the final model are given in Table 5.1. It can be seen that lower doses were prescribed for empiric (compared to targeted) treatment and for children with a lower weight. All pairwise comparisons for the variable Reason are reported in Table 5.2. These comparisons show that, as expected, higher doses were prescribed for more severe infections.

Table 5.1: Parameter estimates and standard errors for the fixed effects in the final model for prescribed doses of parenteral ceftriaxone.

| Effect | Estimate | Std. error |
|---|---|---|
| Intercept | 87.7490 | 2.9389 |
| Type_treat (empiric) | −4.9700 | 2.0765 |
| Weight | −0.1627 | 0.0681 |
| Weight$^2$ | −0.0121 | 0.0008 |
| Reason (different) | −11.5444 | 2.3526 |
| Reason (mild) | −16.2708 | 2.8619 |
| Reason (moderate) | −6.4410 | 1.6168 |

Table 5.2: Estimates, standard errors and Tukey-adjusted p-values for pairwise comparisons for Reason in the final model for prescribed doses of parenteral ceftriaxone.

| A versus B | | Estimate | Std. error | P-value |
|---|---|---|---|---|
| Different | Mild | 4.7264 | 3.1946 | 0.4503 |
| Different | Moderate | −5.1033 | 2.1183 | 0.0760 |
| Different | Severe | −11.5444 | 2.3526 | <0.0001 |
| Mild | Moderate | −9.8297 | 2.6585 | 0.0013 |
| Mild | Severe | −16.2708 | 2.8619 | <0.0001 |
| Moderate | Severe | −6.4410 | 1.6168 | 0.0004 |

### 5.4.2 Reasons for prescribing ceftriaxone independent of weight

Likelihood ratio tests indicated that the random effect for hospital could be removed from the model. The random effects for country and department had to be retained. The only explanatory variable that was associated with the indicator was either age or weight, depending on which variable was included in the starting model. Parameter estimates for both models are reported in Table 5.3. These results show that the higher the age (or weight), the higher the odds of receiving a dose prescribed independent of weight. This finding is illustrated in Figure 5.2.

Table 5.3: Estimates and standard errors for the fixed effect in the final model for receiving a prescription independent of weight.

| Parameter | Estimate | Std. error |
|---|---|---|
| Intercept | −0.9386 | 0.2033 |
| Age | 0.1576 | 0.0204 |
| Intercept | −1.0978 | 0.2164 |
| Weight | 0.047 01 | 0.0064 |

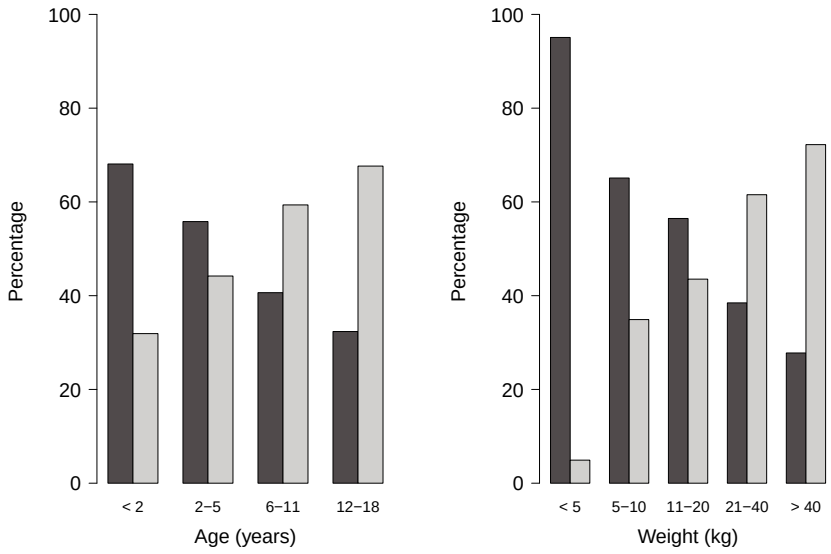



Figure 5.2: Percentage of children receiving a dose of ceftriaxone dependent (black) or independent (grey) of weight according to their age (left) and weight (right).

The need for a random intercept for country (and department) indicates that the preference for one of the two prescribing styles is fairly different between included countries (and departments), as can be seen in Figure 5.3.

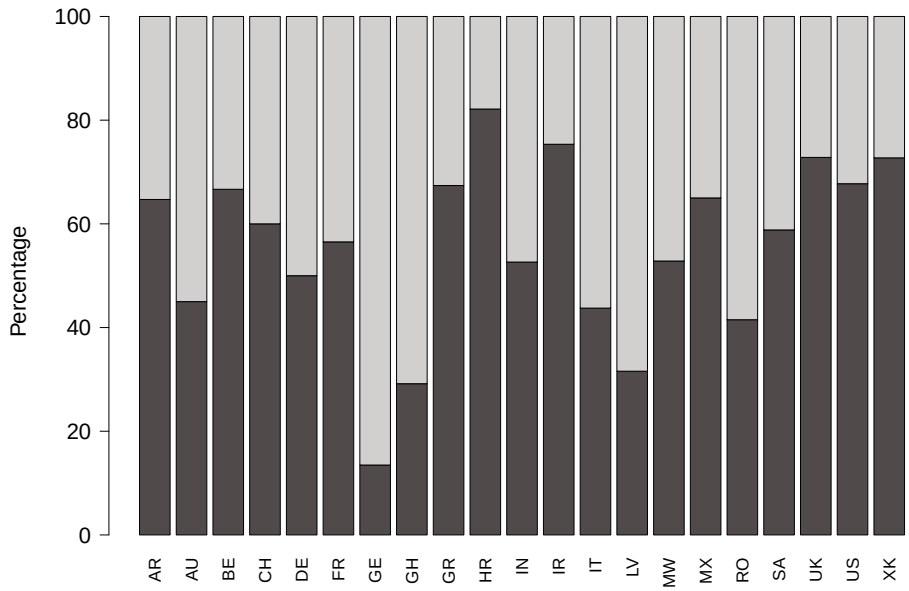



Figure 5.3: Percentage of children receiving a dose of ceftriaxone dependent (black) or independent (grey) of weight in participating countries for which at least 10 ceftriaxone prescriptions were recorded.

### 5.4.3 Causes of variation in prescribed doses for the class of $\beta$-lactam antibiotics

Likelihood ratio tests indicated that none of the random effects could be removed from the model. The mean structure was reduced using backwards model building. There was disagreement on the removal of the interaction between presence of a primary underlying diagnosis and type of antibiotic. The permutation test under REML however showed that the interaction needed to be kept in the model. Significance of the fixed effects in the final meta-model is reported in Table 5.4.

Table 5.4: Significance of the fixed effects in the final meta-model for 12 $\beta$-lactam antibiotics.

| Variable | P-value | Variable | P-value |
|---|---|---|---|
| Type_hosp | 0.0452 | Type_treat | 0.0011 |
| Atb | $< 0.0001$ | | |
| Beds | 0.1174 | Beds*atb | 0.0044 |
| Gender | 0.0096 | Gender*atb | 0.0298 |
| Type_dep | 0.0141 | Type_dep*atb | 0.0005 |
| Vent | 0.7299 | Vent*atb | 0.0466 |
| Ud1 | 0.9953 | Ud1*atb | 0.0624 |
| W1 | 0.2026 | W1*atb | 0.0495 |
| W2 | 0.3364 | W2*atb | $< 0.0001$ |
| Prev | 0.1423 | Prev*atb | 0.0024 |
| Reason | 0.4265 | Reason*atb | $< 0.0001$ |
| Region | 0.0054 | Region*atb | $< 0.0001$ |
| Indic | 0.0006 | Indic*atb | $<0.0001$ |
| W2*indic | 0.1931 | W2*indic*atb | 0.0072 |
| Prev*region | 0.1891 | Prev*region*atb | 0.0384 |
| W1*reason | 0.0876 | W1*reason*atb | $<0.0001$ |

The residual plot does not show a clear structure, indicating that the variables in the model explain most of the variation in the data (Figure 5.4). The final meta-model has an $R^2$ value of 0.8326 and an adjusted $R^2$ value of 0.8175, which implies that the model explains about 82% of the variability in the data and hence verifies that the model fits the data reasonably well.

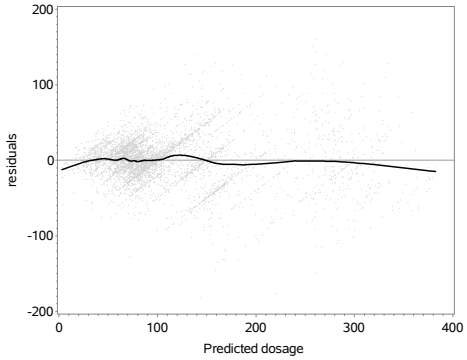



Figure 5.4: Scatter plot of the residuals (dots) and smoothed average trend (solid line) from fitting the final meta-model.

Table 5.4 shows that most variables have an antibiotic-specific influence. Only type of treatment and type of hospital the child was treated in influenced all $\beta$-lactam antibiotics in a similar fashion, with higher $\beta$-lactam doses prescribed for targeted treatment and in tertiary or specialized hospitals. Parameter estimates for these variables are reported in Table 5.5.

Table 5.5: Parameter estimates and standard errors for the effect of type of treatment and type of hospital on the 12 included $\beta$-lactam antibiotics.

| Category | Estimate | Std. error |
|---|---|---|
| Primary or secondary versus tertiary or specialized hospital | −4.5839 | 2.2811 |
| Targeted versus empiric treatment | 5.2327 | 1.6073 |

### 5.4.4 Difference in prescribed doses of $\beta$-lactam antibiotics according to reason for treatment

Using the meta-model, we assessed whether higher doses of $\beta$-lactam antibiotics were prescribed to children treated for a severe infection compared to children treated for a mild or moderate infection. Because the interaction between reason for treatment and type of antibiotic was significant, this question was answered for each antibiotic separately (Table 5.4). None of the children included in the PPS was treated for a severe infection with oral amoxicillin. Therefore, this question could not be answered for this antibiotic. Because the interaction between weight, antibiotic and reason for treatment was significant as well, this question was answered for a child with average weight (i.e. 20kg). The indicator value was set to 0.5.

The results show that doses given for treatment of severe infections were significantly higher than doses given for treatment of mild infections, only when using parenteral ampicillin, benzylpenicillin, cefotaxime or ceftriaxone. Doses given for treatment of severe infections were significantly higher than doses given for treatment of moderate infections, only when using parenteral ampicillin or cefotaxime.

Table 5.6: Parameter estimates, standard errors and p-values for antibiotic-specific assessment of the question whether higher doses are prescribed to children treated for a severe infection than to children treated for a mild or moderate infection.

| Antibiotic | Estimate | Std. Error | P-value | Estimate | Std. Error | P-value |
|---|---|---|---|---|---|---|
| Parenteral ampicillin | 41.3295 | 10.0652 | 0.0001 | 37.4757 | 6.7813 | 0.0001 |
| Parenteral benzylpenicillin | 41.3334 | 8.2605 | 0.0001 | −9.7707 | 6.4236 | 0.9358 |
| Oral amoxicillin with inhibitor | −4.1627 | 16.6268 | 0.5988 | −2.7184 | 15.8908 | 0.5679 |
| Parenteral amoxicillin with inhibitor | −1.0586 | 11.6537 | 0.5362 | −1.2257 | 10.1288 | 0.5482 |
| Parenteral piperacillin with inhibitor | 5.0391 | 14.4334 | 0.3635 | −2.3326 | 4.3991 | 0.7020 |
| Parenteral cefazolin | 1.5004 | 26.9452 | 0.4778 | −2.2599 | 25.0178 | 0.5360 |
| Parenteral cefuroxime | −13.9941 | 9.8635 | 0.9220 | −21.1983 | 8.0528 | 0.9957 |
| Parenteral cefotaxime | 21.4795 | 7.9345 | 0.0034 | 20.9060 | 3.8064 | 0.0001 |
| Parenteral ceftazidime | 2.5141 | 20.9412 | 0.4522 | −4.4752 | 6.5265 | 0.7535 |
| Parenteral meropenem | −10.6687 | 15.7443 | 0.7510 | 2.8163 | 3.8043 | 0.2295 |
| Parenteral ceftriaxone | 14.5776 | 5.6287 | 0.0048 | 4.3872 | 2.8768 | 0.0637 |

None of the children in the PPS was treated for a severe infection with oral amoxicillin.

## 5.5 Discussion

In this chapter, we illustrated that a $\beta$-lactam antibiotic is prescribed according to or independent of weight. We demonstrated that the choice between both styles of prescribing depends solely on weight itself (or age), with a higher chance of getting a prescription independent of weight for children with a higher weight (or age). We identified variables that cause variation in prescribed ceftriaxone doses using a hierarchical model and showed that lower doses were prescribed to children receiving empiric treatment, with a lower weight and with a less severe reason for treatment.

Although the model presented in this chapter has identified several reasons for the observed variation in prescribed ceftriaxone doses, a large proportion of the variation remains unexplained (adjusted $R^2 = 0.5524$). As very diverse variables, playing at different levels of the hierarchy (child-department-hospital-country) have been included in the starting model, this finding suggests that ceftriaxone doses might be prescribed based on a physician's personal idea rather than based on existing guidelines.

When combining antibiotic-specific models for 12 $\beta$-lactam antibiotics, the meta-model explained a large proportion of the variation in the data (adjusted $R^2 = 0.8175$). We showed that the variation in $\beta$-lactam antibiotics is attributed to a large subset of variables. Although we would expect the reasons for deviating from the average, which should correspond to the recommended dose, to be common for the 12 included $\beta$-lactam antibiotics, most variables appeared to act antibiotic-specifically. The variables type of treatment and type of hospital did affect all 12 included $\beta$-lactam antibiotics in a similar fashion.

# Chapter 6

# Modelling outpatient antibiotic use in defined daily doses and packages

Analyses of data on antibiotic consumption collected by ESAC between 1997 and 2009 revealed that total outpatient antibiotic use expressed in DID increased significantly over time while showing a significant seasonal fluctuation that decreased over time (Adriaenssens *et al.*, 2011*a*). Analyses of the eight pharmacological subgroups reached similar conclusions (Adriaenssens *et al.*, 2011*b*,*c*; Coenen *et al.*, 2011; Faes *et al.*, 2011; Versporten *et al.*, 2011*a*,*b*).
Expressing outpatient antibiotic use in DID is however not always optimal, e.g. when the number of DDD per package differs substantially between countries or within a country over time. For this reason ESAC proposed PID as an additional outcome measure (Adriaenssens *et al.*, 2011*a*; Coenen *et al.*, 2014). Recently, Coenen *et al.* (2014) showed that PID is a good proxy for the number of treatments, and a more appropriate measure than DID when assessing antibiotic use over time in Belgium.

In this chapter, we will use the Outpatient Antibiotic Use Data (Section 1.2.4) to complement analyses of European outpatient antibiotic consumption expressed in DID by analyses of data expressed in PID, assess the agreement between both measures and study changes in the number of DDD per package over time.

## 6.1 Analysis of DID and PID separately

Because measurements for each country were taken quarterly, within-country correlation has to be accounted for and hence mixed models are an adequate tool to study the trends in the data. Mixed models include fixed effects, which here represent the average trend in Europe based on the countries in the sample, and random effects, which here represent the deviation of individual countries from this average trend (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2009). The seasonal fluctuation in the data can be modelled using a sinusoidal component which can be included when using a non-linear mixed model.

As a starting model, we used the non-linear mixed model previously applied in the analyses of the ESAC data (Minalu *et al.*, 2011). The model is defined as:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + (\beta_2 + b_{2i} + (\beta_3 + b_{3i})t_{ij})sin(\omega t_{ij} + \delta) + \epsilon_{ij},$$

where $Y_{ij}$ represents the antibiotic consumption (expressed in DID or PID) in country $i$ at time point $t_{ij}$ ($j = 1, 2, ..., n_i$), $n_i$ is the number of observations in country $i$, $t_{ij} = 1$ corresponds to the first measurement (first quarter of 2000), $\beta_0$ and $b_{0i}$ are fixed and random intercepts (reflecting antibiotic consumption for $t_{ij} = 1$), $\beta_1$ and $b_{1i}$ are fixed and random slopes (reflecting the change in antibiotic consumption over time), $\beta_2$ and $b_{2i}$ are fixed and random amplitudes for the sine function (reflecting the height of the upward winter and downward summer peaks), $\beta_3$ and $b_{3i}$ are fixed and random changes in the amplitude over time, $\omega$ is the frequency in which the sine function repeats itself ($= 2\pi/T$ with $T = 4$ quarters), $\delta$ is the phase shift which is an unknown parameter and $\epsilon_{ij}$ is the measurement error. We assumed that the vector of the random effects ($\boldsymbol{b_i}$) follows a normal distribution with mean zero and covariance matrix D(4x4) and that the error terms are independent and normally distributed with mean zero and covariance matrix $\Sigma_i$, which was taken equal to $\sigma^2 \boldsymbol{I}_{n_i}$ with $\boldsymbol{I}_{n_i}$ the $n_i$-dimensional identity matrix.
Because convergence could not be obtained when fitting the model with an unstructured covariance matrix, we set the covariances between $b_{3i}$ and the other random effects equal to zero.

The need for random effects was tested with a likelihood ratio test based on the comparison of the maximized likelihoods for the model with and without the random effect of interest. As the null hypothesis for this test was situated on the boundary of the parameter space, classical likelihood inference based on a single $\chi^2$ distribution could not be used and a mixture of two equally weighted (weight= 0.5) $\chi^2$ distri-

butions with $k$ and $k+1$ degrees of freedom had to be used instead (Verbeke and Molenberghs, 2009). After removal of each random effect we tested whether the accompanying fixed effect could be removed using a likelihood ratio test based on a single $\chi^2$ distribution.

## 6.2 Analysis of DID and PID jointly

The agreement between DID and PID was investigated by combining their final models into a joint non-linear mixed model. To avoid convergence problems, the variance estimates for PID were rescaled (by multiplication with factor 10). Correlations between matching random effects, used to assess the agreement between DID and PID, were estimated using the covariance matrix. To ease convergence, a covariance matrix containing only covariances between matching random effects was used. A squared version of the Wald test was used to test for a perfect correlation (Kutner *et al.*, 2005). As a correlation is restricted to lie between $-1$ and $+1$, the null hypothesis was again situated on the boundary of the parameter space and an equally weighted mixture of a $\chi^2$ distribution with one and zero degrees of freedom had to be used instead of a single $\chi^2$ distribution.

## 6.3 Change in dose per package

The dose per package was calculated by dividing DID by PID and expressed in number of DDD per package. As DID and PID measure the same seasonal fluctuation, dividing both measures cancelled out most of the seasonality. For this reason, a non-linear term was no longer required to model the change in dose per package over time and a linear mixed model was used. This model is defined as:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij},$$

where $Y_{ij}$ represents the dose per package in country $i$ at time point $t_{ij}$ ($j = 1, 2, ..., n_i$), $n_i$ is the number of observations in country $i$, $t_{ij} = 1$ corresponds to the first measurement (first quarter of 2000), $\beta_0$ and $b_{0i}$ are fixed and random intercepts (reflecting the dose per package for $t_{ij} = 1$), $\beta_1$ and $b_{1i}$ are fixed and random slopes (reflecting the change in dose per package over time) and $\epsilon_{ij}$ is the measurement error. We assumed that the vector of random effects ($\boldsymbol{b_i}$) follows a normal distribution with mean zero and covariance matrix D(2x2) and that the error terms are independent and normally distributed with mean zero and covariance matrix $\Sigma_i$. A first order autoregressive (AR(1)) covariance matrix was used with the covariances equal to $\sigma^2\rho^{|td|}$, where $\sigma^2$

is the error variance, $\rho$ is the autocorrelation and $td$ reflects the time (in number of quarters) between two time points.

The need for random effects was assessed with likelihood ratio tests based on a mixture of two equally weighted $\chi^2$ distributions with $k$ and $k+1$ degrees of freedom. The need for inclusion of fixed effects was tested using likelihood ratio tests based on a single $\chi^2$ distribution.

## 6.4 Results

In this section, a detailed description of the results for J01 will be given. Results for other antibiotic subgroups will be summarized briefly.

### 6.4.1 Analysis of DID and PID separately

Likelihood ratio tests indicated that the random change in amplitude ($b_{3i}$) could be removed from the starting model for both DID and PID. Other random and fixed effects had to be retained. Parameter estimates for the final models are given in Table 6.1.

Table 6.1: Parameter estimates for the fixed effects in the non-linear mixed model for J01 use in Europe; $^*: P < 0.05$,$^{**}: P < 0.0001$.

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|---|---|---|---|---|
| DID | $14.2404^{**}$ | $0.1037^{**}$ | $3.0858^{**}$ | $0.0187^{*}$ |
| PID | $3.1006^{**}$ | $-0.0061$ | $0.6608^{**}$ | $-0.0033$ |

In DID, antibiotic use increased significantly over time with a significant seasonal fluctuation that increased significantly over time. In PID, antibiotic use decreased non-significantly over time with a significant seasonal fluctuation that did not change significantly over time. These results indicate that conclusions based on DID and PID could be contradictory.

The estimates for the variance components are:

$$\text{DID: } \begin{pmatrix} 46.1383 & -0.1754 & 12.0934 \\ & 0.0323 & 0.0667 \\ & & 3.8900 \end{pmatrix} \text{ and } \sigma^2 = 2.9932$$

$$\text{PID: } \begin{pmatrix} 2.2574 & -0.0133 & 0.4839 \\ & 0.0013 & -0.0029 \\ & & 0.1193 \end{pmatrix} \text{ and } \sigma^2 = 0.1152$$

The correlation between random effects in the model for antibiotic use expressed in DID was estimated to be $-0.1438$ (between random intercept and random slope), 0.9027 (between random intercept and random amplitude) and 0.0570 (between random slope and random amplitude) (Figure 6.1 left panel). The correlation between random effects in the model for antibiotic use expressed in PID was estimated to be $-0.2455$ (between random intercept and random slope), 0.9325 (between random intercept and random amplitude) and 0.2329 (between random slope and random amplitude) (Figure 6.1 right panel).

The high correlation between random intercept and random amplitude, for both DID and PID, indicates that countries with a high antibiotic intake at baseline (in 2000) tend to have a stronger absolute seasonal effect.

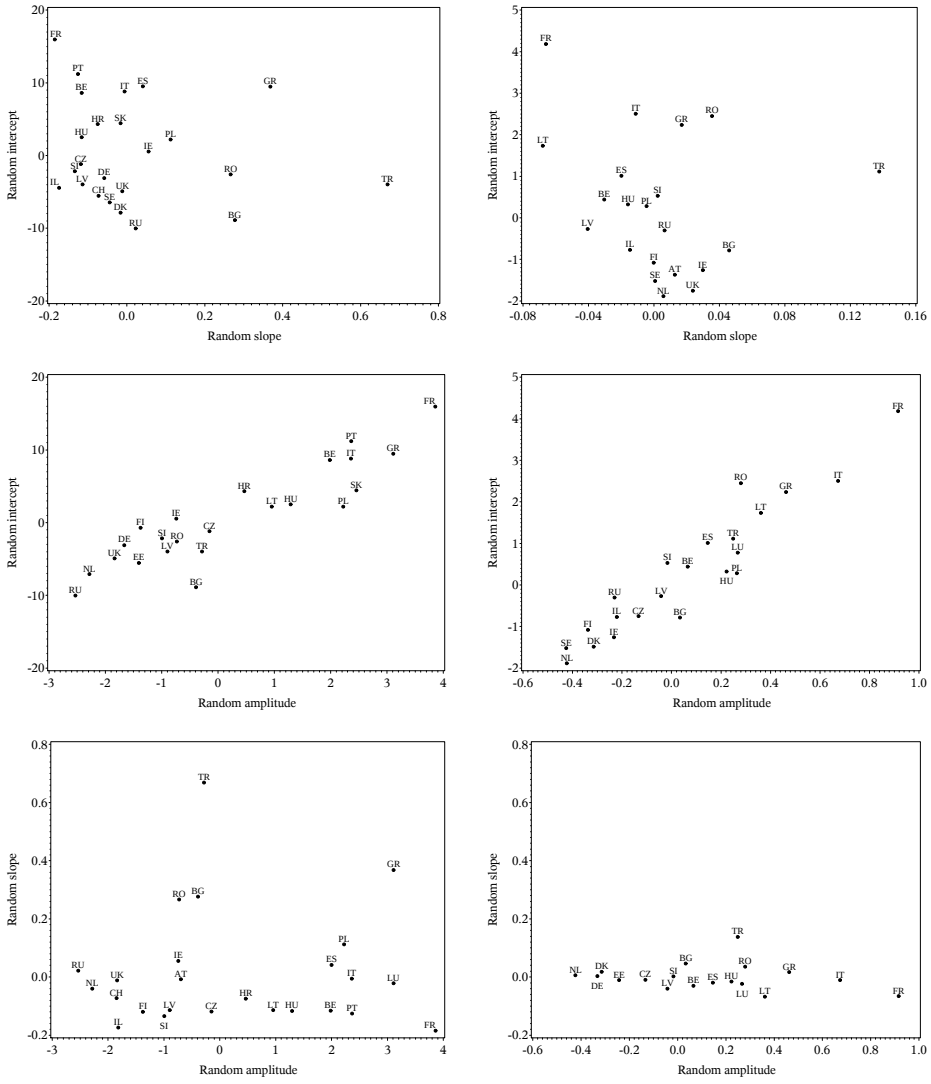



Figure 6.1: Correlation between random intercepts and random slopes (top), between random intercepts and random amplitudes (middle) and between random slopes and random amplitudes (bottom) obtained from the final non-linear mixed model for antibiotic use expressed in DID (left) and PID (right).

Figure 6.2 shows that the models for both DID and PID fit the data reasonably well since observed and predicted outcomes are close together.

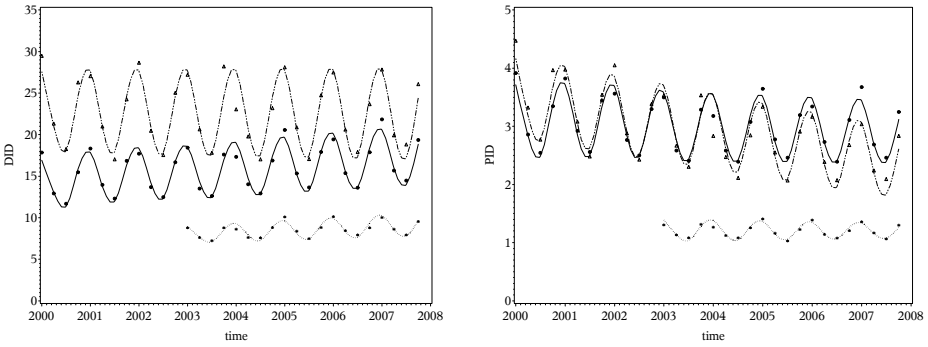



Figure 6.2: Predicted average (full line) and country-specific (Belgium (dashed line) and Netherlands (dotted line)) and observed (filled circles, triangles and stars, respectively) outpatient antibiotic use expressed in DID (left) and PID (right).

Plots of the residuals for both models show no clear structure so we assume that the random effects and the sine wave explain most of the variation in the data (Figure 6.3).

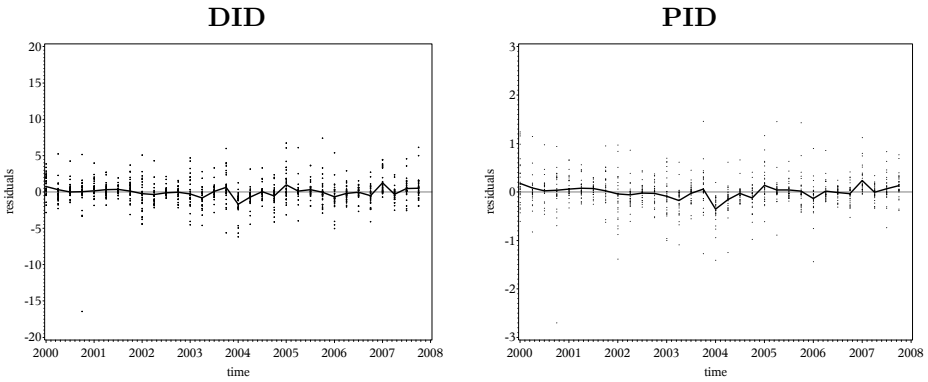



Figure 6.3: Scatter plot of residuals (dots) and smoothed average trend (solid line) from fitting the final non-linear mixed model for outpatient antibiotic use expressed in DID (left) and PID (right).

Parameter estimates for the final models for the other antibiotic subgroups are given in Table 6.2. The seasonal fluctuation with its upward winter peak is significant for all but one subgroup (J01X) whether expressing the consumption in DID or PID. Both in DID and PID the use and seasonal fluctuation increased significantly over time for J01CR and J01M, and decreased significantly for J01A and J01E. The use, but not the seasonal fluctuation, increased significantly over time for J01X in DID and PID. The use and seasonal fluctuation increased significantly over time in DID but not in PID for J01C and J01F while use and seasonal fluctuation decreased significantly over time in PID but not in DID for J01CA and J01BGR. For J01D, no significant changes over time are observed in DID or PID. These results verify that conclusions based on DID and PID can be contradictory.

Table 6.2: Parameter estimates for the fixed effects in the non-linear mixed models for outpatient antibiotic use in Europe expressed in DID and PID; $^*$ : $P < 0.05$,$^{**}$ : $P < 0.0001$.

| | DID | | | | PID | | | |
|---|---|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| J01C | $5.9694^{**}$ | $0.0649^{**}$ | $1.3531^{**}$ | $0.0116^{*}$ | $1.3529^{**}$ | $-0.0037$ | $0.2866^{**}$ | $-0.0016$ |
| J01CA | $3.6967^{**}$ | 0.0113 | $0.9633^{**}$ | - | $0.6533^{**}$ | $-0.0040^{*}$ | $0.1645^{**}$ | $-0.0015^{*}$ |
| J01CR | $2.0284^{**}$ | $0.0506^{**}$ | $0.4515^{**}$ | $0.0098^{**}$ | $0.3486^{**}$ | $0.0037^{*}$ | $0.0797^{**}$ | $0.0007^{*}$ |
| J01F | $2.1834^{**}$ | $0.0270^{*}$ | $0.5906^{**}$ | $0.0114^{**}$ | $0.4543^{**}$ | 0.0006 | $0.1334^{**}$ | - |
| J01M | $1.1847^{**}$ | $0.0178^{*}$ | $0.0858^{*}$ | $0.0021^{*}$ | $0.2012^{**}$ | $0.0026^{*}$ | $0.0153^{*}$ | $0.0003^{*}$ |
| J01D | $1.8794^{**}$ | 0.0125 | $0.5247^{**}$ | 0.0007 | $0.5268^{**}$ | 0.000014 | $0.1446^{**}$ | $-0.0009$ |
| J01A | $1.8911^{**}$ | $-0.0106^{*}$ | $0.3660^{**}$ | $-0.0047^{*}$ | $0.2371^{**}$ | $-0.0026^{*}$ | $0.0442^{**}$ | $-0.0008^{**}$ |
| J01E | $0.9900^{**}$ | $-0.0089^{*}$ | $0.1560^{**}$ | $-0.0024^{*}$ | $0.1876^{**}$ | $-0.0021^{**}$ | $0.0331^{**}$ | $-0.0006^{**}$ |
| J01X | 0.0357 | $0.0018^{*}$ | $-0.0029$ | - | $0.0187^{*}$ | $0.00056^{*}$ | $-0.00054$ | - |
| J01BGR | $0.1179^{*}$ | $-0.0017$ | $0.0072^{*}$ | - | $0.1302^{*}$ | $-0.0018^{*}$ | $0.0179^{*}$ | $-0.0005^{*}$ |

### 6.4.2 Analysis of DID and PID jointly

A joint model was constructed based on the final models for antibiotic use expressed in DID and PID. The correlation between matching random effects for DID and PID was estimated to be 0.7743 (between random intercepts), 0.8647 (between random slopes) and 0.9571 (between random amplitudes) (Figure 6.4). All correlations were high and positive, indicating that there is an agreement between the random effects.

This means that when a random effect is above average in DID, it will also be above average in PID, and vice versa. The correlations between random intercepts and slopes were imperfect while the correlation between the random amplitudes seemed close to perfect.

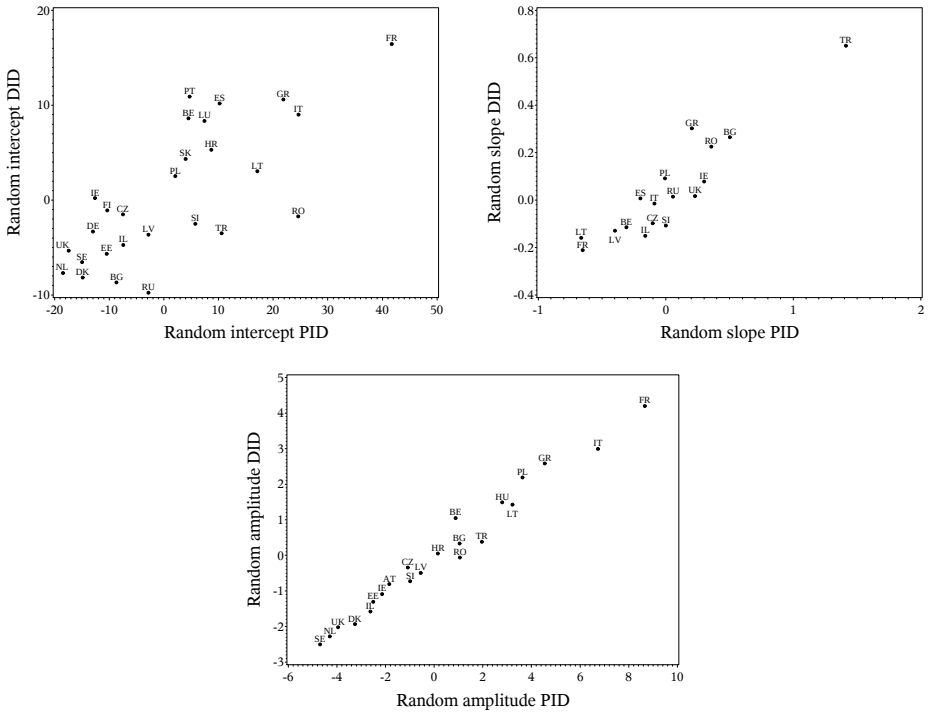



Figure 6.4: Correlation between matching intercepts (top left), slopes (top right) and amplitudes (bottom).

For the other antibiotic groups, the correlation between random intercepts for DID and PID and between random slopes for DID and PID were both positive but imperfect. The correlation between the amplitudes for DID and PID was positive for all but one subgroup (i.e. J01BGR) and seemed close to perfect for J01CR, J01F and J01BGR.

### 6.4.3 Change in dose per package

Likelihood ratio tests indicated that none of the random effects could be removed from the starting model for all but one antibiotic group (J01X). In the model for J01X, likelihood ratio tests indicated that both random effects were redundant. The need for random effects in all but one subgroup suggests that the average dose per package in 2000 and the change in dose per package over time differed substantially between the countries. Parameter estimates for all final models are given in Table 6.3.

Table 6.3: Parameter estimates for the fixed effects in the linear mixed models for the dose per package; $^{*}: P < 0.05$,$^{**}: P < 0.0001$.

| | $\beta_0$ | $\beta_1$ |
|---|---|---|
| J01 | 4.9149** | 0.0436** |
| J01C | 4.8265** | 0.0617** |
| J01CA | 6.6670** | 0.0573** |
| J01CR | 5.8891** | 0.0693** |
| J01F | 4.9924** | 0.0390** |
| J01M | 6.1181** | 0.0105 |
| J01D | 3.9906** | 0.0428* |
| J01A | 9.2237** | 0.0433* |
| J01E | 5.2627** | 0.0083* |
| J01X | 1.0179* | 0.0086* |
| J01BGR | 1.1447* | 0.0778* |

The average number of DDD per package in 2000 ranged between 1 and 9. Over time, the number of DDD per package increased with the size of this increase varying between 0.01 and 0.08 DDD which translates to a yearly increase between 0.04 and 0.32 DDD per package. This increase was significant for all but one antibiotic group studied (J01M).

## 6.5 Discussion

In this chapter, we used a non-linear mixed model to complement analyses of the trend in antibiotic use for 31 countries expressed in DID, by analyses of data expressed in PID. We showed that antibiotic use and its seasonal fluctuation increased significantly over time in DID while they did not change significantly in PID and hence demonstrated that conclusions based on both measures can be contradictory. For all antibiotic groups studied, the correlation between the random intercept and the random amplitude was high, indicating that countries with a high antibiotic use at baseline tend to experience strong absolute seasonal fluctuations.

A strong correlation between measurements in DID and the number of prescriptions at one time point has previously been described by Monnet *et al.* (2004). This finding was corroborated here, as strong correlations between the random intercepts for DID and PID were found for all antibiotic groups studied. The correlations did however not seem perfect, implying that when the average antibiotic consumption expressed in DID and PID is known, country-specific information on either measure is not sufficient to obtain information on the other measure. The correlations between random slopes in DID and PID and between random amplitudes in DID and PID were also strong, with only the latter seeming to be close to perfect for J01, J01CR, J01F and J01BGR.

The number of DDD per package increased significantly over time for all antibiotic groups studied, except for J01M for which the dose per package did not change significantly over time. Both the number of DDD per package in 2000 and the increase of this number over time differed substantially between countries and between studied antibiotic groups.

### 6.5.1 Non-linear mixed models under REML or ML

Estimates for mixed models are usually ML-based, where the likelihood is maximized jointly for fixed effects and variance components. The ML estimators can however be biased downwards and for this reason REML estimates are often preferred. Rather than maximizing the joint likelihood, REML maximizes a set of error contrasts that are independent of the fixed effects.
Likelihood ratio tests can be based on ML or REML when checking if a random effect can be left out of the model. When checking the need for a fixed effect, the likelihood ratio test should however always be ML-based. Here REML is no longer valid as

a different mean structure goes together with different error contrasts (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2009).
To determine whether it was required to use REML rather than ML in reducing the covariance structure, a likelihood ratio test for the removal of random amplitudes ($b_{2i}$) was conducted under both REML and ML. Parameter estimates and standard errors for fixed effects were the same for both methods. Variance components were slightly different with ML estimates being consistently smaller than REML estimates hence confirming the downward bias. However, we concluded that the bias was so small that it was not necessary to use REML rather than ML. The likelihood ratio tests under REML and ML resulted in the same conclusions and the test statistics were very similar. For the reasons mentioned above and for our convenience, as SAS proc NLMIXED does not contain the option to specify REML, all models in this chapter were fitted under ML.

# Chapter 7

# Exploring the association between resistance and antibiotic use expressed in defined daily doses or packages

In the previous chapter, we have shown that expressing outpatient antibiotic use in DID or PID could lead to contrasting conclusions. Therefore, we recommended to consider both when monitoring outpatient antibiotic use over time. Although the number of DDD is internationally used to quantify antibiotic use, the number of packages seems to be a better proxy for the number of antibiotic treatments given and the number of individual patients treated (Coenen *et al.*, 2014). To date, it is not clear which measurement unit correlates best with antibiotic resistance. DID is used most often, but a study that assessed the association between proportions of ENSP and the use of tetracycline, macrolide, lincosamide and streptogramin (TMLS) in Belgium found that expressing use in PID and including a time lag between use and resistance provided the best-fitting model (Van Heirstraeten *et al.*, 2012).

In this chapter, we will use the Yearly Antimicrobial Resistance Data (Section 1.2.5) to explore the association between European outpatient antibiotic use, expressed in

either DID or PID, and non-susceptibility against penicillin or erythromycin. We will assess whether both DID and PID should be accounted for when modelling antibiotic resistance and whether including a time lag would improve model fit.

## 7.1 Analysis of the association between antibiotic use and resistance

Because country-specific information on resistance is gathered annually, within-country correlation has to be accounted for and hence mixed models are an adequate tool to study the trends in the data. The mixed model used here consists of a fixed component, which represents the average trend in Europe based on the included countries, and a random component, which represents the country-specific deviation from this average trend (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2009). Because resistance status is a binomial response, a generalized linear mixed model employing a logit link was used (Molenberghs and Verbeke, 2005). This model can be presented as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2 X_{ij}^{DID} + \beta_3 X_{ij}^{PID},$$

where $\pi_{ij}$ represents the proportion of PNSP or ENSP in country $i$ at time-point $t_{ij} (j = 1, ..., n_i)$, $n_i$ is the number of observations from the $i$th country, $t_{ij} = 1$ corresponds to the start of the study (year 2000), $\beta_0$ is the global intercept, $\beta_1$ is the global slope, $\beta_2$ and $\beta_3$ are the effects of DID and PID in country $i$ at time-point $t_{ij}$ and $\boldsymbol{b_i} = (b_{0i}, b_{1i})$ is a vector of country-specific random effects (for intercept and slope) for which we assume $\boldsymbol{b_i} \backsim N(0, \boldsymbol{D})$. The matrix $\boldsymbol{D}$ is an unstructured matrix with $d_{11}$ the variance of the random intercept $b_{0i}$, $d_{22}$ the variance of the random slope $b_{1i}$ and $d_{12}$ the covariance between the random intercept and the random slope.

In a first step, we selected the most appropriate time lag (time lag $= 0, 1$ or 2) by comparing goodness-of-fit statistics for the model fitted to the common part of the three datasets. Statistics that were used in this selection procedure included the Akaike Information Criterion (AIC), the Pearson Chi-square statistic and the pseudo-$R^2$ statistic which is defined as:

$$1 - \frac{L(\hat{\theta})}{L(0)},$$

with $L(\hat{\theta})$ the likelihood for the model containing all parameters and $L(0)$ the likelihood for the model containing only a general intercept.

A good model fit is reflected by a low AIC value, a high pseudo-$R^2$ value and a low Chi-square value.

In a second step, we used the same goodness-of-fit statistics to determine whether to include both DID and PID, only DID or only PID. Because these statistics were not always in agreement, we continued by using backwards model selection to obtain a final model. The need for random intercepts and slopes was assessed with likelihood ratio tests based on a mixture of equally weighted $\chi^2_1$ and $\chi^2_2$ distributions. A likelihood ratio test based on a $\chi^2_1$ distribution was used to assess whether the covariance structure could be simplified from an unstructured to a simple structure and to test the need for inclusion of the fixed effects.

Using the full model, we predicted the proportion of PNSP and ENSP if $\beta$-lactam or TMLS use expressed in DID, PID or both were to be lower ($0 - 70\%$ in steps of 5%) than the value reported for the last observed year (2007 for time lag = 0, 2008 for time lag = 1 and 2009 for time lag = 2).

## 7.2 Results

### 7.2.1 $\beta$-Lactam use and PNSP

When exploring the association between PNSP and $\beta$-lactam use, the best model fit was found for a model without time lag (Table 7.1).

Table 7.1: Goodness-of-fit statistics for the model for PNSP including a time lag.

| Time lag | AIC | Pseudo-$R^2$ | Chi-square |
|---|---|---|---|
| 0 | 889.99 | 0.79 | 167.46 |
| 1 | 897.38 | 0.79 | 171.65 |
| 2 | 896.80 | 0.79 | 177.63 |

While AIC suggested using PID alone, Pearson Chi-square suggested using DID alone (Table 7.2). Likelihood ratio tests indicated that none of the random effects could be removed while a simplification of the covariance matrix was allowed. DID was removed from the model while PID had to be retained.

Table 7.2: Goodness-of-fit statistics for the model for PNSP including both DID and PID, only DID and only PID.

| | AIC | Pseudo-$R^2$ | Chi-square |
|---|---|---|---|
| DID & PID | 1146.55 | 0.82 | 207.37 |
| DID | 1157.03 | 0.82 | 201.67 |
| PID | 1144.64 | 0.82 | 207.28 |

From the final model containing only PID, it can be concluded that the odds of PNSP increased significantly with increasing $\beta$-lactam use expressed in PID while it did not change significantly over time (Table 7.3).

Table 7.3: Parameter estimates for the final model for PNSP.

| | Estimate | Std.error | Odds ratio (95% CI) |
|---|---|---|---|
| Time | −0.003 | 0.013 | 0.997 (0.970, 1.025) |
| PID | 0.673 | 0.111 | 1.959 (1.574, 2.439) |

### 7.2.2 TMLS use and ENSP

When exploring the association between ENSP and TMLS use, the best model fit was found for a model with a 1 year time lag (Table 7.4).

Table 7.4: Goodness-of-fit statistics for the model for ENSP including a time lag.

| Time lag | AIC | Pseudo-$R^2$ | Chi-square |
|---|---|---|---|
| 0 | 893.70 | 0.80 | 95.08 |
| 1 | 889.28 | 0.80 | 94.66 |
| 2 | 890.94 | 0.80 | 95.31 |

Both AIC and Pearson Chi-square suggested using both DID and PID (Table 7.5). Likelihood ratio tests indicated that none of the random effects could be removed. A simplification of the covariance matrix was not allowed. Neither DID nor PID could be removed from the model.

Table 7.5: Goodness-of-fit statistics for the model for ENSP including both DID and PID, only DID and only PID.

| | AIC | Pseudo-$R^2$ | Chi-square |
|---|---|---|---|
| DID & PID | 1216.22 | 0.80 | 170.35 |
| DID | 1220.09 | 0.80 | 176.89 |
| PID | 1221.84 | 0.80 | 179.07 |

From the final model, it can be concluded that the odds of ENSP increased significantly with increasing TMLS use expressed in PID and with decreasing TMLS use in DID, while it did not change significantly over time (Table 7.6).

Table 7.6: Parameter estimates for the final model for ENSP.

| | Estimate | Std.error | Odds ratio (95%CI) |
|---|---|---|---|
| Time | 0.005 | 0.016 | 1.005 (0.973, 1.039) |
| DID | −0.250 | 0.091 | 0.779 (0.650, 0.933) |
| PID | 1.304 | 0.540 | 3.683 (1.265, 10.719) |

### 7.2.3 Predictions of antibiotic resistance

The average predicted proportion of PNSP isolates associated with $\beta$-lactam use was based on all countries that had resistance data in 2007 (all but Slovakia; Figure 7.1, top). The average predicted proportion of ENSP isolates associated with TMLS use was based on all countries that had resistance data in 2008 (all but Slovakia and Switzerland; Figure 7.1, bottom). Figure 7.1 illustrates that PNSP proportions are predicted to decrease substantially with a decrease in $\beta$-lactam use expressed in PID alone or in both DID and PID, but are predicted to remain stable when $\beta$-lactam use expressed in DID decreases. ENSP proportions are predicted to decrease if TMLS use expressed in PID decreases, but are predicted to increase with a decrease in TMLS use expressed in DID alone or in both DID and PID.

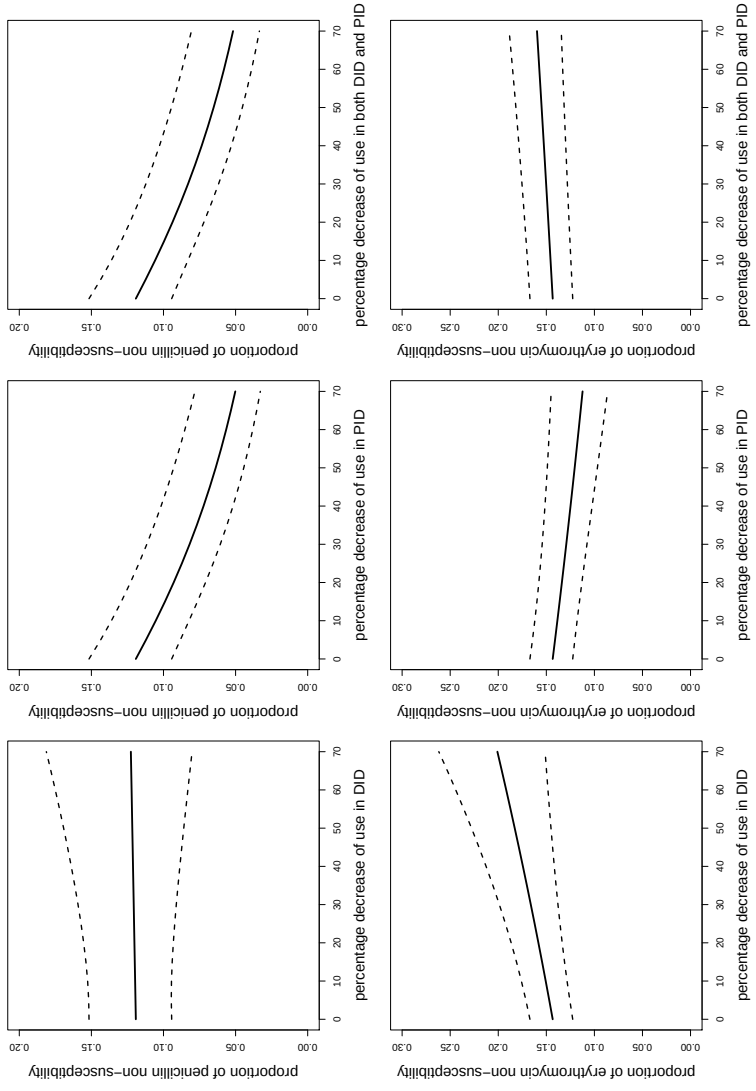



Figure 7.1: Average predicted proportion of non-susceptible *Streptococcus pneumoniae* isolates if outpatient antibiotic use had been lower than reported in 2007 (for $\beta$-lactam; top) or 2008 (for TMLS; bottom) due to a decrease of use in DID (left), PID (middle) or both (right).

## 7.3 Discussion

Exploring the association between outpatient antibiotic use and resistance in Europe revealed that including a time lag might improve model fit. While the association between $\beta$-lactam use and PNSP was modelled best without a time lag, the association between TMLS use and ENSP benefited from including a one year time lag between antibiotic use and resistance. This difference corresponds to the differences in persistence of resistance, which is much longer after exposure to TMLS (e.g. clarithromycin or azithromycin) than after exposure to $\beta$-lactams (e.g. ampicillin) (Chung *et al.*, 2007; Malhotra-Kumar *et al.*, 2007).
The results also demonstrated that use data expressed in DID alone might not provide the best-fitting models. To assess ENSP proportions, TMLS use expressed in both DID and PID was needed, while for PNSP proportions, $\beta$-lactam use expressed in PID was sufficient. Also predictions of resistance after a decrease in use of TMLS were driven by both DID and PID while predictions after a decrease in $\beta$-lactam use were driven by PID alone. Due to these findings, we recommend to adopt the number of packages as an additional outcome to better understand outpatient antibiotic use and its relation to resistance.

The predictions discussed in Section 7.2.3 illustrate that PNSP decreases when decreasing $\beta$-lactam use in PID (alone or together with a decrease in DID), while it is not altered when decreasing $\beta$-lactam use in DID alone. A possible explanation lies in the difference in dosing schedules with $\beta$-lactams generally being dosed higher (both in PID and in DID) than TMLS (Table 6.1). If $\beta$-lactam use decreases both in PID and DID, this will most likely reflect a decrease in the number of patients exposed to a consistent (and appropriate) antibiotic dose. However, if $\beta$-lactam use decreases only in PID (with use expressed in DID remaining constant), this would suggest that fewer patients are exposed to antibiotics, but that these patients are receiving a higher dose per treatment due to an increase in DDD per package. In contrast, if $\beta$-lactam use decreases only in DID (with use expressed in PID remaining constant), this is likely to be due to a decrease in dose per treatment (decrease in DDD per package) while this lower dose may remain appropriate to prevent emergence of resistance.
ENSP decreases only when decreasing TMLS use in PID alone, while it increases when decreasing TMLS use in DID (alone or together with a decrease in PID). A possible explanation could be that if TMLS use decreases only in PID (with use expressed in DID remaining constant), fewer patients are exposed to a higher dose of TMLS. However, if TMLS use decreases both in PID and DID, this would suggest

that fewer patients are exposed to a consistent (and inappropriate) antibiotic dose. When TMLS use decreases in DID alone (with use expressed in PID remaining constant), this would suggest that the same number of patients are exposed to a lower (and even more inappropriate) dose.

### 7.3.1 Disagreement between goodness-of-fit statistics

When assessing the need to include $\beta$-lactam use expressed in both DID and PID in the model for PNSP, there was a disagreement in the goodness-of-fit statistics. While the Pearson Chi-square statistic suggested using DID alone, the AIC value suggested using PID alone. This disagreement can be explained by the nature of the goodness-of-fit statistics used in this chapter. The Pearson Chi-square statistic is constructed by summing all Pearson residuals and hence assesses the fit of the conditional distribution. The AIC value on the other hand assesses the overall fit of the model. This implies that in general, including $\beta$-lactam use expressed in PID should be used to study penicillin resistance. However, when the distribution of random effects (intercept and slope) around the average is known, $\beta$-lactam use expressed in DID could be used to study penicillin resistance.

# Chapter 8

# Persistence of antimicrobial resistance in respiratory streptococci

In the previous chapter, we illustrated that resistance is associated with antibiotic consumption. Malhotra-Kumar *et al.* (2007, 2016) and Chung *et al.* (2007) showed that persistence of resistance can differ based on the antimicrobial that was used. They have illustrated that persistence of resistance in oropharyngeal streptococci after exposure to macrolides lasts for more than six months (Malhotra-Kumar *et al.*, 2007), while it is estimated to be much shorter after exposure to penicillins (Chung *et al.*, 2007; Malhotra-Kumar *et al.*, 2016). In this chapter, we will assess the difference in persistence of resistance after exposure to penicillins or cephalosporins (CD) and macrolides or tetracyclins (AF) using the Bacterial Susceptibility Data (Section 1.2.6).

Getting a thorough understanding of persistence of resistance for different combinations of bacteria and antibiotics would require a huge number of RCTs. Because this is rather expensive and time consuming, we will assess whether routinely collected data on resistance and antibiotic use at the level of the individual patient confirm the conclusions reached in the studies conducted by Malhotra-Kumar *et al.* (2007, 2016) and Chung *et al.* (2007) and hence could serve as a proxy to study other drug-bug combinations.

## 8.1 Persistence of resistance

To assess persistence of resistance (i.e. non-susceptible resistance status), we constructed a multiple logistic regression model using resistance status as the binary outcome and a logit link function. We used an automatic forward selection procedure (with $\alpha = 0.15$) to reach a model including all significant explanatory variables and two-way interactions.

Because multiple samples were taken from some patients, information obtained from the same patient was expected to be correlated. Ignoring correlation would typically result in underestimation of the standard errors and hence wrongfully retaining variables in the model (Agresti, 2002; Molenberghs and Verbeke, 2005). Therefore, we constructed a GEE model (Liang and Zeger, 1986) including explanatory variables that were present in the final logistic regression model and an independence working correlation. Note that although this working correlation might be incorrect, parameter estimates and empirical standard errors are deemed consistent due to the use of a sandwich estimator in the GEE approach (Molenberghs and Verbeke, 2005).

The data used in this study contain information on 14 samples taken from patients that did not survive 2005. Because the studies conducted by Malhotra-Kumar *et al.* (2007) and Chung *et al.* (2007) reported no deaths, we repeated the analysis described above after exclusion of these samples to optimize comparability.

## 8.2 Baseline resistance

Both the logistic regression model and the GEE model implicitly assume that the proportion of non-susceptible isolates in the population falls back to zero when the timespan between antibiotic consumption and sampling becomes large enough. Several authors however found a non-zero proportion of non-susceptible isolates at baseline (i.e. baseline resistance (BR)) (Malhotra-Kumar *et al.*, 2007; Putnam *et al.*, 2005; Raum *et al.*, 2008; Shackcloth *et al.*, 2004). Therefore, we relaxed this assumption and adjusted the link function accordingly:

$$log\left(\frac{p}{1-p}\right) \quad \rightarrow \quad log\left(\frac{p}{g-p}\right) \qquad with \qquad g = 1 - BR,$$

where $p$ represents the proportion of susceptible individuals. We allowed this baseline resistance to differ by treatment (CD or AF) and by type of bacteria (PY or PN) to avoid wrongfully concluding significant differences between treatments and bacteria types caused by the difference in BR.

## 8.3 Results

### 8.3.1 Baseline resistance rates

Because estimates for BR in this specific setting were not available, we calculated BR as the proportion of non-susceptible samples within a sliding 6 month time frame between $t - 186$ and $t$ (with $t = 186, 187, \ldots, 372$) and studied the evolution of this estimate for BR over time (Figure 8.1). Because the BR estimates did not stabilize, we conducted a sensitivity analysis in which we fitted the adjusted GEE model using different BR estimates (for $t = 248$ (BR 3-8 months), for $t = 310$ (BR 5-10 months) and for $t = 372$ (BR 7-12 months)) (Table 8.1). After dropping samples from patients that did not survive 2005, BR estimates were recalculated (Table 8.2).

Table 8.1: Estimates (95% confidence intervals) of baseline resistance (BR) based on all samples ($n = 451$).

| | Frame 3-8 months | Frame 5-10 months | Frame 7-12 months |
|---|---|---|---|
| | BR8 | BR10 | BR12 |
| PY AF | 0.2414 (0.1222, 0.4211) | 0.0625 (0.0032, 0.2833) | 0.0909 (0.0047, 0.3774) |
| PY CD | 0.0400 (0.0137, 0.1111) | 0.0000 (0.0000, 0.0664) | 0.0000 (0.0000, 0.0989) |
| PN AF | 0.2727 (0.1315, 0.4815) | 0.2500 (0.1119, 0.4687) | 0.2632 (0.1181, 0.4879) |
| PN CD | 0.2083 (0.1405, 0.3157) | 0.1081 (0.0429, 0.2471) | 0.1304 (0.0454, 0.3213) |

AF: treatment with macrolides or tetracyclines
CD: treatment with penicillins or cephalosporins.
PY: *Streptococcus pyogenes*; PN: *Streptococcus pneumoniae*

Table 8.2: Estimates (95% confidence intervals) of baseline resistance (BR) based on samples from patients surviving 2005 ($n = 437$).

| | Frame 3-8 months | Frame 5-10 months | Frame 7-12 months |
|---|---|---|---|
| | BR8 | BR10 | BR12 |
| PY AF | 0.2414 (0.1222, 0.4211) | 0.0625 (0.0032, 0.2833) | 0.0909 (0.0047, 0.3774) |
| PY CD | 0.0400 (0.0137, 0.1111) | 0.0000 (0.0000, 0.0664) | 0.0000 (0.0000, 0.989) |
| PN AF | 0.2727 (0.1315, 0.4815) | 0.2222 (0.0900, 0.4521) | 0.1875 (0.0659, 0.4301) |
| PN CD | 0.1618 (0.0928, 0.2669) | 0.1081 (0.0429, 0.2471) | 0.1364 (0.0475, 0.3333) |

AF: treatment with macrolides or tetracyclines
CD: treatment with penicillins or cephalosporins.
PY: *Streptococcus pyogenes*; PN: *Streptococcus pneumoniae*

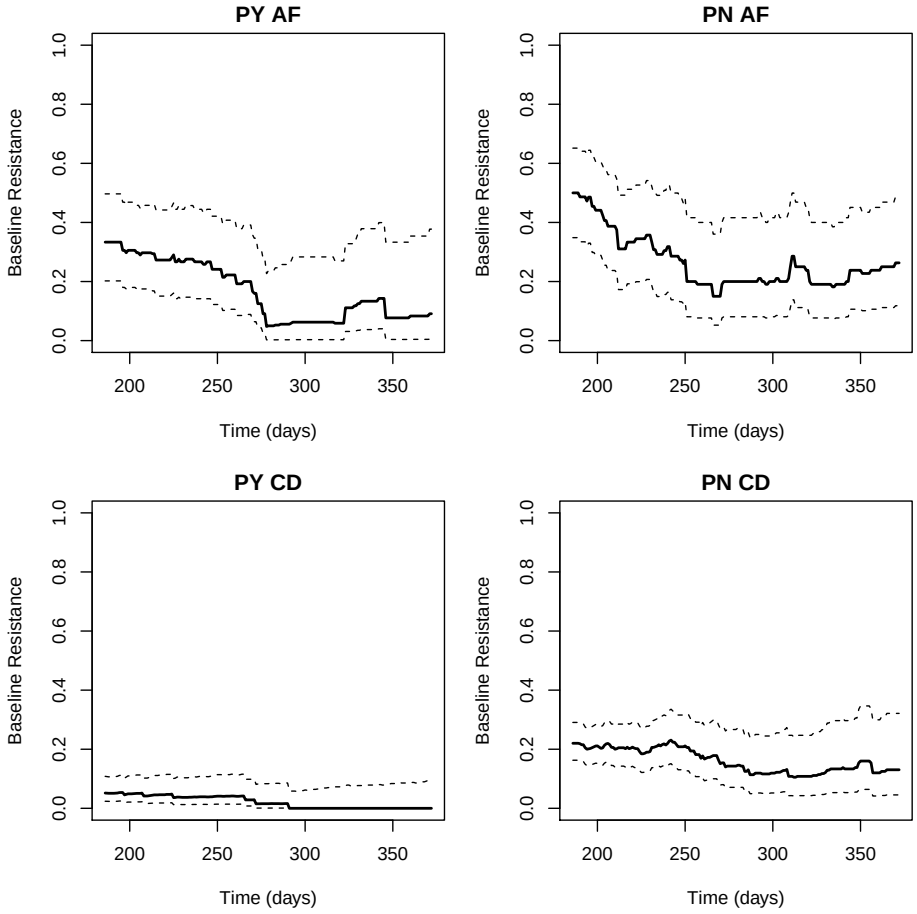



Figure 8.1: Evolution of the proportion of non-susceptible samples within a sliding 6 month time frame between $t - 186$ and $t$ (with $t = 186, 187, \ldots, 372$) over time based on all samples (bold line: estimate, dashed lines: 95% confidence interval) for isolates of *Streptococcus pyogenes* (PY; left) and *Streptococcus pneumoniae* (PN; right) after treatment with macrolides or tetracyclines (AF; top) and penicillins or cephalosporins (CD; bottom).

### 8.3.2 Difference in persistence of resistance by treatment

The final logistic regression model contained treatment (AF or CD), bacteria (PY or PN), survival status (did the patient die in 2005 or not), log(time) with time the distance between consumption and sampling, the interaction between survival status and log(time) and the interaction between treatment and bacteria type. These variables were used in fitting a GEE model, unadjusted for BR. Based on this model, we can conclude that the odds of being susceptible was significantly lower for treatment with AF and significantly higher for PY isolates (Table 8.3). After correcting for baseline resistance, parameter estimates changed and the odds of being susceptible no longer differed significantly by bacteria type (for any of the three BR estimates used). Parameter estimates differed when using different BR estimates, which stresses the need for the sensitivity analysis.

Table 8.3: Parameter estimates for the GEE models on persistence of overall resistance using different estimates for baseline resistance (BR) obtained by forward model building.

| Parameter | Parameter estimate | | | |
|---|---|---|---|---|
| | no BR | BR8 | BR10 | BR12 |
| | | 3-8 months | 5-10 months | 7-12 months |
| Intercept | 1.1015 | 1.0448 | 1.1622 | 1.1460 |
| Treat AF | -1.3204** | -2.3710 | -1.3231** | -1.4705** |
| Bacteria PY | 1.6892** | 2.9179 | 0.8955 | 0.6923 |
| Death no | -1.9274 | -3.5075 | -2.1480 | -2.0767 |
| Log(time) | -0.2928 | -0.0649 | -0.2409 | -0.2142 |
| Log(time) * death no | 0.8777 | 1.7286 | 1.0982 | 1.1159 |
| Treat AF * bacteria PY | -1.1129 | -1.9666 | -1.0507 | -0.7902 |

AF: treatment with macrolides or tetracyclines; PY: *Streptococcus pyogenes*.
*: p-value $< 0.05$; **: p-value $< 0.01$

Further backward model reduction ($\alpha = 0.05$) resulted in a model including treatment, log(time) and survival status (Table 8.4). Based on the final models, we can conclude that the odds of being susceptible was significantly lower after treatment with AF while it was significantly higher when surviving 2005 and with increasing time since antibiotic consumption. The evolution of resistance over time however did not differ by treatment. Including an interaction between log(time) and treatment in the final models resulted in p-values of 0.1147 (BR8), 0.3394 (BR10) and 0.3104 (BR 12). These findings are illustrated in Figure 8.2.

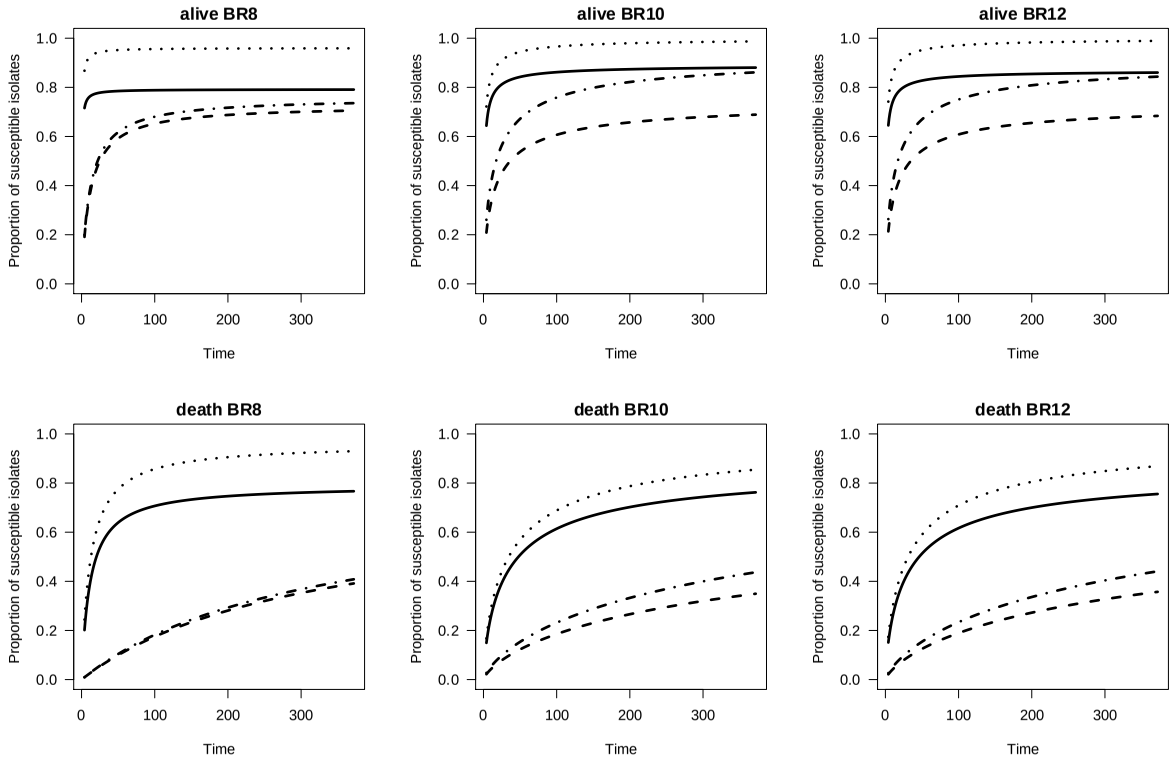



Figure 8.2: Evolution of the predicted proportion of susceptible isolates over time for patients that survived (alive; top) or did not survive 2005 (death; bottom) based on the final adjusted GEE model on persistence of overall resistance using different estimates for baseline resistance (left: BR8, middle: BR10, right: BR12).
solid line: *Streptococcus pneumoniae* isolates after treatment with penicillins or cephalosporins.
dashed line: *Streptococcus pneumoniae* isolates after treatment with macrolides or tetracyclines.
dotted line: *Streptococcus pyogenes* isolates after treatment with penicillins or cephalosporins.
dot-dashed line: *Streptococcus pyogenes* isolates after treatment with macrolides or tetracyclines.

Table 8.4: Parameter estimates for the GEE models on persistence of overall resistance using different estimates for baseline resistance (BR) after final backward model reduction.

| | Parameter estimate | | |
|---|---|---|---|
| Parameter | BR8 | BR10 | BR12 |
| Intercept | -2.4517 | -2.6359 | -2.6159 |
| Treat AF | -3.2705** | -1.9071** | -1.9512** |
| Death no | 3.3138* | 2.5573* | 2.6152* |
| Log(time) | 0.9926* | 0.7445** | 0.7613** |

AF: treatment with macrolides or tetracyclines.
*: p-value $< 0.05$; **: p-value $< 0.01$

Figure 8.2 shows that there was a difference between patients that did and those that did not survive 2005, which was found to be significant (Table 8.4). Because of this significant difference and because the studies conducted by Malhotra-Kumar *et al.* (2007) and Chung *et al.* (2007) reported no deaths, we repeated the analysis on samples from patients that did survive 2005 to optimize comparability.

Based on the final models for patients surviving 2005, we can conclude that the odds of being susceptible was significantly lower for treatment with AF while it was significantly higher with increasing time since antibiotic consumption (Table 8.5). The evolution of resistance over time however did not differ by treatment. Including an interaction between log(time) and treatment in the final models resulted in p-values of 0.1035 (BR8), 0.7040 (BR10) and 0.7095 (BR 12). These findings are illustrated in Figure 8.3.

Figure 8.3 also shows that the proportion of susceptible isolates stabilized more quickly after treatment with CD than after treatment with AF. We considered the proportion of susceptible isolates to be stable when it increased with less than 0.05% per day. Table 8.6 demonstrates that resistance after treatment with AF persisted about three times as long as after treatment with CD.

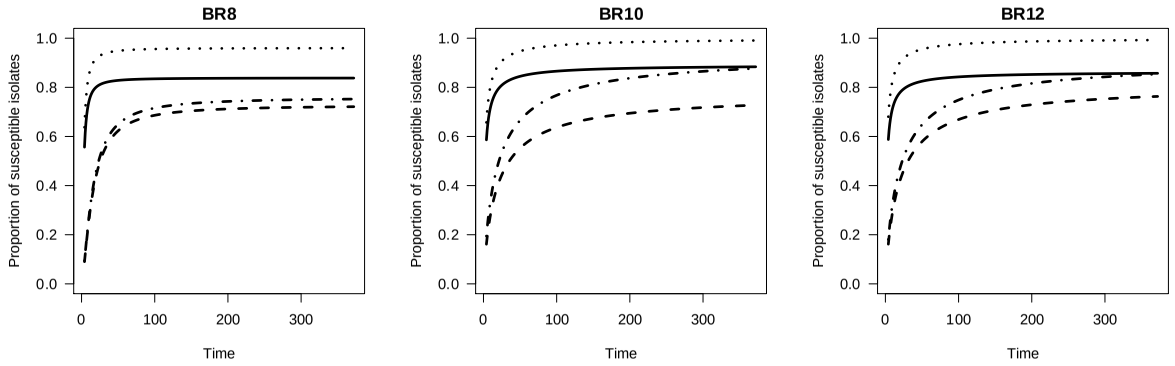



Figure 8.3: Evolution of the predicted proportion of susceptible isolates over time based on the final adjusted GEE model on persistence of resistance for patients surviving 2005 using different estimates for baseline resistance (left: BR8, middle: BR10, right: BR12).

solid line: *Streptococcus pneumoniae* isolates after treatment with penicillins or cephalosporins.

dashed line: *Streptococcus pneumoniae* isolates after treatment with macrolides or tetracyclines.

dotted line: *Streptococcus pyogenes* isolates after treatment with penicillins or cephalosporins.

dot-dashed line: *Streptococcus pyogenes* isolates after treatment with macrolides or tetracyclines.

Table 8.5: Parameter estimates for the GEE models on persistence of resistance for patients surviving 2005 using different estimates for baseline resistance (BR) after final backward model reduction.

| | Parameter estimate | | |
|---|---|---|---|
| Parameter | BR8 | BR10 | BR12 |
| Intercept | -1.3733 | -0.5764 | -0.5080 |
| Treat AF | -2.6361** | -1.9909** | -2.1492** |
| Log(time) | 1.4817* | 0.8850** | 0.9111** |

AF: treatment with macrolides or tetracyclines.
*: p-value $< 0.05$; **: p-value $< 0.01$

Table 8.6: Number of days needed for the proportion of susceptible isolates to stabilize based on the final adjusted GEE model on persistence of resistance for patients surviving 2005 using different estimates for baseline resistance (BR).

| | BR8 | BR10 | BR12 |
|---|---|---|---|
| PY CD | 44 | 71 | 66 |
| PN CD | 44 | 71 | 66 |
| PY AF | 125 | 195 | 194 |
| PN AF | 125 | 195 | 194 |

AF: treatment with macrolides or tetracyclines
CD: treatment with penicillins or cephalosporins.
PY: *Streptococcus pyogenes*; PN: *Streptococcus pneumoniae*

## 8.4 Discussion

In this chapter, we used routinely collected data on antibiotic use and resistance at the level of the individual patient to assess persistence of resistance after exposure to macrolides and tetracylines or penicillins and cephalosporins. The use of routinely collected data is both a strength, as individuals were not exposed to an additional intervention (e.g. new treatment or placebo control), and a limitation of the study, as we had no control over prescribed dose, duration of treatment and treatment adherence. Major advantages of studying routinely collected data are that real field conditions are met (e.g. co-morbidities) and that it is less labour-intensive and expensive than conducting e.g. a RCT. Ethical and insurance concerns are also of another dimension given the retrospective concept of analysing routinely collected data. This analysis

would however benefit from including more samples, which would make BR estimates more reliable and might make a sensitivity analysis redundant.

In the GEE analysis, we acknowledged that the proportion of non-susceptible samples taken from a population not treated with antibiotics does not necessarily equal zero and adjusted the model to account for a baseline resistance level (using BR). We calculated BR as the proportion of non-susceptible isolates within a sliding 6-month time frame and recognized that nominating one time frame would be extremely challenging. We would argue against a BR estimate based on the early time frames, as resistance due to antibiotic use might still persist and distort the estimate. However, we would also argue against a BR estimate based on the late time frames, as confidence intervals got wider at later time points and estimates less reliable. Since BR estimates were unstable and confidence intervals wide, a sensitivity analysis using three different BR estimates (for time frames $3-8$, $5-10$ and $7-12$ months) was conducted.

BR estimates for *Streptococcus pneumoniae* calculated using samples from patients treated with AF and CD were slightly lower than the resistance rates reported by EARS-Net. A possible explanation is that we calculated BR based on a population of individuals that did not take any antibiotics during the last $3-8$, $5-10$ or $7-12$ months (for BR8, BR10 and BR12 respectively) while EARS-Net calculated resistance rates based on the population of individuals that have their resistance against antibiotics tested in practice. In the future, it would be interesting to report both an estimate for resistance rate with and without previous antibiotic treatment to get a more complete picture of the resistance rate in the general population. Baseline resistance rates, together with a measure of the persistence of this resistance, could guide clinicians in prescribing antibiotics at low risk of treatment failure.

The analysis conducted in this chapter showed that the odds of being susceptible increased significantly when time between consumption and sampling increased for the respiratory streptococci under study. We also found that the rate of this increase did not differ significantly between isolates from patients treated with CD or AF. This implies that when the proportion of susceptibility directly after treatment is comparable, persistence of resistance to these antimicrobial agents will likely not differ. A reasonable assumption would be that the proportion of susceptibility directly after treatment equals 0%. Malhotra-Kumar *et al.* (2007) however found 18% susceptible isolates after treatment with macrolides. Therefore, we did not make this assumption

and allowed for different variables to influence the adjusted GEE models intercept. The analysis including all patients showed that the intercept differed significantly by treatment group and survival status. The analysis focussing on patients surviving 2005 revealed that the intercept differed significantly by treatment group.

While Malhotra-Kumar *et al.* (2007) showed that persistence of resistance to macrolides lasts for up to six months, Chung *et al.* (2007); Malhotra-Kumar *et al.* (2016) indicated a much shorter persistence of resistance to penicillins. We showed that the rate of increase in the odds of being susceptible did not differ between the two studied treatment groups while the proportion of susceptible isolates directly after treatment was significantly lower for AF than for CD. Therefore, it would take longer for the proportion of susceptible isolates to recover after treatment with AF than after treatment with CD hence confirming the differences found by Malhotra-Kumar *et al.* (2007, 2016) and Chung *et al.* (2007). The difference in persistence of resistance is illustrated in Figure 8.2 and Table 8.6.

The findings reported in this chapter suggest the equivalence of the use of routinely collected data and carefully designed studies to answer research questions on the persistence of resistance after antibiotic consumption. However, before drawing such a strong conclusion, there is need for additional validation of these findings using other drug-bug combinations.

# Chapter 9

# Using change-points to study the impact of Belgian policies on antimicrobial use

The main goal of the Belgian Antibiotic Policy Coordination Committee (BAPCOC), founded in 1999, is to reduce resistance by improving antimicrobial consumption (Goossens *et al.*, 2008). One of their well-known initiatives consists of providing hospitals with both financial and technical support in hiring a manager for an antimicrobial management team (AMT). This initiative was piloted in 37 voluntary hospitals in 2002, extended to another 24 hospitals in 2006 and to the final 55 hospitals in 2007. Apart from funding, the intervention included technical guidance and advanced specialist training for the formal establishment and follow-up of AMTs.

In this chapter, we will use the Hospital Stays Data (Section 1.2.7) to assess the impact of the introduction of AMTs in hospitals, the first antibiotic awareness campaign organized by BAPCOC in 2001 and the change in financing mechanisms for hospital-drugs for pneumonia in 2006.

## 9.1 Assessment of the impact of national policies on three selected quality indicators in Belgian hospitals

Longitudinal data are generally modelled using a generalized linear mixed model (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2009). A somewhat abrupt change in the evolution of the outcome over time can be modelled using a change-point component (Muggeo, 2003). A change-point can be presented as:

$$\beta_{CP_k}(t_{ij} - CP_k)_+,$$

where $x_+ = max(x, 0)$, $\beta_{CP_k}$ is the $kth$ global difference in the linear trend before and after a change-point and $CP_k$ is the $kth$ change-point.
The significance of such a change-point can be assessed by testing the hypothesis $\beta_{CP_k} = 0$, which implies that there is no difference in the slope before and after the change-point.

Based on expert advice, three change-points were included in the model. The first change-point is 2001 which is the year the first large antimicrobial awareness campaign was launched in Belgium.
This change-point can be modelled as:

$$\beta_{AC}(t_{ij} - 2001)_+,$$

The second change-point is the year the hospital got funding for its antimicrobial management team (AMT) (2002, 2006 or 2007).
This change-point can be modelled as:

$$\begin{aligned}&(\beta_{AMTIN02}(t_{ij} - 2002)_+)\, X_{AMT02}\\ &+ (\beta_{AMTIN06}(t_{ij} - 2006)_+)\, X_{AMT06}\\ &+ (\beta_{AMTIN07}(t_{ij} - 2007)_+)\, X_{AMT07},\end{aligned}$$

with $X_{AMT02} = 1$ if AMT was implemented in 2002 and 0 otherwise, $X_{AMT06} = 1$ if AMT was implemented in 2006 and 0 otherwise, and $X_{AMT07} = 1$ is AMT was implemented in 2007 and 0 otherwise.

The third change-point is 2006 which is the year a new funding mechanism for hospital-drugs for pneumonia was launched. This change-point was included only for pneumonia-related outcomes and can be modelled as:

$$\beta_{FIN}(t_{ij} - 2006)_+.$$

### 9.1.1 Lower limb surgery

Because compliance to guidelines at patient level is a binary response, we used a logit link to model the data. The starting model contained fixed effects for the patient's gender, age, severity of surgery, stay on an intensive care unit and length of stay, change-point variables AMTIN and AC, all two-way interactions between the change-point variables and the fixed effects, time, all two-way interactions between time and the fixed effects, an indicator for AMT and the interaction between the AMT indicator and time. Because there was a lot of between-hospital variability, random intercepts and slopes were included in the model.

As a first step in building the final model, we used likelihood ratio tests to assess whether the change-point (together with all its two-way interactions) was required in the model. If the change-point was redundant, it was removed from the model together with all its two-way interactions. Afterwards, we used a likelihood ratio test based on an equally weighted mixture of $\chi^2$ distributions to verify the need for the random effects. In a final step, the mean structure of the model was reduced in a backwards hierarchical fashion using F tests. In order to assess whether the size of the hospital (number of patients treated for the same APR-DRG) affects compliance to guidelines, we added size and all its two-way interactions with the included fixed effects to the final model and used a likelihood ratio test to compare the model with and without the size component.

As a measure for goodness of fit, the pseudo-$R^2$ for the final model was calculated as follows:

$$1 - \left( \frac{LR(\hat{\theta})}{LR(0)} \right)$$

with $LR(\hat{\theta})$ the likelihood for the model including all parameters and $LR(0)$ the likelihood for the model containing only the general intercept. The contribution of the fixed effects was determined by calculating the pseudo-$R^2$ for the model containing only fixed effects and dividing it by the pseudo-$R^2$ for the final model. The contribution of the random effects was calculated by subtracting the fixed effects contribution from 100%.

### 9.1.2 Pneumonia

Because both outcomes of interest (DDDhosp and OP) were continuous outcomes, we used an identity link to model the data. The starting model contained fixed effects for the median los and the distribution of severity, of gender, of stay on an intensive care unit, of patient origin, of discharge status and of age group, change-point variables AMTIN, FIN and AC, all two-way interactions between the change-point variables and the fixed effects, time, all two-way interactions between time and the fixed effects, an indicator for AMT and the interaction between the AMT indicator and time. Because there was a lot of between-hospital variability, random intercepts and slopes were included in the model. Because hospitals were observed yearly and we assumed that observations closer in time were more alike than observations further apart, we used an autoregressive (AR(1)) residual structure. Because the outcomes were aggregated by hospital and year, all information on the number of patients treated for pneumonia in the hospital was lost. For this reason, we weighted the observations according to hospital size (i.e. number of patients treated for the same APR-DRG).

We used likelihood ratio tests to assess whether the change-point (together with all its two-way interactions) was redundant and hence could be removed from the model. A likelihood ratio test based on a $\chi^2_1$ distribution was used to test if the AR(1) structure could be simplified. Afterwards, we used a likelihood ratio test based on an equally weighted mixture of $\chi^2$ distributions to verify the need for random intercepts and random slopes. In a final step, the mean structure of the model was reduced in a backwards hierarchical fashion using F tests. In order to assess whether the size of the hospital affects DDDhosp or OP, we added size and all its two-way interactions with the included fixed effects to the final model and used a likelihood ratio test to compare the model with and without the size component.

As a measure for goodness of fit, the adjusted $R^2$ was calculated as follows:

$$\left(1 - \frac{\sum_{i=1}^{n} \left\{size_i (y_i - f_i)^2\right\}}{\sum_{i=1}^{n} \left\{size_i (y_i - \bar{y})^2\right\}}\right) \left(\frac{n-1}{n-m}\right)$$

with $size_i$ being the number of patients treated for the same APR-DRG in the hospital, $y_i$ the observed outcome, $f_i$ the predicted outcome, $\bar{y}$ the weighted average of the observed outcomes, $n$ the number of observations and $m$ the number of parameters. The contribution of the fixed effects was determined by calculating the adjusted $R^2$ for the model containing only fixed effects and dividing it by the adjusted $R^2$ for the

final model. The contribution of the random effects was calculated by subtracting the fixed effects contribution from 100%.

## 9.2 Inclusion of hospital-specific time lags

In order to account for the fact that the impact of policies might already be noticed before implementation or be delayed after implementation, a time lag ($\pm$ 1 year) was included. The time lag was modelled by a hospital-specific shift in time, which could range from $-1$ year before to $+1$ year after the fixed change-point. In order to test the need for a hospital-specific time lag, we fitted a model containing a fixed and random intercept, a fixed and random slope, all change-points and a random time lag for each change-point. Variances were set equal to 1 and covariances to 0 in order to reach convergence. Because this model failed to converge for the data on lower limb surgery, we fitted two models containing a fixed and random intercept, fixed and random slope and one specific change-point with its time lag. The likelihoods for the models with and without time lags were compared in order to assess the need for time lags.

## 9.3 Results

### 9.3.1 Need for a hospital specific time lag

Likelihood ratio tests showed that, when studying compliance to guidelines for limb surgery (at patient level), DDDhosp for pneumonia (at hospital level) or OP for pneumonia (at hospital level), inclusion of hospital-specific time lags was not required. This implies that the fact that not all hospitals are impacted by the studied changes at the exact same moment is negligible.

### 9.3.2 Limb surgery: compliance at patient level

Because likelihood ratio tests indicated that both change-points (AMTIN and AC) and both random effects were required to study the evolution of compliance over time, all were kept in the model. A fixed effect representing the size of the hospital was required and hence was added to the final model. Significance of all fixed effects in the final model is shown in Table 9.1. This model appeared to fit the data reasonably well (Figure 9.1).

Table 9.1: Significance of fixed effects in the final model for compliance at patient level for limb surgery.

| Variable | P-value | Variable | P-value | Variable | P-value |
|---|---|---|---|---|---|
| Age | 0.1324 | Time*age | 0.0010 | AC*icu | $< 0.0001$ |
| Los | $< 0.0001$ | Time*sev | 0.0016 | AMTIN | $< 0.0001$ |
| AMT | 0.6488 | Time*gender | 0.0020 | AMTIN*sev | $< 0.0001$ |
| Sev | $< 0.0001$ | AC | 0.0092 | AMTIN*icu | 0.0001 |
| Gender | 0.0028 | AC*age | 0.0005 | Size | $< 0.0001$ |
| Icu | $< 0.0001$ | AC*los | 0.0009 | Time*size | $< 0.0001$ |
| Time | 0.9848 | AC*sev | 0.0633 | AC*size | $< 0.0001$ |
| | | AC*gender | 0.0012 | AMTIN*size | $< 0.0001$ |

The final model had a pseudo-$R^2$ value of 0.1927. Although $R^2$ values for models on binary outcomes are typically much lower than for models on continuous outcomes, this $R^2$-value indicates that there still is a lot of residual variability that is unexplained by the model. Note that 17% of the explained variability comes from the fixed effects while 83% comes from the random effects.

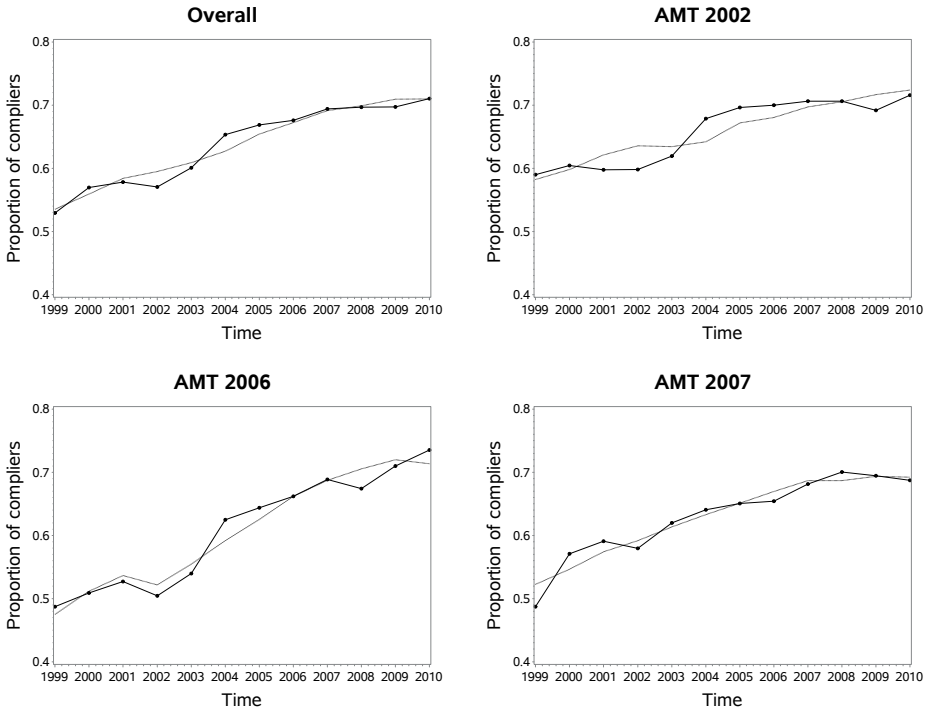



Figure 9.1: Average observed (solid line) and predicted (dotted line) evolution of the proportion of compliers over time: overall (top left), for patients treated in a hospital with AMT in 2002 (top right), with AMT in 2006 (bottom left) and with AMT in 2007 (bottom right).

### 9.3.3 Pneumonia: DDDhosp at hospital

Because likelihood ratio tests showed that change-point AMTIN was redundant ($p = 0.4659$), it was removed from the model. In a next step, likelihood ratio tests indicated that both change-points (FIN and AC) and both random effects were required in the model and that the AR(1) structure could not be simplified. After weighting the observations for the size of the hospital, a fixed effect for size was redundant. Significance of all fixed effects in the final model are shown in Table 9.2. This model appeared to fit the data well (Figure 9.2), which was verified by an adjusted $R^2$ value of 0.8072. Note that 33% of the variability is explained by the fixed effects while 67% is explained by the random effects.

Table 9.2: Significance of fixed effects in the final model for the number of DDD per 100 hospital days.

| Variable | P-value | Variable | P-value |
|---|---|---|---|
| Sev | 0.0258 | FIN | 0.0022 |
| Pt_ori | 0.3328 | FIN*los | 0.0026 |
| Dis_st | 0.4272 | FIN*pt_ori | 0.0736 |
| Age | 0.0202 | FIN*dis_st | 0.0263 |
| Los | 0.4639 | FIN*age | 0.0026 |
| Time | 0.6995 | AC | 0.0010 |
| Time*sev | 0.0792 | AC*pt_ori | 0.0093 |
| Time*dis_st | 0.0432 | AC*age | 0.0054 |
| Time*age | 0.0326 | | |

### 9.3.4 Pneumonia: OP at hospital level

Because likelihood ratio tests indicated that change-point AC was redundant ($p = 0.9884$), it was removed from the model. In a next step, likelihood ratio tests indicated that change-point AMTIN was also redundant ($p = 0.0972$). Therefore, also this change-point was removed from the model. Afterwards, likelihood ratio tests indicated that the remaining change-point (FIN) and both random effects were required to model the evolution of OP over time. The AR(1) residual structure could not be simplified and a fixed effect for size was redundant. Significance of all fixed effects in the final model are shown in Table 9.3. This model appeared to fit the data well (Figure 9.3), which was confirmed by an adjusted $R^2$ value of 0.7559. Note that 14% of the variability is explained by the fixed effects while 86% is explained by the random effects.

Table 9.3: Significance of fixed effects in the final model for the ratio of oral versus parenteral antimicrobial use at hospital level.

| Variable | P-value | Variable | P-value |
|---|---|---|---|
| Sev | 0.1038 | Time*sev | 0.0031 |
| Gender | 0.5360 | Time*los | 0.0111 |
| Dis_st | $< 0.0001$ | FIN | 0.0047 |
| Age | 0.0420 | FIN*gender | 0.0041 |
| Los | 0.2625 | FIN*los | 0.0022 |
| Time | 0.0006 | | |

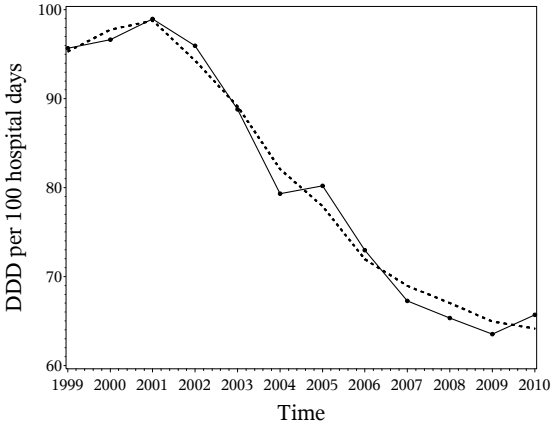



Figure 9.2: Observed (solid line) and predicted (dotted line) evolution in the number of DDD per 100 hospital days.

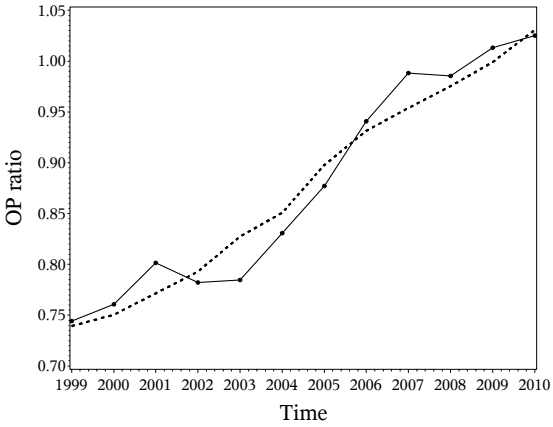



Figure 9.3: Observed (solid line) and predicted (dotted line) evolution in the ratio of oral versus parenteral antimicrobial use at hospital level.

## 9.4 Discussion

In this chapter, we used a change-point model to assess the impact of Belgian policies on antimicrobial consumption in hospitals. We showed that the proportion of compliant antibiotic prescriptions for limb surgery at patient level was changed suddenly by both the establishment of AMTs and the 2001 antimicrobial awareness campaign. The number of DDD per 100 hospital days for pneumonia (at hospital level) was changed suddenly by both the 2001 antibiotic awareness campaign and the new financing mechanism in 2006. The ratio of oral over parenteral DDD use for pneumonia (at hospital level) was changed suddenly by the new financing mechanism in 2006. Although all these sudden changes were statistically significant, only the decrease in the number of DDD per 100 hospital days for pneumonia seen in 2001 was steep enough to be clinically relevant.

There are several possible explanations for the lack of sudden change caused by the year of implementation of an AMT in hospitals. The selected outcomes are topics that were likely addressed at different time points, fully or largely independent from the implementation of the AMT. Also, AMTs were not given specific targets and hence might have chosen other priorities for interventions than the outcomes under study here, although these were the indications with the highest volume of patients receiving an antimicrobial. It is also possible that the AMT funding was not always used directly for AMT implementation. This cannot be verified as no data on the actual use of funding within hospitals exist.
Our findings do not question the need for AMTs, nor the need for continuation of AMT funding, but do suggest the need for transparency in the use of AMT funding and more guidance in terms of identifying priorities for action.

The major strength of this study is the use of exhaustive patient-based data spanning 12 years and including all Belgian hospitals. Although we study the two APR-DRGs accounting for the largest number of patients receiving an antimicrobial in hospital, only three outcomes for two APR-DRGs were studied which does not provide a complete picture of trends in quality of antimicrobial use in Belgian hospitals. This study should therefore ideally be complemented by other APR-DRGs, which might result in different conclusions than the ones reached in this study.

# Chapter 10

# Concluding remarks

In this thesis, we studied antibiotic use in hospitalized children and outpatients, resistance in outpatients, and the link between antibiotic consumption and resistance. Additionally, we developed a prediction rule to assist GPs in accurately identifying patients that would, or would not, benefit from antibiotic treatment, and assessed the impact of policies aimed at reducing resistance rates by optimizing antibiotic consumption.

In Chapter 2, we developed a new prediction rule to predict poor prognosis in patients presenting to primary care with acute cough. We identified the country's baseline risk for poor prognosis, the presence of crackles during physical examination, the severity of phlegm as assessed by the patient, the severity of interference with daily activities, the time since the patient last smoked and the patient's diastolic blood pressure to be important predictors for poor prognosis. CRP and BUN did not improve the discriminative power of the new prediction rule. Cross-validations showed that the new prediction rule is quite stable, and comparison to five existing prediction rules showed that the new prediction rule outperforms all, although there is still room for improvement. Therefore, our recommendation to GPs is not to use any of the existing prediction rules to predict poor prognosis in patients presenting to primary care with acute cough but to use this newly developed prediction rule instead.

In Chapter 3, we conducted a simulation study to investigate the impact of an increasing percentage of singletons at the lowest level of the hierarchy on different aspects of the linear multi-level model. We showed that ignoring and dropping singletons should be avoided as they come with low coverage and wide confidence intervals, respectively.

Regrouping the singletons into an artificial department might be considered, although the regular linear multi-level model performed better even when the percentage of singletons increased. As always, it is recommended to avoid the inclusion of singletons in the sample when designing a study. If, due to circumstances, high proportions of singletons do appear at the lowest level, the linear multi-level model can be used safely. There is no need for, and we would even advise against, grouping of the singletons into an artificial unit. Removing the singletons from the analysis or ignoring the dependency within clusters is not advisable and should be avoided.

In Chapter 4, we conducted a simulation study to investigate the impact of increasing primary unit sparseness on the performance of the F test. We showed that the F test performed well when there is no primary unit sparseness, with REML outperforming ML. In the presence of primary unit sparseness, the performance of the F test was inadequate. We considered dropping, regrouping or splitting the singletons as ways to reduce primary unit sparseness. They could all solve either the problem of a high type I error or the problem of low power, whilst worsening the other, which forces us to conclude that neither method acts as a solution to the poor performance of the F test in the presence of primary unit sparseness. As an alternative to the F test, we studied the performance of the Wald test, the likelihood ratio test and the permutation test. While the performance of the Wald and likelihood ratio test were comparable to the performance of the F test, the permutation test outperformed the F test both under REML and ML. Therefore, we recommend the use of a permutation test (under REML) to determine significance of a fixed effect at the primary level in a multi-level model with a continuous outcome suffering from primary unit sparseness. Ideally, a permutation test should also be used to determine the significance of fixed effects at lower levels. However, if computing time is an issue, it is advisable to check if the F test under REML and ML agree on the significance of the parameters and to conduct a permutation test under REML if there is disagreement.

In Chapter 5, we illustrated that $\beta$-lactam antibiotics are prescribed either according to or independent of weight, with the odds of prescribing independent of weight being higher for children with a higher weight (or age). We showed that lower doses of parenteral ceftriaxone were prescribed to children receiving empiric treatment, with a lower weight and with a less severe reason for treatment. Although we expected reasons to deviate from the average prescribed $\beta$-lactam dose, which should correspond to the recommended dose, to be common for the 12 included $\beta$-lactam antibiotics, most predictors were antibiotic-specific. Only type of treatment and type of hospital affected all included $\beta$-lactam antibiotics in a similar fashion, with lower doses being prescribed in primary or secondary hospitals and for empiric treatment.

Evolution of antibiotic doses prescribed to outpatients over time was studied in Chapter 6. We showed that conclusions based on antibiotic use expressed in DID and in PID are contradictory, with antibiotic use expressed in DID showing a significant increase over time while antibiotic use expressed in PID did not change significantly over time. Matching random effects were shown to be highly correlated, albeit not perfect, which implies that both DID and PID should be reported when studying antibiotic consumption. Additionally, we showed that the number of DDD per package increased significantly over time for all antibiotics except for quinolones.
In Chapter 7, we showed that while the association between $\beta$-lactam antibiotics and resistance could be modelled using PID alone, the association between TMLS use and resistance needed both DID and PID. Additionally, we showed that the association between $\beta$-lactam antibiotics and resistance was modelled best without a time lag, while the the association between TMLS use and resistance needed a one year time lag, reflecting a possible difference in persistence of resistance.
In Chapter 8, we demonstrated that persistence of resistance is longer after treatment with AF than after treatment with CD. We hereby confirmed the differences found by Malhotra-Kumar *et al.* (2007) and Chung *et al.* (2007), suggesting that routinely collected data could serve as a proxy in assessing differences in persistence of resistance.

The impact of policies to limit antimicrobial resistance by reducing antibiotic consumption was assessed in Chapter 9. The proportion of compliant antibiotic prescriptions for limb surgery at patient level was changed suddenly by both the establishment of AMTs and the 2001 antimicrobial awareness campaign. The number of DDD per 100 hospital days for pneumonia (at hospital level) was changed suddenly by both the 2001 antibiotic awareness campaign and the new financing mechanism in 2006. The ratio of oral over parenteral DDD use for pneumonia (at hospital level) was changed suddenly by the new financing mechanism in 2006. Although all these sudden changes were statistically significant, only the decrease in the number of DDD per 100 hospital days for pneumonia seen in 2001 was steep enough to be clinically relevant.

## 10.1 Topics for further research

The prediction rule, developed in Chapter 2, was shown to outperform the existing prediction rules, although there is still room for further improvement. Possible extensions could include addition of aetiology of the underlying infection (viral or bacterial)

and interpretation of a chest radiograph. The prediction rule currently includes both variables assessed by the patient (severity of phlegm) and variables assessed by the GP (presence of crackles). It would however be more objective and practical for the prediction rule to only include variables assessed by the GP. Information on variables assessed by the GP are included in the Acute Cough Data, and could replace variables assessed by the patient when they are shown to be in agreement.
Although the new prediction rule has been validated using internal validation (cross-validation), external validation using an independent dataset on the same subject remains desirable.

In Chapters 3 and 4, we focussed on a multi-level mode with a continuous outcome, which was applied in Chapter 5. Further research could focus on exploring the behaviour of tests for the significance of fixed effects in the presence of singletons when the outcome variable is for example binary or follows a Poisson distribution.
In Chapter 5, we focussed on $\beta$-lactam antibiotics and constructed a meta-model to assess one specific hypothesis. Using the Inpatient Antibiotic Use Data, the same strategy could be used to assess other hypotheses on the use of $\beta$-lactam antibiotics, to study other frequently used antibiotics (e.g. parenteral gentamicin) and to study antibiotic use in neonates.
In Chapters 6 and 7, we studied the use of antibiotics in outpatients and its link with resistance. Further research would benefit from the inclusion of characteristics related to outpatients and prescribers, as this would provide a more clear and thorough image of the variability in antibiotic consumption patterns both between and within countries.

The strategy developed in Chapter 8 could be used to assess other drug-bug combinations for which information is contained in the Bacterial Susceptibility Data, while the strategy developed in Chapter 9 could be used to assess the impact of policies on other APR-DRGs.

# References


3M Health Information Systems (2003), 'All patient refined diagnosis related groups (APR-DRGs) version 20.0 methodology overview', https://www.hcup-us.ahrq.gov/db/nation/nis/APR-DRGsV20MethodologyOverviewandBibliography.pdf. (February 26, 2016, date last accessed).

Adriaenssens N, Coenen S, Versporten A, Muller A, Minalu G, Faes C, Vankerckhoven V, Aerts M, Hens N, Molenberghs G and Goossens H (2011*a*). European surveillance of antimicrobial consumption (ESAC): outpatient antibiotic use in Europe (1997-2009), *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi3–vi12.

Adriaenssens N, Coenen S, Versporten A, Muller A, Minalu G, Faes C, Vankerckhoven V, Aerts M, Hens N, Molenberghs G and Goossens H (2011*b*). European surveillance of antimicrobial consumption (ESAC): outpatient macrolide, lincosamide and streptogramin (MLS) use in Europe (1997-2009), *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi37–vi45.

Adriaenssens N, Coenen S, Versporten A, Muller A, Minalu G, Faes C, Vankerckhoven V, Aerts M, Hens N, Molenberghs G and Goossens H (2011*c*). European surveillance of antimicrobial consumption (ESAC): outpatient quinolone use in Europe (1997-2009), *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi47–vi56.

Agresti A (2002). *Models for discrete longitudinal data*, Springer, New York.

Akram A R, Chalmers J D and Hill A T (2011). Predicting mortality with severity assessment tools in out-patients with community-acquired pneumonia, *Quarterly journal of medicine* **104**, 871–879.

Alanis A J (2005). Resistance to antibiotics: are we in the post-antibiotic era?, *Archives of Medical Research* **36**, 697–705.

Bell B A, Morgan G B, Schoeneberger J A, Kromrey J D and Ferron J M (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models, *Methodology* **10**, 1–11.

Bont J, Hak E, Hoes A W, Macfarlan J T and Verheij T (2008). Predicting death in elderly patients with community-acquired pneumonia: a prospective validation study reevaluating the CRB-65 severity assessment tool, *Archives of internal medicine* **168**, 1465–1468.

Butler C C, Hood K, Verheij T, Little P, Melbye H, Nuttall J, Kelly M J, Mölstad S, Goycki-Cwirko M, Almirall J, Torres A, Gillespie D, Tautakorpi U, Coenen S and Goossens H (2009). Variation in antibiotic prescribing and its impact on recovery in patients with acute cough in primary care: prospective study in 13 countries, *British medical journal* **338**, b2242.

Catry B, Hendrickx E, Preal R and Mertens R (2008). Verband tussen antibioticaconsumptie en microbiële resistentie bij de individuele patiënt.

Chihara L M and Hesterberg T C (2011). *Mathematical statistics with resampling and R*, John Wiley & Sons, Hoboken,NJ.

Chung A, Perera R, Brueggemann A, Elamin A, Harnden A, Mayon-White R, Smith S, Crook D and Mant D (2007). Effect of antibiotic prescribing on antibiotic resistance in individual children in primary care: prospective cohort study, *British Medical Journal* **335**, 429.

Clarke P (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data, *Journal of epidemiology & community health* **62**, 752–758.

Coenen S, Adriaenssens N, Versporten A, Muller A, Minalu G, Faes C, Vankerckhoven V, Aerts M, Hens N, Molenberghs G and Goossens H (2011). European surveillance of antimicrobial consumption (ESAC): outpatient use of tetracyclines, sulphonamides and trimethoprim, and other antibacterials in Europe (1997-2009), *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi57–vi70.

Coenen S, Gielen B, Blommaert A, Beutels P, Hens N and Goossens H (2014). Appropriate international measures for outpatient antibiotic prescribing and consumption: recommendations from a national data comparison of different measures, *Journal of antimicrobial chemotherapy* **69**(2), 529–534.

Cortiñas Abrahantes J, Molenberghs G, Burzykowski T, Shkedy Z and Renard D (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models, *Computational statistics and data analysis* **47**, 537–563.

Costelloe C, Metcalfe C, Lovering A, Mant D and Hay A D (2010). Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis, *British medical journal* **340**, c2096.

Dempster A P, Rubin D B and Tsutakawa R K (1981). Estimation in covariance components models, *Journal of the american statistical association* **76**, 341–353.

Descheemaeker P (2000). Macrolide resistance and erythromycin resistance determinants among Belgian Streptococcus pygogenes and Streptococcus pneumoniae isolates, *Journal of Antimicrobial Chemotherapy* **45**, 167–73.

Dever L A and Dermody T S (1991). Mechanisms of bacterial resistance to antibiotics, *Archives of internal medicine* **151**, 886–95.

Dinant G J, Buntinx F and Butler C C (2007). The necessary shift from diagnostic to prognostic research, *Biomed central family practice* **8**, 53.

Faes C, Molenberhgs G, Hens N, Muller A, Goossens H and Coenen S (2011). Analysing the composition of outpatient antibiotic use: a tutorial on compositional data analysis, *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi89–vi94.

Fine M J, Auble T E, Yealy D M, Hanusa B H, Weissfeld L A, Singer D E, Coley C M, Marrie T J and Kapoor W N (1997). A prediction rule to identify low-risk patients with community-acquired pneumonia, *New england journal of medicine* **23**, 243–250.

Fleming A (1929). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae, *British Journal of Experimental Pathology* **10**, 226–36.

Fleming A (1945). Penicillin, *Nobel lecture* .

Francis N A, Cals J W, Butler C C, Hood K, Verheij T, Little P, Goossens H and Coenen S (2012). Severity assessment for lower respiratory tract infections: potential use and validity of the CRB-65 in primary care, *Primary care respiratory journal* **21**, 65–70.

French G L (2005). Clinical impact and relevance of antibiotic resistance, *Advanced drug delivery reviews* **57**(10), 1514–1527.

Garson D G (2013). *Hierarchical linear modeling: Guide and appliations*, Sage publications, London.

Gibson J, Loddenkemper R, Sibille Y and Lundback B (2013). *The European lung white book: respiratory health and disease in Europe*, European respiratory society, Sheffield.

Goldstein H (2003). *Multilevel statistical models*, 4th edn, John Wiley & Sons, Chichester.

Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, Nuttall D and Thomas S (1993). A multilevel analysis of school examination results, *Oxford review of education* **19**(4), 425–433.

Goossens H, Coenen S, Costers M, De Corte S, De Sutter A, Gordts B, Laurier L and Struelens M (2008). Achievements of the Belgian Antibiotic Policy Coordination Committee (BAPCOC), *Euro Surveillance* **13**, 19036.

Goossens H, Ferech M, Vander Stichele R and Elseviers M (2005). Outpatient antibiotic use in Europe and association with resistance: a cross-national database study, *Lancet* **365**, 579–587.

Harville D A (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the american statistical association* **72**(358), 320–338.

Hosmer D W and Lemeshow S (2000). *Applied logistic regression*, Wiley, New York.

Hothorn T, Hornik K and Zeileis A (2006). Unbiased recursive partitioning: a conditional inference framework, *journal of computational and graphical statistics* **15**(3), 651–674.

Hox J (1998). Multilevel modeling: when and why, *in* I Balderjahn, R Mathar and M Schader, eds, *In: Classification, data analysis and data highways*, Springer Verlag, New York, pp. 147–154.

Hox J (2010). *Multilevel analysis: techniques and applications*, 2nd edn, Routledge, New York.

Ilić K, Jakovljević E and Škodrić V (2012). Social-economic factors and irrational antibiotic use as reasons for antibiotic resistance of bacteria causing common childhood infections in primary healthcare, *European journal of paediatrics* **171**, 767–77.

Kreft I and De Leeuw J (1998). *Introducing multilevel modeling*, Sage publications, London.

Kutner M H, Nachtsheim C J, Neter J and Li W (2005). *Applied linear statistical models*, 5th edn, McGraw-Hill Irwin, New York.

Lee K R and Kapadia C H (1984). Variance components estimators for the balanced two-way mixed model, *Biometrics* **40**(2), 507–512.

Lee V E (2000). Using hierarchical linear modeling to study social contexts: the case of school effects, *Educational psychologist* **35**(2), 125–141.

Lehmann E L (2006). *Nonparametrics: statistical methods based on ranks*, 1st edn, Springer Science+Business media, LLC, New York.

Liang K Y and Zeger S L (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

Lim W S, van der Eerden M M, Laing R, Boersma W G, Karalus N, Town G I, Lewis S A and Macfarlane J T (2003). Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study, *Thorax* **58**, 377–382.

Littell R C, Milliken G A, Stroup W W, Wolfinger R D and Schabenberger O (2006). *SAS ® for mixed models*, 2nd edn, SAS Institute Inc., Cary.

Little P, Stuart B, Moore M, Coenen S, Butler C C, Godycki-Cwirko M, Mierzecki A, Chlabicz S, Torres A, Amirall J, Davies M, Schaberg T, Mölstad S, Blasi F, De Sutter A, Kersnik A, Hupkova H, Touboul P, Hood K, Mullee M, O'Reilly G, Brugman C, Goossens H and Verheij T (2013). Amoxicillin for uncomplicated acute lower respiratory tract infection in primary care: a 12 country randomised placebo controlled trial, *Lancet infectious diseases* **13**, 123–129.

Loke Y K, Kwok C S, Niruban A and Myint P K (2010). Value of severity scales in predicting mortality from community-acquired pneumonia: systematic review and meta-analysis, *Thorax* **65**, 884–890.

Maas C and Hox J (2005). Sufficient sample sizes for multilevel modeling, *Methodology* **1**, 86–92.

Malhotra-Kumar S, Lammens C, Coenen S, Van Herck K and Goossens H (2007). Impact of azithromycin and clarithromycin therapy on pharyngeal carriage of

macrolide-resistant streptococci among healthy volunteers: a randomised, double-blind, placebo-controlled study, *Lancet* **369**, 482–90.

Malhotra-Kumar S, Van Heirstraeten L, Coenen S, Lammens C, Adriaenssens N, Kowalczyk A, Godycki-Cwirko M, Bielicka Z, Hupkova H, Lennering C, Mölstad S, Fernandez-Vandellos P, Torres A, Parizel M, Ieven M, Butler C, Verheij T, Little P and Goossens H (2016). Impact of amoxicillin therapy on resistance selection in patients with community-acquired lower respiratory tract infections: a randomised placebo-controlled study, *Journal of antimicrobial chemotherapy* **In revision**.

McCulloch C E and Searle S R (2001). *Generalized, linear, and mixed models*, John Wiley & Sons, New York.

Meng X and Rubin D (1992). Performing likelihood ratio tests with multiply-imputed data sets, *Biometrika* **79**(1), 103–111.

Minalu G, Aerts M, Coenen S, Versporten A, Muller A, Adriaenssens N, Beutels P, Molenberghs G, Goossens H and Hens N (2011). Application of mixed-effects models to study the country-specific outpatient antibiotic use in Europe: a tutorial on longitudinal data analysis, *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi79–vi87.

Molenberghs G and Verbeke G (2005). *Models for discrete longitudinal data*, Sprringer, New York.

Monnet D L, Mölstad S and Cars O (2004). Defined daily doses of antimicrobials reflect antimicrobial prescriptions in ambulatory care, *Journal of antimicrobial chemotherapy* **53**(6), 1109–1111.

Moore M, Stuart B, Butler C C, Goossens H, Verheij T J and Little P (2014). Amoxicillin for acute lower respiratory tract infection in primary care: subgroup analysis of potential high-risk groups, *British journal of general practice* **64**, e75–e80.

Moulton B R (1986). Random group effects and the precision of regression estimates, *Journal of econometrics* **32**(3), 385–397.

Muggeo V M R (2003). Estimating regression models with unknown break-points, *Statistics in medicine* **22**, 3055–3071.

Neill A M, Martin I R, Weir R, Anderson R, Chereshsky A, Epton M J, Jackson R, Schousboe M, Frampton C, Hutton S, Chambers S T and Town G I (1996). Community acquired pneumonia: aetiology and usefulness of severity criteria on admission, *Thorax* **51**, 1010–1016.

Neu H C and Gootz T D (1996). Antimicrobial chemotherapy, *in* B S, ed., *Medical Microbiology*, 4th edn.

Patterson H D and thompson R (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**(3).

Pickering R and Weatherall M (2007). The analysis of continuous outcomes in multicentre trials with small centre sizes, *Statistics in medicine* **26**, 5445–5456.

Putnam S D, Sanders J W, Tribble D R, Rockabrand D R, Riddle M S, Rozmajzl P J and Frenck R W (2005). Posttreatment changes in Escherichia coli antimicrobial susceptibility rates among diarrheic patients treated with ciprofloxacin, *Antimicrobial agents and chemotherapy* **49**, 2571–2572.

Raudenbush S W and Bryk A S (2002). *Hierarchical linear models: applications and data anlysis methods*, 2nd edn, Sage publications, Thousand Oaks.

Raum E, Lietzau S, von Baum H, Marre R and Brenner H (2008). Changes in Escherichia coli resistance patterns during and after antibiotic therapy: a longitudinal study among outpatients in Germany., *Clinical microbiology and infection* **14**, 41–48.

Renard D, Molenberghs G, Van Oyen H and Tafforeau J (1998). Investigation of the clustering effect in the Belgian Health Interview Survey 1997, *Archives of public health* **56**, 345–361.

Rubin D B (1987). *Mulitple imputation for nonresponse in surveys*, John Wiley & Sons, New York.

Sauzet O, Wright K C, Marston L, Brocklehurst P and Peacock J L (2012). Modelling the hierarchical structure in datasets with very small clusters: a simulation study to explore the effect of the proportion of clusters when the outcome is continuous, *Statistics in medicine* **32**, 1429–1438.

Searle S R, casella G and McCulloch C E (1992). *Variance components*, John Wiley & Sons, Hoboken,NJ.

Shackcloth J, Williams L and Farrell D J (2004). Streptococcus pneumoniae and Streptococcus pyogenes isolated from a paediatric population in Great Britain and Ireland: the in vitro activity of telithromycin versus comparators., *Journal of infection* **48**, 229–235.

Singer J D (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models, *Journal of educational and behavioral statistics* **24**(4), 323–355.

Smith R (1998). Action on antimicrobial resistance. Not easy, but Europe can do it, *British medical journal* **317**(7161), 764–770.

Snijders T A B and Bosker R J (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, Sage Publications, London.

Strobl C, Boulesteix A, Zeileis A and Hothorn T (2007). Bias in random forest variable importance measures: illustrations, sources and a solution, *Biomed central bioinformatics* **8**(25).

Swallow W H and Monahan J F (1984). Monte carlo comparison of ANOVA, MIVQUE, REML and ML estimators of variance components, *Technometrics* **26**(1), 47–57.

Teepe J, Broekhuizen B, Loens K, Lammens C, Ieven M, Goossens H, Butler C, Coenen S, Godycki-Cwirko M and Verheij T (2016). Predicting the presence of bacterial pathogens in the airways of patients with acute cough in primary care, *Canadian medical association journal* ***Accepted***.

Van Buuren S and Groothuis-Oudshoorn K (2011). mice:Multivariate imputation by chained equations in R, *Journal of statistical software* **45**.

Van Heirstraeten L, Coenen S, Lammens C, Hens N, Goossens H and Malhotra-Kumar S (2012). Antimicrobial drug use and macrolide-resistant *Streptococcus pyogenes*, *Emerging infectious diseases* **18**(9), 1515–18.

Van Vugt S F, Broekhuizen B D L, Lammens C, Zuithoff N P A, de Jong P A, Coenen S, Ieven M, Butler C C, Goossens H, Little P and Verheij T J M (2013). Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study., *British medical journal* **346**, f2450.

Verbeke G and Molenberghs G (1997). *Linear mixed models in practice: A SAS-oriented approach*, Springer Verlag, New York.

Verbeke G and Molenberghs G (2009). *Linear mixed models for longitudinal data*, Springer Verlag, New York.

Versporten A, Coenen S, Adriaenssens N, Muller A, Minalu G, Faes C, Vankerckhoven V, Aerts M, Hens N, Molenberghs G and Goossens H (2011*a*). European surveillance of antimicrobial consumption (ESAC): outpatient penicillin use in Europe (1997-2009), *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi13–vi23.

Versporten A, Coenen S, Adriaenssens N, Muller A, Minalu G, Faes C, Vankerckhoven V, Aerts M, Hens N, Molenberghs G and Goossens H (2011*b*). Europen surveillance of antimicrobial consumption (ESAC): outpatient cephalosporin use in Europe (1997-2009), *Journal of antimicrobial chemotherapy* **66**(suppl 6), vi25–vi35.

Versporten A, Sherland M, Bielicki J, Drapier N, Vankerckhoven V and Goossens H (2013). The antibiotic resistance and prescribing in European children project: a neonatal and pediatric antimicrobial web-based point prevalence survey in 73 hospitals worldwide, *Pediatric infectious disease journal* **32**, e242–53.

Vugt S F V, Butler C C, Hood K, Kelly M J, Coenen S, Goossens H, Little P and Verheij T (2012). Predicting benign course and prolonged illness in lower respiratory tract infections: a 13 European country study, *Family practice* **29**, 131–138.

Wang J, Xie H and Fisher J H (2012). *Multilevel models: applications using SAS* ®, De Gruyter, Berlin.

WHO (2011), 'WHO collaborating centre for drug statistics methodology. Anatomical Therapeutic chemical (ATC) classification system: guidelines for ATC classification and DDD assignment 2011.', `http://www.whocc.no/filearchive/publications/2011guidelines.pdf`. (February 26, 2016, date last accessed).

Willemsen I, Bogaers-Hofman D, Winters M and Kluytmans J (2009). Correlation between antibiotic use and resistance in hospital: temporary and ward-specific observations, *Infection* **37**, 432–37.

Youden W J (1950). Index for rating diagnostic tests, *Cancer* **3**(1), 32–35.

Zaletel-Kragelj L and Božikov J (2010). *Methods and Tools in Public Health*, Hans Jabobs, Lage.

# Supplementary information

Table A1: Inclusion criteria for further analysis of the Acute Cough data.

| Inclusion criterion | Number of violations |
|---|---|
| Older than 18 years | 1 |
| Presence of acute cough | 0 |
| First consultation for this episode | 0 |
| Not included in this study before | 2 |
| Capable to fill in study material | 0 |
| Written consent for participation | 0 |
| Immunocompetent | 0 |
| Not treated with antibiotics in the past month | 6 |
| Not pregnant | 1 |

*Note that one patient can violate multiple inclusion criteria.*

Table A2: Variables included in the Acute Cough data.

| Variable | Description |
|---|---|
| **CRF - exclusion criteria** | |
| Alergic_penicill | Presence of penicillin allergy |
| Hist_physical_exam | Physical exam suggestive for pneumonia |
| **CRF - interview by the GP** | |
| Cough | Severity of cough (5-point scale) |
| Phlegm | Severity of phlegm (5-point scale) |
| Breath | Severity of breathlessness (5-point scale) |
| Wheeze | Severity of wheezing (5-point scale) |
| Runn_nose | Severity of running nose (5-point scale) |
| Fever | Severity of fever (5-point scale) |
| Chest_pain | Severity of chest pain (5-point scale) |
| Muscle_ach | Severity of muscle ache (5-point scale) |
| Headache | Severity of headache (5-point scale) |
| Sleep | Severity of sleeping problems (5-point scale) |
| Gen_unwell | Severity of general unwellness (5-point scale) |
| Interf_act | Severity of interference with daily activities (5-point scale) |
| Conf_disor | Severity of confusion (5-point scale) |
| Diarhh | Severity of diarrhoea (5-point scale) |
| Phlegm_colour | Colour of phlegm (no colour, white, yellow, green, red, none produced) |
| Dur_pr_illness | Duration of present illness (in days) |
| Dur_pr_cough | Duration of present cough (in days) |
| **CRF - examination by the GP** | |
| Norm_consc_yn | Normal consciousness |

Table A2 Continued.

| Variable | Description |
|---|---|
| Gen_tox_yn | Sick impression |
| Dim_vesic_breaht_yn | Presence of diminished vesicular breath |
| Wheeze2_yn | Presence of wheezing |
| Crackles2_yn | Presence of crackles |
| Rhonchi2_yn | Presence of rhonchi |
| Beats_min | Number of heart beats per minute |
| Breahts_min | Number of breaths per minute |
| Prol_exp | Prolonged expiration |
| Syst_bp | Systolic blood pressure |
| Diast_bp | Diastolic blood pressure |
| Oral_temp | Oral temperature |
| Suspected_pneumonia | Suspicion pneumonia |
| Prescr_med_yn | Medication prescribed |
| Nr_meds | Number of prescribed drugs |
| Location | Location the patient was seen (practice, home) |
| **CRF - History taking by the GP** | |
| Copd_yn | Presence of chronic obstructive pulmonary disorder |
| Asthma_yn | Presence of asthma |
| Lung_other_yn | Presence of other lung diseases |
| Heart_fail_yn | Presence of heart failure |
| Isch_heart_yn | Presence of ischemic heart disease |
| Other_heart_yn | Presence of other heart diseases |
| Diabetes_yn | Presence of diabetes |
| Prev_hosp_resp_yn | Previously hospitalized for pulmonary disorder |
| Ab_treat_yn | Treated with antibiotics in the past six months |

Table A2 Continued.

| Variable | Description |
|---|---|
| Allergic_disease_yn | Presence of allergies (e.g. hay fever) |
| Other_fine_diseases | Presence of other FINE disease |
| Ethic_backgr | Ethnic background (Caucasian, African, Asian, other) |
| Smoke | Smoking status (never, past, now) |
| Smoke_per_day | Number of cigarettes smoked per day |
| Smoke_years | Number of years smoked (up till now) |
| Inh_bronch_yn | Use of inhaled bronchodilators |
| Inh_ster_yn | Use of inhaled steroids |
| Or_ster_yn | Use or oral steroids |
| Or_agent_diab_yn | Use of oral agents for diabetes |
| Insulin_yn | Use of insulin |
| Ant_hyp_yn | Use of anti-hypertensives/diuretics |
| Non_ster_yn | Use of non-steroidal anti-inflammatory drugs (oral) |
| Benz_ad_yn | Use of benzodiazepines/antidepressants |
| Infl_vacc_yn | Vaccination with flu vaccine |
| **Patient diary - general questions** | |
| Age | Age of the patient |
| Sex | Gender of the patient |
| Longterm_illness_yn | Presence of long-term illness |
| Fever_eczema | Ever experienced hay fever or eczema |
| Asthma_family | Presence of asthma in the family |
| Cough_2wksplus | Number of cough episodes lasting longer than one week (in the past year) |
| Chest_wheez | Presence of wheeze in past year |
| Chest_tightness | Presence of chest tightness in the past year |

Table A2 Continued.

| Variable | Description |
|---|---|
| Attack_coughing | Presence of coughing attacks in the past year |
| Smokestat | Smoking status (current, past, never) |
| Smokeyears | Year smoked (up till now) |
| Smokedays | Number of cigarettes smoked per day |
| Stopsmoke | Number of years stopped smoking |
| Day_unwell | Number of days unwell before consultation |
| Counter_meds | Consumption of over-the-counter drugs |
| Nr_c_meds | Number of different over-the-counter drugs consumed |
| Other_remedies_cough | Use of other cough remedies |
| **Patient Diary - day 1** | |
| Mobility | Inference with mobility (i.e. walking) (3-point scale) |
| Self_care | Interference with self care (e.g. getting dressed) (3-point scale) |
| Usual_activities | Interference with usual activities (e.g. work) (3-point scale) |
| Pain_discomfort | Severity of discomfort (3-point scale) |
| Anxiety_depression | Severity of depression (3-point scale) |
| Day1_cough | Severity of cough on day 1 (7-point scale) |
| Day1_phlegm | Severity of phlegm on day 1 (7-point scale) |
| Day1_sh_breath | Severity of shortness of breath on day 1 (7-point scale) |
| Day1_wheeze | Severity of wheezing on day 1 (7-point scale) |
| Day1_runny_nose | Severity of runny nose on day 1 (7-point scale) |
| Day1_chestpain | Severity of chest pain on day 1 (7-point scale) |

Table A2 Continued.

| Variable | Description |
|---|---|
| Day1_fever | Severity of fever on day 1 (7-point scale) |
| Day1_musc_ache | Severity of muscle ache on day 1 (7-point scale) |
| day1_head_ache | Severity of headache on day 1 (7-point scale) |
| Day1_dist_sleep | Severity of disturbed sleep on day 1 (7-point scale) |
| Day1_gen_unwell | Severity of general unwellness on day 1 (7-point scale) |
| Day1_int_normact | Severity of interference with normal activities on day 1 (7-point scale) |
| day1_int_socact | Severity of interference with social activities on day 1 (7-point scale) |
| Worsenewadmit | Admission to hospital or reconsultation (with new or worse complaints) |
| Country | Country |
| Intervention | Receiving amoxicillin or not |
| Season | Season (winter,summer) |
| CRP | Concentration of C-reactive protein (in mg/l) |
| BUN | Concentration of blood urea nitrogen (in mg/dl) |

GP: general practitioner

3-point scale: no problem, moderate problem, severe problem

5-point scale: absent, no problem, mild problem, moderate problem, severe problem

7-point scale: absent, very small problem, small problem, moderate problem,severe problem, very severe problem, could not be worse

**Step 1**

- Age > 50
- Congestive heart failure
- Other FINE disease (e.g. neoplastic disease, renal disease, liver disease, ...)
- Confusion
- Pulse ≥ 125 beats/min
- Respiratory rate ≥ 30 breaths/min
- Systolic blood pressure < 90 mm Hg
- Oral temperature < 35°C or ≥ 40°C

None of these characteristics present => risk class I
Otherwise → proceed to step 2

**Step 2**

| Characteristic | Points |
|---|---|
| age for men | age (years) |
| age for women | age (years) – 10 |
| nursing home resident | + 10 |
| neoplastic disease | + 30 |
| liver disease | + 20 |
| congestive heart failure | + 10 |
| cerebrovascular disease | + 10 |
| renal disease | + 10 |
| altered mental status | + 20 |
| respiratory rate ≥ 30 breaths/minute | + 20 |
| systolic blood pressure < 90 mm Hg | + 20 |
| temperature < 35°C or ≥ 40°C | + 15 |
| pulse ≥ 125 beats/minute | + 10 |
| arterial pH < 7.35 | + 30 |
| blood urea nitrogen ≥ 11 mmol/l | + 20 |
| sodium < 130 mmol/l | + 20 |
| glucose ≥ 14 mmol/l | + 10 |
| haematocrit < 30% | + 10 |
| partial pressure of arterial oxygen < 60 mm Hg or oxygen saturation < 90% on pulse oximetry | + 10 |
| pleural effusion | + 10 |

| PSI score | Category |
|---|---|
| **0** | Class I – very low mortality |
| **≤ 70** | Class II – low mortality |
| **71 - 90** | Class III – intermediate mortality |
| **91 - 130** | Class IV – high mortality |
| **> 130** | Class V – very high mortality |

Figure A1: The Pneumonia Severity Index.

| Abbreviation | Clinical factor | Points |
|---|---|---|
| **C** | Confusion | 1 |
| **U** | Blood urea nitrogen > 19 mg/dl | 1 |
| **R** | Respiratory rate ≥ 30 breaths/min | 1 |
| **B** | Systolic blood pressure < 90 mm Hg or Diastolic blood pressure ≤ 60 mmHg | 1 |
| **65** | Age ≥ 65 | 1 |

| CRB score | Category |
|---|---|
| **0 or 1** | Non-severe pneumonia |
| **2 or 3** | Severe pneumonia |

| CURB score | Category |
|---|---|
| **0 or 1** | Non-severe pneumonia |
| **2 or 3 or 4** | Severe pneumonia |

| CRB-65 score | Category |
|---|---|
| **0** | Low mortality – likely suitable for home treatment |
| **1 or 2** | Intermediate mortality – likely need for hospitalisation |
| **3 or 4** | High mortality – urgent hospitalisation |

| CURB-65 score | Category |
|---|---|
| **0 or 1** | Low mortality – likely suitable for home treatment |
| **2** | Intermediate mortality –hospital supervised treatment |
| **3 or 4 or 5** | High mortality – manage in hospital as severe pneumonia |

Figure A2: The CRB, CURB, CRB-65 and CURB-65 scores.

Table A3: Pooled parameter estimates and standard errors for the full general model over three cross-validations.

| Parameter | Cross-validation 1 | Cross-validation 2 | Cross-validation 3 |
|---|---|---|---|
| Group A | -0.652 (0.623) | -0.454 (0.603) | 0.266 (0.591) |
| Group B | -0.006 (0.616) | 0.012 (0.597) | 0.727 (0.586) |
| Group C | 0.346 (0.641) | 0.350 (0.631) | 0.982 (0.618) |
| Crackles2_yn | -0.381 (0.191) | -0.310 (0.193) | -0.562 (0.189) |
| Day1_phlegm 1 | -0.625 (0.310) | -0.588 (0.340) | -0.616 (0.268) |
| Day1_phlegm 2 | -0.426 (0.271) | 0.239 (0.217) | -0.016 (0.257) |
| Day1_phlegm 3 | 0.087 (0.222) | 0.197 (0.193) | -0.173 (0.206) |
| Day1_phlegm 4 | 0.263 (0.222) | 0.473 (0.223) | 0.063 (0.226) |
| Day1_phlegm 5 | 0.299 (0.253) | 0.364 (0.258) | 0.049 (0.268) |
| Day1_phlegm 6 | -0.156 (0.359) | -0.289 (0.329) | -0.365 (0.300) |
| Usual_acitvities 2 | 0.253 (0.145) | 0.388 (0.134) | 0.286 (0.135) |
| Usual_activities 3 | 0.736 (0.242) | 0.884 (0.239) | 1.018 (0.250) |
| stopsmoke | 0.005 (0.003) | 0.006 (0.002) | 0.007 (0.003) |
| Diast_BP | -0.012 (0.006) | -0.015 (0.006) | -0.016 (0.006) |

Table A4: Pooled parameter estimates and standard errors for the reduced general models over three cross-validations.

| Parameter | Cross-validation 1 | Cross-validation 2 | Cross-validation 3 |
|---|---|---|---|
| Group A | -1.364 (0.500) | -1.034 (0.481) | -0.920 (0.454) |
| Group B | -0.729 (0.485) | -0.573 (0.471) | -0.492 (0.437) |
| Group C | -0.352 (0.528) | 0.229 (0.519) | -0.220 (0.489) |
| Day1_phlegm 1 | -0.611 (0.307) | -0.590 (0.338) | - |
| Day1_phlegm 2 | -0.412 (0.269) | 0.238 (0.217) | - |
| Day1_phlegm 3 | 0.105 (0.221) | 0.208 (0.192) | - |
| Day1_phlegm 4 | 0.289 (0.221) | 0.480 (0.222) | - |
| Day1_phlegm 5 | 0.308 (0.253) | 0.369 (0.257) | - |
| Day1_phlegm 6 | -0.117 (0.356) | -0.274 (0.328) | - |
| Usual_acitvities 2 | 0.251 (0.154) | 0.389 (0.133) | 0.289 (0.134) |
| Usual_activities 3 | 0.741 (0.242) | 0.903 (0.239) | 1.011 (0.248) |
| stopsmoke | 0.005 (0.003) | 0.006 (0.002) | 0.007 (0.003) |
| Diast_BP | -0.012 (0.006) | -0.015 (0.006) | -0.015 (0.005) |

Table A5: Performance characteristics for the fixed effect $Age_{ij}$ in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | -2.030 | 0.288 | -1.5 | 1.1 | 1.144 | 94.9 |
| 5 | -2.041 | 0.280 | -0.9 | 3.8 | 1.143 | 96.6 |
| 10 | -2.034 | 0.283 | -1.3 | 2.3 | 1.139 | 95.2 |
| 15 | -2.046 | 0.287 | -0.7 | 1.0 | 1.140 | 95.5 |
| 20 | -2.044 | 0.284 | -0.8 | 1.3 | 1.133 | 95.9 |
| 25 | -2.046 | 0.296 | -0.7 | -2.5 | 1.135 | 94.7 |
| 30 | -2.038 | 0.275 | -1.1 | 5.2 | 1.138 | 96.4 |
| 35 | -2.020 | 0.294 | -1.9 | -2.0 | 1.133 | 93.8 |
| 40 | -2.046 | 0.283 | -0.7 | 1.5 | 1.129 | 95.1 |
| 45 | -2.024 | 0.283 | -1.7 | 0.8 | 1.124 | 95.6 |
| 50 | -2.035 | 0.292 | -1.2 | -2.0 | 1.127 | 94.1 |
| 55 | -2.038 | 0.287 | -1.1 | -0.5 | 1.125 | 95.4 |
| 60 | -2.024 | 0.274 | -1.8 | 4.5 | 1.126 | 94.6 |
| 65 | -2.048 | 0.289 | -0.6 | -1.4 | 1.119 | 95.4 |
| 70 | -2.036 | 0.274 | -1.2 | 2.8 | 1.107 | 95.5 |
| 75 | -2.044 | 0.278 | -0.8 | 0.6 | 1.102 | 94.6 |
| 80 | -2.026 | 0.269 | -1.6 | 4.8 | 1.110 | 95.4 |
| 85 | -2.032 | 0.268 | -1.4 | 4.0 | 1.098 | 96.4 |
| 90 | -2.037 | 0.285 | -1.1 | -1.5 | 1.104 | 94.6 |
| 95 | -2.029 | 0.275 | -1.5 | -0.7 | 1.075 | 93.9 |

SES: simulation standard error

RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error

Table A6: Performance characteristics for the fixed effect $Reason_{2ij}$ in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | -21.724 | 7.006 | -1.2 | -0.1 | 27.552 | 95.2 |
| 5 | -21.851 | 6.927 | -0.7 | 0.4 | 27.353 | 94.8 |
| 10 | -21.889 | 7.150 | -0.5 | -2.6 | 27.411 | 94.6 |
| 15 | -21.922 | 7.059 | -0.3 | -0.6 | 27.605 | 95.5 |
| 20 | -21.968 | 7.092 | -0.1 | -1.3 | 27.554 | 95.1 |
| 25 | -21.595 | 7.003 | -1.8 | -0.1 | 27.529 | 94.0 |
| 30 | -21.874 | 7.108 | -0.5 | -2.8 | 27.193 | 94.7 |
| 35 | -21.657 | 6.960 | -1.5 | -1.0 | 27.103 | 94.9 |
| 40 | -21.551 | 7.078 | -2.0 | -1.9 | 27.320 | 95.3 |
| 45 | -21.965 | 6.761 | -0.1 | 1.7 | 27.039 | 95.9 |
| 50 | -21.842 | 6.716 | -0.7 | 4.1 | 27.505 | 95.5 |
| 55 | -22.068 | 7.268 | 0.3 | -4.2 | 27.386 | 93.5 |
| 60 | -22.106 | 6.655 | 0.5 | 2.9 | 26.954 | 96.4 |
| 65 | -21.733 | 6.805 | -1.2 | -1.0 | 26.493 | 94.4 |
| 70 | -21.932 | 6.778 | -0.3 | 0.4 | 26.766 | 95.1 |
| 75 | -21.867 | 6.717 | -0.6 | 0.9 | 26.671 | 95.1 |
| 80 | -21.647 | 6.935 | -1.6 | -0.0 | 27.284 | 95.4 |
| 85 | -21.718 | 6.758 | -1.3 | 2.9 | 27.352 | 96.2 |
| 90 | -21.707 | 6.245 | -1.3 | 3.9 | 25.535 | 96.0 |
| 95 | -21.677 | 6.012 | -1.4 | 0.4 | 23.760 | 95.2 |

SES: simulation standard error

RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error

Table A7: Performance characteristics for the fixed effect $Reason_{3ij}$ in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | -5.362 | 3.057 | -4.0 | 1.1 | 12.164 | 94.8 |
| 5 | -5.329 | 3.041 | -4.6 | 1.5 | 12.140 | 95.5 |
| 10 | -5.703 | 3.123 | 2.1 | -1.4 | 12.120 | 95.2 |
| 15 | -5.699 | 3.102 | 2.0 | -0.8 | 12.109 | 94.1 |
| 20 | -5.634 | 3.109 | 0.9 | -1.3 | 12.072 | 94.8 |
| 25 | -5.424 | 3.071 | -2.9 | -0.1 | 12.070 | 94.1 |
| 30 | -5.494 | 3.076 | -1.6 | -0.2 | 12.074 | 94.9 |
| 35 | -5.515 | 3.108 | -1.3 | -1.5 | 12.050 | 94.4 |
| 40 | -5.587 | 3.126 | 0.0 | -2.2 | 12.026 | 94.8 |
| 45 | -5.604 | 3.118 | 0.3 | -2.1 | 12.009 | 94.9 |
| 50 | -5.438 | 3.102 | -2.6 | -1.7 | 11.997 | 94.9 |
| 55 | -5.398 | 3.119 | -3.4 | -2.6 | 11.954 | 94.5 |
| 60 | -5.658 | 2.982 | 1.3 | 1.8 | 11.941 | 95.8 |
| 65 | -5.498 | 3.158 | -1.6 | -4.2 | 11.900 | 93.8 |
| 70 | -5.597 | 2.958 | 0.2 | 2.2 | 11.892 | 95.4 |
| 75 | -5.668 | 3.079 | 1.5 | -2.1 | 11.855 | 95.1 |
| 80 | -5.493 | 3.027 | -1.7 | 0.1 | 11.920 | 94.8 |
| 85 | -5.534 | 3.137 | -0.9 | -4.2 | 11.821 | 93.9 |
| 90 | -5.482 | 2.995 | -1.8 | -0.1 | 11.775 | 95.2 |
| 95 | -5.619 | 3.074 | 0.6 | -3.5 | 11.670 | 93.8 |

SES: simulation standard error

RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error

Table A8: Performance characteristics for the fixed intercept in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | 82.856 | 4.624 | -0.1 | -0.7 | 18.267 | 94.6 |
| 5 | 82.878 | 4.555 | -0.1 | 1.4 | 18.380 | 95.4 |
| 10 | 83.071 | 4.718 | 0.1 | -0.8 | 18.633 | 95.3 |
| 15 | 83.022 | 4.698 | 0.1 | 0.9 | 18.887 | 95.6 |
| 20 | 83.140 | 4.652 | 0.2 | 0.8 | 18.691 | 94.9 |
| 25 | 82.960 | 4.788 | 0.0 | -0.5 | 19.010 | 95.3 |
| 30 | 82.807 | 5.025 | -0.2 | -3.9 | 19.270 | 93.3 |
| 35 | 82.840 | 5.024 | -0.1 | -2.7 | 19.528 | 94.9 |
| 40 | 82.752 | 4.929 | -0.2 | 1.7 | 20.038 | 96.1 |
| 45 | 82.854 | 5.246 | -0.1 | -3.1 | 20.336 | 94.2 |
| 50 | 82.758 | 5.148 | -0.2 | -2.0 | 20.213 | 94.4 |
| 55 | 82.804 | 5.330 | -0.2 | -3.5 | 20.638 | 94.2 |
| 60 | 82.689 | 5.437 | -0.3 | -3.6 | 21.045 | 94.5 |
| 65 | 82.938 | 5.267 | -0.0 | 2.7 | 21.755 | 95.3 |
| 70 | 82.992 | 5.647 | 0.0 | -1.5 | 22.357 | 95.1 |
| 75 | 82.880 | 6.177 | -0.1 | -7.1 | 23.101 | 94.5 |
| 80 | 82.834 | 5.976 | -0.1 | -3.6 | 23.298 | 94.9 |
| 85 | 82.530 | 6.318 | -0.5 | -6.1 | 24.125 | 92.6 |
| 90 | 82.741 | 6.784 | -0.3 | -7.6 | 25.402 | 93.9 |
| 95 | 82.914 | 6.901 | -0.0 | -1.1 | 27.151 | 94.1 |

SES: simulation standard error

RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error

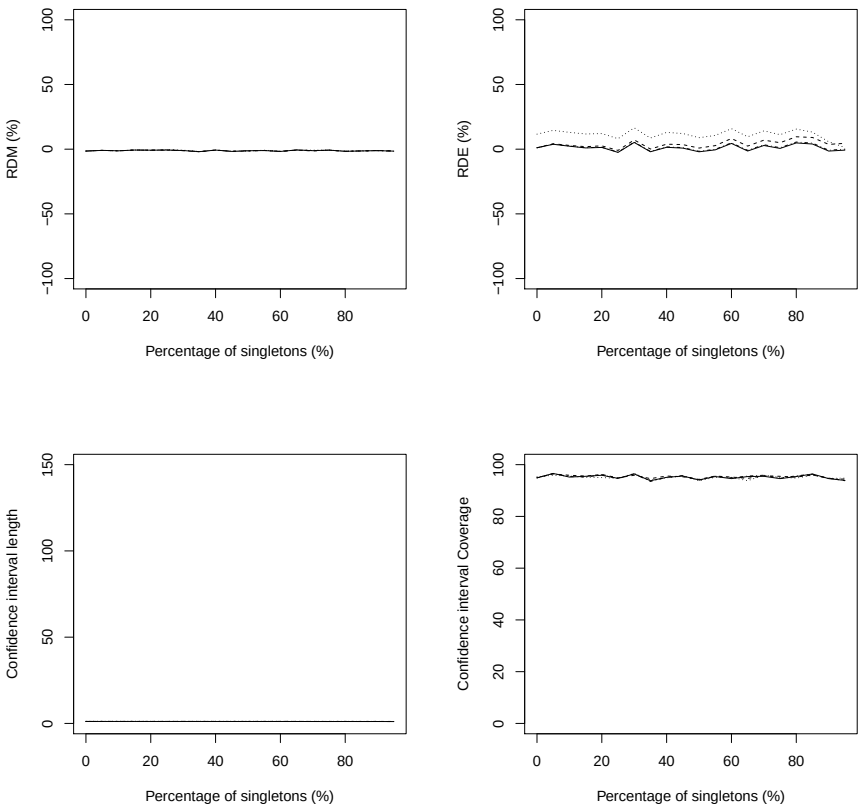
Table A9: Performance characteristics for the fixed effect $Size_{2j}$ in the three covariates model with an increasing percentage of singletons.

| Singletons (%) | Mean | SES | RDM (%) | RDE (%) | Length CI | Coverage |
|---|---|---|---|---|---|---|
| 0 | 4.401 | 5.352 | -1.5 | 1.5 | 21.840 | 95.9 |
| 5 | 4.408 | 5.603 | -1.4 | -1.5 | 22.190 | 95.7 |
| 10 | 4.412 | 5.458 | -1.3 | 1.8 | 22.367 | 94.8 |
| 15 | 4.630 | 5.687 | 3.6 | -0.6 | 22.758 | 94.6 |
| 20 | 4.327 | 5.546 | -3.2 | 1.5 | 22.687 | 95.6 |
| 25 | 4.227 | 5.960 | -5.4 | -3.6 | 23.185 | 94.0 |
| 30 | 4.472 | 6.012 | 0.1 | -2.8 | 23.575 | 94.8 |
| 35 | 4.522 | 6.061 | 1.2 | -2.2 | 23.949 | 95.1 |
| 40 | 4.875 | 5.895 | 9.1 | 2.6 | 24.468 | 96.0 |
| 45 | 4.519 | 6.156 | 1.1 | -0.2 | 24.871 | 95.3 |
| 50 | 4.725 | 6.427 | 5.7 | -4.1 | 25.013 | 93.6 |
| 55 | 4.206 | 6.614 | -5.9 | -4.5 | 25.659 | 94.2 |
| 60 | 4.850 | 6.864 | 8.5 | -6.0 | 26.241 | 93.3 |
| 65 | 4.411 | 6.584 | -1.3 | 1.6 | 27.243 | 95.8 |
| 70 | 4.497 | 6.889 | 0.6 | -1.6 | 27.661 | 95.2 |
| 75 | 4.589 | 7.460 | 2.7 | -6.1 | 28.635 | 94.1 |
| 80 | 4.838 | 7.255 | 8.2 | -1.3 | 29.436 | 95.7 |
| 85 | 4.742 | 8.013 | 6.1 | -7.9 | 30.729 | 93.7 |
| 90 | 4.334 | 8.209 | -3.0 | -4.9 | 32.493 | 93.6 |
| 95 | 4.709 | 8.798 | 5.4 | -4.7 | 34.498 | 94.2 |

SES: simulation standard error

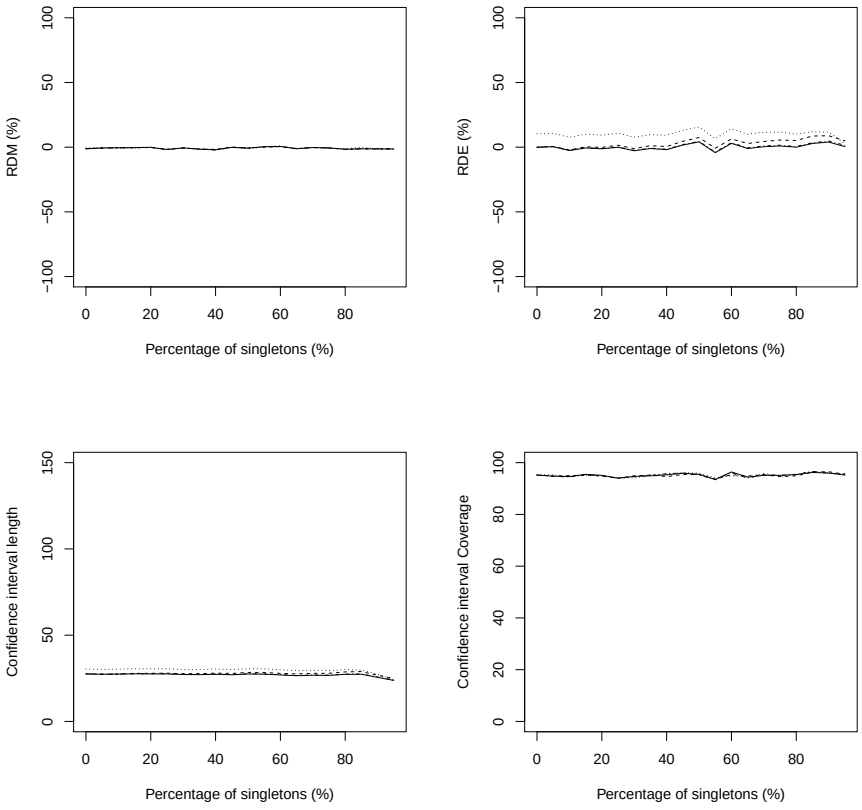
RDM: relative difference between estimated and true mean

RDE: relative difference between estimated and true standard error



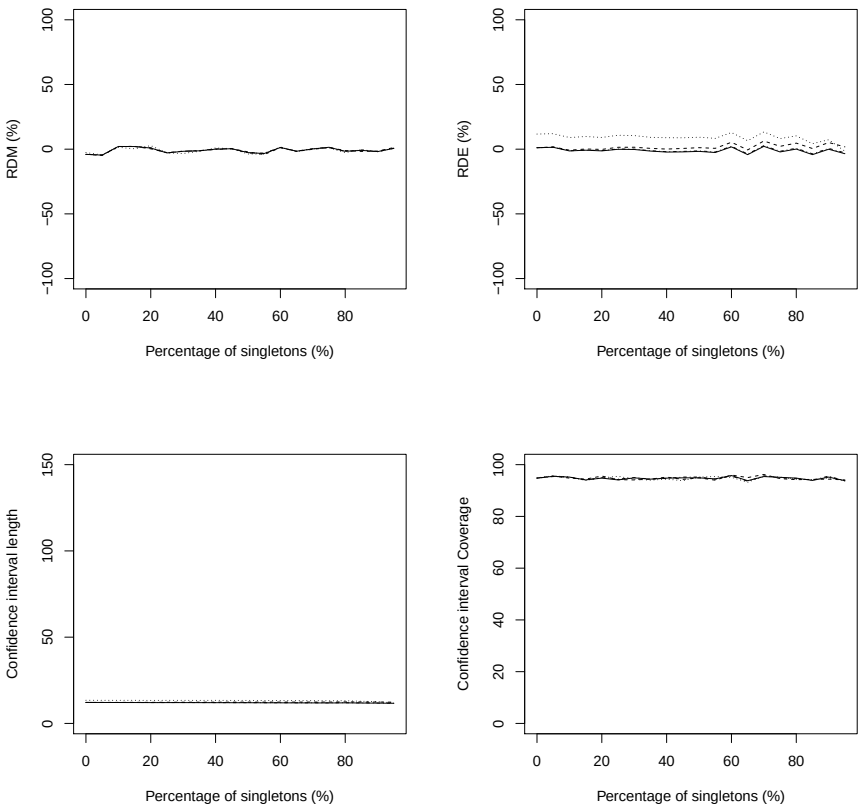

RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure A3: Performance measures for the fixed effect $Age_{ij}$ when ignoring (dotted lines), deleting (dashed lines) or grouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model on the original data (full lines).



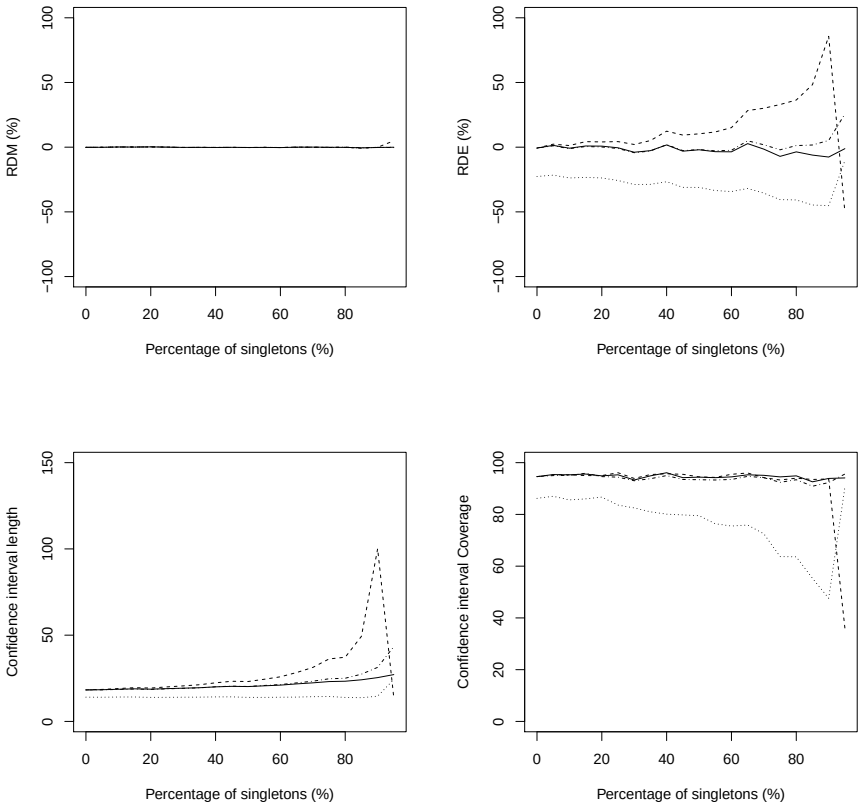

RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure A4: Performance measures for the fixed effect $Reason_{2ij}$ when ignoring (dotted lines), deleting (dashed lines) or grouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model on the original data (full lines).



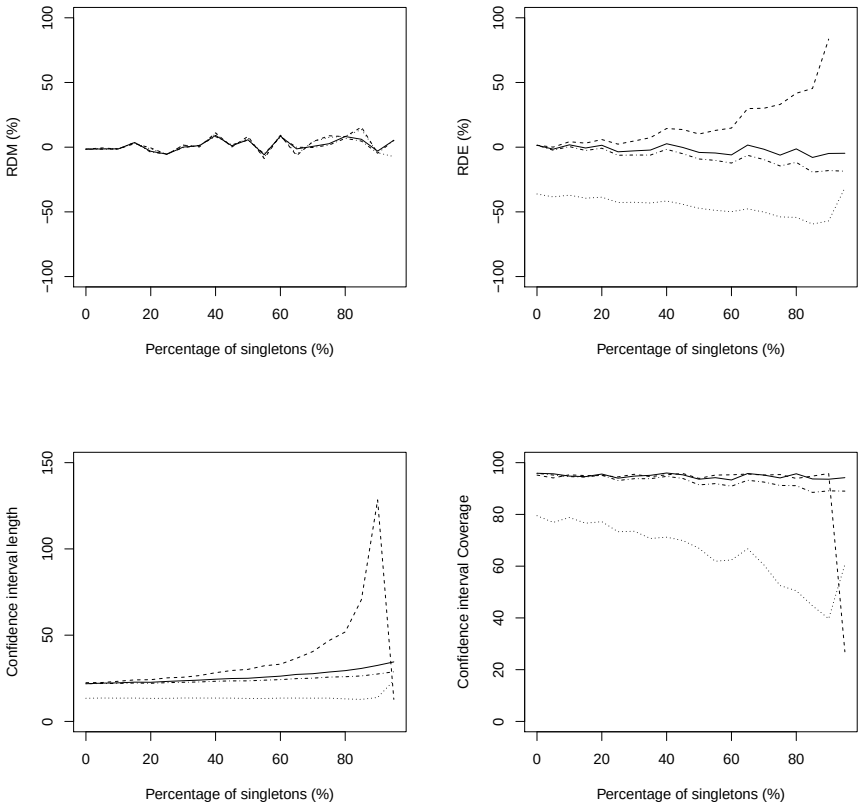

RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure A5: Performance measures for the fixed effect $Reason_{3ij}$ when ignoring (dotted lines), deleting (dashed lines) or grouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model on the original data (full lines).




RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure A6: Performance measures for the fixed intercept when ignoring (dotted lines), deleting (dashed lines) or grouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model on the original data (full lines).




RDM: relative difference between estimated and true mean
RDE: relative difference between estimated and true standard error

Figure A7: Performance measures for the fixed effect $Size_{2j}$ when ignoring (dotted lines), deleting (dashed lines) or grouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model on the original data (full lines).

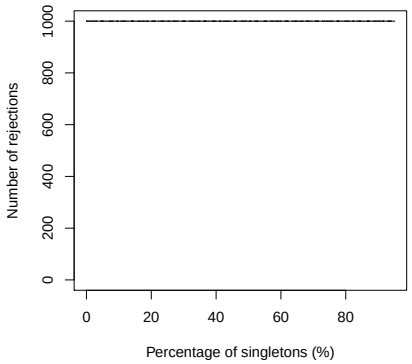



Figure A8: F test rejection rates for fixed effect *Age* in the model when ignoring (dotted lines), dropping (dashed lines) or regrouping (dot-dashed lines) the singletons, compared to performance measures for the three covariates model fitted to the original data (full lines).

Table A10: True and average parameter estimates (standard deviations) of the fixed effects estimates in Scenarios 1-9 under $H_0$ and $H_a$.

| | | True values | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
|---|---|---|---|---|---|---|---|
| under $H_0$ | Intercept | 70.68 (15.01) | 72.23 (14.35) | 71.68 (14.09) | 71.75 (14.39) | 71.87 (14.34) | 71.38 (13.45) |
| | Region 1 | 0 (21.23†) | -0.89 (19.98) | -0.19 (17.41) | 0.06 (18.04) | -0.07 (17.48) | 0.45 (17.87) |
| | Region 2 | 0 (21.23†) | -0.09 (20.11) | 0.32 (19.56) | 0.50 (17.65) | 0.20 (17.89) | 0.63 (17.17) |
| | Region 3 | 0 (21.23†) | -0.28 (19.45) | 0.10 (19.97) | -0.31 (20.86) | -0.05 (17.39) | 0.07 (16.99) |
| | Region 4 | 0 (21.23†) | -0.64 (19.66) | -0.05 (19.91) | 0.18 (20.36) | -0.09 (20.69) | 0.89 (16.85) |
| under $H_a$ | Intercept | 98.78 (15.01) | 98.86 (15.60) | 98.19 (15.02) | 99.03 (15.22) | 99.11 (15.75) | 99.69 (15.17) |
| | Region 1 | -41.70 (21.23†) | -39.79 (20.96) | -39.06 (17.69) | -39.44 (18.03) | -38.97 (18.76) | -39.61 (18.31) |
| | Region 2 | -35.92 (21.23†) | -34.65 (20.54) | -33.83 (20.33) | -34.75 (18.42) | -34.34 (18.63) | -35.37 (18.00) |
| | Region 3 | -28.10 (21.23†) | -27.24 (21.44) | -26.72 (20.71) | -28.19 (20.22) | -27.38 (18.86) | -27.99 (18.41) |
| | Region 4 | -14.11 (21.23†) | -14.04 (21.38) | -13.29 (21.26) | -14.37 (21.88) | -14.95 (21.88) | -14.81 (18.69) |

| | | True values | Scenario 6 | Scenario 7 | Scenario 8 | Scenario 9 |
|---|---|---|---|---|---|---|
| under $H_0$ | Intercept | 70.68 (15.01) | 71.99 (10.89) | 72.16 (8.61) | 71.54 (7.51) | 71.56 (6.31) |
| | Region 1 | 0 (21.23†) | -0.68 (14.46) | 0.35 (12.70) | 0.64 (11.02) | -0.28 (8.75) |
| | Region 2 | 0 (21.23†) | -0.35 (15.25) | -0.13 (12.72) | -0.09 (10.37) | 0.31 (8.85) |
| | Region 3 | 0 (21.23†) | -0.58 (15.43) | -0.33 (12.34) | -0.02 (10.82) | 0.34 (8.86) |
| | Region 4 | 0 (21.23†) | -0.65 (15.44) | -0.59 (12.41) | -0.19 (10.94) | 0.08 (8.87) |
| under $H_a$ | Intercept | 98.78 (15.01) | 98.60 (11.33) | 99.18 (9.47) | 98.52 (8.60) | 98.73 (6.62) |
| | Region 1 | -41.70 (21.23†) | -38.93 (14.99) | -39.70 (12.67) | -39.12 (11.52) | -38.95 (8.89) |
| | Region 2 | -35.92 (21.23†) | -34.05 (15.20) | -34.46 (12.77) | -34.44 (11.52) | -33.95 (8.95) |
| | Region 3 | -28.10 (21.23†) | -26.69 (15.35) | -28.15 (13.16) | -26.98 (11.74) | -27.34 (9.10) |
| | Region 4 | -14.11 (21.23†) | -14.02 (15.83) | -14.35 (13.01) | -13.20 (12.14) | -14.06 (9.13) |

† *True standard deviations are reported for regions containing one country. True standard deviations for regions with other compositions can easily be calculated by replacing $N_{Int}$ and $N_{Reg}$ with the desired number of countries in the following formula:* $SD_{new} = \sqrt{\left(15.01^2/N_{Int}\right) + \left(15.01^2/N_{Reg}\right)}$
*where $N_{Int}$ represents the number of countries in Region 5 and $N_{Reg}$ represents the number of countries in Region i with i ranging from 1 to 4.*

Table A11: True and average parameter estimates (standard deviations) of the fixed effects estimates in Scenario 10 for ML and REML under $H_0$ and $H_a$.

| | Under $H_0$ | | | Under $H_a$ | | |
|---|---|---|---|---|---|---|
| | True values | ML | REML | True values | ML | REML |
| Intercept | 70.68 (15.01) | 71.58 (9.20) | 71.55 (8.86) | 62.86 (15.01) | 64.75 (8.92) | 64.73 (8.54) |
| Region 1 | 0 (16.44) | -0.11 (21.44) | -0.08 (21.30) | 6.67 (16.44) | 5.32 (20.61) | 5.33 (20.46) |
| Region 2 | 0 (16.44) | -0.22 (17.22) | -0.19 (17.07) | 6.31 (16.44) | 5.84 (17.02) | 5.86 (16.79) |
| Region 3 | 0 (16.44) | -0.37 (17.21) | -0.34 (16.94) | 35.92 (16.44) | 34.68 (17.63) | 34.69 (17.45) |
| Region 4 | 0 (16.44) | 0.80 (18.34) | 0.83 (18.31) | -5.78 (16.44) | -4.12 (17.10) | -4.10 (16.80) |
| Region 5 | 0 (10.96) | 0.48 (14.39) | 0.42 (14.01) | -2.89 (10.96) | -2.53 (13.47) | -2.58 (13.09) |
| Region 6 | 0 (16.44) | 0.51 (17.34) | 0.54 (17.15) | 8.75 (16.44) | 7.40 (17.00) | 7.41 (16.82) |
| Region 7 | 0 (10.07) | -0.16 (13.26) | -0.11 (12.56) | 2.62 (10.07) | 1.93 (12.76) | 1.98 (12.17) |
| Region 8 | 0 (16.44) | 0.09 (17.45) | 0.12 (17.27) | 21.81 (16.44) | 20.96 (18.03) | 20.98 (17.86) |
| Region 9 | 0 (12.56) | 0.68 (14.98) | 0.54 (14.26) | 8.10 (12.56) | 6.95 (14.96) | 6.92 (13.93) |

Table A12: Average parameter estimates and standard deviations of the fixed effects estimates in the three additional scenarios for Scenario 3 under ML and REML (under $H_0$).

| | | Scenario 3 | Scenario 3b | Scenario 3d | | | Scenario 3c |
|---|---|---|---|---|---|---|---|
| Intercept | ML | 71.75 (14.39) | 72.26 (10.37) | 71.75 (14.39) | Intercept | ML | 71.71 (8.08) |
| | REML | 71.75 (14.39) | 72.26 (10.37) | 71.75 (14.39) | | REML | 71.71 (8.08) |
| Region 1 | ML | 0.06 (18.04) | -0.44 (14.97) | 0.06 (18.04) | Group 1 | ML | 0.10 (13.59) |
| | REML | 0.06 (18.04) | -0.44 (14.97) | 0.06 (18.04) | | REML | 0.10 (13.59) |
| Region 2 | ML | 0.50 (17.65) | | 0.50 (17.65) | Group 2 | ML | 0.55 (13.24) |
| | REML | 0.50 (17.65) | | 0.50 (17.65) | | REML | 0.55 (13.24) |
| Region 3 | ML | -0.31 (20.86) | | -0.31 (20.86) | | | |
| | REML | -0.31 (20.86) | | -0.31 (20.86) | | | |
| Region 4 | ML | 0.18 (20.36) | | 0.18 (20.36) | | | |
| | REML | 0.18 (20.36) | | 0.18 (20.36) | | | |

*Note: The intercept in Scenario 3b corresponds to Region 2 from Scenario 3.* .

*The intercept in Scenario 3c contains all singletons from Scenario 3* .

Table A13: Average parameter estimates and standard deviations of the fixed effects estimates in the three additional scenarios for Scenario 3 under ML and REML (under $H_a$).

| | | Scenario 3 | Scenario 3b | Scenario 3d | | | Scenario 3c |
|---|---|---|---|---|---|---|---|
| Intercept | ML | 99.03 (15.22) | 64.28 (10.20) | 99.03 (15.22) | Intercept | ML | 84.84 (8.60) |
| | REML | 99.03 (15.22) | 64.28 (10.20) | 99.03 (15.22) | | REML | 84.84 (8.60) |
| Region 1 | ML | -39.44 (18.03) | -4.69 (13.85) | -39.44 (18.03) | Group 1 | ML | -25.25 (12.91) |
| | REML | -39.44 (18.03) | -4.69 (13.85) | -39.44 (18.03) | | REML | -25.25 (12.91) |
| Region 2 | ML | -34.75 (18.42) | | -34.75 (18.42) | Group 2 | ML | -20.56 (13.18) |
| | REML | -34.75 (18.42) | | -34.75 (18.42) | | REML | -20.56 (13.18) |
| Region 3 | ML | -28.19 (20.22) | | -28.19 (20.22) | | | |
| | REML | -28.19 (20.22) | | -28.19 (20.22) | | | |
| Region 4 | ML | -14.37 (21.88) | | -14.37 (21.88) | | | |
| | REML | -14.37 (21.88) | | -14.37 (21.88) | | | |

*Note: The intercept in Scenario 3b corresponds to Region 2 from Scenario 3.* .

*The intercept in Scenario 3c contains all singletons from Scenario 3* .

Table A14: Average parameter estimates and standard deviations of the fixed effects estimates in the three additional scenarios for Scenario 10 under ML and REML (under $H_0$).

| | | Scenario 10 | Scenario 10b | Scenario 10d | | | Scenario 10c |
|---|---|---|---|---|---|---|---|
| Intercept | ML | 71.58 (9.20) | 71.57 (8.97) | 71.59 (9.16) | Intercept | ML | 71.55 (8.89) |
| | REML | 71.55 (8.86) | 71.55 (8.86) | 71.56 (8.91) | | REML | 71.53 (8.84) |
| Region 1 | ML | -0.11 (21.44) | | -0.12 (21.41) | Group 1 | ML | 0.12 (10.90) |
| | REML | -0.08 (21.30) | | -0.09 (21.32) | | REML | 0.16 (10.83) |
| Region 2 | ML | -0.22 (17.22) | | -0.23 (17.20) | | | |
| | REML | -0.19 (17.07) | | -0.20 (17.09) | | | |
| Region 3 | ML | -0.37 (17.21) | | -0.38 (17.16) | | | |
| | REML | -0.34 (16.94) | | -0.35 (16.98) | | | |
| Region 4 | ML | 0.80 (18.34) | | 0.79 (18.34) | | | |
| | REML | 0.83 (18.31) | | 0.82 (18.31) | | | |
| Region 5 | ML | 0.48 (14.39) | 0.45 (14.17) | 0.47 (14.34) | Group 2 | ML | 0.45 (13.94) |
| | REML | 0.42 (14.01) | 0.44 (14.01) | 0.44 (14.06) | | REML | 0.42 (13.87) |
| Region 6 | ML | 0.51 (17.34) | | 0.51 (17.33) | | | |
| | REML | 0.54 (17.15) | | 0.54 (17.20) | | | |
| Region 7 | ML | -0.16 (13.26) | -0.11 (12.82) | -0.16 (13.17) | Group 3 | ML | -0.09 (12.62) |
| | REML | -0.11 (12.56) | -0.11 (12.57) | -0.13 (12.68) | | REML | -0.06 (12.51) |
| Region 8 | ML | 0.09 (17.45) | | 0.08 (17.44) | | | |
| | REML | 0.12 (17.27) | | 0.11 (17.29) | | | |
| Region 9 | ML | 0.68 (14.98) | 0.59 (14.44) | 0.63 (14.91) | Group 4 | ML | 0.54 (14.33) |
| | REML | 0.54 (14.26) | 0.55 (14.26) | 0.60 (14.31) | | REML | 0.53 (14.26) |

*Note: The intercept in Scenario 10b corresponds to the intercept from Scenario 10.*
*Group 1 in Scenario 10c contains all singletons from Scenario 10* .

Table A15: Average parameter estimates and standard deviations of the fixed effects estimates in the three additional scenarios for Scenario 10 under ML and REML (under $H_a$).

| | | Scenario 10 | Scenario 10b | Scenario 10d | | | Scenario 10c |
|---|---|---|---|---|---|---|---|
| Intercept | ML | 64.75 (8.92) | 64.73 (8.70) | 64.74 (8.86) | Intercept | ML | 64.75 (8.47) |
| | REML | 64.73 (8.54) | 64.74 (8.55) | 64.74 (8.60) | | REML | 64.75 (8.44) |
| Region 1 | ML | 5.32 (20.61) | | 5.33 (20.58) | Group 1 | ML | 12.40 (10.49) |
| | REML | 5.33 (20.46) | | 5.33 (20.47) | | REML | 12.23 (10.44) |
| Region 2 | ML | 5.84 (17.02) | | 5.86 (17.00) | | | |
| | REML | 5.86 (16.79) | | 5.86 (16.83) | | | |
| Region 3 | ML | 34.68 (17.63) | | 34.69 (17.59) | | | |
| | REML | 34.69 (17.45) | | 34.69 (17.49) | | | |
| Region 4 | ML | -4.12 (17.10) | | -4.10 (17.06) | | | |
| | REML | -4.10 (16.80) | | -4.11 (16.85) | | | |
| Region 5 | ML | -2.53 (13.47) | -2.55 (13.21) | -2.52 (13.41) | Group 2 | ML | -2.56 (13.02) |
| | REML | -2.58 (13.09) | -2.59 (13.09) | -2.56 (13.13) | | REML | -2.58 (13.05) |
| Region 6 | ML | 7.40 (17.00) | | 7.41 (16.96) | | | |
| | REML | 7.41 (16.82) | | 7.41 (16.85) | | | |
| Region 7 | ML | 1.93 (12.76) | 1.96 (12.43) | 1.94 (12.68) | Group 3 | ML | 1.95 (12.12) |
| | REML | 1.98 (12.17) | 1.98 (12.17) | 1.95 (12.20) | | REML | 1.93 (12.12) |
| Region 8 | ML | 20.96 (18.03) | | 20.98 (18.00) | | | |
| | REML | 20.98 (17.86) | | 20.97 (17.90) | | | |
| Region 9 | ML | 6.95 (14.96) | 6.91 (14.29) | 6.98 (14.79) | Group 4 | ML | 6.94 (13.83) |
| | REML | 6.92 (13.93) | 6.92 (13.93) | 6.94 (14.05) | | REML | 6.92 (13.80) |

*Note: The intercept in Scenario 10b corresponds to the intercept from Scenario 10.*
*Group 1 in Scenario 10c contains all singletons from Scenario 10* .

# Summary

Antibiotics are drugs that are used to treat bacterial infections. Over time, the use and misuse of antibiotics has lead to resistance of bacteria to several antibiotics. One part of the solution to this worldwide public health problem is to gather trustworthy information on antibiotic use and its relationship with resistance.

With an incidence of 30 to 50 cases per 1000 patients per year, acute cough is one of the main reasons for consulting in primary care. Although antibiotic treatment for acute cough cases has been shown to have little or no effect, antibiotics are prescribed to over 50% of patients. In Chapter 2, we used data on adults presenting to primary care with acute cough in six European countries (i.e. Belgium, the Netherlands, the UK, Germany, Poland and Spain) to develop a prediction rule for poor prognosis (i.e. admission to hospital or reconsultation with new or worsened complaints), which will enable general practitioners to reassure patients at low risk and provide appropriate treatment for patients at high risk.
In order to account for the fact that different countries have a different baseline probability to experience poor prognosis, they were grouped according to this baseline risk (in three groups: $< 15\%$, $15 - 25\%$ and $> 25\%$). Missing values were imputed using multiple imputations by chained equations, and a combination of group-specific conditional inference trees and logistic regression were used to construct a new prediction rule. Important predictors were the country's baseline risk for poor prognosis, the presence of crackles during physical examination, the severity of phlegm as assessed by the patient, the severity of interference with daily activities, the time since the patient last smoked and the patient's diastolic blood pressure. Including measurements of C-reactive protein or blood urea nitrogen in this prediction rule did not improve its discriminative performance. The new prediction rule was shown to be quite stable in cross-validation, and to outperform available prediction rules, although there is room for further improvement.

Data that are used to study antibiotic consumption in hospitalized patients often have a hierarchical structure, with patients nested in departments, nested in hospitals, nested in countries. Such a complex multi-level structure automatically gives rise to sparseness issues caused by the low number of subunits at different levels of the hierarchy. Whenever a higher-level unit contains only one subunit, this unit is referred to as a singleton. In Chapter 3, we used a simulation study to evaluate the performance of the linear multi-level model in the presence of singletons at the lowest level (i.e. the child). Performance was assessed using four different performance characteristics which revealed that neither the relative difference between the estimated and true mean nor the relative difference between estimated and true standard error were affected by the percentage of singletons in the data. The width of the confidence interval fluctuated for an effect at the level of the child, while it increased with the percentage of singletons for an effect at the level of the department. The coverage approached 95% for explanatory variables at both levels and for varying percentages of singletons. Because some of the simulated datasets contain a fairly high proportion of singletons, one might decide to either ignore the dependency within clusters (ignoring singletons), remove the singletons from the data (dropping singletons) or group them into an artificial department (regrouping singletons). Using a simulation study we showed that ignoring and dropping the singletons should be avoided as they come with low coverage and wide confidence intervals, respectively. Regrouping the singletons into an artificial department might be considered, although the regular linear multi-level model performs better even when the percentage of singletons increases. If the data at hand contain a high percentage of singletons at the lowest level of the hierarchy, the linear multi-level model can be used safely.

In Chapter 4, we used a simulation study to evaluate the performance of the F test in the presence of singletons at the highest level (i.e. the macro-geographical region). Performance was assessed using type I error rate, power and corrected power for scenarios simulated under the null or a specific alternative hypothesis, and showed that the F test performs well when there are no singletons at the highest level, while it performs inadequately when there are. We studied the impact of deleting, regrouping or splitting the singletons and demonstrated that they could solve either the problem of a high type I error rate or the problem of low power, while worsening the other. As alternatives to the F test, we considered the Wald test, likelihood ratio test and permutation test. While performance of the Wald and likelihood ratio test were comparable to the performance of the F test, the permutation test outperformed the F test. We therefore recommend to use the permutation test to determine significance of a fixed effect at the primary level in a multi-level model with a continuous outcome

suffering from primary unit sparseness. Ideally, a permutation test should also be used to determine the significance of fixed effects at lower levels.
In Chapter 5, we used the findings from the two previous chapters and data on antibiotic doses prescribed to hospitalized children to determine which factors are causing variation in doses of $\beta$-lactam antibiotics prescribed to hospitalized children. We used a linear multi-level model, correcting for the presence of two different prescribing styles, to determine the causes of variation in prescribed ceftriaxone doses and showed that doses of ceftriaxone are lower when prescribed to children receiving empiric treatment, with lower weight and with less severe reason for treatment. Using a meta-model for the $\beta$-lactam antibiotics that includes a fixed effect for the type of antibiotic and an interaction between each predictor and the type of antibiotic, we showed that most predictors acted on an antibiotic-specific basis. Variables that influenced the 12 included $\beta$-lactam antibiotics in a similar fashion were type of treatment and type of hospital, with lower doses prescribed in primary or secondary hospitals (compared to tertiary and specialized hospitals) and for empiric treatment (compared to targeted treatment).

One factor influencing antibiotic dosing that could not be assessed in Chapter 5 is time. In Chapter 6, we used information on outpatient antibiotic consumption reported quarterly and expressed in the number of defined daily doses or packages per 1000 inhabitants per day to assess the evolution of outpatient antibiotic use over time. Using a non-linear mixed model, that includes a sine wave to model the seasonal variation of antibiotic consumption, we showed that conclusions based on both measures can be contradictory, with antibiotic use expressed in defined daily doses showing a significant increase over time while antibiotic use expressed in packages did not change significantly. The agreement between random effects for both models was studied by combining the final models into one joint model, which showed that matching intercepts, slopes and amplitudes were highly correlated, albeit not perfectly. Using a linear mixed model, we showed that the number of defined daily doses per package increased significantly over time.
In Chapter 7 we linked information on outpatient antibiotic use reported yearly with information on resistance levels reported yearly to investigate whether the number of defined daily doses or the number of packages best explains the association. Using a generalized linear mixed model with different time lags, we showed that the association between $\beta$-lactam use and resistance is modelled best without a time lag and using the number of packages alone, while the association between the use of tetracycline, macrolide, lincosamide and streptogramin and resistance needs a one year time

lag and both the number of defined daily doses and the number of packages.
In Chapter 8, we focussed on resistance profiles in streptococci which reside asymptomatically in the oropharynx and compared the persistence of resistance after exposure to penicillins and cephalosporins or macrolides and tetracyclines using data on individual antibiotic consumption and resistance status. Using a generalized estimating equations model, corrected for antibiotic- and bacteria-specific baseline resistance, we showed that the rate of increase in the odds of being susceptible did not differ after treatment with penicillins and cephalosporins or macrolides and tetracyclines. The proportion of susceptible isolates directly after treatment was significantly different, which implies that it would take longer for the proportion of susceptible isolates to recover after treatment with marcolides and tetracyclines than after treatment with penicillins and cephalosporins.

In Belgium, several attempts have been made to lower resistance rates by optimizing antibiotic consumption through e.g. the introduction of antimicrobial management teams in hospitals. In Chapter 9, we fitted a generalized linear mixed model containing change-points to data on yearly antibiotic consumption within Belgian hospitals to determine the impact of these policies on three selected quality indicators. We showed that the proportion of compliant antibiotic prescriptions for limb surgery at patient level was changed suddenly by both the establishment of antimicrobial management teams and the 2001 antimicrobial awareness campaign. The number of defined daily doses per 100 hospital days for pneumonia (at hospital level) was changed suddenly by both the 2001 antibiotic awareness campaign and the new financing mechanism in 2006. The ratio of oral over parenteral use for pneumonia (at hospital level) was changed suddenly by the new financing mechanism in 2006. No additional change-points were required in the models and a time lag for implementation of these changes in different hospitals was redundant. Although all these sudden changes were statistically significant, only the decrease in the number of defined daily doses per 100 hospital days for pneumonia seen in 2001 was steep enough to be clinically relevant.

# Samenvatting

Antibiotica zijn geneesmiddelen die gebruikt worden om bacteriële infecties te bestrijden. Doorheen de jaren heeft het gebruik, en misbruik, van antibiotica ertoe geleid dat vele bacteriën resistent geworden zijn tegen bepaalde antibiotica. Een deel van de oplossing voor dit probleem is de verzameling van betrouwbare informatie omtrent antibioticagebruik en de relatie met resistentie.

Met een incidentie van 30 tot 50 gevallen per 1000 patiënten per jaar is acute hoest een van de voornaamste reden om een huisarts te consulteren. Hoewel een behandeling met antibiotica in dit geval geen tot weinig effect heeft, worden antibiotica voorgeschreven in meer dan 50% van de gevallen. In Hoofdstuk 2, gebruikten we data rond volwassenen die met acute hoestklachten naar de huisarts gingen in zes Europese landen (nl. België, Nederland, het Verenigd Koninkrijk, Duitsland, Polen en Spanje) om een predictieregel voor slechte prognose (zijnde opname in het ziekenhuis of een tweede consultatie met nieuwe of erger geworden klachten) te ontwikkelen, die huisartsen zal helpen om patiënten met laag risico op slechte prognose gerust te stellen en een gepaste behandeling op te starten bij patiënten met hoog risico op slechte prognose.
Omdat het basisrisico op slechte prognose verschilt per land, werden landen gegroepeerd volgens dit basisrisico (in drie groepen: $< 15\%$, $15 - 25\%$ en $> 25\%$). Ontbrekende observaties werden geïmputeerd, en de nieuwe predictieregel werd opgebouwd met behulp van groep-specifieke modellen. Predictoren die belangrijk zijn in het voorspellen van slechte prognose zijn het basisrisico op slechte prognose, de aanwezigheid van crepitaties, de ernst van ophoesten van slijmen beoordeeld door de patiënt, de hinder die ondervonden wordt bij het uitvoeren van dagelijkse activiteiten (vb. werk, huishouden), de tijd sinds de patiënt stopte met roken en de diastolische bloeddruk. Het in rekening brengen van de concentratie aan C-reactief proteïne of ureum in het bloed verbeterden de predictieregel niet. De nieuwe predictieregel was stabiel, en een verbetering ten opzichte van bestaande

predictieregels, hoewel er nog ruimte is voor verdere verbetering.

Gegevens die worden gebruikt om antibioticagebruik bij gehospitaliseerde patiënten te bestuderen hebben vaak een hiërarchisch karakter, waarbij patiënten genest zijn binnen departementen, die genest zijn binnen hospitalen, die genest zijn binnen landen. Zulke complexe hiërarchische structuren brengen automatisch problemen met zich mee, veroorzaakt door het lage aantal subeenheden binnen eenheden op verschillende niveaus in de hiërarchie. Wanneer een eenheid maar één subeenheid bevat wordt deze eenheid een singleton genoemd. In Hoodstuk 3 evalueerden we de stabiliteit van het multi-level model in aanwezigheid van singletons op het laagste niveau (zijnde het kind) door middel van een simulatiestudie. Stabiliteit werd beoordeeld door vier verschillende karakteristieken, die aantoonden dat noch het relatieve verschil tussen het geschatte en ware gemiddelde, noch het verschil tussen de geschatte en ware standaardfout beïnvloed werden door het percentage singletons. De breedte van het betrouwbaarheidsinterval fluctueerde lichtjes voor een predictor op het niveau van het kind, maar steeg met het percentage singletons voor een effect op het niveau van het departement. De dekking van het betrouwbaarheidsinterval benaderde 95% voor predictoren op beide niveaus en voor verschillende percentages singletons. Omdat sommige van de gesimuleerde datasets een redelijk hoog percentage singletons bevatten, zou men kunnen besluiten om de afhankelijkheid binnen een unit (zijnde het departement) te negeren, de singletons weg te laten uit de analyse of de singletons te groeperen in een artificiële unit. Wij toonden aan dat het negeren van de afhankelijkheid binnen een unit of het weglaten van de singletons uit de analyse ten allen tijde moet vermeden worden. Deze opties gaan namelijk gepaard met een lage dekkingsgraad van het betrouwbaarheidsinterval of brede betrouwbaarheidsintervallen. Het hergroeperen van singletons in een artificiële unit kan overwogen worden, hoewel het algemene multi-level model het beter doet, zelfs bij een toegenomen percentage singletons. Als de beschikbare data een hoog percentage singletons op het laagste niveau bevatten, kan het multi-level model dan ook veilig worden toegepast.
In Hoofdstuk 4, gebruikten we een simulatiestudie om de stabiliteit van de F test te evalueren wanneer er singletons zijn op het hoogste niveau (zijnde de macro-geografische regio). Stabiliteit werd beoordeeld aan de hand van type I fout, power en gecorrigeerde power, voor scenarios die werden gesimuleerd onder de nul of een specifieke alternatieve hypothese. We toonden aan dat de F test een betrouwbare test is wanneer er geen singletons zijn, hoewel hij in de aanwezigheid van singletons ondermaats presteert. We bestudeerden het effect van het weglaten,

hergroeperen of opsplitsen van singletons en zagen dat deze opties een daling of een stijging veroorzaakten van zowel type I fout als power. Als alternatieven voor de F test bekeken we de Wald test, likelihood ratio test en permutatietest. Zowel de Wald test als de likelihood ratio test waren vergelijkbaar met de F test, terwijl de permutatietest veel beter deed. We raden dan ook aan om voor het testen van een fixed effect op het hoogste niveau van het multi-level model, een permutatietest te gebruiken. Idealiter wordt ook voor het testen van effecten op andere niveaus een permutatietest gebruikt.
In Hoofdstuk 5, maakten we gebruik van gegevens rond antibioticagebruik bij gehospitaliseerde kinderen om factoren die variabiliteit in voorgeschreven dosissen van $\beta$-lactam antibiotica veroorzaken te bepalen. We gebruikten een multi-level model, gecorrigeerd voor aanwezigheid van twee verschillende stijlen van voorschrijven, om te bepalen welke factoren variabiliteit in voorgeschreven ceftriaxone dosissen veroorzaakten en toonden aan dat lagere dosissen ceftriaxone worden voorgeschreven wanneer de behandeling empirisch is, het kind minder weegt of de reden voor behandeling niet ernstig is. Aan de hand van een meta-model voor $\beta$-lactam antibiotica, dat een fixed effect bevat voor het type antibiotica en een interactie tussen elke predictor en het type antibiotica, toonden we aan dat de meeste predictoren op een antibiotica-specifieke manier werken. Predictoren die de 12 $\beta$-lactam antibiotica op eenzelfde manier beïnvloeden zijn type van behandeling en type van hospitaal, waarbij een lagere dosis wordt voorgeschreven in primaire of secundaire ziekenhuizen (in vergelijking met tertiaire of gespecialiseerde ziekenhuizen) en voor empirische behandeling (in vergelijking met doelgerichte behandeling).

Een factor die we niet konden bestuderen in Hoofdstuk 5 is tijd. In Hoofdstuk 6 gebruikten we informatie rond antibioticagebruik bij ambulante patiënten, uitgedrukt in aantal dagelijks aanbevolen dosissen of aantal pakketjes, om de evolutie van antibioticagebruik over de tijd te bestuderen. We gebruikten een niet-lineair mixed model, dat een sinusfunctie bevat om seizoensvariatie te modelleren, en toonden aan dat conclusies op basis van dosissen en pakketjes contradictorisch zijn, aangezien gebruik uitgedrukt in aantal dosissen significant steeg terwijl gebruik uitgedrukt in aantal pakketjes niet significant veranderde. Aan de hand van een lineair mixed model toonden we aan dat het aantal dosissen per pakket over de tijd significant toegenomen is.
In Hoofdstuk 7 linkten we jaarlijkse gegevens rond antibioticagebruik bij ambulante patiënten met gegevens rond jaarlijkse resistentieniveaus. Met een generalized linear

mixed model toonden we aan dat de associatie tussen gebruik van $\beta$-lactam antibiotica en resistentie best gemodelleerd wordt met gegevens rond antibioticagebruik in hetzelfde jaar uitgedrukt in aantal pakketjes, terwijl de associatie tussen gebruik van tetracyclines, macrolides, licosamides en streptogramins en resistentie best gemodelleerd wordt met gegevens rond antibioticagebruik van het jaar voordien uitgedrukt in zowel aantal dosissen als aantal pakketjes.
In Hoofdstuk 8 gebruikten we gegevens rond individueel antibioticagebruik en resistentie om de persistentie van resistentie bij streptokokken die asymptomatisch aanwezig zijn in de oropharynx na behandeling met penicillines en cefalosporines of macroliden en tetracyclines te vergelijken. We gebruikten een generalized estimating equations model, gecorrigeerd voor antibiotica- en bacterie-specifieke basisresistentie, om aan te tonen dat de snelheid waarmee de proportie susceptibele isolaten zich herstelt niet verschilt na behandeling met penicillines en cefalosporines of macrolides en tetracyclines. Wat wel verschilt is de proportie susceptibele isolaten onmiddellijk na behandeling, die lager ligt na behandeling met macroliden en tetracyclines dan na behandeling met penicillines en cefalosporines, waardoor het na behandeling met macroliden en tetracyclines langer duurt om het basisniveau opnieuw te bereiken.

In België werden tal van pogingen gedaan om resistentie te verlagen door antibioticagebruik te optimaliseren, onder andere door introductie van antimicrobiële management teams in hospitalen. In Hoofdstuk 9 gebruikten we een generalized linear mixed model met change-points en data rond jaarlijks antibioticagebruik in Belgische ziekenhuizen om het effect van enkele initiatieven te bepalen. We toonden aan dat de proportie voorschriften bij operaties aan de ledematen die gebeuren volgens richtlijnen plots veranderden na introductie van antimicrobiële management teams en na een eerste bewustmakingscampagne in 2001. Het aantal dosissen per 100 hospitaaldagen (op hospitaalniveau) veranderde plots na de eerste bewustmakingscampagne in 2001 en de herziening van financiering in 2006. De ratio oraal over parenteraal gebruik voor pneumonie (op hospitaalniveau) veranderde plots door de herziening van financiering in 2006. In geen van de andere jaren waren er additionele change-points nodig, en het instellen van een tijdskader van $+1$ en $-1$ jaar rondom de change-points bleek overbodig. Hoewel al deze plotse veranderingen statistisch significant waren, is enkel de daling in het aantal dosissen per 100 hospitaaldagen voor pneumonie in 2001 fors genoeg om klinisch relevant te zijn.