Multilevel modelling in the analysis of observational datasets in the health care setting

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Matthias Michael Schwenkglenks

aus Geislingen an der Steige - Deutschland

Basel, 2007

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von Herrn Prof. Dr. Marcel Tanner, Herrn Prof. Dr. Thomas D. Szucs und Herrn Prof. Dr. Martin Schumacher

Basel, den 19. September 2006

Prof. Dr. Hans-Peter Hauri
Dekan

Meiner Familie und meinen Eltern gewidmet.

Table of contents

List of tables

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Appendix I

List of figures

Chapter 1

Chapter 6

Chapter 7

Appendix I

Acknowledgements

In the first place, I would like to thank all the contributors and all the financial supporters of the studies on which this work is based. This includes the individuals and companies explicitly acknowledged in the peer-reviewed publications which form chapters 4-6 of this thesis, but also many unnamed individuals. In particular, I would like to thank my co-authors for all their contributions, input, and support.

Moreover, I would like to thank all study participants, patients as well as insurance beneficiaries, who made this work at all possible.

Many stimulating discussions around the topic of chemotherapy-induced neutropenia involved Prof Gary Lyman and Dr Nicole Kuderer, Rochester, USA; Prof Leo Auerbach, Vienna, Austria; the members of the Impact of Chemotherapy in Neutropenia - European Study Group (INC-EU); and, above all others, Dr Ruth Pettengell, London, UK. I honour the memory of Mr Rhys Roberts, Hook, UK, organiser of the INC-EU, who died suddenly in February 2004. Without his immense efforts, the neutropenia part of this work would not have become reality.

I am very grateful to Prof Fritz R. Bühler, the founder and chairman of my institution European Center of Pharmaceutical Medicine (ECPM), University of Basel, Switzerland, for creating an excellent working environment and for all his encouragement and support. I am most grateful to Prof Thomas D. Szucs, co-chairman of ECPM and my superior for more than six years, for all his encouragement, continuous and unstinting support, and outstanding leadership qualities. Thank you also to my collaborators Mrs Majbrit Holm and Dr Nick Draeger, to Dr Annette Mollet, ECPM Programme Director, and to Mrs Amanda Pinto, ECPM Administrator, for all their support.

I would like to thank Dr Gorana Capkun-Niggli, Basel, Switzerland, for reviewing and discussing a draft version of this thesis, for enduring my questions about mathematical subtleties, and for her encouragement. Thank you also to Prof Sophia Rabe-Hesketh, Berkeley, USA, author of the *gllamm* multilevel modelling procedure, for her readiness to clarify technical details in a last minute situation.

My sincere thanks go to Prof Marcel Tanner, Basel, Switzerland; to Prof Thomas D. Szucs, Basel and Zürich, Switzerland (who is rightly mentioned a second time); and to Prof Martin Schumacher, Freiburg, Germany, the members of the thesis committee, for all their encouragement, feedback, suggestions, and support. A special thank you goes to Prof Schumacher for his critical questions which have helped me to detect a hidden problem in time.

Finally, my most heartfelt thanks go to my wife Sabine and my children Jonathan and Amélie for their love, endurance and support, and to my parents, to whom I owe everything.

Preface

This work is based on three health care-related observational studies conducted and/or analysed by the author. All three were primarily analysed using conventional multivariate regression methods, where the term conventional is used in the sense of not fully taking into account the hierarchical or multiple membership structure of the underlying datasets. In an additional step, re-analyses of all three datasets were performed using multilevel modelling, a novel statistical technique taking hierarchical data structures into account.

The thesis is structured as follows. An introduction (chapter 1) sketches the fields of research addressed by the author and establishes the rationale for taking hierarchical (multilevel) data structures into account when doing research work in these fields. Different ways of dealing with multilevel data are briefly compared. The introduction is followed by a brief section summarising overall objectives (chapter 2).

A general methods section (chapter 3) identifies the common statistical principles applied to the conventional analyses performed, and describes the approach to multilevel re-analysis. The issue of predictive ability is addressed in this context and a framework for comparing the results of the conventional regression analyses vs. multilevel analyses is described.

The designs, methodological details and conventional-based results of the empirical studies which form the basis of this work have been reported and discussed in peer-reviewed publications. These publications constitute the first three out of four results chapters (chapters 4-6). The fourth (chapter 7) summarises the results of multilevel re-analysis and compares these with the conventional-based findings.

Adopting a broader viewpoint, the overall discussion (chapter 8) addresses the contribution of the three observational studies reported, and of the multilevel modelling-based results in particular, to current knowledge in the respective fields of research. Multilevel vs. conventional results are put into perspective and some conclusions are drawn with respect to the use of multilevel modelling in health care-related research.

Three appendices describe details of the multilevel re-analyses, of the cross-validation techniques applied, and of the characteristics, advantages and disadvantages of the statistical software used for these purposes.

As chapters 4-6 of this thesis were previously published, it was decided to leave their counting of tables and figures untouched. Consequently, tables and figures are numbered chapter-wise. A list of tables and a list of figures are contained after the table of contents. All references are listed in alphabetical order at the end of the thesis. In addition, the references lists of the peer-reviewed publications were retained, but adjusted to the same format and numbering as in the main references list. Journal names were abbreviated according to the *List of Journals Indexed for MEDLINE* (2005; formerly *Index Medicus*). Journal names not contained in this list were not abbreviated.

Summary

In health care-related research, many studies circle around the problem of identifying risk factors for clinical events of interest, with a potential for economic consequences, or risk factors for increased health care costs. Multivariate regression methods are typically used to analyse such studies and have become central for an efficient control of confounding and assessment of effect modification.

However, most of the data used for this type of research are characterised by hierarchical (multilevel) data structures (e.g., patients are frequently nested within treating physicians or study centres). Standard multivariate regression methods tend to ignore this aspect and it has been shown that this may lead to a loss of statistical efficiency and, in some cases, to wrong conclusions. Multilevel regression modelling is an emerging statistical technique which claims to correctly address this type of data, and to make use of their full potential.

The author conducted and/or analysed three observational studies of factors associated with clinical events or cost endpoints of interest. In all cases, conventional regression methods were primarily used. In a second step, multilevel re-analyses were performed and the results were compared.

The first study addressed the effect of exacerbation status, disease severity and other covariates on the disease-specific health care costs of adult Swiss asthma patients. Among other factors, the occurrence of asthma exacerbations was confirmed to be independently associated with higher costs, and to interact with disease severity.

The second study addressed the impact of gatekeeping, a technique widely used to manage the use of health care resources, on the health care costs accrued by a general Swiss population. In a situation characterised by ambiguous research findings, the author's study indicated substantial cost savings through gatekeeping as opposed to fee-for-service based health insurance.

Finally, a combined dataset of six retrospective audits of breast cancer treatment from several Western European countries was used to estimate, for common chemotherapy regimen types, the frequency of chemotherapy-induced neutropenic events and to identify or confirm potential neutropenic event risk factors. Neutropenic events were shown to occur frequently in routine clinical practice. Several factors, including age, chemotherapy regimen type, planned chemotherapy dose intensity, and planned number of chemotherapy cycles, were shown to be potentially important elements of neutropenia risk models.

Multilevel re-analysis showed higher level variation (i.e., variation at the level of the treating physicians or study centres) to be present in the asthma dataset and in the neutropenia dataset, but not in the gatekeeping dataset. In the first-mentioned cases, multilevel modelling allowed to quantify the amount of higher level variation; to identify its sources; to identify spurious findings by analysing influential higher level units; to achieve a gain in statistical precision; and to achieve a modest gain in predictive ability for out-of-sample observations whose corresponding higher level units contributed to model estimation. The main conclusions of the conventional analyses were confirmed.

Based on these findings and in conjunction with published sources, it is concluded that multilevel modelling should be used systematically where hierarchical data structures are present, except if the higher level units must be regarded as distinct, unrelated entities or if their number is very small. Erroneous inferences will thus become more unlikely. Moreover, multilevel modelling is the only technique to date which allows to efficiently test hypotheses at different hierarchical levels, and hypotheses involving several levels, simultaneously. In the authors opinion, multilevel analysis is of particular interest where characteristics of health care providers, and clinical practice patterns in particular, may impact on health outcomes or health economic outcomes. It is only another facet of the same argument that multilevel modelling should also be used in multi-centre studies (including randomised clinical trials) to take into account study centre-specific characteristics and behaviours.

In many instances, the use of the technique will be tentative and rule out the presence of substantial higher level variation. If so, simpler methods can again be used.

Besides some technical issues, the main disadvantage of multilevel modelling is the complexity involved with the modelling process and with correctly interpreting the results. A careful approach is therefore needed. Multilevel modelling can be applied to datasets *post hoc*, as the author has done, but superior results can be expected from studies which are planned with the requirements of multilevel analysis (e.g., appropriate sample size, collection of relevant covariates at all hierarchical levels) in mind.

Chapter 1

Introduction

In modern epidemiology and health care-related research, many studies circle around the problem of identifying factors associated with the occurrence of clinical events of interest, which will often have a potential for severe medical and/or substantial economic consequences. Goals may include to describe and identify such factors as a basis for further analysis, to establish causality between these factors and the events of interest (i.e. to establish them as true risk factors), to develop statistical models predicting individual or group-level risk, and ultimately to design new interventions for the betterment of the underlying health problems [100:10-3·176]. In other studies using very similar methodology, the endpoint may not be clinical event occurrence but cost in a population with a given disease of interest, or in a general population. In populations with chronic diseases, identifying correlates of cost, together with clinical findings, may help targeting further research and resources efficiently and, moreover, provide important input information for health economic evaluation studies (i.e., cost-effectives and cost utility analyses) [208]. Where general populations are regarded, the interest typically is in the ability of health care financing or health systems interventions to contain or reduce health care costs (without affecting quality of care).

Where randomised trials are not feasible for ethical or practical reasons, e.g. in the study of treatments which are generally believed to be effective, of health risks and health behaviours, or of large-scale health system interventions, such research is often based on prospective or retrospective observational data collections. This thesis is based on three distinct studies falling into this group.

Cost of illness of asthma

The first study is a cost of illness analysis of adult Swiss asthma patients. Numerous studies have been conducted to describe the absolute and relative contribution of different cost items (e.g., medication costs, inpatient costs, indirect costs due to lost working days, etc.) to the cost of asthma, mostly in industrialised countries [11·79]. In contrast, the patient, disease, and treatment characteristics associated with high asthma costs were rarely and less systematically addressed. Based on a retrospective, medical charts-based dataset collected in 1996/1997, the author assessed correlates of direct medical asthma costs, with a special focus on the impact of disease severity and asthma control (chapter 4).

Gatekeeping vs. fee-for-service based health insurance

In the last three decades, most if not all industrialised countries were confronted with a situation where health care expenditures grew faster than gross domestic products [152]. This pronounced rise, caused by a mix of demand-side factors (e.g., demographic change, increased standard of living, enhanced access to information) and supply-side factors (new products and technologies) was rapidly perceived as a problem and a search for cost-saving interventions in the fields of health care financing, regulation, and organisation began [17·37]. The rise of managed care in the USA is the most widely known result of this process [155]. Various types of financial incentives, such as capitation, and techniques of utilisation management, such as utilisation review and gatekeeping, were newly introduced in a variety of countries [17·219]. Research addressing the impact of such interventions on health care costs and on quality of care gained in importance in parallel, but had to deal with several methodological challenges. In general populations, person-level health care costs vary widely, show a complex distribution, and are influenced by a multitude of health plan beneficiary characteristics (resulting in substantial differences in case-mix) and provider characteristics [51]. Where studies focus on a limited set of diseases or types of service, generalisability may be substantially reduced. Moreover, in most real-life settings, several interventions and techniques are applied jointly in various combinations [194]. This makes it difficult to separate out the effects of single techniques of interest and confronts the researchers with changing comparators.

Gatekeeping is one the most frequently used techniques of utilisation management [17·194]. In typical health plan-level gatekeeping arrangements, patients select a primary care physician who must authorise specialist referrals, expensive diagnostic procedures, and hospital admissions [194]. The primary aims are to improve coordination, to avoid wasteful use of resources, to protect patients from redundant or harmful treatments, and to foster continuity of care [68·73·137]. The author had the opportunity to compare a Swiss gatekeeping solution, one of the few worldwide where gatekeeping is used as stand-alone technique, with a classical fee-for-service plan (chapter 5).

Neutropenic events in breast cancer chemotherapy

Chemotherapy-induced neutropenia (CIN) and febrile neutropenia (FN) are regarded as the most common dose-limiting toxicities of modern anticancer chemotherapy. They may impact on short-term as well as long-term treatment outcomes [23·119·153·191]. Research addressing the incidence and risk factors of CIN, FN, and subsequent events (i.e., chemotherapy dose delays, chemotherapy dose reductions, hospitalisations) has gained in importance during the last 15 years. Related sources of information include

- clinical trials of anticancer treatments where CIN and FN are reported as adverse events [47];
- clinical trials of myelopoietic growth factors (colony-stimulating factors; CSFs), undertaken to assess the anti-neutropenic effects of this class of substances and typically performed under tightly controlled conditions (defined chemotherapy regimens; narrow eligibility criteria) [88·102·157·213·218];
- observational studies, undertaken prospectively or retrospectively, with the primary aims of assessing neutropenic event incidence in clinical practice and related risk factors, and with the ultimate goal of developing clinically applicable risk models [38·123·132·185];
- clinical trials specifically undertaken (very rarely) to confirm neutropenic event risk factors or validate risk models [171].

Despite some substantial efforts, no final conclusions have yet been reached in this field. (See Lyman et al. for a recent review [132]). The author used six retrospective audits of breast cancer treatment from several Western European countries to estimate neutropenic event incidences for the chemotherapy regimen types used, and to extract any information considered suitable to contribute to the discussion around neutropenic event risk factors. The audits used were essentially reflecting routine clinical practice, but had some limitations in terms of data availability (chapter 6).

Multivariate regression methods and multilevel data structures

Various types of multivariate regression techniques, appropriate for different types of response variables, have become standard methods in the analyses of observational and, increasingly, experimental data. Since computers have become available to rapidly execute estimation algorithms, their importance for an improved control of confounding and assessment of effect modification has become paramount. Consequently, such regression methods were also used in the primary analysis of the above-introduced studies.

However, most conventional regression methods, inclusive of multiple least squares regression, logistic regression, and generalised linear modelling, assume independence of the observations on which the regression is performed [105:3,5·116:116]. In health care-related research as well as in many other fields, this assumption is conceptually violated in a wide range of situations [83:1/1-2·99:1-2]. For example, there may be repeated measurements across time, nested within observed persons. Observed persons may be nested within families or geographical units, within physicians' offices, or within hospitals. Participants in multi-centre clinical trials are nested within study centres. More complex situations of multi-membership or cross-classification can, e.g., occur where persons are independently nested within physicians' offices and, at the same time, within insurance companies.

Ignoring such hierarchical (multilevel) data structures was and probably still is the rule. Typically, analysis uses *"conventional […] regression analysis with one dependent variable at the lowest (individual) level and a collection of explanatory variables from all available levels"* [105:4], which involves disaggregation of higher level variables [58]. To what extent this approach influences research findings cannot be answered in general but will depend on the *de facto* degree of violation of the independence assumption. Potential consequences of ignoring hierarchical data structures include the following.

- A decrease in statistical efficiency and inflated standard errors may occur, as no full use is made of the available information [83:1/2·169: 42-3]. On the other hand, violation of the independence assumption can lead to false low standard error estimates [9·58·83:1/2·105:3,5-6]. For both reasons, there is a risk of incorrect inferences regarding the existence of statistical associations and of incorrect decisions regarding the inclusion or exclusion of model parameters.

- Effects on the response variable occurring at different hierarchical levels cannot be appropriately identified and explained [83:1/11·105:3-4,6-7]. Related difficulties to interpret regression results may lead to an arbitrary choice of models and impact negatively on predictive ability [220].
- Variation of the response variable of interest occurring at the higher level(s) cannot be quantified at the population level, i.e. any statements are at best possible for the higher level units directly observed [83:1/3].
- Ultimately, wrong conclusions may occur [1·83:1/1-2].

Approaches to multilevel data structures

The problem of multilevel data structures has been discussed for at least two and a half decades [83:1/2] and there are a variety of ways to address it (Table 1). However, most of the proposed techniques provide only partial solutions and suffer from sub-optimal use of the available information. Some of them are suitable if the researchers' interest is restricted to either the lowest or the highest level in the hierarchy of observation, and allow to estimate unbiased standard errors. However, a meaningful quantification of higher level variation in the total population of interest is only achieved by random effects analysis or multilevel modelling, and only the latter technique is suitable to satisfactorily assess effects occurring at different levels or involving several levels [9·52·54].

Multilevel modelling

Multilevel modelling has been developed since the early 1980s [1·186:49]. It was first used in the educational sciences [83:1/1-3·105:8] but disseminated to other fields rapidly [54·105:8]. Recent advances in multilevel modelling software have further accelerated this development [91].

In health care-related research, the number of applications is still small if compared to the huge overall number of studies published [91], but growing exponentially (Figure 1).

Table 1. Statistical techniques to address multilevel and multi-membership data structures

Statistical technique	Advantages	Disadvantages	Remarks
Conventional regression techniques applied to the lowest-level observations, ignoring the multilevel data structure [54·58]	Familiar to most researchers.	Subject to all potential problems described on pp. 20-1. Ignores the potential importance of effects occurring at higher levels or involving several levels. Standard error estimates may be incorrect. May lead to "atomistic fallacy" [2·52].[a]	Ignoring multilevel data structures will not always have a negative impact.
Conventional regression techniques applied to data aggregated at the highest level [54·58]	Problem of non-independence of observations removed, hence standard error estimates unbiased.	Loss of information and, consequently, statistical power. No possibility to assess effects occurring at lower levels or involving several levels. May lead to "ecological fallacy" [52·173].[b]	May be suitable if the interest is only in the highest level.
Stratified regression (estimating a separate regression model for each higher level unit) [54]	Problem of non-independence of observations removed, hence standard error estimates unbiased.	No useful quantification of higher level variation. No possibility to assess effects occurring at the higher levels or involving several levels.	Inefficient if the number of higher level units is large or if the number of lower level units per higher level unit is small.
Use of dummy variables to represent the higher level units [9·54]	Allows for a limited assessment of effects occurring at a higher level or involving several levels.	No useful quantification of higher level variation. Only limited possibilities to assess effects occurring at higher levels or involving several levels. Residual non-independence is not taken into account, i.e. standard errors may still be biased.	Models a separate intercept for each higher level unit. Modelling separate covariate coefficients (slopes) requires additional interaction terms. May be suitable if the number of higher level units is small, but inefficient if the number of higher level units is large.
Sample survey techniques [9·105:5-6]	Incorporate the "design effect" (effect of clustered sampling) into the analysis, thus allowing for unbiased standard error estimates	No useful quantification of higher level variation. No possibility to assess effects occurring at higher levels or involving several levels.	May be suitable if the interest is only in the lowest level.

Table 1 ctd.

Statistical technique	**Advantages**	**Disadvantages**	**Remarks**
Population-average (marginal) models using the generalised estimating equations (GEE) approach [52·54·85]	Takes into account non-independence of observations, thus allowing for unbiased standard error estimates	No useful quantification of higher level variation. No possibility to assess effects occurring at higher levels or involving several levels. Random parameters treated as a "nuisance".	May be suitable if the interest is only in the lowest level.
ANOVA [58·99:2-3]	Distinguishes within- and between-higher level variance and estimates separate intercepts for the higher level units.	Fixed effects approach to higher level units. Therefore, no useful quantification of higher level variation in the total population of interest is achieved. Only limited possibilities to assess effects occurring at higher levels or involving several levels.	Can only be used with continuous, normally distributed responses. No possibility to model separate covariate effects (slopes). Inefficient if the number of higher level units is large.
Contextual analysis [54]	Allows for a limited assessment of effects occurring at the higher levels or involving several levels.	Fixed effects approach to higher level units. Therefore, no useful quantification of higher level variation in the total population of interest is achieved. Only limited possibilities to assess effects occurring at higher levels or involving several levels. Residual non-independence is not taken into account, i.e. standard errors may still be biased.	The term "contextual analysis" refers to extended conventional regression models which allow to assign separate fixed effects to separate higher level units, for the higher level predictors. Inefficient if the number of higher level units is large.

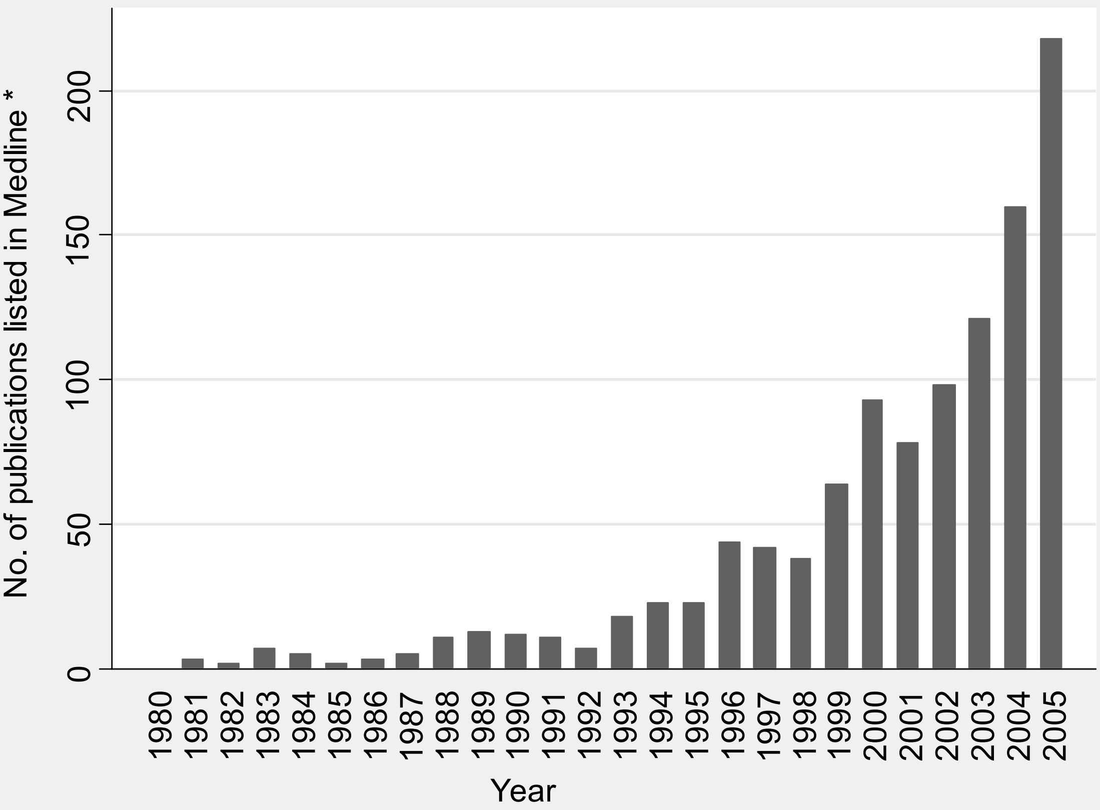
Table 1 ctd.

Statistical technique	Advantages	Disadvantages	Remarks
Random effects analysis (in the strict sense) [52]	Allows to partially assess effects occurring at different levels or involving several levels. Higher level units are treated as random samples (parts of a distribution). Therefore, a quantification of higher level variation in the total population of interest is achieved.	Only limited possibilities to assess effects occurring at the higher levels or involving several levels. Residual non-independence may not be fully taken into account, i.e. standard errors may still be biased.	Random effects analysis in the strict sense is equivalent to multilevel modelling allowing for random intercepts, but not for random covariate effects (slopes). Random effects analysis in the wider sense encompasses all regression techniques allowing for random effects, inclusive of full-scale multilevel models.
Multilevel analysis or multilevel modelling [58·83]	Allows to assess effects occurring at different levels or involving several levels. Higher level units are treated as random samples (parts of a distribution). Therefore, a quantification of higher level variation in the total population of interest is achieved. Allows for unbiased standard error estimates.	Modelling process and interpretation may be complex [83:1/11].	

[a] The term "atomistic fallacy" refers to wrong conclusions that may occur when inferences regarding associations at the group level are drawn from individual level data.

[b] The term "ecological fallacy" refers to the reverse situation, i.e. to wrong conclusions that may occur when inferences regarding associations at the individual level are drawn from group level (aggregate) data.

Figure 1. Frequency of use of multilevel modelling in health care-related research



* Search term: "multilevel modeling" OR "multilevel modelling" OR "multilevel model" OR "multilevel models" OR "hierarchical modeling" OR "hierarchical modelling" OR "hierarchical model" OR "hierarchical models" OR "multilevel assessment"

Multilevel modelling aims at combining and analysing information from different hierarchical levels within a single statistical model [105:7]. It can be regarded as a multilevel extension of multiple linear regression and other conventional multivariate regression techniques [105:8]. Fixed effects are estimated for all model covariates. In addition, random effects are estimated for the higher level covariates as needed, i.e. the variance-covariance structure of the data is analysed. Higher level effects are thus summarised by very few distribution parameters, instead of estimating separate, unrelated effects for each higher level unit [58]. In a second step, the higher level units are characterised by sets of higher level residuals derived using the empirical Bayes approach [33:57-8·52·91·99:7-9·161:27·186:221-35,245,247]. In other words, the values of the random effects (and the uncertainty around them) are estimated for each higher level unit, making use of the information directly available from this unit, but also of the information provided by all other units contributing to estimate the model (see chapter 3, pp. 36-7) [83:2/9-10·91·99:7-9·168·186:228]. This allows to compare the higher level units involved and may allow, for their corresponding lowest level observations, to predict the response of interest with increased precision [82·83:2/9-10·161:27·168·186:245]. Influences on the response of interest occurring at different levels, or involving several levels, can be identified and characterised [9·54·58]. The key underlying assumption is that the observed higher level units are not distinct and unrelated, but a random sample drawn from a wider population of higher level units [58·83:2·99:3].

A well-structured, relatively non-technical introduction to multilevel modelling, although no longer reflecting today's technical possibilities, has been published by Duncan, Jones and Moon [58]. As a starting point, the authors use the concepts of contextual effects (i.e., higher level effects such as geographical area effects impacting on individual-level health outcomes of interest) and compositional effects (i.e., apparent higher level effects which are in fact caused by an unequal distribution of influential individual-level factors across higher level units.)

Rationale to use multilevel modelling with the above-introduced studies

The datasets underlying the above-introduced cost of asthma, gatekeeping, and neutropenia studies are all characterised by multilevel or multi-membership data structures. Primary analysis as reported in chapters 4-6 did not take this into account, with the partial exception of the neutropenia study where robust standard errors were estimated using the generalised estimating equations (GEE) approach. Obvious reasons not to use multilevel modelling, such as very small numbers of higher level units [163:95], did not apply. Some systematic reviews of the use of multilevel modelling in health care-related research were available [54·58·151·167·168], but did not help to decide if a benefit of using multilevel modelling with the datasets under study could be expected.

This situation gave rise to the question if multilevel modelling would extend, confirm, or contradict the results of the conventional regression analyses performed on the cost of asthma, gatekeeping, and neutropenia datasets. Further, would multilevel modelling make a substantial contribution to the research questions under study, and would comparison of the multilevel and conventional-based results allow to draw more general conclusions regarding the use and usefulness of multilevel modelling?

Chapter 2

Objectives

The first objective of this thesis is to present the methodology and results of three observational studies conducted and/or analysed by the author, all addressing the impact of potential risk factors on clinical events of interest (neutropenia study), or on endpoints representing health care costs (cost of asthma study, gatekeeping study). This includes to describe and put into perspective any substantial contributions to the current body of knowledge in the fields of research addressed. (The objectives of the individual studies are detailed in chapters 4-6).

All three underlying datasets are characterised by multilevel or multi-membership data structures. They were primarily analysed using conventional multivariate regression techniques, then re-analysed using multilevel modelling. In other words, the multilevel approach was used to perform a sensitivity analysis with a tool that was assumed to make more realistic assumptions on the nature of the data. The second objective of this thesis is to compare the findings established, and to assess if the results of the conventional approach are supplemented and refined, confirmed, or contradicted by the findings of the multilevel approach.

Making use of the above, a third objective is to assess if multilevel modelling, by its application to the author's studies or to other studies, has influenced, or can be expected to influence, current knowledge in the fields addressed.

The fourth, most general objective is to describe implications for the use of multilevel modelling in health care-related research: Under what circumstances is the multilevel approach promising or even a requirement, and what additional contributions can be expected?

Chapter 3

Methods

This chapter identifies the common statistical principles applied to the conventional analyses reported in chapters 4-6, and describes the approach to multilevel re-analysis. The issue of predictive ability is addressed in this context and a framework for comparing the results of conventional regression analysis and multilevel analysis in a structured way is described.

All statistical tests were carried out two-sided at a 5% significance level, and all confidence intervals (CIs) were two-sided 95% CIs, except where otherwise stated.

Approach to conventional analysis - descriptive and univariate statistics

At the descriptive level, total numbers of observations were reported and the numbers, percentages and reasons of non-evaluable observations were assessed. The occurrence of missing values was assessed for the endpoints and covariates of major interest. Where appropriate, missing values were evaluated for any relationship with these endpoints and covariates.

Endpoints and covariates were assessed based on the following descriptive statistics: number of observations, mean, standard deviation, median, quartiles and range for discrete numerical and continuous variables; number of observations, counts and percentages for categorical and ordered categorical variables. Graphical analyses (histograms, boxplots) were added as needed. Where appropriate, CIs were calculated for incidences, durations, etc.

Univariate statistical tests comprised the chi-squared test or Fisher's exact test for categorical data. Where one ordered categorical variable was involved, a chi squared test for trend was added [4:261]. Where two ordered categorical variables were involved, Spearman's correlation coefficient and its p value were added [4:265]. Where one discrete numerical or continuous variable was involved, parametric tests (e.g., T-test, ANOVA) or non-parametric tests (e.g., Mann-Whitney U test, Kruskal-Wallis test) were performed as appropriate, depending on the underlying distributions. For two discrete numerical or continuous variables, Pearson's or Spearman's correlation coefficients and their p values were calculated, depending on the underlying distributions and shape of the corresponding scatter plots [4:279,286-7].

Measures of effect and their CIs were calculated as appropriate. They comprised absolute and relative differences, relative risks (RRs), and odds ratios (ORs).

Approach to conventional analysis - regression analyses

The choice of regression techniques was based on the scaling and distribution of the dependent variables of interest (responses). In the asthma study, multiple linear least squares regression (multiple regression) was used on logarithmically transformed health care cost data containing no zero values. In the gatekeeping study, two-part regression models were used to analyse heavily skewed health care cost data containing a substantial amount of zero values. These two-part models consisted of logistic models of any costs, and of generalized linear models (GLMs) of the amount of costs in the persons with non-zero costs. The GLMs assumed a gamma distribution of the response and used a logarithmic link function [20·51]. In the neutropenia study, logistic regression was used to model a binary response in a combined dataset composed of several retrospective audits of neutropenic event occurrence in breast cancer patients. As heterogeneity between these audits was found with respect to some variables [205], robust standard errors allowing for clustering of observations were estimated using the GEE approach [175·186:259-60]. In alternative model specifications, the individual audits were represented by dummy variables, or the clustering option was used at the level of study centres, for comparison purposes.

Potential predictor variables qualified as candidate predictors if an association with the response variable (dependent variable) of interest seemed realistic on logical (i. e., biological, clinical, etc.) grounds or on statistical grounds ($p \leq 0.25$ in univariate analysis) [15·104:95·142]. Potential direct correlates of the dependent variables of interest (such as resource use variables in the cost models) were not used to rule out circularity effects. In the model building process, main effects were identified by exploring all plausible combinations of covariates manually. Decisions to include or eliminate variables were based on the significance of the individual predictors (based on the T statistic in the case of multiple regression and on Wald tests in the cases of logistic regression and general linear modelling [105:45·116:141-3,647]), and on their ability to significantly improve the model (namely where groups of variables were addressed; based on multiple partial F tests in the case of multiple regression and on likelihood ratio tests in the other cases [104:12-6·116:143-5,649-53]). Formalised variable selection procedures such as stepwise regression may yield implausible results and were used for control purposes only [90]. In a second step, statistically significant first order interaction terms were added.

Additional relative fit criteria [186:262] taken into account when comparing competing models were the unadjusted and adjusted R squared statistic in the case of multiple regression [4:345-6], the pseudo R squared statistic in the case of logistic regression [104:164-7·143], and minimisation of the deviance and the Akaike Information Criterion (AIC) in generalised linear modelling [96:38,45].

In the assessment of model adequacy, Skrondal and Rabe-Hesketh distinguish global absolute fit criteria, which are indicators of any misspecification being present, and local absolute fit criteria, which are used to identify the sources of misspecification [186:267-8,272-3]. Local absolute fit criteria include, among others, graphical residuals diagnostics and the identification of overly influential observations.

Global absolute fit was assessed using tests of model summary statistics in the case of multiple regression (using F tests [116:137-8]) and in the case of logistic regression (using likelihood ratio tests [104:12-6]). It was not specifically assessed in the GLM case. A number of model summary measures have been proposed for GLMs, but none of these could be assumed to have satisfactory properties [223]. Local absolute fit of the models was assessed and near-collinearity issues were addressed as shown in Table 1. The impact of any influential points identified was assessed by tentatively omitting them or absorbing them into dummy variables, and re-estimating the models.

Apparent prediction error was assessed for the final multiple regression models and gamma generalised linear models using the root mean squared error (RMSE) and the mean absolute error (MAE) [51·127]. The RMSE was calculated by taking the absolute difference of each observation's predicted and observed outcome, squaring the difference, and taking the square root of the mean squared differences [51]. The mean absolute error was calculated by averaging the absolute difference of each observation's predicted and observed outcome [51]. In the case of logistic regression,

Table 1. Local absolute fit criteria (measures of goodness-of-fit and to identify influential data points) and techniques applied to assess near-collinearity problems, by type of regression

Measures and techniques used	Multiple regression	Logistic regression	Generalized linear modelling
• to assess goodness-of-fit	Plots of residuals against predictor variables and predicted values, inverse normal plots of residuals [4:346-7]	Hosmer-Lemeshow goodness-of-fit test [104:147-56], plots of mean observed against mean predicted event probabilities by deciles of the linear predictor	Residual diagnostics using studentised Deviance residuals and Anscombe residuals [96:40-4·141:37-40,396-9]
• to identify influential data points	Tentative exclusion of observations with high values of the dependent variable, of observations with large studentised residuals, and of data points with Cook's distance > 1 [116:228-33]	--	Tentative exclusion of observations with high values of the dependent variable, of observations with large studentised residuals, and of data points with Cook's distance > 1 [141:406-7]
• to assess near-collinearity problems	Check for high correlation coefficients between independent variables, inflated standardised regression coefficients, and high variance inflation factors (VIFs) [116:241-2]	Check for high correlation coefficients between independent variables, and inflated standardised regression coefficients	Check for high correlation coefficients between independent variables, check for inflated standardised regression coefficients, comparison of models using centred vs. non-centred independent variables, assessment of collinearity in "parallel" multiple regression models

classification tables indicating the proportion of correct predictions were used [104:156-60]. As addressed in the discussion (chapter 8), split-sample and cross-validation methods were not primarily used in the conventional regression analyses. However, predictive ability was an important issue when comparing conventional and multilevel models. For reference purposes, cross-validation was performed, at a later stage, on those conventional models with multilevel counterparts. Details are described on pp. 39-40 below.

Approach to multilevel modelling

Multilevel analyses were based on the same responses and distributional assumptions as the conventional regression analyses. In the asthma study, multilevel modelling of the continuous, normally distributed logarithm of direct medical asthma costs was used as the multilevel equivalent of multiple regression analysis.

The binary endpoints of any health care costs being accrued (in the gatekeeping study) and of any neutropenic event occurrence (in the neutropenia study) were addressed by using multilevel GLMs of binomial responses and with a logit link function [83:7/1-3·163:101,127]. For the neutropenia dataset, the approach to estimate GEE-based robust standard errors was retained (see Appendices I and III). A possible impact on decisions regarding the inclusion or exclusion of model parameters was assessed by alternatively using conventional standard error estimates. Dummy variables representing the individual audits contributing to the neutropenia dataset were also tried (see Appendix I).

In the gatekeeping study, multilevel modelling of health care costs in those with non-zero costs, a heavily skewed cost response, was either based on multilevel GLMs with a logarithmic link function and assuming a gamma distribution, which is directly equivalent to the approach chosen in conventional analysis. Alternatively, the cost data were interpreted as discrete count data and a negative binomial distributional assumption was used in combination with a (canonical) logarithmic link function [83:7/10·141:373]. Under this assumption and conditionally on the fitted explanatory variables and higher level terms, the mean count for each level 1 unit has a gamma distribution. It was expected that this approach would allow to reproduce the original gamma GLM results fairly adequately, and perform subsequent multilevel analyses as applicable. (The adequacy of this approach was also supported by the fact that in the original analysis, tentative negative binomial regression and the gamma approach yielded very similar results.)

The multilevel analyses used an analysis strategy similar to the one described by Hox [105:49-54]. They were based on the same candidate predictors as the conventional regression analyses. In a first step, the hierarchical structure of each underlying dataset was analysed and hierarchical levels were assigned to the candidate predictors. In a second step, using the conventional regression models as a starting point, the intercept terms were allowed to vary at random at the higher levels, i.e. variance component models were estimated [105:52]. Variance partition coefficients indicating the proportion of the total variance which was due to higher level variance, were calculated [83:2/4-5·105:51]. For logistic multilevel models, the method proposed by Snijders and Bosker was used [164:114·192].

If substantial and significant higher level variation was found to be present in the dataset, the variance structure was analysed, i.e. potential higher level predictors in the strict sense as well as level 1 variables with a potential for higher level variation were tentatively allowed to vary at random, on a variable-by-variable basis at first [105:52-3]. Interactions between predictors from different levels which seemed plausible on logical grounds or based on the results of the conventional analyses, were assessed in the same way [105:53-4]. In order to achieve this, it was sometimes necessary to construct group level predictors by aggregating lower level variables (typically, level 1 variables) within their higher level units [52·105·163:63-75]. In the resulting variance-covariance matrices, relevant and significant coefficients were retained, but near-zero or non-significant coefficients dropped, i.e. fixed to zero.

The fixed parts of the models were modified where this was suggested by parameter or standard error changes occurring when the variance structure was taken into account [105:53-4]. In this context, all candidate predictors excluded from the original fixed effects models were tentatively re-assessed. In borderline situations, fixed effects parameters contained in the main conventional models were retained in the multilevel models, in order to facilitate comparison. For the same reason, covariates derived by aggregating lower level variables were not included in the final multilevel models if they did not show random variation, non-regarding their significance at the fixed effects level.

In order to maintain comparability with the conventional regression analyses, against recommendations, predictor variables were not primarily centred. However, centring was used where estimation problems occurred and predictor variables with random slopes were tentatively centred to assess resulting changes in the estimated intercept variance [105:57-8,70-1].

Goodness-of-fit assessments and assessments of influential higher level units were performed on the resulting multilevel models. Subsequently, these models were modified and re-assessed as applicable [105:26]. The criteria and techniques used in this process were the same as for the conventional regression models where possible. Some particularities and additions are described in the subsequent sections.

Following a widely used practice, the significance of the fixed parameter estimates, and the ability of groups of variables (i.e., sets of predictor variables or interaction terms) to significantly improve the model, was assessed using Wald tests [105:45]. (Wald tests have been reported to perform sub-optimally, particularly in the case of non-normal responses [186:261], but software-related issues restricted the use of likelihood ratio tests - see Appendix III for details.)

The significance of the random effects (variance components) was primarily assessed by likelihood ratio tests and the resulting p-values were divided by two [16·105:45·186:261]. Wald tests were reported to be less appropriate in this case, because random effects estimators are not normally distributed (although approximately so in the case of normally distributed responses) and because the null value is at the left boundary of the parameter space of the expected distributions (as variances are expected to be ≥ 0) [161:18·164:32-3]. However, one-sided [105:43] Wald tests were additionally used where likelihood ratio tests were unavailable for technical reasons (see Appendix III) and in the logistic situation [164:113·169:39], where the appropriateness of likelihood ratio tests has been questioned [105:45]. Non-parametric bootstrap-based interval estimates for the variance parameters were also calculated to confirm the results obtained [164:113].

The main additional relative fit criterion taken into account in the modelling process was minimization of the AIC [105:45-6·186:262-7,352]. Global absolute fit and, consequently, internal validity was assumed for the multilevel models, based on the corresponding assessments performed on the conventional regression models. Skrondal and Rabe-Hesketh mention several global absolute fit criteria for multilevel models, and Hox addresses efforts to derive R squared-like statistics, but all these are reported to be only partially satisfactory [105:63-71·186:268-71].

Assessments of local absolute fit were largely based on graphical residuals diagnostics. Multilevel modelling distinguishes level 1 residuals from higher level residuals. For example, in a two-level situation, the level 1 residuals describe how far the individual observations depart from the regression line representing the level 2 unit they belong to. The level 2 residuals describe the variation which is present in the regression lines representing the level 2 units. In a random intercept random slope model, there would thus be two level 2 residuals, one describing the departure of the intercept of any given level 2 regression line from the intercept of the overall regression line, the other describing the departure of the slope. In a 3-level situation there would be additional level 3 residuals, etc. Higher level residuals are always shrunken residuals, i.e. their value is always lower than the mean of the raw residuals corresponding to the respective higher level unit [99:7-9]. (Raw residuals, in this context, are defined as the difference between the predicted and observed values of the response variable [186:228].) Following the principle of empirical Bayes estimation [33:57-8·52·91·161:27·186:221-35,247], the shrinkage occurs due to the fact that any given higher level unit is assumed to belong to a random distribution. Thus, the information available from the other units is taken into account [58·83:2/9-10·168]. The fewer within-unit observations are available and the higher the within-unit variance is, the smaller is the "credibility" of a given unit's deviation from the overall regression line, i.e. the more pronounced is the shrinkage effect towards the mean of all higher level units [83:2/9-10·99:7-9·164:35-7·186:228-9]. Expressed in more Bayesian terms, the estimated random effects parameters derived during the main stage of the multilevel modelling process are used as the prior distribution. The likelihood for each higher level unit is then derived from the covariate values and responses observed for this unit. Based on Bayes Theorem [33:17], posterior distributions and means are estimated, and the latter represent the unit-specific higher level residuals or empirical Bayes estimates [162]. The term empirical Bayes is used to denote situations where the prior distribution is estimated, using maximisation methods, from the same observed data that are to used to derive the likelihood. In contrast, in full Bayesian estimation, establishing the prior distribution would involve the use of a hyperprior distribution and integration [33:57-8·186:225-6].

Residuals-based goodness-of-fit assessments were performed on the level 1 residuals first, and on the higher level residuals thereafter [186:273].

- In the case of a continuous, normally distributed response, level 1 assessments were based on standard residual diagnostics similar to those applied in conventional multiple regression analysis. These included inverse normal plots of the studentised level 1 residuals and plots of studentised level 1 residuals against the fixed part predicted values [105:23·116:222-224]. (In the latter case, the residuals should be grouped around the x-axis in a parallel band.)
- In the binary case, plots of mean predicted against mean observed event probabilities, by deciles of the linear predictor, were used, as in the non-multilevel case. The assumption of an underlying binomial distribution was assessed by tentatively allowing extra-binomial variation. Where the binomial assumption is part of the model specification itself, the level 1 residual variance component is set to one by definition [83:7/2]. Where extra-binomial variation is allowed, the level 1 variance is estimated, and it is expected to be very close to one [83:7/2·169:35-6]. If this criterion is not met, extra-binomial variation is observed, i.e. the data exhibit more or less variation than a binomial distribution. This may be due to true over- or underdispersion, or due to model misspecification (e.g., omission of levels, important observed or unobserved explanatory variables, or interaction terms), or it could be due to use of an incorrect link-function (i.e., the use of a logistic model would be inappropriate in this case).
- In the GLMs using a logarithmic link function and assuming a gamma distribution of the conditional response, residual diagnostics taking into account GLM-specific particularities were planned to be used at level 1, using studentised Deviance residuals and Anscombe residuals (which are not expected to average to zero) [96:40-4·141:37-40,396-9·186:276].

With respect to the higher level residuals, in all cases, the assumption of a multivariate normal distribution was assessed graphically by inverse normal plots. [83:2/7-8·105:23·122]. In a binary response situation, this assumption may be violated when there are few lower level units per higher level unit, or when the underlying probabilities are close to zero or close to one [163:127]. However, no practical impact of such a violation was assumed, if at all occurring, as it has been shown that parameter estimates in mixed (i.e., part fixed, part random) effects logistic regression are robust to misspecification of the distribution of the random effects [150].

The identification of influential observations followed a downwards approach, starting at the highest level [105:25-6·163:181-2]. Large higher level residuals were identified using caterpillar plots indicating statistically significant deviations from the average intercept, or average coefficient, at the 95% level [105:25]. Influence values combining residuals and leverage values to measure the impact of each unit or observation on the coefficients estimated at the respective level, were additionally used [121·163:168-72]. They are a multilevel equivalent of the difference in fit (DFIT) measure used in conventional regression [14], which describes the change in the predicted value of a data point which is induced by excluding this point before the model is estimated. The characteristics of any influential higher level units identified were explored further by tentatively absorbing into dummy variables either these units as a whole, or (groups of) observations with large lower level residuals nested within these units [105:27]. Model re-estimations were then performed and further steps were taken as appropriate [83:1/11·163:182].

Observations being influential at level 1 could have been identified and analysed using the same technique, non-regarding if they were nested in an influential level 2 unit or not. However, this approach was not pursued further, as the issue of lowest level influential observations had already been addressed in the conventional analyses described in chapters 4-6.

Collinearity issues were not explicitly re-addressed. The assessments performed when the conventional regression models were fitted were assumed to be sufficient, as the same sets of potential predictors were used.

Predictive power of multilevel vs. conventional models

In order to judge predictive power, the final multilevel models were compared with the main conventional models, and with conventional models using same fixed effects predictors as the final multilevel models. The last-mentioned, additional comparison was introduced to distinguish gains in predictive power which were due to multilevel modelling-induced refinements at the fixed effects level, or due to the modelling of the higher level variation itself.

As the indicator of prediction error, the MSE was used in multiple regression, and was intended to be used in generalised linear modelling. Logistic models were compared using the Brier score, an MSE-like measure derived by subtracting, for a given observation, the predicted probability from the binary endpoint value (i.e., from 0 or 1), squaring it, and averaging it [97]. In addition, classification tables showing the percentage of incorrect predictions were used [104:156-60].

In a first step, the apparent prediction error of the above-mentioned models was calculated. For additional graphical illustration, overlaid plots of predicted vs. observed values and of raw residuals (observed minus predicted values) were drawn in the case of multiple regression. In the case of logistic regression, overlaid plots of predicted vs. observed event probabilities, by deciles of the linear predictor, were used.

Apparent error-based assessments will typically overestimate predictive ability [116:401-3·170·186:271-2]. Apparent gains in predictive ability due to multilevel modelling might not be repeatable in cross-validation. Cross-validation techniques were therefore used for further assessment. These were modified to allow for differences in the predictive ability of multilevel models in observations whose corresponding higher level units did, or did not, contribute to model estimation. The empirical Bayes approach [33:57-8·52·91·161:27-8·186:221-35] only allows to produce meaningful random effects estimates for the former group of higher level units [162]. Here, the available information is used efficiently and gains in predictive ability are expected. In contrast, for the latter group of higher level units, informative (i.e., non-zero) random effects estimates are unavailable, as estimating the likelihood function would require some knowledge not only of covariate values but also of the corresponding observed responses. This would contradict the notion of prediction for observations from additional higher level units with all responses unknown. Consequently, gains in predictive ability can only be expected in this situation if the multilevel modelling process leads to an improved modelling of the fixed effects also.

Ten-fold cross-validation and bootstrapping were used to assess prediction error for the conventional regression models (see Appendix II for details) [29·60]. In order to allow for appropriate comparisons with the performance of the multilevel models, three situations were regarded. First, the hierarchical structure of the data was ignored and all observations were treated as independent. Second, the test set (the subset of observations used to assess predictive ability) was restricted to those observations which were not contained in the training set (the complementary subset of observations used for model estimation) but whose corresponding higher level units, through other observations, contributed to model estimation. Finally, the test set was restricted to those observations which were not contained in the training set and whose corresponding higher level units did neither contribute to model estimation, not even through other observations.

Cross-validation of the multilevel models was restricted to ten-fold cross-validation, due to computation time requirements. It was performed for the second and third of the above-described situations. Consequently, the results were indicative of the predictive ability of the multilevel models "within" and "outside" the higher level units contributing to model estimation. The option of ignoring the hierarchical structure of the data was not used here, as the results would have been difficult to interpret.

Comparison of multilevel analyses and conventional regression analyses

The main conventional regression models and the final multilevel models were compared on the basis of a set of predefined questions as listed hereafter.

- Did the multilevel models reveal substantial and significant higher level variation?
- If yes,
  - which proportion of the total variance was explained by between-higher level unit variation?
  - did the modelling of the variance structure provide substantial additional insights?
  - did the (fixed parts of the) final multilevel models comprise the same predictor variables as the main conventional models, or did the modelling of the variance structure necessitate modifications at level 1?
  - were the fixed parameter estimates similar to those found in the original regression analyses, or did they differ substantially? Were standard error changes observed?
  - was the relative fit of the multilevel models substantially improved according to the AIC criterion?
  - was the predictive power of the multilevel models increased?
- As a summary, how do the key findings of the original analyses and of the multilevel analyses compare (brief qualitative description)?
- Can the results of the multilevel analyses, compared to those of the original analyses, be described as
  - contradictory
  - making a major additional contribution
  - making an additional contribution
  - largely unchanged, confirmatory

Results

The results section consists of four chapters. The first three report and discuss the designs, methodological details, and non multilevel-based results of the studies introduced in chapter 1. To recapitulate, the first study (chapter 4) addressed factors associated with the amount of direct medical asthma costs in an adult Swiss patient population. The second (chapter 5) focused on the question if gatekeeping plan membership, as opposed to fee-for-service plan membership, was associated with reduced health care costs in a general Swiss population. The third (chapter 6) used five retrospective European audits of breast cancer chemotherapy to identify risk factors of the occurrence of chemotherapy-induced neutropenic events. Multivariate regression methods were used to establish associations between these outcomes of interest and potential influences. The fourth results chapter (chapter 7) reports if and which additional insights were gained by replacing conventional multivariate regression methods with multilevel modelling, in order take the hierarchical structure of the underlying datasets into account. Findings from both approaches are compared.

Chapter 4

Costs of asthma in a cohort of Swiss adults: associations with exacerbation status and severity

Matthias Schwenkglenks MA[1], Adam Lowy MB ChB[2], Hanspeter Anderhub MD[3], Thomas D. Szucs MD MBA MPH[1]

Affiliations at the time of manuscript preparation:

[1] Hirslanden Research, Zürich, Switzerland

[2] Division of Medical Economics, University Hospital, Zürich, Switzerland

[3] Freiestrasse 211, Zürich, Switzerland

Financial support:

This study has in part been financed by an unrestricted educational grant from Novartis Pharma AG, Basel, Switzerland

Published in Value In Health:

Value in Health 2003;6(1): 75-83.

ABSTRACT

Background. A retrospective chart-based study examined the health economics of asthma in Switzerland in 1996/97.

Objective. To address the effect of exacerbation status, disease severity (defined by medication required) and other variables on resource use and costs.

Methods. A sample of 422 adults was analysed. Target variables were stratified by disease severity and exacerbation status. Bivariate associations were assessed. Multiple linear regression was performed on the logarithm of direct medical costs.

Results. The probability of exacerbations was positively associated with disease severity. Resource use and costs were associated with both these variables. Multiple linear regression identified age, presence of asthma-related comorbidities, degree of severity, exacerbation status, quick reliever versus controller therapy, and diagnosis or treatment by a pulmonologist as independent influences on direct costs. An interaction between severity and exacerbation status was also noted. Regression identified direct costs in the highest severity group to be 2.5 times higher than in the lowest, if there were no exacerbations. If exacerbations were present, costs were 5.7 times higher.

Conclusions. Due to its high prevalence, asthma has a high impact on public health. This impact depends on disease severity and, according to these findings, may also depend on the extent to which exacerbations are avoided or at least controlled.

INTRODUCTION

Asthma seriously affects both children and adults, and the prevalence of asthma has increased considerably during the last three decades [149·181]. The disease affects between 4% and 8% of the population in industrialised countries [87·98·187·210]. In Switzerland, prevalence has been reliably estimated by Leuenberger, at 6.7% in adults [124].

Diseases with such high prevalence require detailed knowledge not only of their clinical and medical aspects, but also of their economic implications. Information on the costs and cost structure of asthma is essential for sound health policy decisions in the field of respiratory diseases. Studies on the costs of asthma have therefore been performed in the U.S. as well as, more rarely, in European and other countries. An overview was given in a recent article by Weiss and Sullivan [216].

However, data from Switzerland on this topic are sparse. In 1996/97 Szucs and colleagues published an analysis of the resource use and cost structure of asthma in Switzerland [203·204]. They demonstrated that asthma has a large economic impact, with total costs of CHF 1'250 Million per year, and a structure of direct medical costs dominated by medication costs in children and by hospitalisation costs in adults. No further Swiss data have been published.

The current analysis was based on the dataset collected by Szucs and colleagues. The goal was to identify independent determinants of asthma-related resource use and of direct medical asthma costs in adult Swiss patients. Particular importance was given to the impact of asthma severity and exacerbation status on direct medical costs.

MATERIALS AND METHODS

Patient sample and study design. The original dataset by Szucs and colleagues comprised 589 patients who were treated by 120 primary care physicians [203]. There were 472 adults and 117 children aged 14 years or younger. The latter were excluded from this analysis as their clinical and economic characteristics differ considerably from those of the adult population.

Szucs et al. collected data by retrospective chart review. There was a one year reference period in 1996/97. Assessment of asthma-specific resource use comprised physician visits, hospital care and medication. The decision if a particular resource use was asthma-specific or not was left to the physician's judgement. Data on community nursing were not collected, as this kind of service provision does not play a major role in Swiss asthma management. The possibility cannot be ruled out that some patients received unrecorded asthma-specific services from other physicians. These facts may have led to a modest underestimation of absolute direct medical costs.

Loss of work due to personal illness or caring for relatives was recorded as a basis for calculating indirect costs, which are not considered in this analysis. Additional information was collected on disease duration, demographics and physiological variables, comprising height, weight, body mass index (BMI), one second forced expiratory volume (FEV1), and forced vital capacity (FVC). Asthma-related comorbidities were assessed by explicitly asking for the presence of chronic bronchitis, emphysema, cor pulmonale, reflux disease, and others.

Reliable health-care related cost estimates are difficult to obtain in Switzerland, as there are no large administrative databases allowing for access to claims data. Most health insurance companies are reluctant to provide case-specific cost information. In this situation, Szucs and colleagues proceeded as follows to calculate direct medical costs in a health system perspective: Asthma-specific resource units were derived from the physician's medical charts as stated above. In the case of medication, prescriptions were used as a proxy of use. Unit costs were estimated as follows: (1) The average charge per inpatient day on the general ward of a public acute care hospital was calculated from a tariff list covering all Swiss hospitals [117]. Cantonal subsidies were added to this average charge, to obtain a proxy of hospitalization costs. (2) Costs of physician visits were calculated individually, based on the services performed. Mean charges per service were calculated from seven regional tariff lists (representing German and French speaking Switzerland) and used as proxy measures for costs. (3) Medication unit costs were assumed to be represented by the Swiss public prescription prices as stated in the 1997 Swiss Drug Compendium [6].

All costs are indicated in Swiss Francs (CHF). In mid-1997, at the end of the reference period of the data collection, CHF 1 equaled US-$ 0.68.

Assessment of disease severity was not an explicit goal of the original study. Therefore, degree-of-severity classification was based on the 1995 Global Initiative for Asthma (GINA) recommendations on medication use, which were in effect at the time of data collection [80]. This kind of procedure has been recommended if clinical information is insufficent [41]. Physicians' prescriptions and dosage instructions were used as a proxy of real medication use. Patients who, according to this information, regularly used short-acting β_2-agonists only were classified as 'mild intermittent'. Patients regularly using inhaled corticosteroids, alone or in combination, were classified as 'mild persistent'. Patients regularly using long-acting β_2-agonists or systemic corticosteroids, alone or in combination, were classified as 'moderate persistent' or 'severe persistent', respectively. Complementary information was available on treatment type (quick reliever versus controller therapy as defined by the treating physician, without any pre-given reference to guidelines) and on the presence or absence of asthma attacks/exacerbations (as defined by the treating physician) during the reference period. The remaining sample comprised 422 patients; 50 patients were excluded due to lack of data for one or more of these variables. Exacerbations were not necessarily associated with a resource use episode. They could be self-reported to the treating physician at a later point in time.

Statistical methods. Demographic and disease characteristics are presented in comparison with the complete adult study sample.

Exacerbation frequencies were stratified by degree of severity. A Chi-squared trend test was used to evaluate the differences observed. Main resource use and cost variables were stratified by degree of severity and exacerbation status.

Bivariate associations were assessed between direct medical costs, degree of severity, exacerbation status and other possible influences. Costs were heavily rightskewed, but logarithmic transformation achieved an approximately normal distribution. Non-parametric tests of the untransformed and T-tests of the transformed cost variable were used. In general, Chi-squared tests were used to compare two categorical variables. If one variable was continuous, Mann-Whitney U tests/Kruskall-Wallis tests or T-tests/ANOVA were used as appropriate [4]. If both variables were continuous, Spearman's or Pearson's correlation coefficients were calculated.

Least-squares regression was used to evaluate interaction between degree of severity and exacerbation status in their effect on the number of physician visits, specialist referrals and hospitalisations.

Finally, multiple least squares regression was performed on the logarithm of direct medical costs. Independent variables qualified as possible influence factors if an association with direct medical costs could reasonably be assumed ($p \leq 0.2$ in bivariate analysis). Resource use variables that were direct contributors to costs (e.g., emergency room visits, days spent in hospital) were excluded. Nevertheless there may be a problem of circularity, particularly because disease severity, in the absence of other options, was defined by medication use and thus directly linked to (medication) costs. Similar correlations may affect the treatment type variable and other parameters, albeit to a much lesser extent. To assess the scope of this problem, an additional regression was performed on the logarithm of direct medical costs excluding medication costs, and the proportion of the variance of medication costs explained by the degree-of-severity and treatment type variables was calculated.

In all cases $p = 0.05$ was used as the level of statistical significance and p-values were two-tailed.

All calculations were performed using STATA 6.0® and SPSS 10.0®.

RESULTS

Patient characteristics. The demographic and disease-specific characteristics of the remaining patient sample are shown in Table 1. In comparison with the complete adult study population, there were no unexpected differences. The percentage of patients with exacerbations was slightly higher, but the mean number of exacerbations was lower in the analysed sample. Fewer French speaking patients were included due to a higher number of missing degree-of-severity or exacerbation status data. Over 50% of the FEV1, FVC and absence-from-work data were missing.

Table 1. Demographic and disease-specific characteristics of study population

	Adults, information on severity and exacerbation status available (N = 422)[a]		**Complete adult study sample (N = 472)[a]**	
	Value (mean ± standard deviation or %)			
Age (years)	53.4 ± 20.6		52.5 ± 20.8	
German / French speaking (%)	78.0 / 22.0		75.4 / 24.6	
Height (cm)	167.0 ± 8.9	(N = 219)	167.3 ± 9.3	(N = 246)
Weight (kg)	72.0 ± 15.0	(N = 240)	72.0 ± 15.7	(N = 268)
BMI (kg/m^2)	25.8 ± 5.4	(N = 204)	25.6 ± 5.3	(N = 231)
FEV1 (liters/second)	2.4 ± 1.0	(N = 193)	2.5 ± 1.1	(N = 216)
FVC (liters)	3.3 ± 1.3	(N = 173)	3.3 ± 1.3	(N = 192)
Duration of asthma (years)	12.4 ± 12.8	(N = 268)	11.3 ± 12.5	(N = 301)
Employed (%)	44.2	(N = 410)	45.6	(N = 458)
Absences from work (% of employed)	25.8	(N = 163 of 181)	23.9	(N = 188 of 209)
Type of treatment:				(N = 461)
Quick reliever therapy (%)[b]	23.7		24.9	
Controller therapy (%)[b]	76.3		75.1	
Degree of severity (GINA):				(N = 432)
Mild intermittent (%)	10.4		10.7	
Mild persistent (%)	26.0		26.2	
Moderate persistent (%)	32.0		31.9	
Severe persistent (%)	31.5		31.3	
Presence of exacerbation(s) during observation period (%)	37.7		36.4	
Number of exacerbations[c]	1.6 ± 1.3		1.8 ± 2.5	

[a] Differing sample sizes due to missing values are indicated separately.

[b] Quick reliever therapy: treatment only when symptoms occur. Controller therapy: prophylactic treatment.

[c] Base: patients with exacerbations.

Degree of severity and exacerbation status. The distributions of the medication-derived degree of severity groups and the occurrence and number of exacerbations are shown in Table 1. Spearman's correlation between FEV1 and medication-based degree of severity was –0.22 ($p = 0.002$).

The proportion of patients who experienced at least one exacerbation during the reference period was almost constant in the mild intermittent (31.8%) to moderate persistent (31.1%) groups, with a minimum in the patients classified as mild persistent (28.2%). This percentage was distinctly higher in the severe persistent group (54.1%), leading to a highly significant Chi squared trend test ($p < 0.0005$).

Resource use. The presence of exacerbations during the reference period was significantly associated with more physician visits (7.2 versus 4.9, $p < 0.005$), specialist referrals (0.33 versus 0.17, $p = 0.048$), and hospitalisations (0.19 versus 0.015, $p < 0.005$) per year. Similar associations with degree of severity were observed; there were 7.9 versus 3.8 physician visits (p for trend < 0.005), 0.3 versus 0.02 specialist referrals (p for trend = 0.13), and 0.16 versus 0.02 hospitalisations (p for trend = 0.05) per year in the highest versus lowest severity groups. Stratification by both exacerbation status and severity is shown in Table 2. Again, the number of resource units consumed increased with severity, and higher levels were reached in the presence of exacerbations. These tendencies were evident in all subgroups, albeit somewhat less unambiguously in the specialist referrals. Regression analysis demonstrated interaction between degree of severity and exacerbation status in their effect on the number of physician visits ($p = 0.03$ for a set of three dummy variables representing interaction) and hospitalisations ($p = 0.04$), but not the number of specialist referrals ($p = 0.53$).

Table 2. Resource omitted use by degree of severity and exacerbation status in units per patient-year

Group	N	Physician visits	Specialist referrals	No. of hospita-lisations
		Mean ± standard deviation		
Total sample	422	5.8 ± 5.0	0.23 ± 0.80	0.083 ± 0.34
Exacerbations absent	263	4.9 ± 4.5	0.17 ± 0.68	0.015 ± 0.12
By degree of severity:				
Mild intermittent	30	3.7 ± 4.5	0.03 ± 0.18	0.033 ± 0.18
Mild persistent	79	4.1 ± 4.7	0.18 ± 0.53	0.013 ± 0.11
Moderate persist.	39	5.2 ± 4.0	0.16 ± 0.52	0.011 ± 0.10
Severe persistent	61	6.3 ± 4.7	0.25 ± 1.10	0.016 ± 0.13
Exacerbations present	159	7.2 ± 5.4	0.33 ± 0.96	0.19 ± 0.51
By degree of severity:				
Mild intermittent	14	4.0 ± 2.7	0 ± 0	0 ± 0
Mild persistent	31	4.0 ± 3.4	0.29 ± 0.78	0.10 ± 0.30
Moderate persist.	42	7.1 ± 5.3	0.45 ± 1.30	0.19 ± 0.40
Severe persistent	72	9.3 ± 5.7	0.35 ± 0.86	0.27 ± 0.65

Costs. The presence of exacerbations during the reference period was associated with higher direct medical costs (CHF 3'202 versus CHF 1'029, p = 0.0001), physician costs (CHF 269 versus CHF 207, p < 0.00005), medication costs (CHF 724 versus CHF 901, p = 0.056), and hospitalisation costs (CHF 2'031 versus CHF 99, p < 0.00005) per year. There also was a steady increase with degree of severity. In the highest versus lowest severity groups, direct medical costs amounted to CHF 3'075 versus CHF 627, physician costs to CHF 284 versus CHF 109, medication costs to CHF 1'122 versus CHF 336, and hospital costs to CHF 1'669 versus CHF 182 (p for trend < 0.005 in all cases). Stratification by exacerbation status as well as severity revealed further details (Table 3). In the patients with no exacerbations, there was a clear positive association of severity with direct medical costs and medication costs, but less of an association with physician costs and no association with hospitalisation costs. In absolute terms, hospitalisation costs were minimal here. In the patients who experienced exacerbations, a positive association with severity was seen in all cost categories. Absolute hospitalisation costs were important here. In the moderate and severe patients with hospitalisations, mean hospitalisation costs were CHF 10'333 (SD 7'018) and CHF 13'875 (SD 11'039) respectively.

Table 3. Costs by degree of severity and exacerbation status in CHF per patient-year

Group	N	Direct medical costs	Physician costs	Medication costs	Hospitalisation costs
			Mean ± standard deviation		
Total sample	422	1'848 ± 4'134	230 ± 257	791 ± 746	827 ± 3'887
Exacerbations absent	263	1'029 ± 1'274	207 ± 266	724 ± 654	99 ± 1'010
By degree of severity:					
Mild intermittent	30	708 ± 1'509	99 ± 101	342 ± 481	267 ± 1'461
Mild persistent	79	804 ± 676	225 ± 362	541 ± 380	38 ± 338
Moderate persist.	93	1'187 ± 1'617	211 ± 217	826 ± 516	151 ± 1'452
Severe persistent	61	1'238 ± 1'089	229 ± 230	992 ± 978	16 ± 128
Exacerbations present	159	3'202 ± 6'315	269 ± 237	901 ± 868	2031 ± 6'019
By degree of severity:					
Mild intermittent	14	452 ± 476	130 ± 75	322 ± 433	0 ± 0
Mild persistent	31	1'275 ± 3'249	153 ± 134	380 ± 329	742 ± 3'151
Moderate persist.	42	3'089 ± 4'947	297 ± 224	912 ± 789	1'881 ± 4'964
Severe persistent	72	4'631 ± 8'058	331 ± 272	1'231 ± 969	3'069 ± 7'717

Consequently, the relative proportions of cost categories differed greatly between those with and without exacerbations: In the latter, physician costs accounted for 20.1% of total direct medical costs, medication costs accounted for 70.4%, and hospitalisation costs for 9.6%. In those with exacerbations, physician costs contributed 8.4%, medication costs contributed 28.1%, but hospitalisation costs contributed 63.4%. The absolute levels of physician and medication costs in those with exacerbations were only slightly higher than in those without exacerbations. To a very large extent, hospitalisation costs accounted for the differences observed.

Spearman's correlation coefficients of annual direct medical costs with the number of physician visits (0.69), the number of specialist referrals (0.14), and the number of hospitalisations (0.44) were significant at the 5% level ($p < 0.005$ in all three cases). However, Spearman's correlation coefficients with possible non-resource use influence factors were weak, except in the case of age (0.25, $p < 0.005$), FEV1 (-0.19, $p = 0.008$) and disease duration (0.13, $p = 0.033$). Pearson's coefficients of the same variables with the logarithm of direct costs were almost identical. There appeared to be no relevant correlations between costs and BMI or FVC.

Man Whitney-U or Kruskall-Wallis tests revealed significantly higher costs for the following influences: controller therapy compared to quick reliever therapy (p <

0.005); involvement of a pulmonologist in diagnosis or regular treatment ($p < 0.005$); non-employment at the beginning of the reference period in patients aged 65 or younger ($p = 0.005$); and presence of asthma-related comorbidities ($p = 0.029$). The use of T-tests or ANOVA with the logarithm of direct costs led to the same results. Absences from work in the employed, German language region, and rural versus urban dwelling were associated with higher costs and had non-significant p-values below 0.2, whereby these factors qualified as candidate predictors in multivariate analysis. Evaluation by insurance coverage did not reveal any existing associations.

There were no unexpected associations between possible influence factors on direct costs. Degree of severity and treatment type correlated significantly ($p < 0.00005$), but Spearman's correlation coefficient was only 0.33.

Multivariate analysis of direct medical costs. Multiple regression analysis on direct medical costs was based on the following possible influence factors derived from bivariate analysis: degree of severity; presence of exacerbations; quick reliever versus controller therapy; involvement of a pulmonologist (in diagnosis or regular treatment); age; duration of disease; FEV1; presence of asthma-related comorbidities; employment status; absences from work; language region; and urban or rural dwelling. Other potential influences such as height, weight, BMI and FVC were also explored.

Resource use variables (e.g., number of physician visits, number of hospital days) were not taken into account, as they were direct contributors to costs.

Models using the logarithm of direct medical costs identified degree of severity, exacerbation status, quick reliever versus controller therapy, age and age squared, presence of asthma-related comorbidities, and involvement of a pulmonologist (in diagnosis or regular treatment) as relevant and significant influence factors, allowing for an adjusted R squared of 0.34. Influences of language region and urban versus rural dwelling were not confirmed.

A four-level ordinal variable, represented by three dummy variables, was introduced in the model to allow for interaction between medication-based degree of severity and exacerbation status. Thus, the effect of exacerbations could be described separately for each degree of severity. This resulted in a partial F-test with p-value of 0.0008 for the set of dummy variables and increased the adjusted R squared to 0.36, showing a greater effect of exacerbations on costs in the more severe asthma patients.

Terms representing employment status in the patients aged 65 or younger, and absences from work in the employed, were not included in the final analysis albeit significant or near-significant, as they altered the model only slightly (adjusted R squared = 0.38).

Age and age squared were centered to avoid a colinearity problem with these variables. After this procedure, variance inflation factors showed a mean of 3.91. The highest value was seen in the exacerbation status variable (VIF 10.60), with the dummy variables representing interaction between degree of severity and exacerbation status showing VIFs of 8.28, 5.26 and 4.40. Other criteria were clearly non-critical: There were no standardized regression coefficients larger than 1. After inclusion of the interaction terms, the parameter estimates and standard errors for the other variables changed very little, except of course for those terms which were bound to change because their meaning is different in the model with the interaction terms.

Details of the main model (N = 420) are shown in Table 4. In larger models including other potential confounders, the degree of severity and exacerbation variables, and the respective interaction terms, had very similar coefficients. (Details not shown.) Relevant coefficient changes only occurred when FEV1 and body height or BMI were included. These models, though, had to rely on less than 100 observations.

Residual analysis (based on scatterplots of residuals versus predicted values and age, boxplots of residuals grouped by non-continuous influence factors and normality plot of residuals) gave satisfactory results. Exclusion of influential points identified by Cook's distance and the covariance ratio (resulting N = 393) did not affect significance or greatly alter coefficients, but increased R squared to 0.43. Alternative versions of the model, e.g. using the frequency of exacerbations rather than their presence or absence, gave very similar results.

Table 4. Multiple linear regression on the logarithm of direct medical costs (N = 420)

F(12,407) = 20.63 Prob > F = 0.0000 R-squared = 0.38 Adj R-squared = 0.36

Covariates	**Coefficient**	**Std. Err.**	**t**	**p > |t|**	**95% Conf. Interval**	
Degree of severity:						
Mild persistent	0.82[a]	0.20	4.19	< 0.001	0.44	1.21
Moderate persistent	1.01[a]	0.19	5.25	< 0.001	0.63	1.39
Severe persistent	0.90[a]	0.21	4.36	< 0.001	0.49	1.30
Exacerbations present[f]	0.31[d]	0.30	1.04	0.299	-0.27	0.89
Interaction variable, ordinal:						
Level 1	-0.49[b]	0.35	-1.37	0.172[e]	-1.18	0.21
Level 2	0.23[b]	0.34	0.68	0.496[e]	-0.44	0.90
Level 3	0.53[b]	0.34	1.56	0.118[e]	-0.14	1.19
Age (centered)	0.00541	0.00240	2.26	0.025	0.00070	0.01013
Age squared (centered)	-0.00026	0.00011	-2.36	0.019	-0.00047	-0.00004
Asthma-related comorbid. present[f]	0.37	0.13	2.94	0.003	0.12	0.62
Involvement of pulmonologist[f]	0.28	0.09	3.00	0.003	0.10	0.46
Controller therapy[f]	0.60[c]	0.12	5.17	< 0.001	0.37	0.83
Intercept	5.15	0.18	27.84	< 0.001	4.79	5.51

[a] Compared to mild intermittent.

[b] Compared to level 0.

[c] Compared to quick reliever therapy.

[d] This coefficient indicates the effect of exacerbations being present in a patient with the lowest degree of severity. The effect of exacerbations at higher degrees of severity is the sum of this coefficient and that for the relevant interaction term.

[e] Partial F test for this set of variables: p < 0.001.

[f] Dichotomous variables, values coded '0' or '1'. Presence of exacerbations, presence of comorbidities, involvement of a pulmonologist in diagnosis or treatment, and controller therapy are coded '1'.

Costs increased with age, but the effect was mild, non-linear and less pronounced in older age groups. Assuming constant other parameters, the presence of asthma-related comorbidities was associated with a 50% increase in direct medical asthma costs, controller therapy versus quick reliever therapy with an 80% increase and involvement of a pulmonologist in diagnosis or treatment with a 30% increase. Greater degrees of severity and the presence of exacerbations during the reference period were also associated with higher costs (Table 5). As the effect of these variables is greater than multiplicative, patients with a greater degree of asthma severity who experienced exacerbations were particularly expensive. Patients classed as severe persistent who experienced exacerbations cost more than 5 times as much as mild intermittent patients who did not.

Table 5. Effect of degree of severity and exacerbation status on direct medical costs according to the estimated regression model

Degree of severity:	Exacerbations absent		Exacerbations present	
	Multiplication factor (95% Confidence Interval)[a]			
Mild intermittent	1	(reference)	1.4	(0.8 - 2.4)
Mild persistent	2.3	(1.5 - 3.3)	1.9	(1.2 - 3.0)
Moderate persistent	2.8	(1.9 - 4.0)	4.7	(3.1 - 7.3)
Severe persistent	2.5	(1.6 - 3.7)	5.7	(3.8 - 8.4)

[a] Calculations performed using Stata's *lincom* command.

Regression analysis on the logarithm of direct medical costs excluding medication costs resulted in a model comprising the same influence factors as described above, except for age and the interaction term between degree of severity and exacerbation status (N = 420, R squared = 0.22). These variables themselves remained highly significant. Degree of severity and treatment type accounted for 19% of the variance seen in the medication costs.

DISCUSSION

This analysis is concerned with the determinants of the direct medical costs of adult asthma, which are most relevant from a third party payer perspective. An inclusion of indirect costs, whose analysis may require a different set of predictors, was not undertaken.

Stratification by exacerbation status and degree of severity reveals these variables to be positively associated with all subcategories of direct medical costs. However, there is no association between severity and hospitalisation costs in the patients without exacerbations. In those with exacerbations, such an association is clearly present. In absolute terms, the cost difference observed between those with and without exacerbations mainly stems from dramatically increased hospitalization costs, while the levels of physician and medication costs are only slightly higher.

The multiple regression model presented identifies a set of factors which, while easy to measure, explain a considerable percentage of the variance seen in the direct medical costs of asthma. Age, comorbidity status, and several factors linked to the concept of disease severity (primarily degree of severity defined by medication, and exacerbation status) influence the direct medical costs of asthma. Treatment type (quick reliever versus controller therapy, as judged by the treating physician) reflects treatment habits, but also comprises aspects of disease severity. The same is probably true for the involvement of a pulmonologist, which can be assumed to be more frequent in cases that are more difficult to treat. If so, the higher cost associated with such an involvement can be expected to be partly due to patient or disease characteristics not covered by our degree-of-severity variable. Different treatment patterns in similar patients would be an additional or alternative explanation. However, there is some evidence that these patterns, albeit more costly at the outpatient level, are cost-saving overall by reducing complication costs [204].

It may seem improper to include several explanatory variables linked to the concept of disease severity that directly impact on cost, which is the outcome variable. In particular, the use of medication as a proxy measure of severity may overestimate the impact of this factor on costs and introduces circularity which may exaggerate statistical significance. It also assumes that empirical treatment follows guidelines, at least to a certain extent. On the other hand, the results obtained demonstrate that the economic impact of the degree-of-severity and treatment type variables was not restricted to medication costs (representing 43% of direct medical costs), and that the latter were not to a very large extent explained by these variables. In clinical terms, a higher degree of severity, defined by medication use, is probably associated with higher baseline medication costs (drug costs per usual daily dose), but also with considerably greater amounts of medication needed and, as confirmed by resource use analysis, other health care resources consumed. Restriction of regression analysis to the non-medication costs would exclude the first of these two aspects and therefore lead to an incomplete picture.

The finding that greater degrees of asthma severity lead to higher medical costs confirms the results of other studies [11·86·135·182·209·216]. The findings of our multivariate analysis are in accordance with data obtained by Hoskins and colleagues [103] on the influence of disease severity on the frequency of asthma exacerbations, and the influence of exacerbations on costs. However, detailed comparisons are hindered by the different health systems and methodologies used.

The modest impact of age on direct medical costs in this analysis does not contradict other studies [159·209]. Plaza et al. reported the costs of asthma patients aged 65 or older to be twice as high as those of adults under 65, but the investigators did not correct for disease severity or apply multivariate methods [159]. The current analysis could not demonstrate a significant effect of insurance status. Such an effect was described in other studies, but insurance systems may induce greater cost differences in the U.S. and in Canada than in Switzerland [135·209]. In the U.S., health plans differ greatly in terms of coverage, and situations of underinsurance have been reported [7·144]. In Canada, drug plan participation is an issue [209]. In Switzerland, participation in the statutory health insurance, which is obligatory, guarantees a high level of medical care to everybody.

Possible effects of BMI or FEV1 were difficult to assess due to missing values. Inclusion of these variables led to models based on less than 100 observations. Relevant coefficient changes of the medication-based degree of severity and exacerbation variables occurred, which supports the idea that these variables and FEV1 contain competing information on disease severity.

The current analysis was based on a data collection primarily targeted at resource use structures and costs, not at identifying influence factors on costs. For this reason it was not possible to take into account some potentially important variables such as patient compliance, inhaler technique, or smoking status [11·103·177·209]. Future studies addressing associations between costs and medical correlates of asthma should assess these. Severity classification should be based on clinical parameters. A clear definition of exacerbations would allow a distinction between different kinds of events with different economic impacts, thus leading to more precise results.

Despite the limitations discussed, the findings of this multivariate analysis show that severity and the presence of exacerbations have considerable and interacting effects on direct medical costs of asthma. Due to its high prevalence, asthma has a high impact on public health. This impact depends on disease severity and may also depend on the extent to which exacerbations are avoided or at least controlled. A prospective study would be needed to finally clarify this issue. If the findings presented here are confirmed, further efforts at preventing exacerbations may well be repaid in reduced treatment costs, as well as reduced patient suffering.

REFERENCES

4 Altman DG (1991). Practical statistics for medical research. Chapman & Hall/CRC, London

6 Anonymous (1997). Arzneimittelkompendium der Schweiz (1997). Documed, Basel

7 Anonymous (1998). Age- and state-specific prevalence estimates of insured and uninsured persons – United States, 1995-1996. *MMWR Morb Mortal Wkly Rep* **47**:529-532

11 Barnes PJ, Jonsson B, Klim JB (1996). The costs of asthma. *Eur Respir J* **9**:636-642

41 Cockcroft DW, Swystun VA (1996). Asthma control versus asthma severity. *J Allergy Clin Immunol* **98**:1016-1018

80 Global Initiative for Asthma (GINA) (2002). Global Strategy for Asthma Management and Prevention. (Updated from: NHLBI/WHO Workshop Report: Global Strategy for Asthma Management and Prevention issued January, 1995). NIH Publication No. 02-3659. National Institutes of Health. National Heart, Lung, and Blood Institute, Bethesda

86 Graf von der Schulenburg JM, Greiner W, Molitor S, Kielhorn A (1996). Kosten der Asthmatherapie nach Schweregrad. Eine empirische Untersuchung. *Med Klin* **91**:670-676

87 Grant EN, Wagner R, Weiss KB (1999). Observations on emerging patterns of asthma in our society. *J Allergy Clin Immunol* **104**:S1-9

98 Hartert TV, Peebles RS, Jr. (2000). Epidemiology of asthma: the year in review. *Curr Opin Pulm Med* **6**:4-9

103 Hoskins G, McCowan C, Neville RG et al (2000). Risk factors and costs associated with an asthma attack. *Thorax* **55**:19-24

117 Konkordat der Schweizerischen Krankenversicherer (1999). Tagestaxen in Heilanstalten (1999). Konkordat der Schweizerischen Krankenversicherer, Bern

124 Leuenberger P (1995). Pollution de l'air en Suisse et maladies respiratoires chez l'adulte. Resultats preliminaires de la partie transversale de l'etude Sapaldia. *Schweiz Rundsch Med Prax* **84**:1096-1100

135 Malone DC, Lawson KA, Smith DH (2000). Asthma: an analysis of high cost patients. *Pharm Pract Manage Q* **20**:12-20

144 Morgan RO, Virnig BA, DeVito CA, Persily NA (1997). The Medicare-HMO revolving door – the healthy go in and the sick go out. *N Engl J Med* **337**:169-175

149 Neri M, Spanevello A (2000). Chronic bronchial asthma from challenge to treatment: epidemiology and social impact. *Thorax* **55**:S57-58

159 Plaza V, Serra-Batlles J, Ferrer M, Morejon E (2000). Quality of life and economic features in elderly asthmatics. *Respiration* **67**:65-70

177 Rutten-van Molken MP, Van Doorslaer EK, Jansen MC, Kerstjens HA, Rutten FF (1995). Costs and effects of inhaled corticosteroids and bronchodilators in asthma and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* **151**:975-982

181 Sears MR (1997). Descriptive epidemiology of asthma. *Lancet* **350**:SII1-4

182 Serra-Batlles J, Plaza V, Morejon E, Comella A, Brugues J (1998). Costs of asthma according to the degree of severity. *Eur Respir J* **12**:1322-1326

187 Sly RM (1999). Changing prevalence of allergic rhinitis and asthma. *Ann Allergy Asthma Immunol* **82**:233-248

203 Szucs TD, Anderhub H, Rutishauser M (1999). The economic burden of asthma: direct and indirect costs in Switzerland. *Eur Respir J* **13**:281-286

204 Szucs TD, Anderhub HP, Rutishauser M (2000). Determinants of health care costs and patterns of care of asthmatic patients in Switzerland. *Schweiz Med Wochenschr* **130**:305-313

209 Ungar WJ, Coyte PC (2001). Prospective study of the patient-level cost of asthma care in children. *Pediatr Pulmonol* **32**:101-108

210 Valdivia G (2000). Asma bronquial y enfermedades atopicas como problema emergente de Salud Publica: nuevas hipotesis etiologicas. La experiencia de sociedades desarrolladas. *Rev Med Chil* **128**:339-346

216 Weiss KB, Sullivan SD (2001). The health economics of asthma and rhinitis. I. Assessing the economic impact. *J Allergy Clin Immunol* **107**:3-8

Chapter 5

Economic efficiency of gatekeeping versus fee-for-service plans – a Swiss example

Matthias Schwenkglenks, M.P.H. [+], Georges Preiswerk, M.D. [++], Roman Lehner, M.D. [+++], Fritz Weber, M.D. [+++], Thomas D. Szucs, M.D. [+]

Affiliations at the time of manuscript preparation:

[+] European Center of Pharmaceutical Medicine (ECPM), University of Basel, Basel, Switzerland

[++] SanaCare AG, Winterthur, Switzerland

[+++] Office-based physicians, Aarau, Switzerland

Financial support:

This work was partially funded by unrestricted, educational grants from CONCORDIA Schweizerische Kranken- und Unfallversicherung, Lucerne; Helsana Versicherungen AG, Zürich; Krankenkasse SBB, Zürich; Wincare Versicherungen, Winterthur; Novartis Pharma Schweiz AG, Berne, all Switzerland.

Published in The Journal of Epidemiology and Community Health:

J Epidemiol Community Health 2006;60:24-30. Used as part of this thesis based on the journal's online information regarding permissions granted to authors.

Remark: This publication is based on the author's Master of Public Health thesis (Interuniversitäres Weiterbildungsprogramm Public Health, Universities of Basel, Bern and Zürich).

ABSTRACT

Study objective. The impact of isolated gatekeeping on health care costs remains unclear. We aimed to assess to what extent lower costs in a gatekeeping plan compared to a fee-for-service plan were due to more efficient resource management, or explained by risk selection.

Design. Year 2000 costs to the Swiss statutory sick funds and potentially relevant covariates were assessed retrospectively from beneficiaries participating in an observational study, their primary care physicians and insurance companies. To adjust for casemix, two-part regression models of health care costs were fitted, consisting of logistic models of any costs occurring, and of generalized linear models of the amount of costs in persons with non-zero costs. Complementary data sources were used to identify selection effects.

Setting. A gatekeeping plan introduced in 1997 and a fee-for-service plan, in Aarau, Switzerland.

Participants. Of each plan, 905 randomly selected adult beneficiaries were invited. The overall participation rate was 39%, but was unevenly distributed between plans.

Main results: The characteristics of gatekeeping and fee-for-service beneficiaries were largely similar. Unadjusted total costs per person were CHF 231 (8%) lower in the gatekeeping group. After multivariate adjustment, the estimated cost savings achieved by replacing fee-for-service based health insurance with gatekeeping in the source population amounted to CHF 403-517 (15-19%) per person. Some selection effects were detected but did not substantially influence this result. An impact of non-detected selection effects cannot be ruled out.

Conclusions: This study hints at substantial cost savings through gatekeeping which are not due to mere risk selection.

KEYWORDS

Economics, health care costs, managed care programs, gatekeeping, Europe

INTRODUCTION

In the early 1990s Switzerland was among the first European countries to introduce managed care solutions [21]. Health insurance is mandatory in Switzerland and these solutions were offered to the population as an alternative to traditional fee-for-service plans. In 2001, managed care organizations had a market share of 5% [8]. Unlike in the US, Swiss managed care lacks strong incentives to restrict the consumption of medical services [13:126-32·118].

Early efforts to evaluate the medical and financial impact of managed care in Switzerland indicated reduced costs, but casemix adjustment was incomplete [31:11,136-41·67·160:39-58].

Swiss gatekeeping plans report cost savings of 10-25% compared to fee-for-service based health insurance [13:136-7]. It remains unclear to what extent these savings are independent of risk selection mechanisms. Various studies have tried to answer similar questions for the US and Europe, but findings were ambiguous [24·49·77·110·137·178·212].

This study compares two local health plans, a gatekeeping and a fee-for-service plan, offered by the same group of health insurance companies in Aarau, Switzerland. These companies report costs to be about 10% lower in the gatekeeping plan, after adjusting for age and gender. We sought to assess to what extent this difference is due to more efficient resource management, or can be explained by risk selection.

METHODS

Health plans

In the region of Aarau, a group of four companies provides health insurance to about 31'250 fee-for-service beneficiaries and 12'500 gatekeeping beneficiaries. The terms of fee-for-service insurance are uniformly defined by Swiss law. Free access to primary care physicians and medical specialists is guaranteed. The gatekeeping plan is managed by a single intermediary company. Its beneficiaries pay reduced insurance premiums. They are required to choose a primary care physician who will also act as a care coordinator and help avoiding unnecessary use of medical resources such as duplicate diagnostic tests. Specialist visits, except in emergencies, require referral by that coordinating physician. However, there is direct access to ophthalmologic and gynaecological care. General coverage of medical services does not differ between plans. Gatekeeping physicians have no financial incentives to limit the use of medical services. They receive a minor administrative fee of CHF 12 per inscribed patient per year. Any additional time spent on their coordinating function is reimbursed at normal rates, i.e. it is reflected in the gatekeeping beneficiaries' cost to the Swiss statutory sick funds. Gatekeeping physicians have varying proportions of fee-for-service patients.

Study population

The population studied were 18 years or older in 1996 and either fee-for-service or gatekeeping beneficiaries throughout 2000, whether they consumed medical services or not. (According to the intermediary company managing the gatekeeping plan, there were hardly any beneficiaries who switched between health plans during the year, except for persons who moved into or out of the area.) Cohorts of 905 beneficiaries of each plan were randomly selected from the enrolment files. In early 2001, they were mailed an informed consent form including a self-administered questionnaire. Only beneficiaries returning the questionnaire became known to the investigators. Further data were provided on consenting participants by their insurance companies and physicians. Three weeks after the first letter was sent, non-responders were mailed a reminder. Data collection was completed in June 2001. Elaborate procedures were applied to ensure a maximum of data protection.

Study outcome and covariates

The primary outcome was gross cost to the Swiss statutory sick funds in 2000. Collection of covariates aimed to allow comprehensive casemix adjustment [179]. It included cost and morbidity data between January and December 1996, the year before the gatekeeping plan was first offered to the population. Cost data and health insurance contract details were provided by the insurance companies, the latter reflecting the beneficiaries' decisions taken in 1999 and defining their status in 2000. The physicians provided morbidity scores for 1996 and 2000 (Index of Co-Existent

Diseases - ICED) [89], physiologic data, and their own practice characteristics. Data collection from the study participants comprised demographic and socio-economic covariates; subjective health status (self-administered SF-36); health behaviour; inclination to over- or under-use medical services; and medical resource use. As discussed later, some of these covariates were time-dependent and documented the situation at the time of data collection, i.e. in the first half of 2001. Analyses were performed including and excluding these covariates.

In addition to the main dataset, anonymous age and gender data for all randomly selected potential participants and year 2000 cost data for the total source population, aggregated by age, gender and health plan, were available. These were used to identify selection effects.

All costs are in Swiss Francs (CHF). On Dec 31, 2000, CHF 1 equalled EUR 0.66.

Statistical methods

Multiple logistic regression was used to model plan membership as a function of beneficiary characteristics identified in univariate analysis.

Expectedly, health care costs included a substantial proportion of zero values and were heavily left-skewed and heteroskedastic (their variance increasing with increasing cost). Two-part regression models of total and outpatient costs were fitted. In the first part we modelled whether any costs were accrued using logistic regression and in the second, generalized linear models (GLMs) were used to analyse the amount of costs in the persons with non-zero costs [57·145]. The GLMs used a logarithmic link function and assumed a gamma distribution of errors [20·51]. Potential covariates were assessed if an association with costs seemed plausible on logical or on statistical grounds ($p \leq 0.25$ in univariate analysis). First, all time-dependent covariates primarily describing the situation in 2000 or 2001 were excluded (reduced models). In a second step (extended models), such covariates were allowed. Resource use variables were not used as covariates. As detailed in Tables 4 and 5, some few observations with costs over CHF 20'000 in 2000 were excluded from the main analysis, in order to reduce the impact of chance effects in this small sample. Complementary analyses included all available observations.

Total predicted values were calculated by multiplying the predicted values of both sub-models [20]. To estimate the marginal (population-level) cost impact of gatekeeping, all participants were assumed to be gatekeeping beneficiaries, or fee-for-service beneficiaries. Both sets of predicted values were calculated and their difference was taken. The result estimates the cost impact of replacing fee-for-service based health insurance with gatekeeping in the source population.

Two-sided p values of 0.05 were used to determine statistical significance. Confidence intervals (CIs) shown are at the 95% level. CIs for the marginal effects were calculated by bias-corrected bootstrapping using 1'000 repetitions.

RESULTS

Participation and data availability

In total, 700 (39%) of the randomly selected persons returned the mailed questionnaire, 433 (48%) of the gatekeeping beneficiaries and 267 (30%) of the fee-for-service beneficiaries. In both groups, 86% of these consented to have additional data collected from their insurance companies and physicians. Full data inclusive of year 2000 cost data and year 1996 cost and morbidity data were finally available from 466 (26%) of the randomly selected persons, 317 (35%) of the gatekeeping beneficiaries and 149 (16%) of the fee-for-service beneficiaries.

Data completeness among respondents was at least 90%. Data provided by a total of 82 participating physicians were near complete, but some physicians who only treated fee-for-service beneficiaries refused to participate, which reduced the number of fee-for-service beneficiaries with full data available. Data provided by the insurance companies were complete.

Beneficiary characteristics and health status

Demographic characteristics, health insurance contract details and indicators of socio-economic status were similar between plans (Table 1). The age range was 23-92 years in the gatekeeping group and 23-96 years in the fee-for-service group. However, the gatekeeping beneficiaries were on average 3.2 years older than the fee-for-service beneficiaries and the proportion of women was lower by 7%. The gatekeeping beneficiaries appeared to be slightly less mobile, less professionally active and had a lower household income.

Table 1. Selected beneficiary characteristics by plan

Characteristic		**Gate-keeping (N = 433)**[a]	**Fee-for-service (N = 267)**[a]	**p value**
Age (mean years ± SD)		56.8 ± 17.1	53.6 ±16.3	0.014[e]
Female gender (%)		53.2	60.3	0.068[f]
Duration of health insurance with the same company (mean years ± SD)		29.7 ± 18.2	28.8 ± 18.0	0.555[e]
Complementary insurance contracts (mean number ± SD)		0.94 ± 0.75	1.0 ± 0.70	0.088[g]
Importance assigned to low insurance premiums (mean score ± SD)[b]		2.8 ± 1.1	2.5 ± 1.2	0.003[g]
Professional status (%)[c]	professionally active	48.7	60.1	0.004[f]
	housework and childcare	47.1	52.4	0.171[f]
	unemployed	1.9	2.6	0.492[f]
	retired	41.6	32.2	0.013[f]
Marital status (%)	unmarried	10.0	16.7	0.055[f]
	married	71.6	64.0	
	widowed	11.2	10.6	
	divorced	7.2	8.7	
Household size (mean number ± SD)	adults	2.0 ± 0.7	1.9 ± 0.8	0.208[g]
	children ≤ 18 years	0.40 ± 0.82	0.35 ± 0.73	0.631[g]
Nursing home residency (%)		1.9	1.9	0.988[f]
Residency in the Aarau area in 1996 (%)		97.2	90.6	< 0.001[f]
BMI (mean kg/m^2 ± SD)		25.2 ± 4.4	24.9 ± 4.4	0.498[e]
Physically active or doing sports (%)		38.2	44.3	0.113[f]
Current smoking (%)		22.2	30.0	0.022[f]
Current alcohol consumption		87.0	87.2	0.801[f]
Importance assigned to healthy nutrition (mean score ± SD)[d]		3.09 ± 0.67	3.04 ± 0.62	0.267[g]
Subjective health status (mean score ± SD)	SF-36 General health scale	70.6 ± 18.5	71.3 ± 19.9	0.648[e]
	SF-36 Physical health summary scale	49.6 ± 9.9	50.0 ± 10.2	0.648[e]
	SF-36 Mental health summary scale	52.1 ± 8.9	51.2 ± 9.4	0.247[e]
ICED (mean score ± SD)	in 2000	2.2 ± 3.1	2.0 ± 2.9	0.470[g]
	in 1996	1.8 ± 2.7	1.5 ± 2.4	0.378[g]
History of mental illness (%)		17.1	18.8	0.635[f]

[a] N is slightly smaller at the individual variable level due to missing values.
[b] Score on a 5-point Likert scale.
[c] Several answers could be ticked.
[d] Score on a 4-point Likert scale.
[e] Unpaired t test.
[f] Chi squared test.
[g] Man-Whitney U test.

Both groups were similar with respect to health behaviour and health status (Table 1). However, the gatekeeping group had a lower proportion of current smokers, especially in the younger age groups. The proportion of physically active persons was higher in the fee-for-service group (non-significant).

Medical resource use and cost

Medical resource use and cost to the Swiss statutory sick funds are detailed in Table 2. Fewer consultations with medical specialists and fewer hospitalisations were reported in the gatekeeping group and their year 2000 total costs per person were CHF 231 lower. Outpatient costs were CHF 377 lower (consultation costs, CHF 7 lower; medication costs, CHF 130 lower; other outpatient costs CHF 239 lower). Inpatient costs were CHF 145 higher in the gatekeeping group, but this difference was annulled when 9 observations with costs over CHF 20'000 were excluded. Year 1996 costs were CHF 762 lower in the gatekeeping group, but this difference was reduced to CHF 199 when 6 observations with costs over CHF 20'000 were excluded. None of the differences observed were statistically significant.

Table 2. Resource use and cost to the Swiss statutory sick funds by plan

Characteristic	Gatekeeping (N = 433)[a]	Fee-for-service (N = 267)[a]	p value[b]
Primary care physician consultations in 2000[c]	3.2 ± 4.4; 2	3.2 ± 4.2; 2	0.440
Medical specialist consultations in 2000[c]	1.0 ± 2.2; 0	1.6 ± 4.0; 0	0.083
Hospitalisations in 2000[c]	0.16 ± 0.44; 0	0.21 ± 0.58; 0	0.664
Total costs > CHF 0 in 2000 (%)	83.6	83.6	0.984
Total costs in 2000[d]	2'496 ± 4'870; 1'120	2'727 ± 4'311; 1'344	0.407
Outpatient costs in 2000[d]	1'815 ± 2'287; 1'102	2'192 ± 3'113; 1'261	0.382
Inpatient costs in 2000[d]	680 ± 3'821; 0	535 ± 1'948; 0	0.994
Total costs > CHF 0 in 1996 (%)	80.7	79.1	0.634
Total costs in 1996[d]	1'674 ± 2'991; 731	2'436 ± 5'466; 824	0.646
Outpatient costs in 1996[d]	1'284 ± 1'702; 670	1'648 ± 2'726; 811	0.469
Inpatient costs in 1996[d]	390 ± 1'934; 0	789 ± 3'849; 0	0.375

[a] N is smaller at the individual variable level due to missing values.

[b] Man-Whitney U test.

[c] Mean number ± SD; median. Self reported values, in good accordance with physician-reported values.

[d] Mean CHF ± SD; median. Observations with zero values included.

Complementary data sources

Comparison with complementary data sources was undertaken to identify selection effects. Among all randomly selected persons, the observation of a higher mean age and a lower proportion of women on the gatekeeping side was confirmed, but less distinct. Participation rates by age and gender group revealed moderate deviations (1-17%) from mean plan-specific participation rates.

After adjusting for resulting differences in the age and gender distribution and after excluding all cases with costs over CHF 20'000, year 2000 study-level costs and the corresponding aggregated costs for the source population were similar. In the gatekeeping plan, study-level total costs per person were CHF 86 lower than population-level costs, and in the fee-for-service plan, they were CHF 69 lower. Within the strata defined by an age cut-off of 65 years and gender, some of the differences seen were more distinct, but still moderate. The female fee-for-service beneficiaries above 65 years of age were the only exception. Their study-level total costs were CHF 740 lower than in the source population, compared to CHF 18 lower in the corresponding gatekeeping beneficiaries.

Predictors of plan membership

Logistic regression indicated that gatekeeping plan membership in 2000 was positively associated with lower 1996 total health care costs; higher 1996 ICED score; having complementary dental insurance; having a higher importance assigned to healthy nutrition; having a lower household income; having more children in the household; living in the Aarau area in 1996; and having a primary care physician with a higher number of consultations per year. The explanatory power of the model remained low (pseudo R squared 0.10, 71% correct predictions).

Predictors of cost

In the reduced model, non-zero total costs in 2000 were associated with higher 1996 outpatient costs; higher 1996 ICED score; lower age (note: after correction for morbidity). The effect of plan membership was modified by age, hinting at a reduced probability of non-zero costs in younger gatekeeping beneficiaries and vice versa. The extended model is shown in Table 3.

Table 3. Logistic regression model of non-zero total health care costs in 2000 (part 1 of two-part model)

N = 418[a]	**Log likelihood -99.05[b]**	**Pseudo R squared of the model = 0.43[c]**	
Independent variable		**Coefficient (95% CI)**	**p value**
Fee-for-service plan membership		-3.43 (-5.76- -1.12)	0.004
Fee-for-service plan membership divided by age (females)[d]		138.51 (35.37-241.65)	0.008
Fee-for-service plan membership divided by age (males)[d]		143.83 (41.24-246.42)	0.006
Age (females)		0.01 (-0.08-0.11)	0.762
Age squared (females)		-0.001 (-0.002- -0.000)	0.020
Age (males)		-0.06 (-0.12- -0.01)	0.021
1996 outpatient costs (log scale)		0.31 (0.18-0.43)	< 0.001
ICED score in 1996		0.67 (0.12-1.22)	0.017
ICED score increase between 1996 and 2000		0.73 (0.14-1.31)	0.016
SF-36 General Health Scale score		-0.03 (-0.06- -0.01)	0.018
Fixed beneficiary co-payment[e]	CHF 400	-0.85 (-1.76-0.06	0.067
	CHF 600	-0.92 (-2.23-0.39)	0.171
	≥ CHF 1'200	-1.92 (-3.04- -0.79)	0.001
Importance assigned to low insurance premiums[f]		0.41 (0.07-0.75)	0.019
Self-reported low aversion of consulting a doctor[f]		-0.43 (-0.81- -0.06)	0.024
Being retired		1.75 (0.35-3.14)	0.014
Constant		5.42 (1.96-8.89)	0.002

a N < 466 due to missing values.

b Uncritical Hosmer-Lemeshow goodness-of-fit test (p = 0.54).

c Predictions correct in 89%.

d Term representing effect modification.

e Compared to lawful minimum of CHF 230.

f Per increase by 1 on a 5-point Likert scale.

In the study participants with non-zero costs, the reduced GLM showed higher total costs to be significantly associated with fee-for-service plan membership (likelihood ratio test, borderline p = 0.066); higher 1996 outpatient costs; higher 1996 ICED score; higher age; choice of lower self-payments but higher insurance premiums; having complementary semi-private insurance; living in the Aarau area in 1996 (modified by age in females). The effect of 1996 ICED score was modified by age and 1996 outpatient costs. The extended model is shown in Table 4.

Table 4. Generalized linear model of total health care costs in 2000, per person with non-zero costs[a] (part 2 of two-part model)

N = 347	**Log likelihood -2988.97**	**Deviance 249.15**	
Independent variable		**Coefficients (95% CI)**	**p value**
Fee-for-service plan membership		0.24 (0.04-0.44)	0.021
Age (females)		-0.03 (-0.09-0.02)	0.246
Age squared (females)		0.001 (0.000-0.001)	0.007
Age (males)		-0.13 (-0.21- -0.04)	0.003
Age squared (males)		0.003 (0.001-0.005)	0.002
Age ^ 3 (males)		-0.000 (-0.000- -0.000)	0.007
1996 outpatient costs (log scale)		0.06 (0.01-0.10)	0.010
ICED score in 1996		0.61 (0.28-0.94)	< 0.001
ICED score in 1996 * age (female)[b]		-0.01 (-0.01- -0.00)	< 0.001
ICED score in 1996 * age (male)[b]		-0.01 (-0.01- -0.00)	< 0.001
ICED score in 1996 * SF-36 Item 2[b]		-0.07 (-0.12- -0.01)	0.025
ICED score in 1996 * 1996 outpatient costs (log scale)[b]		0.02 (0.00-0.05)	0.040
ICED score increase between 1996 and 2000		0.13 (0.06-0.20)	0.001
SF-36 Item 2		-0.13 (-0.29-0.03)	0.121
SF-36 General Health Scale score		-0.01 (-0.02- -0.01)	< 0.001
Complementary semi-private insurance		-0.29 (-0.50- -0.09)	0.005
Importance assigned to low insurance premiums[c]		0.10 (0.02-0.19)	0.017
Living in a partnership		0.60 (0.19-1.01)	0.004
Marital status	married[d]	0.27 (-0.15-0.68)	0.208
	widowed[d]	0.00 (-0.48-0.49)	0.989
	divorced[d]	-0.63 (-1.16- -0.10)	0.19
Household size	2 adults[e]	-1.25 (-1.67- -0.82)	< 0.001
	≥ 2 adults[e]	-1.39 (-1.88- -0.90)	<0.001
Integration	Swiss born or Swiss citizen[f]	-0.39 (-0.88-0.11)	0.127
	Swiss born and Swiss citizen[f]	0.11 (-0.33-0.54)	0.624
Aarau area residency in 1996		2.04 (1.14-2.95)	< 0.001
Aarau area residency in 1996 * age (female)[b]		-0.02 (-0.05-0.00)	0.069
Constant		7.36 (5.69-9.03)	< 0.001

[a] Three gatekeeping observations and 4 fee-for-service observations with health care costs over CHF 20'000 in 2000 not used.

[b] Term representing effect modification.

[c] Per increase by 1 on a 5-point Likert scale.

[d] Compared to unmarried. Combined likelihood ratio test, p = 0.054.

[e] Compared to 1 adult.

[f] Compared to neither Swiss born nor Swiss citizen. Combined likelihood ratio test, p = 0.034.

Comparison of predicted vs. observed costs per person revealed an overestimation in the fee-for-service group (difference CHF 283 when regarding the extended total cost model), but not in the gatekeeping group (difference CHF -5). When the female fee-for-service beneficiaries above 65 years of age (17 observations) were excluded, the difference seen in the fee-for-service group was reduced to CHF -86.

Association of gatekeeping and cost

Direct parameter estimates derived from the conditional cost models as well as the estimated marginal (population) effects, comparing exclusive gatekeeping plan membership to exclusive fee-for-service plan membership, indicated costs savings through gatekeeping at the total and outpatient levels (Table 5). Estimated savings per person were in the range of CHF 403-517 (15-25% of the costs incurred by the fee-for-service source population). Some of the bootstrap-based confidence intervals for the marginal effects overlapped the null, but there was a strong and uniform tendency towards savings by gatekeeping. The reduced and extended models yielded consistent results at the total costs level, but the effect estimate derived from the reduced outpatient cost model appeared high.

Table 5. Estimated cost impact of gatekeeping plan membership compared to fee-for-service plan membership (based on two-part models)

	N[a]	Cost difference (fee-for-service - gate-keeping) in persons with non-zero costs[c]	N[b]	Cost difference (fee-for-service - gatekeeping) in all persons[d]	Cost difference (fee-for-service - gatekeeping) in all persons (%)[d,e]
Total costs					
Reduced model	372	498 (-77-1'072)	439	403 (-120-1027)	14.5
Extended model	347	513 (53-973)	395	517 (-11-1254)	18.9
Outpatient costs					
Reduced model	377	544 (76-1'014)	444	453 (28-973)	24.6
Extended model	354	394 (23-765)	402	372 (-4-813)	17.9

a N available for GLM fitting. Reduced and extended total cost models, 3 gatekeeping observations and 4 fee-for-service observations with health care costs over CHF 20'000 in 2000 not used. Reduced outpatient cost model, 1 fee-for-service observation with outpatient costs over CHF 20'000 in 2000 not used. In the extended outpatient cost model, this observation was not contained due to a missing value in one of the additional predictor variables.

b N available for estimation of marginal (population) effects.

c Conditional effect in persons with non-zero costs as derived from GLM coefficients. Mean CHF per person (CI).

d Marginal (population) effect (combined effect estimate of two-part regression, comparing the assumptions of exclusive gatekeeping plan membership vs. exclusive fee-for-service plan membership). Mean CHF per person (bootstrapped CI).

e Expressed as a percentage of the costs incurred by the fee-for-service source population in the year of reference.

Re-fitting the models and recalculating the marginal effects after inclusion of up to 7 observations with costs over CHF 20'000 led to higher effect estimates (e.g. CHF 773 instead of CHF 517 when using the extended total cost model). In contrast, decreasing the cut-off point further to CHF 15'000 changed the effect estimates only marginally (CHF 486 instead of CHF 517). Exclusion of the female fee-for-service beneficiaries above 65 years of age yielded higher effect estimates (CHF 645 instead of CHF 517). Exclusion of the persons who joined the gatekeeping plan later than in 1997 yielded results in the range of the main results (CHF 481 instead of CHF 517).

DISCUSSION

This study of a gatekeeping and a fee-for-service plan in Aarau, Switzerland, hints at relevant cost savings through gatekeeping which are not due to mere risk selection. Adjustment for casemix was achieved by performing two-part multivariate analyses of year 2000 costs to the Swiss statutory sick funds, taking into account a wide variety of beneficiary and physician characteristics. The characteristics of gatekeeping beneficiaries and fee-for-service beneficiaries were largely similar. A considerable difference in the proportion of current smokers was concentrated on the younger study participants where a substantial impact on health care costs would not yet be expected. Whether physicians treated fee-for-service beneficiaries only, or beneficiaries from both plans, was not a significant predictor of cost on the fee-for-service side.

The result of casemix-adjusted gatekeeping-associated savings of around 20% confirms earlier Swiss reports and earlier, mostly trial-based findings from the US that gatekeeping may be an efficient technique of utilisation management [74·77·107·136·137·178]. However, non-randomised US studies found no or only marginal costs savings associated with gatekeeping [24·92·109·128·212]. A European study using country-level aggregate data found no gatekeeping effect on total costs, but significant savings in the outpatient setting [49].

In this study, constraints on planned sample size in conjunction with a low response rate and incomplete information from some participants led to a small number of observations usable. Thus, the power to detect differential plan member characteristics may have been limited. Furthermore, response rates differed considerably between plans, hinting at the possibility of selection bias.

Accounting for a wide range of potential confounders reduced the probability of strong selection bias. Moreover, external data allowed to assess in part to what extent selection effects were present in the study dataset. Essentially, comparison with aggregated cost data for the source population revealed that this study observed very low costs in the female fee-for-service beneficiaries from age 65 onwards.

Comparison of predicted vs. observed costs at the GLM level revealed an isolated over-estimation of costs in the fee-for-service group, implying a possible exaggeration of the gatekeeping effect. Exclusion of the observations representing female fee-for-service beneficiaries from age 65 onwards diluted this over-estimation. Re-estimation of the two-part cost models after exclusion of this same group of observations did not reduce the combined estimates of the gatekeeping effect. This latter finding may suggest that the identified deviation of observed costs from population-level costs at the subgroup level induced no strong distortion of the main study results. However, additional influences of non-detected selection effects alongside unmeasured covariates cannot be ruled out.

Cost and morbidity data for 1996, the year before the gatekeeping plan was first offered to the population, defined a baseline that could not be influenced by plan-specific mechanisms.

The finding of cost savings through gatekeeping is not invalidated by the observation that the unadjusted increase in total health care costs between 1996 and 2000 was more pronounced on the gatekeeping side. As confirmed by logistic regression modelling of plan membership, persons with higher health care costs in 1996 but not persons with a higher 1996 ICED score were reluctant to join the gatekeeping plan. This must have led to a regression-to-the-mean effect as it has been described before for similar settings [67]. A higher mean age in the gatekeeping group and the fact that more persons in this group reached the age threshold of 65 years between

1996 and 2000 may also have contributed to the more pronounced increase seen here.

Data collection from the study participants occurred in the first half of 2001 and some time-dependent covariates (describing subjective health status, health behaviour and resource use preferences) documented the situation at this point in time, but were nevertheless used in the regression models on year 2000 costs. The intention was not to assess cause-effect relationships between these covariates and the target variable, but merely to reduce the amount of unexplained variance and thus to achieve more precise estimates of the gatekeeping effect. Moreover, most of the covariates in question tend to change slowly over time. The differences between the health plans under study are fairly limited and unlikely to cause differential changes of attitudes, behaviours or even health status in the mid-term. Thus, the 2001 values of these covariates can be assumed to represent the situation directly before and in 2000 fairly adequately, except for random changes of health status.

All regression analyses on cost were performed excluding as well as including these covariates and the resulting estimates of the gatekeeping effect were similar. However, the reduced model on outpatient costs yielded effect estimates which were higher than those seen in the total costs. This was due to insufficient adjustment and it turned out that the difference in explanatory power between the reduced and extended models occurred mainly because no subjective health status variable was available for the former. The level of self-payments chosen (an insurance contract detail) tended to be higher in healthier persons and could be used as an, albeit unsatisfactory, proxy. Variables indicating a high importance assigned to low insurance premiums, or a low household income, behaved near-identical in all cost models, but the former was less affected by missing values and therefore preferred.

The gatekeeping plan under study does not incorporate additional utilisation management practices such as prospective utilisation review [219]. It focuses on the avoidance of duplicate diagnostic tests and unnecessary specialist consultations. Some case management occurs informally. The exact mechanisms behind the cost savings observed could not be identified, as cost and resource use data were not detailed enough for a refined analysis of the medical services provided. Therefore, we cannot contribute to ongoing discussions whether the gatekeeping approach could be optimised, e.g. by allowing direct specialist access for particular subgroups of persons or under special circumstances [65·69]. In our case, most savings were realized in the outpatient setting. The number of consultations was less important for the overall result than the amount of services performed per consultation and the amount of medications prescribed.

This study supports that utilisation management through gatekeeping may be associated with relevant savings in health care costs.

COMMENT: What this paper adds

- The aim of gatekeeping is to reduce the cost of health care without affecting its quality, primarily by avoiding duplicate diagnostic tests and unnecessary consultations with specialists.

- Studies of the impact of gatekeeping in mixed settings, where other techniques of utilisation management were also in place, have been inconclusive. According to this study, isolated gatekeeping may be an efficient technique of utilisation management.

COMMENT: Policy implications

- Policy decisions that have an impact on national reimbursement systems, or on the type of products offered by health insurers, should consider gatekeeping as an option which avoids strong incentives to restrict the use of medical services and which may, therefore, be largely uncontroversial for a public concerned with quality.

ACKNOWLEDGEMENTS

The authors wish to thank Beat Gasser, Atupri Krankenkasse, Berne; Asy Durrer, CONCORDIA Kranken- und Unfallversicherer, Lucerne; Dr René Kühne, Helsana Versicherungen AG, Zürich; Denise Fitzi, Winterthur Versicherungen, Winterthur; and Dr Arthur Vogt, Steinen, all Switzerland, for their contribution during study preparation and conduct. Thank you to Dr Susan Ettner, UCLA Schools of Medicine and Public Health, Los Angeles, USA, and Prof Christoph Minder, ISPM, Berne, Switzerland, for their input on statistical issues. Thank you to Dr Adam Lowy, SIAK, Berne, Switzerland, and Prof Sebastian Schneeweiss, Boston, USA, for their most valuable comments on the manuscript.

FUNDING

This work was partially funded by unrestricted, educational grants from CONCORDIA Schweizerische Kranken- und Unfallversicherung, Lucerne; Helsana Versicherungen AG, Zürich; Krankenkasse SBB, Zürich; Wincare Versicherungen, Winterthur; Novartis Pharma Schweiz AG, Berne, all Switzerland.

CONFLICTS OF INTEREST

Georges Preiswerk is an employee of the company managing the gatekeeping plan under study. Roman Lehner and Fritz Weber are office-based physicians treating beneficiaries of the health plans under study. Matthias Schwenkglenks and Thomas D. Szucs: no potential conflict of interest.

ETHICAL APPROVAL

On February 27, 2001, the data protection registrar of University Hospital, Zürich, Switzerland, the corresponding author's affiliation when this study was planned and data were collected, decided that an ethical approval was not required for this study as the design did not involve an intervention.

REFERENCES

8 Anonymous (2001). Managed-Care-Modelle in der Schweiz. *Managed Care* **5**:37-39

13 Baumberger J (2001). So funktioniert Managed Care. Anspruch und Wirklichkeit der integrierten Gesundheitsversorgung in Europa. Thieme, Stuttgart

20 Blough DK, Madden CW, Hornbrook MC (1999). Modeling risk using generalized linear models. *J Health Econ* **18**:153-171

21 Bohlert I, Adam I, Robra BP (1997). [The Swiss gatekeeper system – a model for improving capacity development and economic effectiveness]. *Gesundheitswesen* **59**:488-494

24 Bonham GS, Barber GM (1987). Use of health care before and during Citicare. *Med Care* **25**:111-119

31 Bundesamt für Sozialversicherung (1998). Neue Formen der Krankenversicherung: Alters- und Kostenverteilungen im Vergleich zu der traditionellen Versicherung. Ergebnisse der Administrativdatenuntersuchung, 2. Teil. Bundesamt für Sozialversicherung, Bern

49 Delnoij D, Van Merode G, Paulus A, Groenewegen P (2000). Does general practitioner gatekeeping curb health care expenditure? *J Health Serv Res Policy* **5**:22-26

51 Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY (1999). Methods for analyzing health care utilization and costs. *Annu Rev Public Health* **20**:125-144

57 Duan N, Manning WGJ, Morris CN, Newhouse JP (1983). A Comparison of Alternative Models for the Deman for Medical Care. *Journal of Business & Economic Statistics* **1**:115-126

65 Escarce JJ, Kapur K, Joyce GF, Van Vorst KA (2001). Medical care expenditures under gatekeeper and point-of-service arrangements. *Health Serv Res* **36**:1037-1057

67 Etter JF, Perneger TV (1998). Health care expenditures after introduction of a gatekeeper and a global budget in a Swiss health insurance plan. *J Epidemiol Community Health* **52**:370-376

69 Ferris TG, Chang Y, Blumenthal D, Pearson SD (2001). Leaving gatekeeping behind – effects of opening access to specialists for adults in a health maintenance organization. *N Engl J Med* **345**:1312-1317

74 Frey W (1996). HMO- und Hausarztmodelle in der Schweiz. *KSK aktuell. Konkordat der Schweizerischen Krankenversicherer* **4**:54-55

77 Galt KA, Rich EC, Kralewski JE et al (2001). Group practice strategies to manage pharmaceutical cost in an HMO network. *Am J Manag Care* **7**:1081-1090

89 Greenfield S, Apolone G, McNeil BJ, Cleary PD (1993). The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement. Comorbidity and outcomes after hip replacement. *Med Care* **31**:141-154

92 Grembowski DE, Martin D, Diehr P et al (2003). Managed care, access to specialists, and outcomes among primary care patients with pain. *Health Serv Res* **38**:1-19

107 Huber-Stemich F, Hees K, Baumann P, Berger D (1996). Sechs Jahre HMO Zürich-Wiedikon. Ein Erfahrungsbericht. *Ars Medici* 18:1079-1082

109 Hurley RE, Paul JE, Freund DA (1989). Going into gatekeeping: an empirical assessment. *QRB. Quality Review Bulletin* **15**:306-314

110 Hurley RE, Freund DA, Gage BJ (1991). Gatekeeper effects on patterns of physician use. *J Fam Pract* **32**:167-174

118 Kuttner R (1998). Must good HMOs go bad? First of two parts: the commercialization of prepaid group health care. *N Engl J Med* **338**:1558-1563

128 Long SH, Settle RF (1988). An evaluation of Utah's primary care case management program for Medicaid recipients. *Med Care* **26**:1021-1032

136 Manning WG, Leibowitz A, Goldberg GA, Rogers WH, Newhouse JP (1984). A controlled trial of the effect of a prepaid group practice on use of services. *N Engl J Med* **310**:1505-1510

142 Mullahy J (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ* **17**:247-281

137 Martin DP, Diehr P, Price KF, Richardson WC (1989). Effect of a gatekeeper plan on health services use and charges: a randomized trial. *Am J Public Health* **79**:1628-1632

160 PROGNOS AG (1998). Evaluation neuer Formen der Krankenversicherung. Synthesebericht. Bundesamt für Sozialversicherung, Bern

178 Schillinger D, Bibbins-Domingo K, Vranizan K et al (2000). Effects of primary care coordination on public hospital patients. *J Gen Intern Med* **15**:329-336

179 Schneeweiss S, Sangha O (2001). [Provider profiling: needs, methodologic requirements and means to increase acceptance]. *Dtsch Med Wochenschr* **126**:918-924

212 Vertrees JC, Manton KG, Mitchell KC (1989). Case-mix adjusted analyses of service utilization for a Medicaid health insuring organization in Philadelphia. *Med Care* **27**:397-411

219 Wickizer TM, Lessler D (2002). UTILIZATION MANAGEMENT: Issues, Effects, and Future Prospects. *Annu Rev Public Health* **23**:233-254

Chapter 6

Neutropenic event risk and impaired chemotherapy delivery in six European audits of breast cancer treatment

Matthias Schwenkglenks[1], Christian Jackisch[2], Manuel Constenla[3], Joseph N. Kerger[4], Robert Paridaens[5], Leo Auerbach[6], André Bosly[4], Ruth Pettengell[7], Thomas D. Szucs[1], Robert Leonard[8] for the Impact of Neutropenia in Chemotherapy - European (INC-EU) Study Group

Affiliations at the time of manuscript preparation:

[1] European Center of Pharmaceutical Medicine, University of Basel, Switzerland
[2] Department of Gynaecology, University Hospital, Marburg, Germany
[3] Complexo Hospitalario de Pontevedra, Pontevedra, Spain
[4] Cliniques Universitaires UCL de Mont-Godinne, Yvoir, Belgium
[5] University Hospital Gasthuisberg, Leuven, Belgium
[6] Vienna General Hospital, Vienna, Austria
[7] St. George's Hospital, London, UK
[8] South West Wales Cancer Centre, Swansea, UK

Financial support:

This work was partially funded by an unrestricted, educational grant from Amgen (Europe) GmbH, Lucerne, Switzerland.

Published in Supportive Care in Cancer:

Supp Care Cancer 2006;14:901-9. © Springer-Verlag 2006. Used as part of this thesis with kind permission of Springer Science and Business Media.

ABSTRACT

Goals of work. The aims of this study were to assess chemotherapy treatment characteristics, neutropenic event (NE) occurrence and related risk factors in breast cancer patients in Western Europe.

Patients and methods. Six retrospective audits of breast cancer chemotherapy were combined into a dataset of 2'860 individuals. NEs were defined as neutropenia-related hospitalisation, dose reduction ≥ 15%, or dose delay ≥ 7 days. Summation dose intensity (SDI) was calculated to compare different types of chemotherapy regimens on a single scale. Risk factors of NE occurrence and of low relative dose intensity (RDI) ≤ 85% were identified by multiple logistic regression.

Main results. Patient populations were comparable between audits. Until 1998 CMF regimens were most frequently used, but thereafter anthracycline-based regimens were most common. NEs occurred in 20% of patients and low RDI in 16%. NE occurrence predicted low RDI and was associated with higher age; bigger body surface area; lower body mass index; regimen type; more chemotherapy cycles planned; normal to high SDI; concomitant radiotherapy; and year of treatment. First cycle NE occurrence predicted NEs from cycle two onwards. A risk score; using age, SDI, number of planned chemotherapy cycles, and concomitant radiotherapy; differentiated patients with increasing NE risk (9-37%). An alternative score version not using concomitant radiotherapy performed moderately less well.

Conclusions. NEs occurred frequently in this combined dataset and they affected treatment delivery. Identifying patients at high NE risk enables targeted prophylaxis and may avoid dose limitations.

KEYWORDS

Breast cancer; adjuvant chemotherapy; adverse effects; neutropenia; Europe

INTRODUCTION

Myelosuppression is a major side-effect of anticancer chemotherapy. Consequences include potentially life-threatening febrile neutropenic episodes, intravenous antibiotic treatment and prolonged hospitalisation [153]. Chemotherapy dose reductions and delays are common sequelae and may affect treatment outcomes adversely [10·126]. In early-stage breast cancer, adjuvant chemotherapy has become an element of standard therapy and reduces the hazard rate of death by about 15% [59]. However, this benefit may be reduced or lost when relative chemotherapy dose intensity (RDI) is reduced [23·30·36·158].

Trial-based reports of chemotherapy-induced neutropenia (CIN) and febrile neutropenia (FN) incidence in breast cancer patients vary widely. During the last decade, cyclophosphamide; methotrexate; fluorouracil (CMF) regimens and fluorouracil; doxorubicin or epirubicin; cyclophosphamide (FAC or FEC) regimens have been most widely used. A systematic review of randomised clinical trials with at least 50 patients per treatment arm published between 1990 and 2000 found grade III-IV CIN rates of 1-78% in CMF chemotherapy and of 3-100% in FAC or FEC chemotherapy [45]. These differences are partially explained by protocol-specific assessment rules and differences in the timing of absolute neutrophil counts (ANC) or white blood cell counts (WBC) [66]. Thus, current evidence does not always enable a specific neutropenia risk to be assigned to commonly used regimens. Detailed information on the impact of neutropenia on chemotherapy delivery in routine practice is also limited.

Prophylactic measures such as colony-stimulating factor (CSF) and anti-infectives administration can be used to avoid neutropenic event (NE) occurrence and maintain RDI. Current US and European guidelines, as well as economic constraints, recommend a limited use of such prophylaxis, and this is reflected in practice [46·66·147·153]. Furthermore, new NCCN myeloid growth factor guidelines recommend that the overall risk of neutropenia should be calculated, taking into account both patient and treatment risk factors, before deciding whether to provide growth factor support [147]. Thus the development of clinically applicable risk models allowing prophylaxis to be targeted at high risk patients is important. Various studies published during the last decade have addressed this and research is currently ongoing [129]. Results are awaited from ongoing American and European prospective observational studies.

We have combined data from six retrospective European audits of breast cancer chemotherapy, enabling us to assess the incidence and extent of chemotherapy dose limitations, the incidence of NEs, and the associations of both with potential risk factors with greater statistical power than the individual audits. From these analyses we propose preliminary NE risk scores.

PATIENTS AND METHODS

Datasets combined

Data on NE occurrence and impaired chemotherapy delivery were obtained from six audits in which patient identification information had been removed. Details of the Chemodose 99 audit conducted in 37 centres in Belgium and Luxembourg; the Optimización del Standard de Quimioterapia Administrada en diferentes Regímenes (OSQAR) audit conducted in 34 Spanish centres; the Audit of Primary Breast Cancer Patients conducted in 15 UK district general and teaching hospitals; and an audit conducted in six German centres were reported earlier [42·111·114·123]. Unpublished data came from two academic centres in Vienna, Austria (principal investigator L. Auerbach), and in Ghent, Belgium (principal investigator S. Van Belle).

In the UK audit, data were collected prospectively in 60% of patients. All other data were collected retrospectively. Patient selection rules were designed to recruit a patient mix as seen in routine practice, but rules regarding the inclusion of stage IV patients differed between audits and the Belgian (Ghent) study only included patients receiving adjuvant CMF regimens.

Variables available from all six audits comprised demographic details; diagnosis and disease stage; prognostic factors and hormone receptor status (except UK audit); planned and administered chemotherapy; toxicities; NEs and related hospitalisations; concomitant radiotherapy and CSF use. Some ANC and WBC values were missing in all audits. Comorbidity and performance status data were not available, although the latter were unlikely to be important in a mostly adjuvant setting. Limited information on long-term outcome was available from the Austrian, German and Spanish audits, but not from the rest.

Coding and grouping criteria were unified. Variable definitions were compared and in discrepant or unclear cases, recalculations were performed from the cycle-specific details available. With respect to chemotherapy dose limitations, cut-off points used were ≥ 15% for reductions and ≥ 7 days for delays. RDI was calculated as administered dose per unit time divided by planned dose per unit time. In the case of combination regimens, the RDIs for each drug were averaged. Low RDI was defined as RDI ≤ 85% [23]. NEs were defined as neutropenia-related dose delay; dose reduction; or hospitalisation. The decision whether events were neutropenia-related was made by the original investigators. Due to a lack of uniform assessment rules, ANC or WBC data were frequently missing and thus could not be used to verify NEs for the combined dataset.

Summation dose intensity (SDI), measuring the planned dose intensity of combination regimens on a single summary scale, was calculated as proposed by Hryniuk et al. [106]. From first-line single-agent trials in metastatic breast cancer, these authors determined the unit dose intensity (UDI) required for each drug to produce a 30% complete plus partial response rate. For each regimen, the dose intensities of the individual agents used were expressed as fractions of their UDIs, and summed. Here, the resulting SDI values were standardised to a recognised ‘standard’ of chemotherapy, namely intravenous CMF with drug administrations on days 1 and 8 of a 28-day cycle, at the following doses: cyclophosphamide, 600 mg m^{-2}; methotrexate, 40 mg m^{-2}; 5-fluorouracil 600 mg m^{-2} (CMF 600-40-600 d1,8 4w). An RDI adjusted to CMF 600-40-600 d1,8 4w was then calculated by multiplication with the standardised SDI values (adjusted RDI).

Statistical methods

Descriptive analysis used standard statistical methods. Univariate relationships between categorical variables were assessed by Chi-squared tests. Where one variable was continuous, t-tests and ANOVA or Mann-Whitney U tests and Kruskal-

Wallis tests were used, depending on the distributions observed. Where both variables were continuous, Spearman's correlation coefficients were employed because of non-normality.

Multiple logistic regression allowing for clustering by audit was used to calculate adjusted odds ratios (OR) with robust standard errors for the following outcomes: any NE occurrence; NE occurrence from cycle two onwards; and low RDI [175].

Chemotherapy delays and dose reductions directly impacted on RDI and they were also part of the NE definition used. As a consequence of this circularity, the coefficient of the NE covariate may have been overestimated in the regression models on low RDI occurrence. Therefore, this was checked by using an alternative NE definition based on the limited ANC and WBC data available.

The independent influences on NE occurrence identified in regression analysis were used as candidate items for the development of tentative NE risk scores. Selection of score items followed the principle of achieving a maximum of predictive and discriminatory power with a minimum of complexity. Cut-off points for continuous variables were chosen empirically to optimise score performance.

Two-tailed p = 0.05 was used to determine statistical significance. Confidence intervals (CIs) shown are at the 95% level.

RESULTS

Patient and treatment characteristics

The Austrian, Belgian and Luxembourg, Belgian (Ghent), German, Spanish, and UK audits contributed 375, 660, 82, 154, 1'167, and 422 patients (total 2'860). Of these, 79% received their chemotherapy treatments during 1995-2001.

Patient and disease characteristics were acceptably similar across audits (Table 1). However, the diagnostic spread in the Belgian (Ghent) audit was unusual with an increased proportion of stage I patients (32% vs. 18% across all audits) and a reduced proportion at stage III (4% vs. 16% across all audits). Post-menopausal patients were under-represented in this dataset, although other patient characteristics were comparable to other audits.

Table 1. Patient, disease and treatment characteristics

Variable		N total	All audits	Inter-audit range		
Age at diagnosis (years; mean ± SD)		2'745[a]	51.5 ± 11.3	48.0 ± 10.9	-	53.1 ± 11.8
Age at diagnosis 60 years or higher (%)		2'745[a]	24.4	13.6	-	30.7
Menopausal status (% postmenopausal)[c]		2'365[b]	51.5	44.0	-	56.8
Hormone receptor status (% positive)		2'179[b]	63.7	56.3	-	70.9
Disease stage[c]	stage I (%)	2'743[a]	18.1	14.6	-	21.6
	stage II (%)		62.5	54.9	-	69.0
	stage III (%)		15.5	10.4	-	28.9
	stage IV (%)		4.0	0.0	-	8.9
Chemotherapy regimen[d]	CMF-based (%)	2'834 [a]	55.7	47.0	-	72.3
	anthracycline-based (%)		40.8	27.3	-	47.2
	taxane-based (%)		1.3	0.0	-	3.0
	other (%)		2.2	0.3	-	7.3
Concomitant radiotherapy (%)		2'606 [a]	30.9	23.7	-	61.6
Colony-stimulating factor use (%)		2'832 [a]	12.9	1.4	-	18.3

[a] N total < 2'860 due to missing values spread over various datasets.

[b] N total < 2'860 due to inavailability of UK data and additional missing values spread over various datasets.

[c] Inter-audit ranges of menopausal status and disease stage do not take into account the Belgian (Ghent) dataset, see text.

[d] Inter-audit ranges of chemotherapy regimens used do not take into account the Belgian (Ghent) audit reporting data on patients receiving CMF chemotherapy only.

Abbreviations: C, cyclophosphamide; F, 5-fluorouracil; M, methotrexate.

In total, 240 different chemotherapy regimens were planned. A comparison of these regimens was facilitated by use of summation dose intensity (SDI) [106], using CMF 600-40-600 d1,8 4w as a standard. Table 2 details the distribution of the most common regimen subtypes and their standardised SDI values, and Figure 1 shows a histogram of standardised SDI. Mean SDI was highest in CMF chemotherapy, followed by anthracycline-based, taxane-based, and other chemotherapy. Older patients generally received lower SDI regimens (correlation coefficient -0.10, p < 0.001); more so in CMF (-0.13, p < 0.001) than in anthracycline-based chemotherapy (-0.07, p < 0.001).

Table 2. Frequency of use and standardised[a] summation dose intensity (SDI) of subtypes of planned chemotherapy regimens

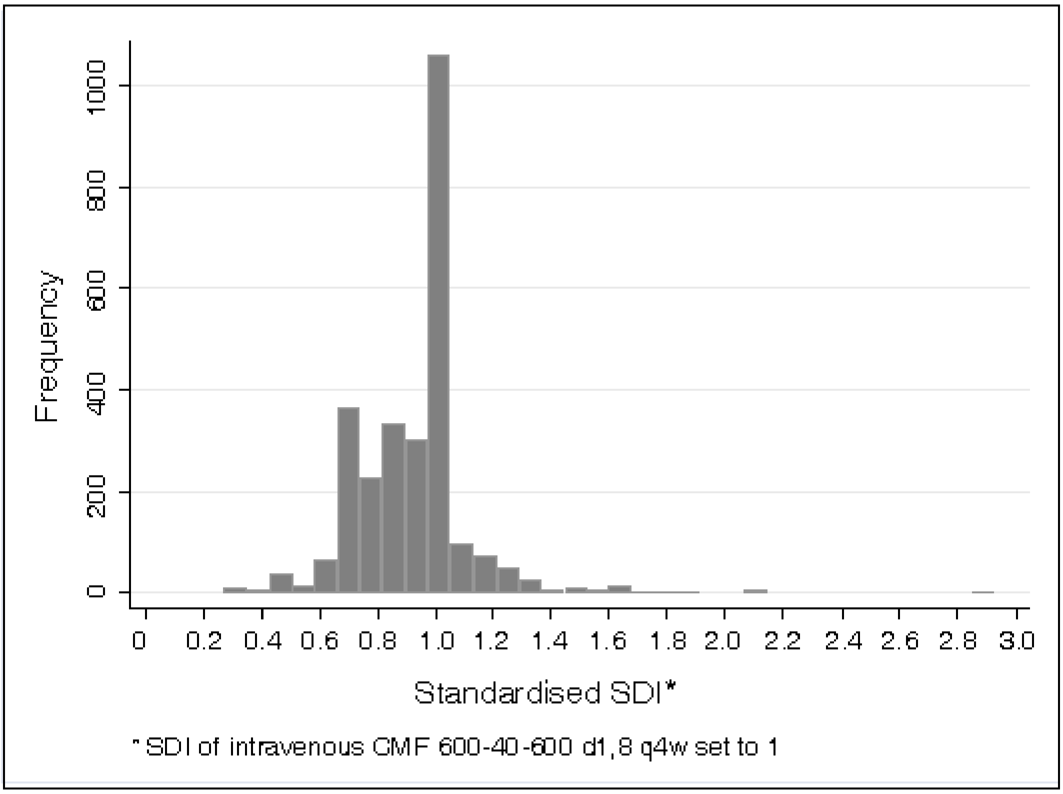
Regimen subtype	N total[b]	Frequency within regimen type (%)	SDI (mean ± SD)	SDI inter-audit range (mean ± SD)
All regimens	2'832	--	0.92 ± 0.19	0.84 ± 0.19 - 1.00 ± 0.00
		CMF-based regimens		
All CMF-based	1'578	--	0.94 ± 0.17	0.85 ± 0.21 - 1.02 ± 0.10
CMF, 28 day cycle	1'012	64.1	0.99 ± 0.11	0.86 ± 0.20 - 1.07 ± 0.00
CMF, 21 day cycle	375	23.8	0.76 ± 0.21	0.69 ± 0.00 - 1.03 ± 0.32
CMF, oral	187	11.9	1.04 ± 0.04	1.03 ± 0.08 - 1.04 ± 0.06
CMF-based sequential	4	0.3	0.66 ± 0.23	0.55 ± 0.07 - 1.00 ± --
		Anthracycline-based regimens		
All anthracycline-based	1'155	--	0.89 ± 0.20	0.76 ± 0.28 - 0.95 ± 0.22
FAC	194	16.8	1.05 ± 0.12	0.97 ± 0.00 - 1.05 ± 0.07
FEC	420	36.4	0.82 ± 0.92	0.61 ± 0.00 - 0.84 ± 0.13
AC	138	12.0	1.00 ± 0.07	0.90 ± 0.20 - 1.02 ± 0.06
EC	154	13.3	0.77 ± 0.29	0.69 ± 0.22 - 1.00 ± 0.46
A→CMF	63	5.5	0.83 ± 0.09	0.79 ± 0.04 - 1.00 ± 0.05
Anthracycline- and taxane-containing	38	3.3	1.21 ± 0.43	0.98 ± 0.03 - 1.25 ± 0.45
Other anthracycline	148	12.8	0.88 ± 0.22	0.79 ± 0.10 - 1.00 ± 0.56
Taxane-based	37	--	0.86 ± 0.34	0.82 ± 0.27 - 1.89 ± --
Other	62	--	0.70 ± 0.08	0.63 ± 0.21 - 0.70 ± 0.07

[a] Standardised to intravenous CMF 600-40-600 d1,8 4w.

[b] N total < 2'860 due to missing values spread over various datasets.

Abbreviations: A, doxorubicin; C, cyclophosphamide; E, epirubicin; F, 5-fluorouracil; M, methotrexate. A→CMF refers to sequential regimens where several courses of A, then several courses of CMF are administered.

Figure 1. Histogram of standardised summation dose intensity (SDI)



* SDI of intravenous CMF 600-40-600 d1,8 q4w set to 1

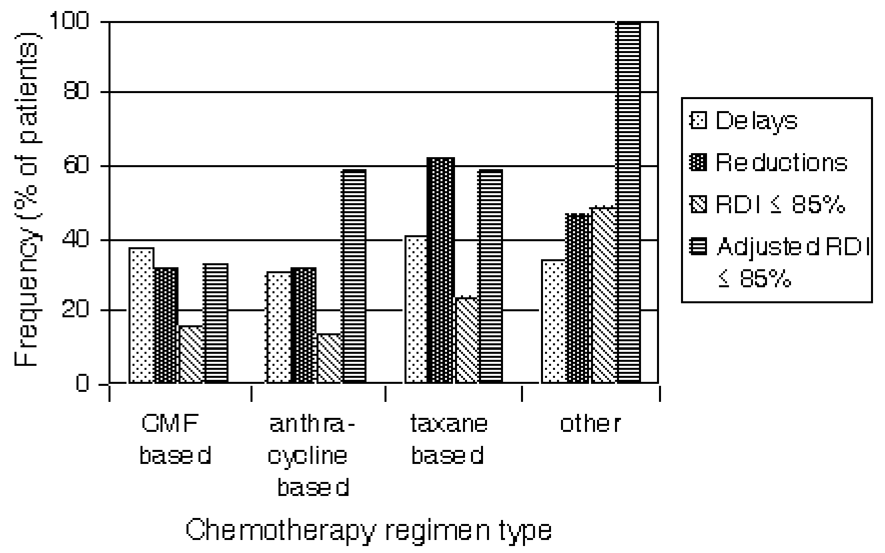
Concomitant radiotherapy was administered to 40% of patients receiving CMF-based chemotherapy and to 22%, 14%, and 16% of those receiving anthracycline-based, taxane-based, and other regimens respectively. CSFs were used in 12% of patients receiving CMF-based chemotherapy and in 14%, 30%, and 17% of those receiving anthracycline-based, taxane-based, and other regimens respectively.

Since 1996, use of CMF regimens and concomitant radiotherapy have decreased, while use of anthracycline-based regimens has increased, becoming the most frequent regimen type after 1998. Use of CSFs increased in the early 1990s and remained relatively constant after 1995.

Chemotherapy dose limitations

Dose delays in at least one cycle occurred in 34% of patients (inter-audit range 16-48%), of which 8% had delays that appeared to be directly related to concomitant radiotherapy administration; whether they were pre-planned chemotherapy interruptions could not be assessed. Dose reductions in at least one cycle occurred in 33% of patients (inter-audit range 14-49%). Forty-seven percent had either dose delays or dose reductions (inter audit range 26-58%). Mean RDI ± SD was 94 ± 11% (inter-audit range 90 ± 16%-96 ± 8%). Figure 2 shows details by regimen type.

Figure 2. Chemotherapy dose limitations by regimen type



For difference between regimen types, p = 0.001 (delays); p < 0.001 (other parameters)

RDI, relative dose intensity; adjusted RDI, product of RDI and summation dose intensity standardised to CMF 600-40-600 d1,8 4w

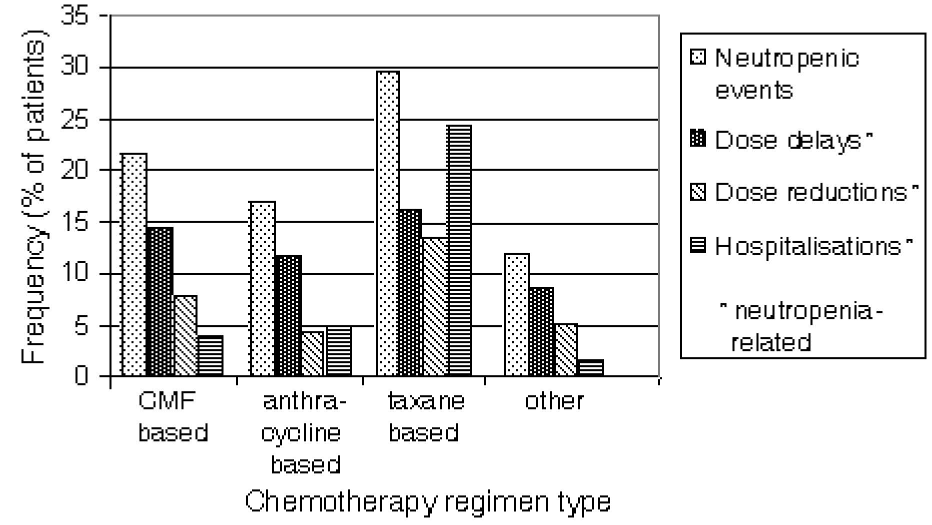
RDI ≤ 85% occurred in 16% of patients (inter-audit range 7-25%). It occurred more frequently in CMF-based than anthracycline-based chemotherapy (16% vs. 14%; p = 0.052). However, anthracycline-based regimens tended to have a lower SDI than CMF-based regimens and when RDI was adjusted to CMF 600-40-600 d1,8 4w chemotherapy, 45% of patients fell short of the 85% threshold (inter-audit range 10-56%).

RDI was slightly lower in older patients (correlation coefficient -0.07, p = 0.001). However, the proportion of patients over age 60 who received RDI ≤ 85% was not significantly higher than for patients below 60 years (17% vs. 15%, p = 0.270). In contrast, the proportion of patients above 60 years who received adjusted RDI ≤ 85% was higher than for younger patients (51% vs. 43% below 60 years; p < 0.001).

Neutropenic event occurrence

NEs (neutropenia-related dose delays, dose reductions or hospitalisations) were observed in 20% of patients (inter-audit range 15-25%), and repeated NEs in 8% (inter-audit range 6-11%). Neutropenia-related dose delays were seen in 13% (inter-audit range 6-22%), dose reductions in 6% (inter-audit range 1-11%), and hospitalisations in 5% (inter-audit range 4-13%). Figure 3 shows details by regimen type.

Figure 3. Neutropenic events by regimen type



For difference between regimen types, p = 0.004 (neutropenic events); p = 0.104 (delays); p = 0.001 (reductions); p < 0.001 (hospitalisations)

Neutropenic event occurrence (regression results)

In logistic regression, NE occurrence was associated with higher age; higher BSA; lower BMI; regimen type; more planned chemotherapy cycles; normal to high SDI (2nd to 4th quartiles); concomitant radiotherapy administration; and year of treatment. Concomitant radiotherapy administration interacted with BSA, number of planned chemotherapy cycles, regimen type, and SDI. The change in NE risk over time was regimen-dependent. Table 3 details the model.

Table 3. Influences on any neutropenic event occurrence (logistic regression allowing for clustering by audit)

N = 2'358 **Pseudo R squared of the model = 0.070**

Independent variable		Odds ratio (95% CI)	p value[a]
Age[b]		1.02 (1.01-1.03)	< 0.001
Body surface area[b]	if concomitant radiotherapy no	3.85 (1.84-8.07)	< 0.001
	if concomitant radiotherapy yes	0.95 (0.47-1.95)	0.895
	if concomitant radiotherapy unknown	13.19 (2.38-73.11)	0.003
Body mass index[b]		0.97 (0.94-0.99)	0.013
Chemotherapy regimen[c]	three weekly CMF	2.75 (2.05-3.67)	< 0.001
	anthracycline-based	1.50 (0.98-2.30)	0.061
	taxane-based	1.68 (1.44-1.97)	< 0.001
	other	0.87 (0.34-2.21)	0.764
Normal to high SDI[d]		1.70 (1.43-2.02)	< 0.001
Planned chemotherapy cycles[b]	if concomitant radiotherapy no	1.43 (1.18-1.74)	< 0.001
	if concomitant radiotherapy yes	1.18 (1.13-1.22)	< 0.001
	if concomitant radiotherapy unknown	0.92 (0.81-1.03)	0.153
Year of treatment[b]	linear, if 4 weekly CMF	0.94 (0.88-1.00)	0.068
	linear, if 3 weekly CMF	1.02 (0.93-1.13)	0.626
	linear, if anthracycline use	0.85 (0.75-0.95)	0.007
	linear, if taxane-based regimen	1.91 (1.69-2.14)	< 0.001
	linear, if other regimen	1.50 (1.44-1.56)	< 0.001
	squared	0.99 (0.98-1.00)	0.075
Concomitant radiotherapy[e]	yes, if 4 weekly CMF	0.65 (0.42-1.00)	0.049
	yes, if 3 weekly CMF	0.37 (0.20-0.70)	0.002
	yes, if anthracycline-based	0.26 (0.21-0.33)	< 0.001
	yes, if taxane-based	--	--
	yes, if other regimen	--	--
	unknown, if 4 weekly CMF	1.01 (0.01-97.00)	0.995
	unknown, if 3 weekly CMF	1.25 (0.01-123.96)	0.925
	unknown, if anthracycline-based	2.27 (0.04-116.86)	0.683
	unknown, if taxane-based	--	--
	unknown, if other regimen[f]	76.27 (2.20-2644.20)	0.017
	yes, if low SDI	41.27 (5.37-316.85)	< 0.001
	yes, if normal to high SDI	128.50 (15.99-1032.9)	< 0.001
	unknown, if low SDI	--	--
	unknown, if normal to high SDI	1.19 (0.14-10.14)	0.876

a Combined Wald tests, for all sets of categorical or ordinal variables and for all sets of interaction terms, $p < 0.05$. Interaction effects are presented as simple effects, not as main effects plus interaction terms, for ease of interpretation.

b Per one unit increase.

c Compared to four weekly CMF.

d Second to 4th quartiles compared to 1st quartile.

e Compared to no concomitant radiotherapy administration.

f Parameter estimate based on 4 observations, assumed to be an artefact

Abbreviations: C, cyclophosphamide; F, 5-fluorouracil; M, methotrexate; SDI, summation dose intensity.

NE risk over time decreased for four-weekly CMF and anthracycline-based regimens, increased for taxane-based and other regimens, and increased then decreased for three-weekly CMF. Three-weekly CMF had a distinctly higher NE risk compared to four-weekly CMF. Taxane-based regimens also posed a greater NE risk than four-weekly CMF, although the sample was rather small (N = 37). For anthracycline-based regimens, the NE risk was only slightly greater than for four-weekly CMF when time trends were taken into account.

A positive association of concomitant radiotherapy and NE occurrence was confirmed for patients receiving four-weekly CMF regimens with normal to high SDI (41% of our sample), with an OR of 2.5 (CI 1.4-4.5; calculated form the linear predictors underlying the ORs shown in Table 3, using mean values for body surface area and number of planned chemotherapy cycles). Use of radiotherapy weakened both the association of NE risk with higher BSA and with the number of planned chemotherapy cycles.

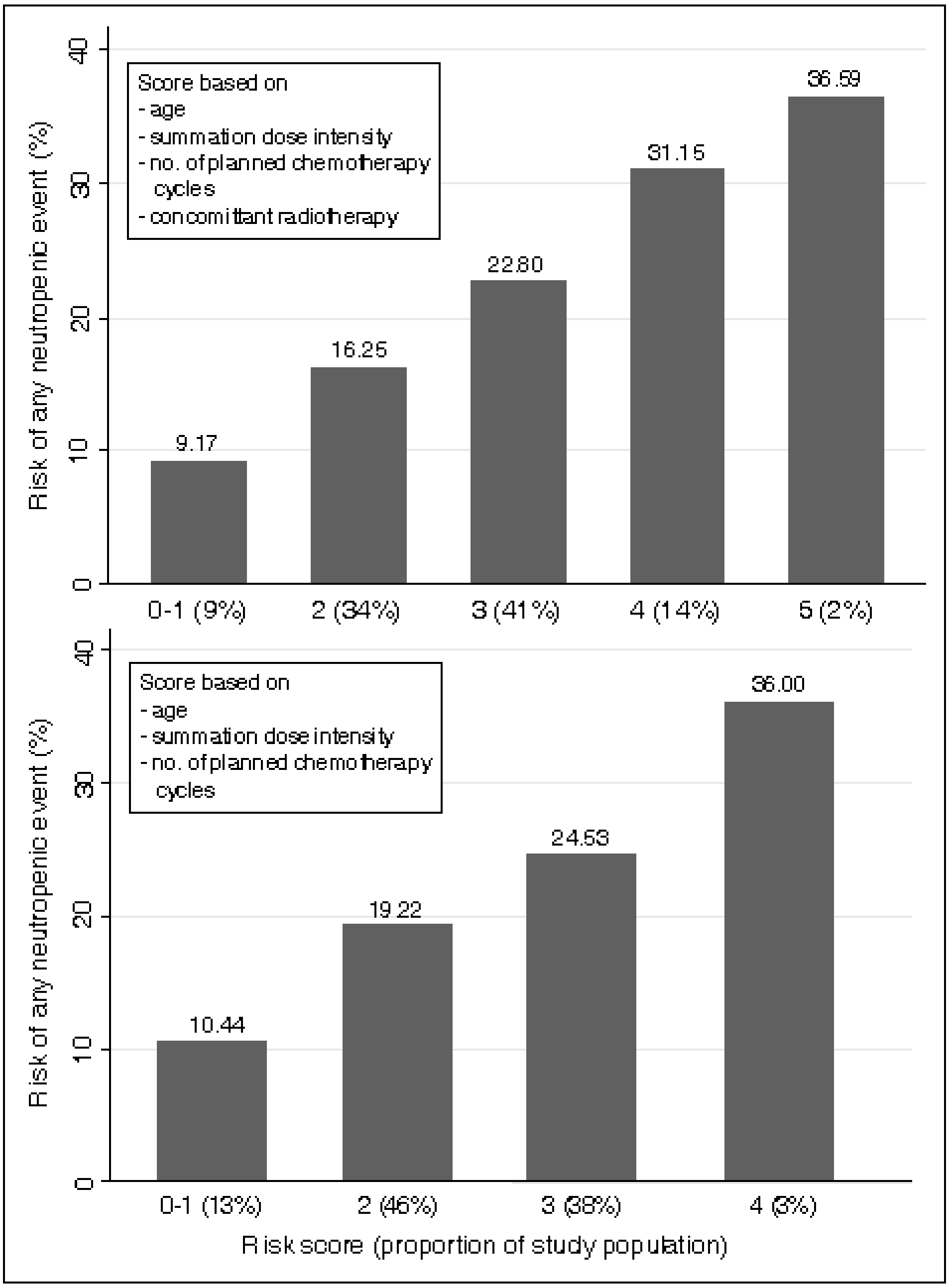
First cycle NE occurrence was a strong predictor of NE occurrence from cycle two onwards, with an OR of 7.7 (CI 4.3-13.9). Otherwise, influence variables and observed interactions were as described for NE occurrence in any cycle, and coefficients were remarkably similar.

When the observations used for model estimation were restricted to patients receiving CMF- or anthracycline-based chemotherapy, the associations and coefficients seen remained largely stable.

Neutropenic event risk scores

A drug-independent risk score based on age (≥ 50 years vs. < 50 years), SDI (2nd to 4th quartiles vs. lowest quartile, corresponding to a standardised SDI cut-off of 0.80), number of planned chemotherapy cycles (≥ 6 vs. < 6), and concomitant radiotherapy administration performed best. The score value was derived by counting the number of risk factors present, and by adding 1 if there were more than 6 chemotherapy cycles planned. Patient groups with an increasing risk of NE occurrence in any cycle of 9-37% were differentiated (Figure 4, upper half). The area under the receiver operating characteristic (ROC) curve was 0.60 (CI 0.58-0.63). When the suggested cut-off point of more than two risk factors being present was used, sensitivity was 69% and specificity 47%. Omitting concomitant radiotherapy administration as a score item reduced the area under the ROC curve to 0.57 (Figure 4, lower half). Sensitivity was reduced to 51% and specificity increased to 61%.

Figure 4. Neutropenic events in any cycle by risk score



When first cycle NE occurrence was included as an additional score item, patient groups with an increasing risk of NE occurrence from cycle two onwards (8-43%) were differentiated with sensitivity 71% and specificity 46%. Omitting concomitant radiotherapy administration from this score led to a similar degradation of performance as described above for any NE occurrence.

Low RDI occurrence (regression results)

Logistic regression showed low RDI occurrence to be significantly associated with NE occurrence; higher stage of disease; regimen type; concomitant radiotherapy administration; and year of treatment. When the alternative NE definition, based on the available blood cell count data, was used, the OR of low RDI for NE occurrence was reduced from 5.1 (CI 4.2-6.2) to 2.5 (CI 2.1-3.1). All other coefficients remained stable.

DISCUSSION

This is the first multi-country study to address the incidence, risk factors, and consequences of neutropenic events induced by adjuvant breast cancer chemotherapy. Six retrospective European audits of breast cancer chemotherapy, with comparable patients (similar to those seen in routine practice) were combined to generate the first transnational European database.

The most commonly followed regimens were CMF- and anthracycline-based, with administration of taxane-based and other regimens being too rare to allow reliable results. CSF use was low and followed no coordinated approach, which, from an analytical perspective, may have been advantageous, because physiological relationships were not hidden by effective prophylaxis.

NEs and low RDI were confirmed to be frequent events and several independent predictors of NE occurrence were identified. SDI was demonstrated to be such a predictor, for the first time according to our knowledge. Using a cut-off point between the 1st and the 2nd quartile of the SDI distribution was optimal for prediction purposes, but treating SDI as a continuous variable also produced significant results. Using SDI was crucial in separating the effects of regimen type from those of planned dose intensity.

A higher NE risk in three-weekly compared to four-weekly CMF was observed in the Spanish and UK audits; i.e. all audits with a substantial proportion of three-weekly CMF patients. This association has been previously reported [64·123], and different centre characteristics have been proposed as a partial explanation [123]. The NE risk of anthracycline-based regimens was only moderately increased compared to four-weekly CMF, but an element of incomplete adjustment may be present in this result, as we corrected for SDI but not for anthracycline dosage directly. Changes in NE risk over time and dependent upon regimen type, may have been due to increasing experience and changes in medical practice.

NE risk was shown to increase linearly with age; an observation supported by earlier studies [10·206]. Correcting for SDI was important for detecting this relationship, which was partially hidden by a tendency to use lower planned dose intensity in older patients. While the association of age and NE risk appears marginal when expressed per year of age (OR 1.02; see Table 3), this corresponds to an OR of around 1.5 for an age difference of 25 years. An increased NE risk in patients with higher BSA and a protective effect of higher BMI were also demonstrated. This seems biologically plausible, particularly if it is acknowledged that BSA-based chemotherapy dosing may not be an optimal solution. The BSA effect was diluted in patients receiving concomitant radiotherapy, which might be because radiotherapy adds a BSA-independent risk component. Concomitant radiotherapy administration itself appeared to be associated with a higher NE risk in univariate analysis; in multivariate analysis this finding was only unequivocal for four-weekly CMF regimens with normal to high SDI.

Of the variables analysed, a combination of age, SDI, number of planned chemotherapy cycles, concomitant radiotherapy administration and first cycle NE (in the case of NEs occurring from cycle two onwards) provided the best indicator of risk. Tentative addition of score items representing BSA, BMI, or chemotherapy regimen type did not improve performance. The observed wide variety of regimen specifications, which is in part reflected in the SDI variable, and differences in practice patterns may have obscured the differences between the main types of chemotherapy regimens involved. In support of our conclusions, age, radiotherapy administration and first cycle NE have previously been used in clinical prediction models [44·185·206]. However, the cut-off found for age, at 50 years, may have been influenced by practice patterns and may not be universally applicable. Compared to scores derived from datasets with more baseline and first cycle haematology parameters available, our tentative scores performed only slightly less well [206]. Omitting concomitant radiotherapy, which is becoming rare, affected the performance moderately. The other parameters used, including SDI, can be expected to become important components of future risk models combining patient-related and treatment-related factors.

An association of NE and low RDI was confirmed even when NE were assessed from the limited ANC and WBC values available. Low adjusted RDI was found to occur more frequently in anthracycline-based compared to CMF-based chemotherapy due to lower SDI values in the former. In contrast, Lyman et al. reported lower SDI values in CMF- than in anthracycline-based chemotherapy in the US [130].

Our findings underline the importance of early prophylaxis. NEs are frequent and they often impact on chemotherapy delivery, with the likely effect of diminished efficacy. The importance of maintaining full chemotherapy dose intensity, in younger and older patients, has been described by several authors [23·30·36·158]. A current tendency toward dose-dense regimens may further exacerbate this problem [40·62·207].

In order to further our current findings on the risk of NE, a prospective study measuring all potential risk factors including first cycle ANC has commenced in Europe. This will allow our findings to be validated against an external dataset, and should allow our risk score to be developed and refined to a model with increased discriminatory power. Such a risk model is becoming fundamentally important as economic constraints and current guidelines require expensive prophylactic treatments to be targeted to those at greatest risk [147].

ACKNOWLEDGEMENTS

The authors wish to thank Amgen (Europe) GmbH, Lucerne, Switzerland, for supporting this work by an unrestricted, educational grant.

Thank you to Dr. Veronique Cocquyt and Prof. Simon van Belle, Department of Medical Oncology, University Hospital Ghent, Ghent, Belgium, for their permission to use the Belgian (Ghent) dataset.

REFERENCES

10 Balducci L (2003). Myelosuppression and its consequences in elderly patients with cancer. *Oncology (Huntingt)* **17**:27-32

23 Bonadonna G, Valagussa P, Moliterni A, Zambetti M, Brambilla C (1995). Adjuvant cyclophosphamide, methotrexate, and fluorouracil in node-positive breast cancer: the results of 20 years of follow-up. *N Engl J Med* **332**:901-906

30 Budman DR, Berry DA, Cirrincione CT et al (1998). Dose and dose intensity as determinants of outcome in the adjuvant treatment of breast cancer. The Cancer and Leukemia Group B. *J Natl Cancer Inst* **90**:1205-1211

36 Chang J (2000). Chemotherapy dose reduction and delay in clinical practice. evaluating the risk to patient outcome in adjuvant chemotherapy for breast cancer. *Eur J Cancer* **36 Suppl 1**:S11-14

40 Citron ML, Berry DA, Cirrincione C et al (2003). Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: first report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. *J Clin Oncol* **21**:1431-1439

42 Constenla M, Bosly A, Jackisch C, et al. (2003). An audit of primary breast cancer management in Spain: the OSQAR study [abstract]. *Proc Am Soc Clin Oncol* **22**:312

44 Crawford J, Ozer H, Stoller R et al (1991). Reduction by granulocyte colony-stimulating factor of fever and neutropenia induced by chemotherapy in patients with small-cell lung cancer. *N Engl J Med* **325**:164-170

45 Dale DC, Crawford J, Lyman CG (2001). Chemotherapy-induced neutropenia and associated complications in randomized clinical trials: an evidence-based review [abstract]. *Proc Am Soc Clin Oncol* **20**:1638

46 Dale DC (2002). Colony-stimulating factors for the management of neutropenia in cancer patients. *Drugs* **62 Suppl 1**:1-15

59 Early Breast Cancer Trialists' Collaborative Group (1998). Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* **352**:930-942

62 Ellis GK, Livingston RB, Gralow JR, Green SJ, Thompson T (2002). Dose-dense anthracycline-based chemotherapy for node-positive breast cancer. *J Clin Oncol* **20**:3637-3643

64 Engelsman E, Klijn JC, Rubens RD et al (1991). "Classical" CMF versus a 3-weekly intravenous CMF schedule in postmenopausal patients with advanced breast cancer. An EORTC Breast Cancer Co-operative Group Phase III Trial (10808). *Eur J Cancer* **27**:966-970

66 ESMO Guidelines Task Force (2001). ESMO recommendations for the application of haematopoietic growth factors (hGFs). *Ann Oncol* **12**:1219-1220

106 Hryniuk W, Frei E, 3rd, Wright FA (1998). A single scale for comparing dose-intensity of all chemotherapy regimens in breast cancer: summation dose-intensity. *J Clin Oncol* **16**:3137-3147

111 Jackisch C, Jaber M, Burkamp U et al (2003). Maintenance of dose intensity in adjuvant chemotherapy of breast cancer in patients treated outside a clinical trial. Results of a retrospective study. *Geburtsh u Frauenheilk* **63**:333-343

114 Kerger JN, Bormans V, Dauwe M (2002). Adjuvant (adj) chemotherapy (CT) delivery in patients (pts) with breast cancer (BC): Results from the Chemodose Working Party Belgium-Luxembourg [abstract]. *Ann Oncol* **13 Suppl 5**:38

123 Leonard RC, Miles D, Thomas R, Nussey F (2003). Impact of neutropenia on delivering planned adjuvant chemotherapy: UK audit of primary breast cancer patients. *Br J Cancer* **89**:2062-2068

126 Link BK, Budd GT, Scott S et al (2001). Delivering adjuvant chemotherapy to women with early-stage breast carcinoma: current patterns of care. *Cancer* **92**:1354-1367

129 Lyman GH (2003). Risk assessment in oncology clinical practice. From risk factors to risk models. *Oncology (Huntingt)* **17**:8-13

130 Lyman GH, Dale DC, Crawford J (2003). Incidence and predictors of low dose-intensity in adjuvant breast cancer chemotherapy: a nationwide study of community practices. *J Clin Oncol* **21**:4524-4531

147 National Comprehensive Cancer Network (NCCN) (2005). Clinical Practice Guidelines in Oncology – v.2.2005. Myeloid Growth Factors in Cancer Treatment. http://www.nccn.org/professionals/physician_gls/PDF/myeloid_growth.pdf. Accessed May 30, 2005

153 Ozer H, Armitage JO, Bennett CL et al (2000). 2000 update of recommendations for the use of hematopoietic colony-stimulating factors: evidence-based, clinical practice guidelines. American Society of Clinical Oncology Growth Factors Expert Panel. *J Clin Oncol* **18**:3558-3585

158 Piccart MJ, Biganzoli L, Di Leo A (2000). The impact of chemotherapy dose density and dose intensity on breast cancer outcome: what have we learned? *Eur J Cancer* **36 Suppl 1**:S4-10

175 Rogers WH (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin* **13**:19-23

185 Silber JH, Fridman M, DiPaola RS et al (1998). First-cycle blood counts and subsequent neutropenia, dose reduction, or delay in early-stage breast cancer therapy. *J Clin Oncol* **16**:2392-2400

206 Talcott JA, Siegel RD, Finberg R, Goldman L (1992). Risk assessment in cancer patients with fever and neutropenia: a prospective, two-center validation of a prediction rule. *J Clin Oncol* **10**:316-322

207 Thatcher N, Girling DJ, Hopwood P et al (2000). Improving survival without reducing quality of life in small-cell lung cancer patients by increasing the dose-intensity of chemotherapy with granulocyte colony-stimulating factor support: results of a British Medical Research Council Multicenter Randomized Trial. Medical Research Council Lung Cancer Working Party. *J Clin Oncol* **18**:395-404

Chapter 7

Multilevel re-analysis

Hierarchical structure of datasets

All three datasets analysed were characterised by hierarchical (multilevel) data structures. In the asthma dataset, patients were nested within physicians [203], which corresponds to a simple two-level situation. The gatekeeping dataset showed a cross-classification structure, i.e. health plan beneficiaries were independently nested within physicians and within insurance companies. However, in the original analysis, no influence of insurance company on cost to the Swiss statutory sick funds was identified and a brief multilevel re-assessment confirmed this finding. Therefore, the gatekeeping dataset was also analysed as a two-level dataset, with health plan beneficiaries at level 1 and physicians at level 2. In the neutropenia dataset, three levels were present: patients were nested within study centres which were nested within audits conducted in different countries. Sample size details are shown in Table 1.

Table 1. Hierarchical structure of datasets

	Asthma dataset	Gatekeeping dataset	Neutropenia dataset
Level 3 description	--	--	audits
N			6
Level 2 description	physicians	physicians	study centres
N	107	59	95
N per level 3 unit (mean; min-max)		--	15.8; 1-38[a]
Level 1 description	patients	health plan beneficiaries	patients
N	422	466[b]	2'860
N per level 2 unit (mean; min-max)	3.9; 1-5	7.9; 1-32	30.1; 1-375[a]

[a] Two of the audits were single-centre studies with 82 and 375 observations, respectively. When these were excluded, the mean number of level 1 units per level 2 unit was 25.8 (range 1-89).

[b] Maximum number of level 1 observations available for estimation of the logistic models of any health care costs being accrued. In some regression models, N was smaller due to missing values in some of the covariates used. The maximum number of level 1 units available for estimation of the GLMs of total health care costs in those with non-zero costs was N = 391.

A number of potential level 2 predictors (e.g., specialty of treating physician) were available from the asthma and gatekeeping datasets. Moreover, in all three cases, some level 1 predictors could be assumed to be influenced by or intertwined with higher level characteristics (e.g., chemotherapy regimen chosen). Details are shown in Appendix I.

Presence of higher level variation

The presence of higher level variation was assessed for the main endpoints of the analyses reported in chapters 4-6. (In the original analysis of the gatekeeping study, a distinction between reduced and extended cost models was made, and multilevel re-analysis was based on the latter, more complete models.)

Random intercept variance components models were estimated by allowing the intercept terms of the conventional regression models described in chapters 4-6 to vary at random. Significant higher level variance was found to be present in the asthma dataset and in the neutropenia dataset at level 2, but not in the neutropenia dataset at level 3 and not in the gatekeeping dataset. Details are shown in Table 2.

In the asthma dataset, 17.5% of the residual variance (i.e., of the variance not explained by the fixed part predictors) was between-physician variance occurring at level 2. Correspondingly, in the neutropenia dataset, 15.1% of the residual variance was found to be between-centre variance.

Given these results, more in-depth multilevel analyses were restricted to the asthma dataset, and to the neutropenia dataset with a focus on levels 1 and 2.

Table 2. Presence of higher level variation according to random intercept variance components models

Dataset / analysis (endpoint)	Intercept	Intercept variance (standard error)	p value[a]	Intra-higher level correlation
Level 2				
Asthma dataset (direct medical cost of asthma, log scale)	4.249	0.140 (0.047)	< 0.001	0.175
Gatekeeping dataset, logistic model of any costs being accrued (total health care costs and outpatient costs)	5.618	0.259 (0.389)	0.201	--
Gatekeeping dataset, GLM of costs in those with non-zero costs (outpatient costs)	4.884	0.000 (0.000)	0.500	--
Gatekeeping dataset, GLM of costs in those with non-zero costs (total health care costs)	7.361	0.000 (0.000)	0.500	--
Neutropenia dataset (any neutropenic event occurrence)[b]	-6.734	0.585 (0.156)	< 0.001	0.151[c]
Level 3				
Neutropenia dataset (any neutropenic event occurrence)[b]	-6.734	0.002 (0.029)	0.473	--

[a] Likelihood ratio test, p-value divided by 2 (see Methods, p. 35).

[b] Based on three-level random intercept variance components model (clustering option not specified).

[c] Based on two-level random intercept variance components model (clustering option not specified); calculated as proposed by Snijders and Bosker [164: 114·192].

Asthma dataset: result of multilevel analysis

In the asthma dataset, significant level 2 variation was observed for the variable distinguishing quick reliever from controller therapy and for the variable describing employment status in those aged 65 years or younger. The random intercept term became non-significant when the first of the afore-mentioned covariates was allowed to vary at random. None of the primary level 2 variables available (representing urban or rural area and language region), or of the level 1 variables aggregated at level 2 and tentatively added to the model, were significant at the fixed effects level or showed significant random variation.

The explanatory variable "involvement of a pulmonologist in diagnosis or treatment", which was used in the conventional regression model, could not be assigned to either level 1 or level 2. For further analysis, it was split up in its original components, i.e. "diagnosis of asthma by a pulmonologist" and "specialty of treating physician" (comprising the categories of general practitioner, specialist in general internal medicine, pulmonologist, and pediatrician [203]). The latter covariate was non-significant and eliminated from the model. Diagnosis of asthma by a pulmonologist was significant at level 1 (but not at level 2) and thus retained.

In a subsequent step, one level 2 unit (physician) with outlying characteristics and three level 1 observations (patients) who turned their corresponding level 2 units into apparent outliers were identified and found to substantially influence the model. As a consequence of absorbing them into dummy variables, the covariate describing employment status in those aged 65 years or younger became non-significant at level 1 and level 2 and was thus removed. The coefficients for age and age squared were somewhat reduced and the latter became non-significant but was retained to facilitate comparison with the conventional regression model reported in chapter 4.

The resulting final multilevel model (Table 3) differed in three respects from the conventional model shown in chapter 4, Table 4. (1) It contained a random term for the variable distinguishing quick reliever from controller therapy. (2) The covariate "involvement of a pulmonologist in diagnosis or treatment" was replaced by a related covariate, "diagnosis of asthma by a pulmonologist". (3) One level 2 unit and three level 1 observations were absorbed into dummy variables.

The coefficient estimates for the fixed effects parameters contained in both the conventional model and the final multilevel model had the same direction. Differences in size were small to moderate.

Concentration of level 2 random variation in the covariate indicating the use of controller therapy, as opposed to quick reliever therapy, indicated physician-specific differences in the cost impact of this choice (see comparison section below and chapter 8, p. 129).

Residuals diagnostics for this final multilevel model of asthma costs showed acceptable properties (Figures 1-3). Some high predicted values with zero residuals, as appearing in Figure 3, were a consequence of absorbing influential observations into dummy variables, as described.

Table 3. Final two-level model of direct medical costs (log scale) induced by asthma, including a random effect for quick reliever versus controller therapy[a]

Number of level 1 units = 420					**Log likelihood = -510.128**	
Number of level 2 units = 107					**AIC 1060.26**	
Variable	**Esti-mates**	**Std. Err.**	**Test stat.[b]**	**p value**	**95% Confidence Interval**	
Fixed part						
Degree of severity:						
Mild persistent[c]	0.917	0.176	5.21	< 0.001	0.572	1.262
Moderate persistent[c]	1.136	0.174	6.53	< 0.001	0.795	1.476
Severe persistent[c]	1.045	0.189	5.52	< 0.001	0.674	1.416
Exacerbations present	0.386	0.262	1.47	0.140	-0.127	0.898
Interaction of degree of severity and presence of exacerbations, ordinal:						
Level 1[d]	-0.517	0.315	-1.65	0.100	-1.134	0.099
Level 2[d]	0.088	0.303	0.29	0.7773	-0.506	0.681
Level 3[d]	0.371	0.304	1.22	0.222	-0.224	0.966
Age (centered)	0.007	0.002	3.22	0.001	0.003	0.012
Age squared (centered)	-0.00014	0.00010	-1.36	0.173	-0.00033	0.00006
Asthma-related comorb. present	0.304	0.116	2.62	0.009	0.076	0.531
Diagnosis by a pulmonologist:						
No[e]	-0.234	0.098	-2.37	0.018	-0.426	-0.041
Unknown[e]	-0.342	0.132	-2.59	0.010	-0.601	-0.083
Controller therapy[f]	0.589	0.108	5.45	< 0.001	0.377	0.801
Dummy for influential observation 1	-2.849	0.872	-3.27	0.001	-4.557	-1.140
Dummy for influential observation 2	3.527	0.857	4.12	< 0.001	1.848	5.206
Dummy for influential observation 3	3.848	0.783	4.91	< 0.001	2.313	5.383
Dummy for influential level 2 unit	-1.279	0.499	-2.56	0.010	-2.257	-0.301
Intercept	5.278	0.173	30.45	< 0.001	4.938	5.618
Random part - level 1						
Residual variance	0.584	0.046			0.494	0.674
Random part - level 2						
Controller therapy variance	0.145	0.052	13.46	< 0.001	0.052	0.252

a Conventional regression model for comparison: chapter 4, Table 4.

b Fixed parameters, Wald test based on z statistic; random variance, likelihood ratio test, p-value divided by 2 (see Methods, p. 35). Standard errors and CIs based on the Wald statistic throughout.

c Compared to mild intermittent. Wald test for this set of variables, p < 0.001.

d Compared to level 0. Wald test for this set of variables, p = 0.002.

e Compared to yes. Wald test for this set of variables, p = 0.007.

f Compared to quick reliever therapy.

Figure 1. Final multilevel cost of asthma model - inverse normal plot of studentised level 1 residuals

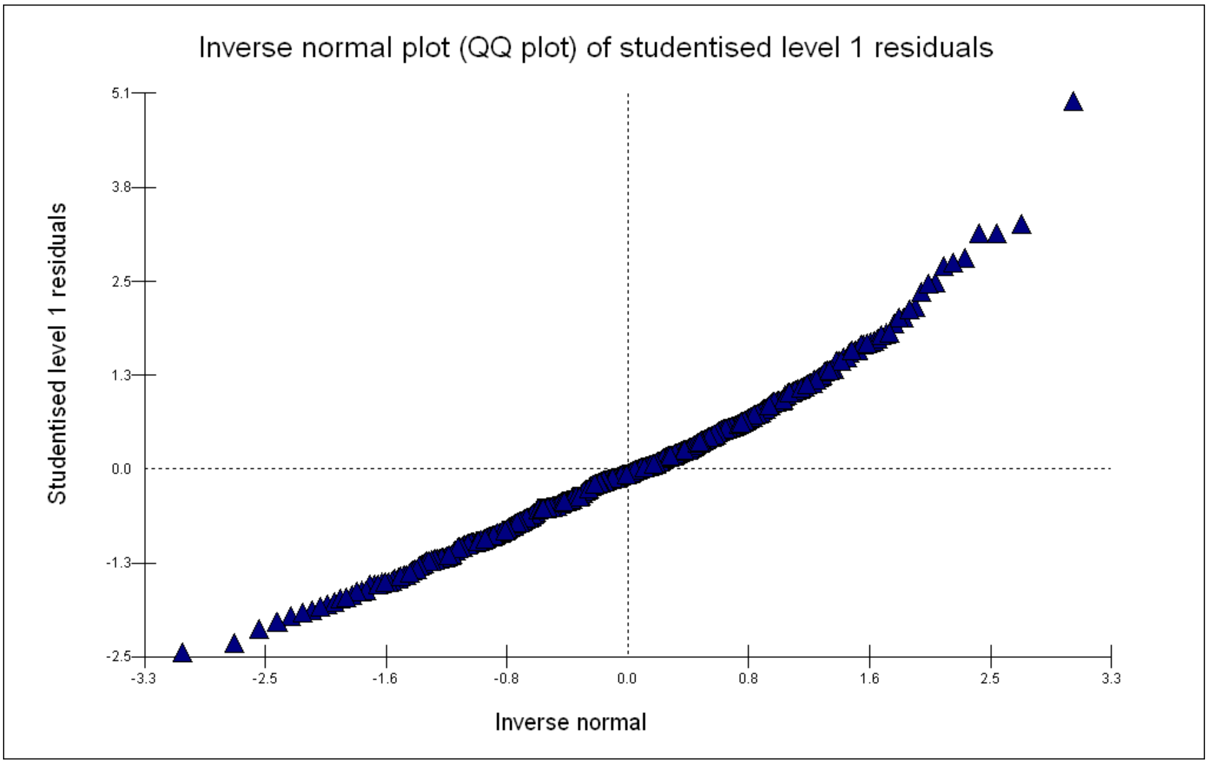


Figure 2. Final multilevel cost of asthma model - plot of studentised level 1 residuals against fixed part predicted values

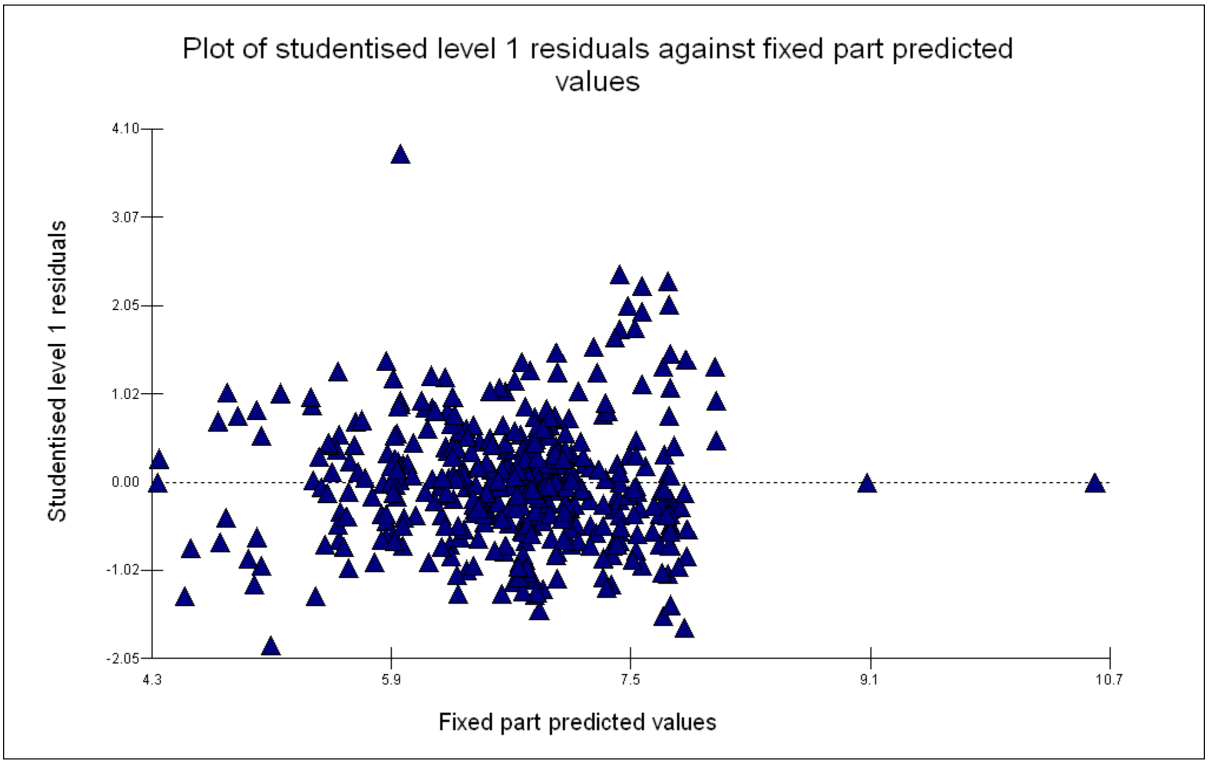
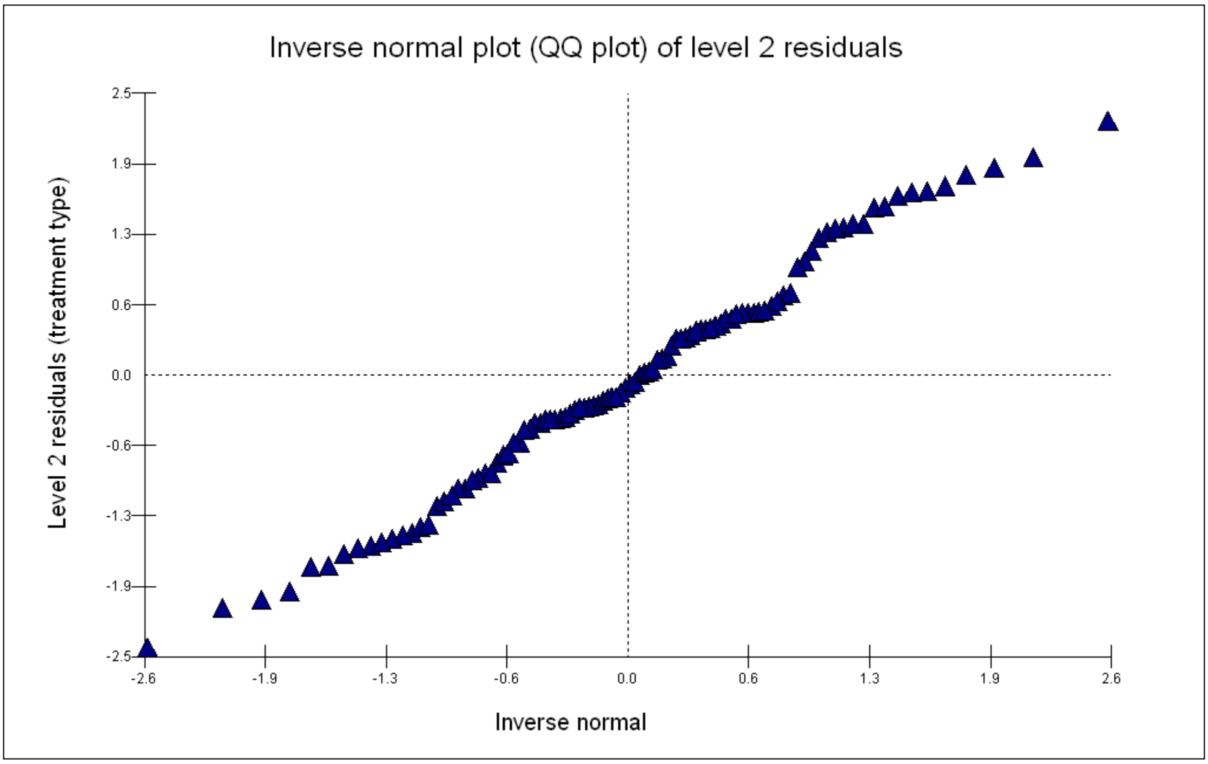


Figure 3. Final multilevel cost of asthma model - inverse normal plots of studentised level 2 residuals



Neutropenia dataset: result of multilevel analysis

In the neutropenia dataset, no significant level 3 variation was detected, but significant level 2 variation was observed for the intercept and for the variable indicating the use of an anthracycline-based chemotherapy regimen. The corresponding covariance term was also significant. In contrast, the fixed effects terms representing year of treatment, year of treatment squared, and interaction between year of treatment and chemotherapy regimen type became clearly non-significant and were removed from the model. All other fixed effects coefficients and their standard errors remained essentially stable.

Some aggregated variables tentatively added to the model were found to be significant at the fixed effects level, but did not show any significant random variation. They represented, at the centre level, the proportion of patients aged 65 or older, the proportion of patients with a BMI ≥ 30 kg/m^2, the proportion of patients receiving an anthracycline-based chemotherapy regimen, and the proportion of patients receiving a three-weekly CMF regimen. As these covariates did not contribute to the analysis of higher level variation or increase predictive ability, and might have introduced over-modelling, it was decided not to include them in the final multilevel model.

However, an alternative model containing them was fully assessed as described in Appendix I.

The resulting final multilevel model of any neutropenic event occurrence (Table 4) differed in two respects from the main conventional model shown in chapter 6, Table 3 (rewritten in Appendix I, Table 6 to facilitate comparison). (1) It contained variance terms for the intercept and for the variable indicating use of an anthracycline-based chemotherapy regimen, plus the corresponding covariance term. (2) The fixed effects terms representing year of treatment, year of treatment squared, and interaction between year of treatment and chemotherapy regimen type were dropped.

Table 4. Final two-level model of any neutropenic event occurrence (logistic regression); including variance and covariance terms for the intercept and use of an anthracycline-based chemotherapy regimen[a]

Number of level 1 units = 2'358
Number of level 2 units = 93
Log likelihood -1115.315

Std. err. adjusted for 6 clusters (audits)
AIC 2284.63

Variable	Estimates	Std. Err.	Test stat.[b]	p value[b]	95% Confidence Interval	
Fixed part						
Age[c]	0.022	0.006	3.52	< 0.001	0.010	0.034
Body surface area[c]	1.357	0.284	4.78	< 0.001	0.800	1.914
BMI[c]	-0.046	0.012	-3.83	< 0.001	-0.069	-0.022
Chemotherapy regimen:[d]						
Three weekly CMF	0.845	0.074	11.38	< 0.001	0.699	0.990
Anthracycline-based	0.323	0.181	1.78	0.075	-0.033	0.680
Taxane-based	1.591	0.072	22.13	< 0.001	1.450	1.731
Other	0.944	0.417	2.26	0.024	0.126	1.762
Normal to high SDI[e]	0.625	0.153	4.10	< 0.001	0.326	0.925
Planned chemotherapy cycles[c]	0.369	0.105	3.52	< 0.001	0.164	0.575
Concomitant radiotherapy administration (Rx):[f]						
Rx yes	2.94	0.936	3.14	0.002	1.105	4.775
Rx unknown	-0.927	1.825	-0.51	0.612	-4.503	2.650
Interaction of chemotherapy regimen and Rx:						
Three weekly CMF, Rx yes	-0.382	0.450	-0.85	0.396	-1.263	0.500
Anthracycline-based, Rx yes	-0.872	0.134	-6.52	< 0.001	-1.134	-0.610
Taxane-based, Rx yes	1.362	0.210	6.48	< 0.001	0.950	1.774
Other, Rx yes	--	--	--	--	--	--
Three weekly CMF, Rx unkn.	0.660	0.505	1.31	0.192	-0.330	1.650
Anthracycline-based, Rx unkn.	0.687	0.494	1.39	0.164	-0.281	1.656
Taxane-based, Rx unknown	--	--	--	--	--	--
Other, Rx unknown[g]	2.047	1.104	1.85	0.064	-0.116	4.210
Interaction of body surface area and Rx:						
Rx yes	-1.235	0.344	-3.59	< 0.001	-1.908	-0.562
Rx unknown	1.678	0.652	2.57	0.010	0.400	2.957
Interaction of planned chemotherapy cycles and Rx:						
Rx yes	-0.194	0.089	-2.19	0.029	-0.368	-0.020
Rx unknown	-0.401	0.083	-4.83	< 0.001	-0.563	-0.238
Interaction of normal to high SDI and Rx:						
Rx yes	1.135	0.327	3.47	0.001	0.493	1.777
Rx unknown	-0.104	1.003	-0.10	0.917	-2.070	1.861
Intercept	-6.895	0.747	-9.22	< 0.001	-8.360	-5.429

Table 4 ctd.

Random part - level 1						
Binomial variance	1[h]	0				
Random part - level 2						
Intercept variance	0.716	0.114	60.01 (2dgf)[i]	< 0.001	0.493	0.939
Anthracycline-based chemotherapy regimen variance	0.802	0.284	18.21 (2dgf)[i]	< 0.001	0.245	1.359
Covariance	-0.433	0.150	4.61	0.016	-0.727	-0.139

a Conventional regression model for comparison: chapter 6, Table 3, re-written to facilitate comparison in Appendix I, Table 6.

b Fixed parameters, Wald test based on z statistic; random parameters, likelihood ratio test, p-value divided by 2 (see Methods, p. 35). Standard errors and CIs based on the Wald statistic throughout. Wald tests for all sets of categorical or ordinal variables and for all sets of interaction terms, p < 0.01.

c Per one unit increase.

d Compared to four weekly CMF.

e Second to 4th quartiles compared to 1st quartile.

f Compared to no concomitant radiotherapy administration.

g Parameter estimate based on 4 observations, assumed to be an artefact.

h Constrained to 1 [169:35].

i Removing any of the level 2 random variance terms also removes the covariance term.

Abbreviations: C, cyclophosphamide; F, 5-fluorouracil; M, methotrexate; SDI, summation dose intensity.

The coefficient estimates for the fixed effects parameters contained in both the conventional model and the final multilevel model had the same direction. Differences in size were small to moderate. Exceptions from these rules were limited to some coefficients which were based on very few observations and thus considered to be doubtful from the beginning (terms involving taxane-based and "other" chemotherapy regimens), or were clearly non-significant in both models but were retained as parts of sets of parameters which were significant as a whole ("unknown" concomitant radiotherapy administration and term representing interaction between normal to high SDI and "unknown" concomitant radiotherapy administration).

The presence of substantial higher level variation in the variable indicating use of an anthracycline-based regimen indicated centre-specific differences in the neutropenia risk associated with this type of chemotherapy (see comparison section below and chapter 8, pp. 129-30).

Although no substantial variation was found at the audit level (level 3), GEE-based robust standard error estimates allowing for clustering by audit were used in order to ensure consistency with the main conventional model of any neutropenic event occurrence. Moreover, the power to correctly estimate variability at this level was presumably low, given a small number of 6 observational units only. Abandoning the approach to use robust standard errors led to increased standard error estimates, as in the corresponding conventional models. The differences were not such that they would have changed the conclusions or affected any decisions regarding the inclusion or exclusion of model parameters (see Appendix I).

Graphical assessment indicated acceptable, albeit not ideal model fit (Figures 4 and 5), and there was no indication of substantial extra-binomial variation.

Figure 4. Main multilevel model of neutropenic event occurrence - mean observed against mean predicted event probabilities, by deciles of the linear predictor

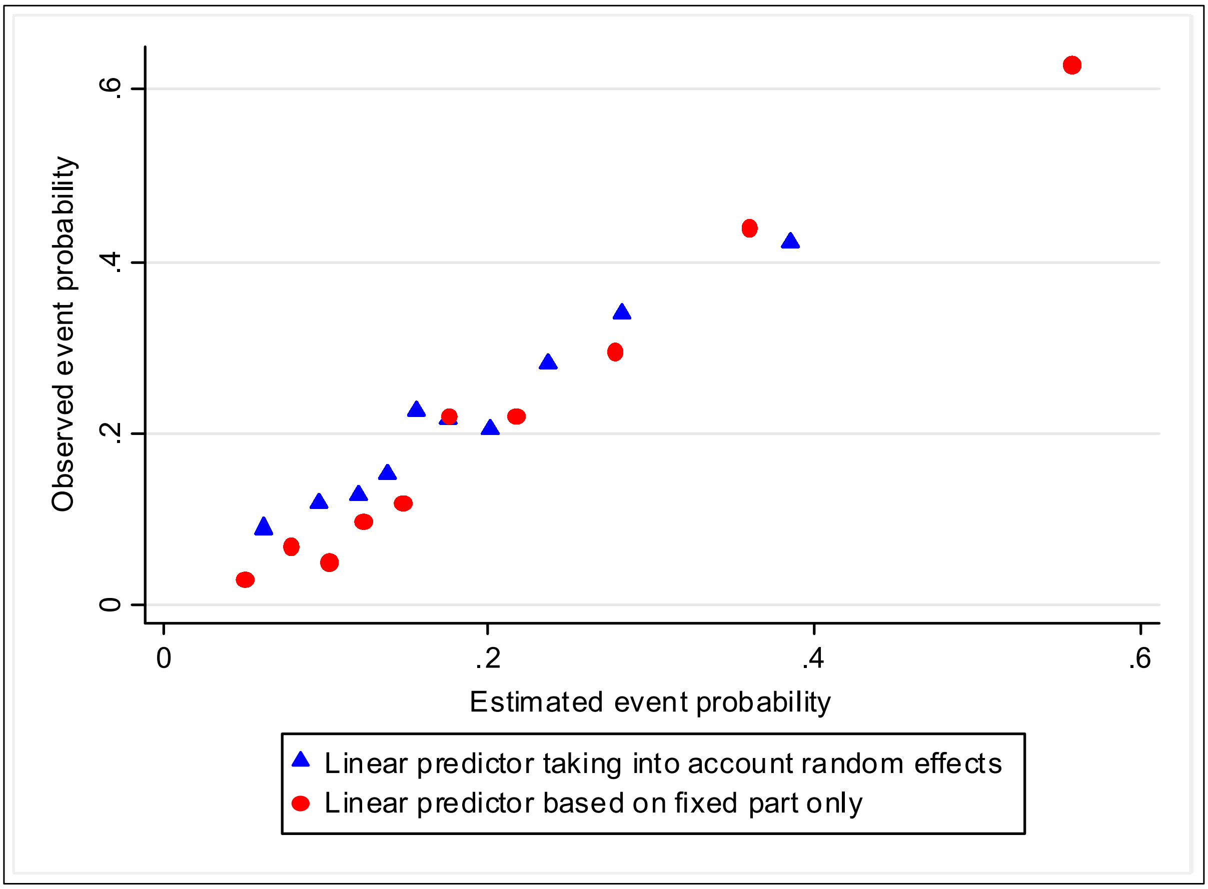
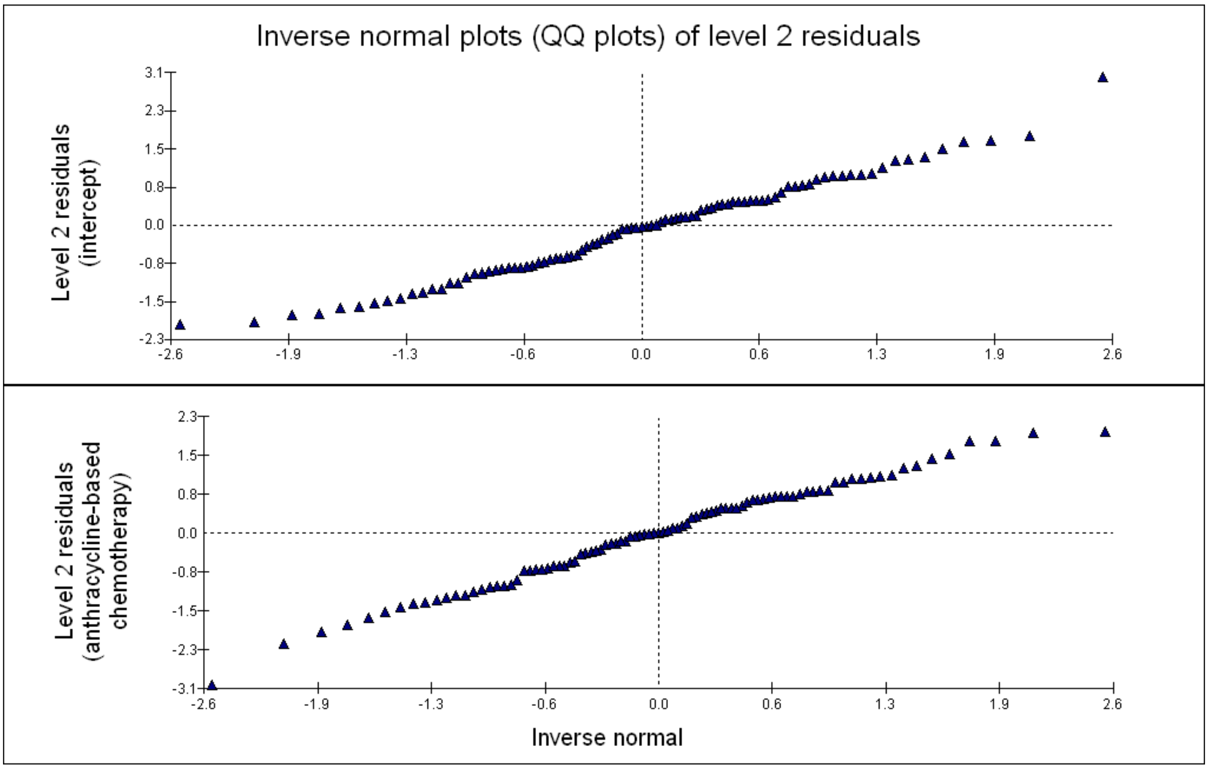


Figure 5. Main multilevel model of neutropenic event occurrence - inverse normal plots of studentised level 2 residuals



Predictive ability of conventional regression models and multilevel models

Conventional regression-based and multilevel-based predictive ability was compared for the asthma and neutropenia studies. In both cases, the predictive ability of three models was regarded. These were the final multilevel models (Tables 3 and 4 above), conventional models using the same fixed effects predictors as the final multilevel models, and the main conventional models reported in chapters 4 and 6.

Predictive ability details for the asthma study are shown in Table 5. The final multilevel model of asthma costs showed a distinct reduction in apparent prediction error, by 34% compared to the main conventional model and by 24% compared to the conventional model using the same fixed effects predictors as the multilevel model. This gain in apparent predictive ability is graphically demonstrated in Figures 6 and 7. The overlaid inverse normal plots of residuals shown in Figure 6 revealed a somewhat improved situation in the tail areas and lower residuals overall in the multilevel model. An overlaid plot of predicted against observed values (Figure 7) hinted at a somewhat better predictive ability of the multilevel model at both ends of the distribution.

Table 5. Apparent prediction error and cross-validation results for the cost of asthma models

Criterion	Main conventional model (chapter 4)	Conventional model using same fixed effects predictors as the multilevel model	Final multilevel model
N	420	420	420
AIC	1120.41	1072.24	1062.78
Apparent prediction error			
MSE	0.793	0.690	0.524
Cross-validation ignoring hierarchical structure			
MSE (e0 bootstrap)	0.880	0.863	--
MSE (.632 bootstrap)	0.848	0.799	--
MSE (10-fold cross-validation)	0.848	0.834	--
Cross-validation only taking into higher level units contributing to model estimation			
MSE (10-fold cross-validation)	0.839	0.825	0.789
Cross-validation only taking into higher level units not contributing to model estimation			
MSE (10-fold cross-validation)	0.857	0.847	0.845

Bootstrap-based and 10-fold cross-validation performed on the conventional models showed a moderate increase in prediction error, compared to the apparent situation. Conventional-based cross-validation results were consistent across methods and only slightly affected by either ignoring the hierarchical structure of the data or restricting the test set to observations whose corresponding higher level units had or had not contributed to model estimation. Ten-fold cross-validation of the final multilevel model revealed a more pronounced increase in prediction error compared to the apparent situation. A modest gain in efficiency was retained compared to the conventional models if the test set was restricted to observations whose corresponding level 2 units had contributed to model estimation. In this case, the multilevel-based prediction error was 6% lower than in the main conventional model and 4% lower than in the conventional model using the same fixed effects parameters as the multilevel model. If the test set was restricted to observations whose corresponding higher level units had not contributed to model estimation, the gain in predictive ability compared to the conventional models was minimal.

Figure 6. Cost of asthma models - overlaid inverse normal plots of level 1 residuals derived from the final multilevel model and of residuals derived from the main conventional model

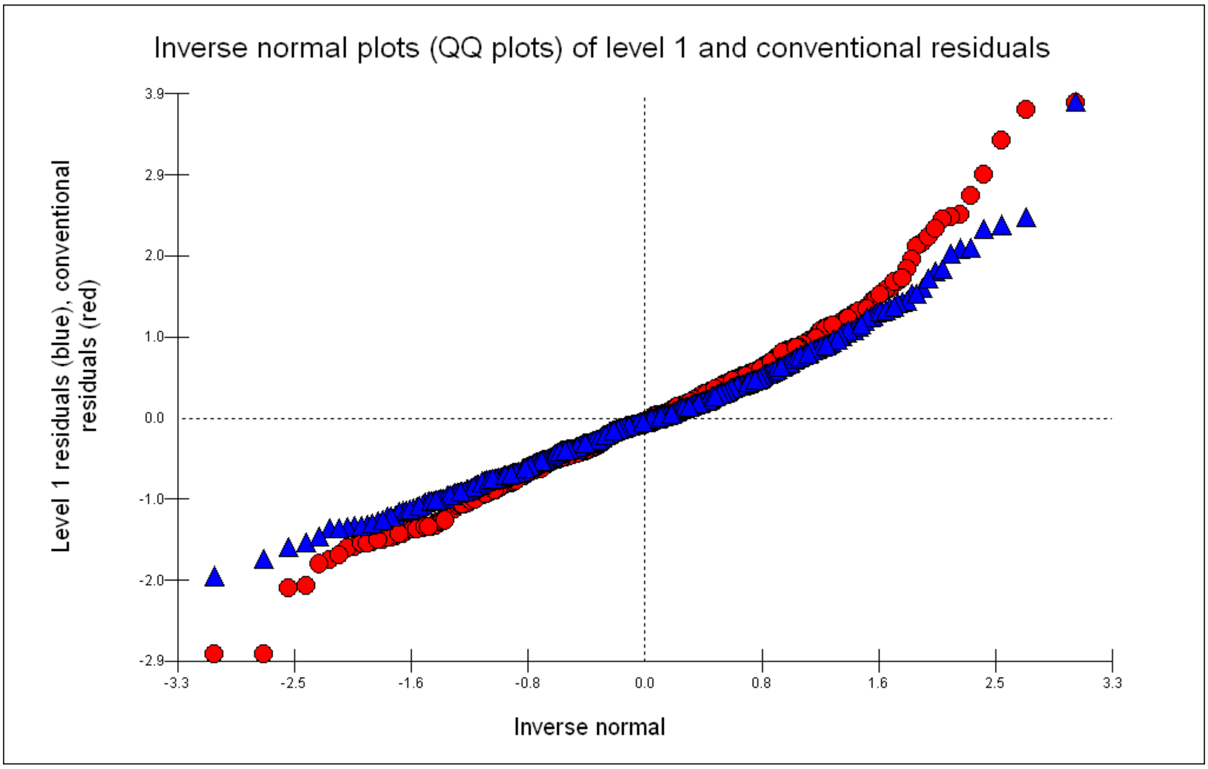
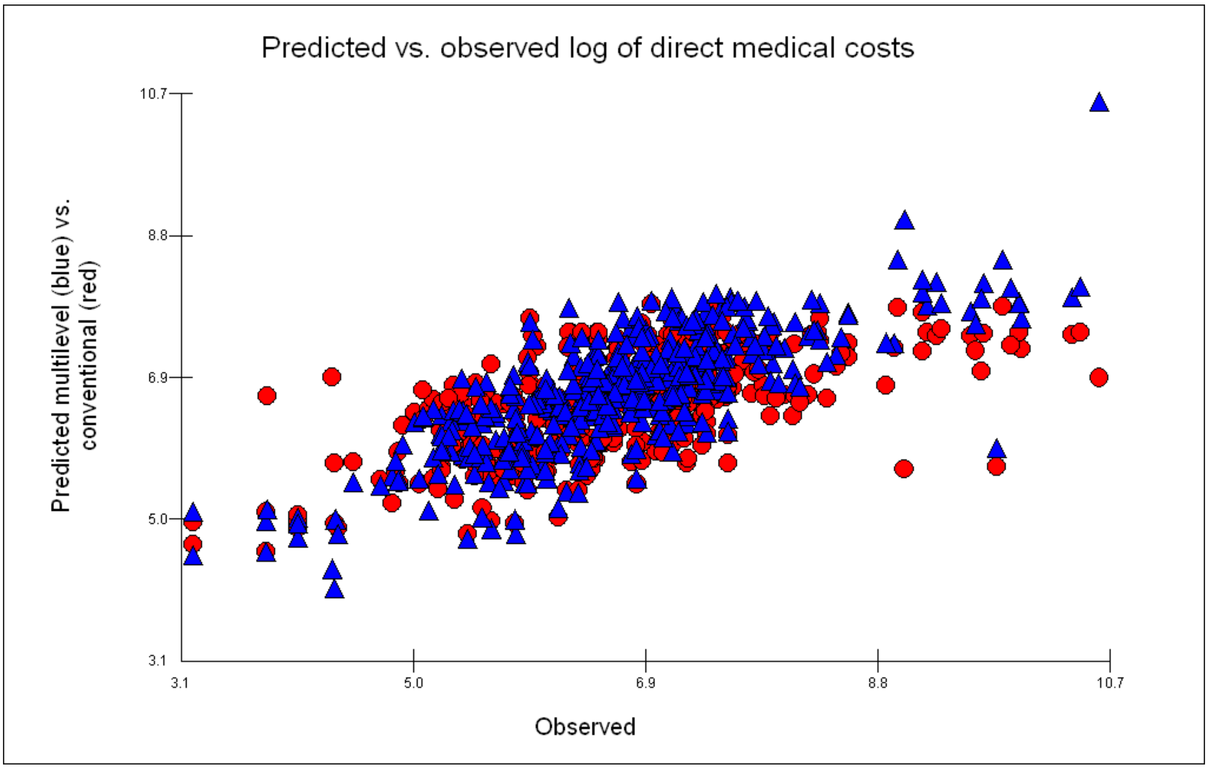


Figure 7. Cost of asthma models - predicted vs. observed log of direct medical costs, predicted values of final multilevel model overlaid with predicted values of main conventional model



In the neutropenia study, the prediction error of the main multilevel model was again distinctly reduced compared to the conventional models in the apparent situation. The MSE indicator was reduced by 13-14% and classification error by 11%. In 10-fold cross-validation, if the test set was restricted to observations whose corresponding level 2 units had contributed to model estimation, the MSE indicator was reduced by 6-7% and classification error by 4-6%. If the test set was restricted to observations whose corresponding level 2 units had not contributed to model estimation, no substantial improvement compared to the conventional models remained. Details are shown in Table 6.

Table 6. Apparent prediction error and cross-validation results for the models of neutropenic event occurrence

Criterion	Main conventional model (chapter 6)	Conventional model using same fixed effects predictors as the multilevel model	Main multilevel model
N	2'358	2'358	2'358
AIC	2355.94	2377.04	2284.63
Apparent prediction error			
MSE / classification error	0.158 / 21.80%	0.160 / 21.80%	0.137 / 19.38%
Cross-validation ignoring hierarchical structure			
MSE (e0 bootstrap)	0.164 / 22.42%	0.164 / 22.00%	--
MSE (.632 bootstrap)	0.162 / 22.23%	0.163 / 21.93%	--
MSE (10-fold cross-validation)	0.163 / 22.43%	0.164 / 22.02%	--
Cross-validation only taking into higher level units contributing to model estimation			
MSE (10-fold cross-validation)	0.162 / 22.49%	0.163 / 21.87%	0.152 / 21.04%
Cross-validation only taking into higher level units not contributing to model estimation			
MSE (10-fold cross-validation)	0.172 / 23.21%	0.169 / 22.65%	0.170 / 22.44%

Comparison of conventional regression-based results and multilevel modelling results

With respect to the response variables of interest, substantial and significant higher level variation was detected in the asthma dataset and in the neutropenia dataset, but not in the gatekeeping dataset. For the latter, the key findings of the original analyses were confirmed, but no additional insights were gained.

In the asthma and neutropenia studies, 18% and 15% of the residual variance, respectively, was between-higher level unit variance. More detailed analysis of the variance structure identified the predictors in which a substantial amount of this higher level variance occurred. In both cases, these predictors were not level 2 predictors in the strict sense (i.e., direct higher level unit characteristics), but level 1 predictors which were influenced by level 2 characteristics.

In the asthma study, most of the higher level random variation occurred in the covariate indicating the use of controller therapy, as opposed to quick reliever therapy, showing physician-specific differences in the cost impact of this choice. In the neutropenia case, some of the higher level variation was concentrated in the covariate indicating use of an anthracycline-based regimen, thus showing centre-specific differences in the neutropenia risk associated with this type of chemotherapy. In contrast, no higher level variation was seen in the other chemotherapy types represented in the dataset. Such additional information could not have been gained using conventional regression methods.

The sets of fixed parameters used in the multilevel models differed slightly from those used in the conventional models, for three types of reasons. In the asthma case, one of the covariates used ("involvement of a pulmonologist in diagnosis or treatment") was composed of a level 1 characteristic ("diagnosis of asthma by a pulmonologist") and a level 2 characteristic ("specialty of treating physician"). In order to achieve clarity in multilevel analysis, this covariate was split up into its components and one of these components was eliminated due to non-significance. The second reason for modification of the set of fixed parameters used in the asthma analysis was the absorption of some influential level 2 units or their corresponding observations into dummy variables. Finally, in the neutropenia analysis, the terms representing year of treatment and year of treatment squared, and their corresponding interaction terms, lost their statistical significance when the variance structure was taken into account. This was not counterintuitive, as the level 2 units (centres) recruited their patients at different points in time, during a period of rapidly growing experience with the use of then novel chemotherapy regimens (involving anthracyclines and taxanes). The time effect may thus have been absorbed into the level 2 residuals.

The coefficient estimates for the remaining fixed parameters had the same direction as their equivalents in the conventional regression models. Some of them differed in size, but not to an extent that would have necessitated different interpretations. Most standard errors were slightly reduced in the multilevel models, confirming the theoretically expected gain in statistical efficiency. Exceptions from these rules were limited to some coefficients in the model of neutropenic event occurrence which were based on very few observations and thus considered to be doubtful from the beginning (terms involving taxane-based and "other" chemotherapy regimens), or were clearly non-significant in both models but were retained as parts of sets of parameters which were significant as a whole.

The AIC, as a criterion of model fit, was reduced in the multilevel models, compared to the corresponding conventional models, in all cases (see Tables 5 and 6 above).

In the asthma study as well as in the neutropenia study, the predictive ability of the multilevel models was improved in the apparent situation. In the asthma study, prediction error as measured using the MSE indicator was reduced by 24-34%. In the neutropenia study, the reduction was 13-14% for the MSE indicator and 11% for the classification error indicator. However, these gains were only partially confirmed under 10-fold cross-validation conditions. If the test set was restricted to observations whose corresponding higher level units had contributed to model estimation, the MSE indicator was reduced by 4-6% compared to conventional in the asthma study. In the neutropenia study, the MSE indicator was reduced by 6-7% and classification error was reduced by 4-6% compared to conventional. If prediction was restricted to those observations whose corresponding higher level units had not contributed to model estimation, the gain in predictive ability was essentially lost.

In summary, multilevel modelling of direct medical asthma costs and of any neutropenic event occurrence made an additional contribution, compared to the conventional models, by providing new insights in the sources of the variance seen in the response variables, and by slightly increasing statistical precision. Moreover, there were modest gains in terms of predictive ability for out-of-sample observations whose corresponding higher level units had contributed to model estimation. The main findings of the conventional analyses remained fully valid.

Chapter 8

Discussion and conclusions

This work is based on three health care-related observational studies conducted and/or analysed by the author (chapters 4-6). All three were primarily analysed using conventional multivariate regression methods. In an additional step, re-analyses of all three datasets were performed using multilevel modelling, a statistical technique taking hierarchical data structures into account (chapter 7). Taking a more distant viewpoint then in the published discussion sections of chapters 4-6, this overall discussion tries to put into a broader perspective the contributions of the cost of asthma, gatekeeping, and neutropenia studies. In a second step, the contribution of multilevel modelling to these studies, and to other studies in the underlying fields of research, is reviewed. Based on this and other available information on the possibilities, disadvantages, and current spread of multilevel modelling, conclusions and recommendations regarding the use of this statistical technique are proposed and associated aspects of study design are briefly addressed.

Cost of illness of asthma

Cost of illness studies in the field of asthma uniformly show a considerable economic impact, with direct medical asthma costs accounting for 1-3% of total health care expenditures in most countries [138]. However, studies vary widely with respect to their choice of methodology, perspective of cost assessment, and patient population regarded. In adults, direct medical costs are typically reported to be responsible for 30-70% of total asthma costs, with indirect costs due to asthma morbidity (i.e., lost working days and early retirement) accounting for the remainder [11·28·39·180·182]. If costs due to asthma mortality are additionally taken into account, the share of indirect costs may even be higher [199]. The key components of direct medical costs, jointly contributing about 70-80% of the total in most studies, are medication costs, followed by emergency room visits and hospitalisations [11·18·39·79·115·148·159·203·217·222]. In the paediatric setting, a substantial number of studies show essentially comparable results but a wider variation in the proportion of emergency room and hospitalisation costs, and a lower importance of indirect costs [11·78·198·215].

Costs due to emergency room visits and hospitalisations are concentrated on a relatively small subgroup of asthma patients and are a major cost component in this group [25·27·146·188·216]. Consequently, correlates of asthma-related hospitalisations have been interpreted as correlates of asthma costs [79]. Higher disease severity has been firmly established as a cost driver [11·81·180·182·211]. The independent impact of asthma control, i.e. of the freedom from vs. occurrence of asthma exacerbations, has also been addressed [11·103·177·195]. The study reported in chapter 4 of this thesis has been the first one in adults to establish this link using multivariate regression methods, and to confirm an interaction between bad asthma control and a high degree of asthma severity. In the meantime, the cost impact of asthma exacerbations has been confirmed by several other studies [101·200]. More insights in the impact of bad asthma control are expected from the ongoing TENOR study, a large three-year multi-centre cohort study conducted in the US [55].

Recent publications also provided additional confirmation for the other patient-level factors the author found to be associated with higher asthma costs, i.e. higher age [199], asthma-related co-morbidity [94], and the use of controller vs. quick reliever therapy. Controller therapy has been strongly advocated in recent years and was often discussed in the context of guidelines adherence and patient as well as physician compliance [3·11·79·93]. Typically, substantial clinical benefits were observed at a moderately increased cost [75·93·189·201], which is consistent with the author's results.

Apart from the author's analysis, no multivariate regression analyses directly addressing person-level asthma costs are available for adults, but some multivariate analyses of key correlates of cost have been published. Using logistic regression analysis, Hoskins et al. found the risk of asthma exacerbations to be associated with the time of day of symptom occurrence, and with exercise-induced symptoms [103]. Schatz et al. found lower median household income, previous emergency room visits

/ hospitalisations, indicators of increased medication use, and a higher number of drug prescribers to be associated with asthma-related hospitalisations.

In children, multivariate analyses confirmed higher asthma severity, the use of peak expiratory flow rate meters, younger age, low-income status, non-white race, and longer duration of asthma to be associated with higher costs [78]. Ungar et al. reported influences of asthma control, age, and season [209]. Van Ganse et al. reported a link between disease severity and emergency rooms visits / hospitalisations [211]. Non-multivariate analyses highlighted the impact of tobacco smoke [140] and low economic status [5].

Although reports of multivariate-based research findings are rare overall, the association of disease severity / asthma control with direct medical asthma costs has been firmly established. In clinical terms, finding ways to improve asthma control may now be of major importance [11·22·79·195]. If it is felt, however, that additional studies are needed to achieve final clarity with respect to other potential influences, or for prediction purposes, the underlying data collections should ideally be prospective and measure all covariates of potential interest currently known, inclusive of provider-level characteristics. Other than in the author's study, disease severity should be assessed using physiologic measures, not medication characteristics, in order to avoid potential circularity.

Gatekeeping vs. fee-for-service based health insurance

The ability of the gatekeeping approach to contain or reduce health care costs in general populations is still under debate and the available evidence remains inconclusive, as briefly mentioned in chapter 5. Some trials and observational studies conducted in general populations in the US, as well as the author's study, hinted at such an ability [77·120·136·137·178], but other observational studies also conducted in general populations didn't [24·109·128·212]. Irrespective of the results, the vast majority of these studies compared gatekeeping with either fee-for-service based health insurance or with other settings lacking strong incentives to restrict the use of medical services. Some authors comparing gatekeeping plans to point-of-service plans, which allow self-referral to specialists but use other strong incentives such as substantial patient co-payments, did not find financial advantages of gatekeeping

[65·112·113]. To complete the picture, some analyses addressed specific health care resources whose use is usually expected to be influenced by gatekeeping (i.e., emergency room visits, re-hospitalisations), and did indeed demonstrate beneficial effects [72·190]. In contrast, a general reduction of specialist referrals by gatekeeping could not be firmly ascertained [70·71].

Among the most recently published studies, a US study using 1996 administrative and survey data from 8'195 adults showed modestly lower health care expenditures in gatekeeping arrangements as compared to fee-for-service based insurance [155]. However, this study was purely cross-sectional and no baseline cost data were available to the authors. A study by Bhat, looking at aggregate data from 24 OECD countries and using econometric methods, found an increased efficiency of health care delivery in countries widely using a gatekeeping approach to control the access to specialist and hospital care [17]. This result is essentially consistent with an earlier study by Delnoij et al., who also used aggregate data [49].

Negative aspects of the gatekeeping approach which may partially explain the overall unclear picture include a potential tendency to merely shift resource use from the higher levels of care to primary care [69]. Concerns about withholding care, and thus, reduced quality of care, have been raised but not demonstrated [174·184]. Moreover, such effects would be unlikely to influence the results of studies concentrating on expenditures, except if they had substantial health consequences in the short- to mid-term.

Unequivocal are the challenges faced when gatekeeping plans are to be assessed. All studies tried to take into account issues of patient self-selection and varying case-mix, usually by applying multivariate regression methods. However, the sets of covariates available were often limited. Some studies were purely cross-sectional and could not address issues of temporal sequence. The fact that in general populations, a substantial proportion of subjects has zero medical resource use, and thus, zero expenditures, was efficiently addressed by estimating two-part regression models [57·65·128·155]. Logistic regression was used to model if the response variable was non-zero, and multiple regression or (in the author's study) GLMs were used to model the response in those subjects with a non-zero response. In order to meet the assumptions of multiple regression, response variables were sometimes transferred to the log scale, which involved a re-transformation problem when combined effect estimates were to be generated [56·155]. In contrast, the GLM approach avoids the re-transformation problem. Beyond these methodological and statistical challenges, a more general problem is the use of gatekeeping in a wide variety of situations, where other techniques of utilization management or incentives are also in place in different combinations. This includes the possibility of interaction and makes it very difficult to isolate effects. Moreover, diverse comparators are involved, which has a negative impact on generalisability. Most studies were conducted in settings where these issues could not be addressed satisfactorily. The author had a rare opportunity to study isolated gatekeeping, and he found cost-savings compared to classical fee-for-service based health insurance. However, caveats remained, due to small sample size and because selection effects could not be ruled out entirely.

Given rapidly changing health care environments, "the gatekeeping question" may not find a generalisable answer. Studies addressing specific applications of the gatekeeping approach remain nevertheless important.

Neutropenic events in breast cancer chemotherapy

In cancer patients, the risk of chemotherapy-induced neutropenia (CIN), febrile neutropenia (FN), and related events is influenced by treatment-related, patient-related, and disease-related factors. Among these, the chemotherapy treatment regimens used are perceived as key factors "setting the stage" [132·153]. For example, in breast cancer, the neutropenic potential of the classical combination of cyclophosphamide, methotrexate, and 5-fluorouracil (CMF) using a cycle length of four weeks was generally lower than that of anthracycline-containing regimens (although in the author's study, this became only visible in multivariate analysis). Modern anthracycline-taxane combinations were found to be even more aggressive [191]. Differences in dosage and schedule can have a very strong impact [191]. Accompanying anti-malignant treatments (e.g., concomitant radiotherapy administration) have been discussed as modifiers of risk [50·185], and the protective role of supportive treatments (i.e., myelopoietic growth factor administration) is generally recognised [40·191].

Current opinion on patient-related and disease-related risk factors has been condensed into comprehensive reviews [48·132]. Proposed patient-related factors include higher age, performance status / functional impairment, presence of severe co-morbidity (including, but not limited to, heart disease, renal disease, and liver disease), and laboratory abnormalities. Laboratory abnormalities of interest may either occur pre-treatment (e.g., low pre-treatment white blood cell count, neutrophil count, or haemoglobin level) [131·132], or they may occur during the first cycle of chemotherapy treatment and predict neutropenic events in subsequent cycles (e.g., low nadir neutrophil count; day 5 lymphopenia; CD-4 lymphopenia) [19·26·132·185]. Laboratory abnormalities may be correlates of advanced disease, co-morbidity, or low chemotherapy tolerance. Proposed disease-related risk factors include tumour type and stage of disease [132]. However, it remains unclear if these factors have a stand-alone importance or if the observed increases in risk are merely due to more aggressive therapies being used in patients with, e.g., haematological malignancies or advanced disease.

The available information has been used to develop and update guidelines of myelopoietic growth factor use [147·153·191]. However, to date, no comprehensive, clinically applicable risk models of neutropenic event occurrence have been presented or validated. There may be several partial explanations for this situation.

- Much of our knowledge of the neutropenic potential of specific chemotherapy regimens stems from clinical trials of chemotherapy efficacy. Such trial-based reports have been shown to vary widely and their reliability and validity may be severely affected by study-specific differences in adverse event reporting [47].
- Various retrospective observational studies (including the author's study reported here) identified candidate risk factors and tentative risk models. However, many of these reports remained spurious and were not confirmed by other studies. No satisfactory, theory-guided integration of the available candidate risk factors into a comprehensive risk model has been proposed so far.
- Cross-validation or external validation were rarely used [132]. Some few efforts were made, but the models assessed were rather limited in scope and concentrated on laboratory abnormalities such as first cycle neutropenia [172·185] or first cycle lymphopenia [26·38·166]. Both these indicators were shown to have an impact. There is a single example of a (non-randomised) clinical trial

conducted to confirm a neutropenic event risk factor, namely the prognostic value of first cycle neutropenia [171].

Future work should re-evaluate the risk factors that have been proposed to date and, on this basis, establish physiologically plausible risk models. These should then be cross-validated and externally validated against as many as possible suitable datasets. It would be an option to combine such datasets to increase statistical power, which would require to appropriately take into account the heterogeneity of the individual datasets. In this context, the multilevel approach to regression modelling might be a helpful technique. Another key requirement would be to adequately model the neutropenic potential of the chemotherapy treatments used. Given the primary importance of this factor, it would presumably be very difficult otherwise to correctly assess weaker influences. (In the past, grouping variables were sometimes used to represent identical or similar combinations of anti-malignant drugs, but with very different dosages and administration schedules. As a refinement, the approach by Hryniuk et al., to express the dose-intensity of a wide variety of breast cancer chemotherapy regimens on a single scale [106], was used to great advantage in the author's regression models reported here, but this can only be seen as a first step. It may prove necessary to move away from summary regimen descriptions altogether. Instead, covariates individually describing the planned dose intensities of the drugs with the highest neutropenic potential could be introduced, and be allowed to interact where physiologically plausible.)

At a more mature stage of risk model development and validation, clinical trials could be used for final assessments. For example, patients could be randomised to receive either primary myelopoietic growth factor prophylaxis if the risk model under study indicated a risk above a certain threshold (treatment arm), or no primary myelopoietic growth factor prophylaxis (control). Such trials would typically be conducted in low- to medium-risk populations where primary growth factor prophylaxis is not recommended by current guidelines. It would be unethical to conduct them in patients receiving high-risk chemotherapy regimens where prophylaxis is indicated anyway.

Character of analyses performed

The analyses reported in this thesis addressed either candidate risk factors for clinical events of interest, with a potential for serious medical and economic sequelae, or potential correlates of high health care costs. The statistical methodology used was structurally the same.

All three underlying datasets had some weaknesses. In the asthma and neutropenia datasets, some known predictors of importance were either unavailable, affected by a substantial amount of missing values (e.g., baseline neutrophil counts in the neutropenia dataset), or of sub-optimal quality (degree of disease severity in the asthma dataset based on medication, not on physiologic measurements). In the gatekeeping dataset, limited sample size and the need to reduce statistical noise were major issues. At least in the asthma and gatekeeping studies, it was not the main goal, and it was clear from the beginning that it would not be possible, to establish comprehensive and final statistical models with a high predictive ability. The contribution of and coefficients for some specific covariates (asthma control; health plan membership) were of particular interest here [116:409]. In the neutropenia study, tentative risk scores were developed, but were not seen as complete solutions for immediate use in clinical practice. Elements of future, more complete models were to be ascertained and an impression of their possible contribution was to be gained. In all three cases, it was obvous that additional studies would be needed for more final results and consequently, the available statistical power was used without splitting the datasets, to identify or confirm influences on the outcomes of interest and to estimate effect sizes. No cross-validation efforts were made at this stage. (However, independently of this, cross-validation was used to assess the benefits of multilevel modelling.) External validation was not considered due to a lack of appropriate validation datasets.

All three modelling processes were based on some concepts but, in essence, limited ascertained knowledge of the underlying processes and causal relationships. The models estimated can thus be viewed as empirical models in the sense of Cox, with some substantive components [43·186:252-4]. The "model generating approach" was used to make the best of the available information, i.e. tentative models were specified based on the available background information and then modified on an empirical basis to achieve the best possible fit of the data [186:254-5]. This approach to modelling may be regarded as less theory-driven then desirable, but appeared to be the most suitable under the given circumstances.

Apart from these issues, disregarding the multilevel (hierarchical) structure of the datasets used was perceived as the potentially most serious deficit of the primary analyses performed. Multilevel re-analyses were therefore performed and became the main focus of this thesis.

Contribution of multilevel modelling to the fields of research addressed

All three datasets analysed had multilevel or multi-membership data structures, but in the gatekeeping study, no higher level variation was detected. This can be interpreted as a confirmation of the appropriateness of the primary analysis performed. The finding of no significant higher level variation being present with respect to the binary response of accruing no or any health care costs, was less surprising than the finding of the amount of costs not being influenced by physician-level characteristics. This is because decisions to consult or not to consult a physician are typically much less influenced by physician-level characteristics than the amount of medical resources used and, consequently, the cost induced once a consultation occurs. With respect to the amount of costs in those with any costs, the invisibility of any higher level variation may be explained by the fact that we studied a general population where inter-person (level 1) variation is huge and determined by a multitude of covariates. Moreover, physician-level treatment styles impacting on cost may differ, and be relatively uncorrelated, across diseases, depending on personal interest, perceived importance of different pathologies, etc. Therefore, physician-level random effects may be more easily visible where a single disease or group of diseases is addressed, as was the case in the asthma study.

In the asthma and neutropenia datasets, substantial higher level variation was present. Multilevel analysis allowed

- to slightly increase statistical precision;
- to assess, for the outcomes of interest, the relative importance of patient-level and provider-level variation;
- to identify the sources of higher level variation;

- to identify spurious findings "dictated" by the impact of few influential observations, by analysing influential higher level units [83:1/11]. In the asthma study, the employment status variable became non-significant at both levels when the impact of as few as seven individual-level observations, of which four were nested within an influential higher level unit, was "neutralised" by using dummy variables;
- to modestly reduce prediction error (as assessed by cross-validation techniques) for out-of-sample observations whose corresponding higher level units contributed to model estimation;
- Some changes in the model parameters used occurred (removal of year of treatment in the neutropenia study), but other than in some published studies [9], these changes were plausible consequences of the multilevel modelling process itself and did not necessitate substantial changes in interpretation.

In both studies, although very different in detail, much or all of the higher level variation occurred in covariates representing clinical practice patterns, which is consistent with expectations from the literature [167·168]. In the asthma study, higher level variation occurred in the covariate indicating the use of controller therapy, as opposed to quick reliever therapy. This appeared plausible, as the value of this covariate was influenced by patient as well as physician characteristics. The random variation seen could be interpreted as being induced by a range of different (physician-specific) treatment intensities or strategies associated with the choice of controller therapy, from simple prescribing to using an enhanced level of diagnostic procedures and accompanying treatments. More specifically, it might reflect different degrees of adherence to treatment guidelines (see p. 121). As an additional aspect, the characteristics of the patients who received controller therapy may have differed across physicians. Therefore, it was not possible to decide if the higher level effect seen was purely contextual (i.e., confined to the provider-level), or if compositional effects (i.e., effects due to an uneven distribution of unmeasured patient-level characteristics) also played a role.

In the neutropenia study, some of the higher level variation was concentrated in the covariate indicating use of an anthracycline-based regimen, thus hinting at centre-specific differences in the neutropenia risk associated with this type of chemotherapy.

In contrast, no higher level variation was seen in the other chemotherapy types represented in the dataset. Generally spoken, it appeared plausible to find higher level variation in this group of covariates, as chemotherapy regimen choice was again patient-specific but higher level unit-dependent (i.e., centre-dependent), and as centres differed with respect to their frequency of use of, and experience with, different regimens. The fact that higher level variation was only confirmed for the anthracycline-based regimens, but not for other regimen types, found several partial explanations. At the time of data collection, clinicians were more experienced with CMF-type regimens then with anthracycline-based regimens, and the former had reached a higher level of standardisation. In contrast, the anthracycline-based group consisted of a wide range of different regimen specifications, which were clustered by centre. In the case of the taxane-based and "other" regimens, the amount of available information was small.

Overall, the main findings of the original analyses were confirmed and some relevant additional information was gained.

Literature searches did not identify any other studies applying multilevel modelling to the immediate research topics addressed by the author. In the wider field of managed care-related studies, one study used multilevel modelling to assess the impact of managed care penetration on patient-level health care costs [32]. A small impact of managed care on costs was visible in this cross-sectional study. Multilevel analysis allowed to identify some provider-level random effects which would not have been found by using conventional multiple regression. Two other studies used multilevel models to confirm provider-level influences on medical resource use [183·196]. In the neutropenia field, some random effects meta-analyses were conducted to address neutropenia-related prophylactic measures, or treatment strategies [12·63·76·156·197·202]. However, these studies are not directly comparable with the type of research reported here, although the models used are formally equivalent to random intercept multilevel models .

Consequently, to date, a major contribution of multilevel modelling to the fields of research addressed by the author cannot be claimed. However, there were some contributions in detail. Moreover, it has been demonstrated that provider-level characteristics may have a substantial influence and should be taken into account in future studies addressing similar topics.

Criteria for the use of multilevel modelling

Information on the current and potential contribution of multilevel modelling to health care-related research, in addition to the author's experience, stems from three types of sources. These are

- theoretical texts describing the technique, its expected usefulness and problems;
- review articles describing the use of multilevel modelling in different fields and listing promising areas of use, but typically without providing systematic assessments of the knowledge gains achieved [35·53·151·167·168]. Single case studies were often used to demonstrate such knowledge gains. Greenland refers to simulation studies which demonstrated advantages of multilevel modelling in a variety of situations including variable selection problems [91]);
- original research articles using multilevel modelling. However, only in some instances, comparisons of conventional-based and multilevel modelling-based results were provided (see [9·32] for examples and [95] for an example where no benefit of multilevel modelling could be demonstrated).

Following these sources, prerequisites of multilevel modelling are that the underlying data are characterised by a hierarchical (multilevel) or multi-membership structure and that regarding the higher level units as a sample drawn from a wider population of higher level units (as opposed to regarding them as totally unrelated) does not appear to be entirely unjustified [83:2·99:3]. Moreover, very small datasets will not allow to successfully use multilevel methods [91].

Beyond these most general criteria, the use of multilevel modelling is questionable in situations with very few higher level units [163:95]. In some cases, it will indeed be reasonable to view these units as distinct, independent entities (e.g., ethnic or religious groups) [9]. In other cases, the fundamental assumption of the higher level units being sampled from an underlying random distribution will be appropriate, but data on very few such units will not usually provide enough information to characterise this distribution. Thus, the power to estimate higher level variability and higher level effects becomes very low and it may be preferable to include the higher level units as dummy variables in a single level model [54]. It would be inappropriate to conclude, though, that no higher level variation is present in these cases.

If the sole interest is in effects occurring at the lowest level of a given hierarchy, alternative methods such as GEE-based estimation or sample survey techniques can be used to ensure unbiased standard errors [9·54·105:5-6]. In all other situations where the above-listed prerequisites are met, ignoring the hierarchical structure of the data may have a negative impact and multilevel modelling should be used at least tentatively, to assess if a substantial amount of higher level variation is present in the data. If this is confirmed, multilevel modelling is the analytic method of choice. Otherwise, it is unnecessary and simpler methods can again be used, or justified in hindsight, as was the case in the gatekeeping study [125].

Advantages and benefits of multilevel modelling

Multilevel modelling is currently regarded as the most advanced technique to deal with multilevel and multi-membership data structures [91]. The technique can be used with all frequent types of responses, including survival data [84:178-81·221], although research is ongoing and some software implementations are still in a developmental stage (see Appendix III). Multilevel modelling is also combinable with two-part or multi-part modelling, as e.g. used to deal with health care cost data from general populations and other skewed data containing a substantial proportion of zero values.

There is a broad consensus that the negative consequences of ignoring hierarchical data structures (loss of information and statistical efficiency, potentially incomplete or incorrect identification of effects, biased standard errors; see chapter 1) can be avoided by sensibly using multilevel modelling. Of the theoretically expected benefits, the analyses performed here confirmed the possibility to assess the presence of higher level variation, quantify its amount, and identify its sources; they confirmed a gain in statistical precision; they allowed to identify spurious findings by analysing influential higher level units [83:1/11]; and they achieved a modest reduction in prediction error.

The use of multilevel modelling may also yield substantially different or extended sets of significant predictor variables and, consequently, model parameters [9]. Erroneous conclusions may thus be avoided [1]. However, this was not the case in the author's studies, and neither in a substantial number of other studies where despite additional insights, the main conclusions of conventional analysis were essentially confirmed (see [32] for an example).

Multilevel modelling allows to simultaneously test hypotheses at different levels of the hierarchy, and hypotheses involving several levels. This is of major importance where contextual and compositional effects are to be separated [58] and, more generally spoken, where substantive explanations involving several levels are to be tested. This would, e.g., be the case if based on prior knowledge, a risk model for a clinical event of interest had been developed, and if provider-level factors (e.g. clinical experience with a certain treatment) and specific relationships between patient-level and provider-level factors (e.g. different approaches to elderly patients) were part of such a model.

The possibility to identify outlying observations at different levels of the hierarchy can contribute to improving overall model validity and performance, but in some situations, the main interest may be in the outlying higher level units themselves. Examples are assessments of the performance of health care providers aiming at direct intervention where quality of care is compromised, or studies of the success factors of newly introduced provider-level or community-level programmes. Distinguishing true outliers from apparent outliers (due to uneven distribution of lowest level characteristics, i.e. patient characteristics etc.) is of paramount importance in such situations.

Multilevel modelling can be combined with the GEE approach to estimate unbiased standard errors in situations where the highest level units of the multilevel model are nested in even higher level clusters [161:26]. (Depending on software used; analysis of the neutropenia dataset made use of this option.)

According to Greenland, multilevel modelling unifies the frequentist and Bayesian approaches to statistics [91]. Least squares-based algorithms to estimate multilevel models should yield the same results as Bayesian estimation methods using non-informative priors. Depending on the software used, use can be made of the "natural" Bayesian way to handle missing values (i.e., these can be estimated together with the other model parameters).

Disadvantages and problems of multilevel modelling

The most frequently mentioned disadvantage of multilevel modelling is its complexity [54·91·168]. This refers to the modelling process itself (where the approach described by Hox may provide some guidance [105:49-54]), but also to the difficulty of correctly interpreting effects occurring at different hierarchical levels or cross-level [58·83:1/11].

Constructing aggregate variables from lower level predictors may be important to make good use of the higher level information contained in a given dataset, but adds to this complexity. The approach taken should certainly be as theory-driven as possible, to ensure interpretability. According to the author's experience, demanding issues can arise here, e.g. if one allows for influences of the age distributions of patient populations nested within different providers, on the clinical practice patterns adopted by these providers.

One of the additional problems of multilevel modelling described in the literature is the issue of sample size and power estimation [54·58]. No simple tools exist to perform such estimates for multilevel models, however some approximate formulae for standard error estimates are available and simulation-based approaches can be used [105:173·193:161-74]. From experience, sample size requirements are quite large at all levels. Paterson and Goldstein, for the two level situation, have suggested a minimum of 25 level 2 units with 25 level 1 units each [154]. Several other rules of thump have been proposed, depending on whether the main interest lies mainly in fixed parameters, in the random part, or in cross-level interactions [105:174-5]. Hox and Snijders give overviews of this topic [105:173-96·193:161-74].

Applying resampling techniques to multilevel situations requires the resampling process to take into account the hierarchical structure of the data [186:260]. The author found it relatively straightforward to modify standard cross-validation and bootstrap techniques to assess the prediction error of a given multilevel model with a defined set of model parameters, but more complex problems arise if cross-validation or resampling techniques are to be used for parameter selection, or if the model estimation process itself should be resampling based [34·186:260]. (A solution for the latter problem using the resampling of residuals approach to bootstrapping is now implemented in one of more important software packages for multilevel modelling [34] - see Appendix III.)

In this work, substantial reductions in apparent prediction error achieved by multilevel modelling were only partially confirmed under cross-validation conditions. When predictions were made for out-of-sample observations whose corresponding higher level units contributed to model estimation, only modest improvements were retained. It should be noted in this context that the number of level 1 observations per level 2 unit was small and below multilevel modelling-specific sample size recommendations in the asthma dataset and in parts of the neutropenia dataset. This may have affected the precision of the level 2 unit-specific random effects estimates. More substantial gains in predictive ability might be achieved in situations with more lower level observations per higher level unit available. When predictions were made for observations whose corresponding higher level units did not contribute to model estimation, the gain in predictive ability was almost entirely lost. This is consistent with expectations, because no meaningful higher level residuals can be estimated for higher level units which did not contribute to model estimation [162]. Only "indirect" gains may occur in this situation, if the multilevel modelling process leads to substantial improvements of the fixed effects parts of the estimated models also. The importance of these limitations and the question if there is room for related methodological improvement may require further empirical study as well as more theoretical, mathematically oriented work beyond the ability of the author.

Design aspects of multilevel studies

This thesis is an example for the *post hoc* use of multilevel modelling, with datasets for which no such analysis was originally planned. However, more satisfactory results can certainly be expected from studies which are planned with the requirements of multilevel analysis in mind.

This requires, during the design and protocol development phase, to describe and define the multilevel structure of the data to be generated. The next step would then be to systematically identify, at all levels, which covariates might have an impact on the outcomes of interest and should therefore be collected. In current research, collecting higher level characteristics such as characteristics of health care providers is often neglected. Efforts should be made to define all covariates such that a level of the hierarchy can be unequivocally assigned to them (see pp. 105-6 for an example). The possibility of cross-level effects or interactions should also be taken into account when defining data collection requirements. For example, a patient characteristic (say, retirement) may have no direct relationship with an outcome of interest (say, direct medical cost induced by a given disease), but physicians (constituting the higher level units) may treat retired patients differently. Even more, the impact of retirement on treatment behaviour may vary across physicians. This cross-level effect could not be assessed if information on retirement was not collected.

Where randomisation or random sampling processes are applied to balance the distribution of covariates, the techniques used should take the multilevel structure of the data into account. In a two-level situation, e.g., randomised sampling would typically be a two step process. Level 2 units would be sampled first, and then the level 1 units would be sampled from within these [193:159-60].

It should be taken into account that sample size requirements may be higher than in conventional analysis. Where simply increasing sample size to a level considered as safe is not an option, sample size and power estimates are a complex issue. As discussed above, no fully satisfactory solutions are currently available.

Current and desirable use of multilevel modelling in health care-related research

A Medline search on May 30, 2006 retrieved 1'202 entries indicating the use of multilevel modelling (and an exponential rise over time; see chapter 1). About 80% of these entries appeared to be true hits.

Reviewing the topics addressed reveals that multilevel modelling was most frequently used where social or geographical units are of strong interest as explanatory factors for the distribution of risk factors or health outcomes (see [53·108·139] for examples). Here, the advantages of multilevel modelling have long been recognised [52·54·58·167]. There is also a multitude of applications to other situations, but no clear pattern is visible. In most of health care-related research, personal interest of the researchers may still be the most important determinant of use of the technique.

Review articles have reported and recommended the use of multilevel modelling

- in social epidemiology [35];
- in public health research and research on health behaviours (where group-level factors may have an influence) [58·167·214];
- in health services research [54];
- in health economics, e.g. in studies addressing inequalities in resource allocation or the provision and utilisation of health services [58·168];
- in studies addressing variations in medical practice patterns, or, closely related, the performance of health care providers [58·167];
- in studies addressing unpaid caregiving by family members or partners [133];
- for longitudinal or repeated measurements data (where several observations spread over time are nested within persons) [54·134·151];
- for multivariate responses (multiple outcomes nested within persons) [54];
- in meta-analysis [105:8·165].
- in variable selection problems. Greenland has pointed out that the advantages of multilevel modelling are *"especially great in studies that search for effects or interactions among many exposures (so-called 'fishing expeditions'), in which standard methods of forcing in all variables or using mechanical variable-selection algorithms easily produce invalid inferences."* [91].

While all these suggestions appear reasonable, the author would like to highlight the importance of using multilevel modelling where characteristics of health care providers, and clinical practice patterns in particular, may impact on health outcomes and health economic outcomes [168]. Such factors were shown to play a role in two out of the three analyses reported here, but they are often neglected in current research practice. It is only another facet of the same argument that multilevel modelling should also be used in multi-centre studies to take into account centre characteristics and behaviours.

As has been highlighted by Rice and Jones, this is not only true for observational studies, but also for multi-centre randomised clinical trials [168]. In such trials, a high degree of regulation and detailed research protocols are hoped to minimise centre-specific effects, particularly in the pharmaceutical setting, but the success of these efforts cannot be taken as granted. *"[…] even with randomization, it can be expected that the site of treatment may have an impact on the outcome regardless of treatment the patient receives. This may result from various sources and is likely to be due to differences in medical practice as administered by individual clinicians or clinical management and resources dictated by provider units. However, it may also be due to differences in subpopulations from which each site recruits and hence although patients may be randomised to treatments at each site, they may not be representative of the general population. Pooling data over sites without regard of such site-specific differences may lead to incorrect inference (for example, inefficient parameter estimates). The inclusion of site as a level in a multilevel analysis will ensure that the clustering effects within sites will be adequately controlled for."* [168] Consequently, multilevel modelling should routinely be used in pharmaceutical development as well as other health care-related research to assess or rule out an impact of higher level variation on the results of national as well as international multi-centre trials.

In conclusion, multilevel regression modelling is an important statistical technique, the only one to date which allows to comprehensively deal with nested (multilevel or multi-membership) data structures. Where such data structures are present, study design should explicitly take them into account and multilevel modelling should be used routinely, irrespective of the observational or experimental character of the research conducted. Advantages in terms of predictive ability may be limited and in many cases, use of the technique will remain tentative and prove unnecessary in hindsight. However, these caveats should not be used as arguments against applying multilevel modelling in the first instance, to cover all possibilities.

References

1 Aitkin M, Anderson D, Hinde J (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society A* **144**:148-61

2 Alker HS (1969). A typology of ecological fallacies. In: Dogan M, Rokkan S (eds) Quantitative Ecological Analysis. MIT Press, Cambridge, MA

3 Allen-Ramey FC, Samet JM, Rand CS, Joseph CL (2004). Trends in use of inhaled corticosteroids for asthma management: 1994-1998. *Ann Epidemiol* **14**:161-7

4 Altman DG (1991). Practical statistics for medical research. Chapman & Hall/CRC, London

5 Amre DK, Infante-Rivard C, Gautrin D, Malo JL (2002). Socioeconomic status and utilization of health care services among asthmatic children. *J Asthma* **39**:625-31

6 Anonymous (1997). Arzneimittelkompendium der Schweiz (1997). Documed, Basel

7 Anonymous (1998). Age- and state-specific prevalence estimates of insured and uninsured persons – United States, 1995-1996. *MMWR Morb Mortal Wkly Rep* **47**:529-32

8 Anonymous (2001). Managed-Care-Modelle in der Schweiz. *Managed Care* **5**:37-9

9 Austin PC, Goel V, van Walraven C (2001). An introduction to multilevel regression models. *Can J Public Health* **92**:150-4

10 Balducci L (2003). Myelosuppression and its consequences in elderly patients with cancer. *Oncology (Huntingt)* **17**:27-32

11 Barnes PJ, Jonsson B, Klim JB (1996). The costs of asthma. *Eur Respir J* **9**:636-42

12 Barza M, Ioannidis JP, Cappelleri JC, Lau J (1996). Single or multiple daily doses of aminoglycosides: a meta-analysis. *BMJ* **312**:338-45

13 Baumberger J (2001). So funktioniert Managed Care. Anspruch und Wirklichkeit der integrierten Gesundheitsversorgung in Europa. Thieme, Stuttgart

14 Belsley DA, Kuh E, Welsch RE (1980). Regression Diagnostics. Wiley, New York

15 Bendel RB, Afifi AA (1977). Comparison of stopping rules in forward regression. *Journal of the American Statistical Association* **72**:46-53

16 Berkhof J, Snijders TAB (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics* **26**:133-52

17 Bhat VN (2005). Institutional arrangements and efficiency of health care delivery systems. *Eur J Health Econ* **6**:215-22

18 Birnbaum HG, Berger WE, Greenberg PE et al (2002). Direct and indirect costs of asthma to an employer. *J Allergy Clin Immunol* **109**:264-70

19 Blay JY, Chauvin F, Le Cesne A et al (1996). Early lymphopenia after cytotoxic chemotherapy as a risk factor for febrile neutropenia. *J Clin Oncol* **14**:636-43

20 Blough DK, Madden CW, Hornbrook MC (1999). Modeling risk using generalized linear models. *J Health Econ* **18**:153-71

21 Bohlert I, Adam I, Robra BP (1997). [The Swiss gatekeeper system – a model for improving capacity development and economic effectiveness]. *Gesundheitswesen* **59**:488-94

22 Bolton MB, Tilley BC, Kuder J, Reeves T, Schultz LR (1991). The cost and effectiveness of an education program for adults who have asthma. *J Gen Intern Med* **6**:401-7

23 Bonadonna G, Valagussa P, Moliterni A, Zambetti M, Brambilla C (1995). Adjuvant cyclophosphamide, methotrexate, and fluorouracil in node-positive breast cancer: the results of 20 years of follow-up. *N Engl J Med* **332**:901-6

24 Bonham GS, Barber GM (1987). Use of health care before and during Citicare. *Med Care* **25**:111-9

25 Bootman JL, Crown WH, Luskin AT (2004). Clinical and economic effects of suboptimally controlled asthma. *Manag Care Interface* **17**:31-6

26 Borg C, Ray-Coquard I, Philip I et al (2004). CD4 lymphopenia as a risk factor for febrile neutropenia and early death after cytotoxic chemotherapy in adult patients with cancer. *Cancer* **101**:2675-80

27 Boston Consulting Group (1993). The costs of adult asthma in Canada. In: Communications Media for Education. Princeton

28 Bousquet J, Bousquet PJ, Godard P, Daures JP (2005). The public health implications of asthma. *Bull World Health Organ* **83**:548-54

29 Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). Classification and regression trees. Wadsworth and Brooks, Belmont

30 Budman DR, Berry DA, Cirrincione CT et al (1998). Dose and dose intensity as determinants of outcome in the adjuvant treatment of breast cancer. The Cancer and Leukemia Group B. *J Natl Cancer Inst* **90**:1205-11

31 Bundesamt für Sozialversicherung (1998). Neue Formen der Krankenversicherung: Alters- und Kostenverteilungen im Vergleich zu der traditionellen Versicherung. Ergebnisse der Administrativdatenuntersuchung, 2. Teil. Bundesamt für Sozialversicherung, Bern

32 Carey K (2000). A multilevel modelling approach to analysis of patient costs under managed care. *Health Econ* **9**:435-46

33 Carlin BP, Louis TA (2000). Bayes and empirical Bayes methods for data analysis. Chapman & Hall/CRC, Boca Raton

34 Carpenter JR, Goldstein H, Rasbash J (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics* **52**:431-43

35 Chaix B, Chauvin P (2002). [The contribution of multilevel models in contextual analysis in the field of social epidemiology: a review of literature]. *Rev Epidemiol Sante Publique* **50**:489-99

36 Chang J (2000). Chemotherapy dose reduction and delay in clinical practice. evaluating the risk to patient outcome in adjuvant chemotherapy for breast cancer. *Eur J Cancer* **36 Suppl 1**:S11-4

37 Chen GJ, Feldman SR (2000). Economic aspect of health care systems. Advantage and disadvantage incentives in different systems. *Dermatol Clin* **18**:211-4

38 Choi CW, Sung HJ, Park KH et al (2003). Early lymphopenia as a risk factor for chemotherapy-induced febrile neutropenia. *Am J Hematol* **73**:263-6

39 Cisternas MG, Blanc PD, Yen IH et al (2003). A comprehensive study of the direct and indirect costs of adult asthma. *J Allergy Clin Immunol* **111**:1212-8

40 Citron ML, Berry DA, Cirrincione C et al (2003). Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: first report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. *J Clin Oncol* **21**:1431-9

41 Cockcroft DW, Swystun VA (1996). Asthma control versus asthma severity. *J Allergy Clin Immunol* **98**:1016-8

42 Constenla M, Bosly A, Jackisch C, et al. (2003). An audit of primary breast cancer management in Spain: the OSQAR study [abstract]. *Proc Am Soc Clin Oncol* **22**:312

43 Cox DR (1990). Role of models in statistical analysis. *Statistical Science* **5**:169-74

44 Crawford J, Ozer H, Stoller R et al (1991). Reduction by granulocyte colony-stimulating factor of fever and neutropenia induced by chemotherapy in patients with small-cell lung cancer. *N Engl J Med* **325**:164-70

45 Dale DC, Crawford J, Lyman CG (2001). Chemotherapy-induced neutropenia and associated complications in randomized clinical trials: an evidence-based review [abstract]. *Proc Am Soc Clin Oncol* **20**:1638

46 Dale DC (2002). Colony-stimulating factors for the management of neutropenia in cancer patients. *Drugs* **62 Suppl 1**:1-15

47 Dale DC, McCarter GC, Crawford J, et al. (2003). Chemotherapy induced neutropenia and associated complications in randomized clinical trials: an evidence-based review. *Journal of the National Comprehensive Cancer Network* **1**:440-54

48 Dang CT, Fornier MN, Hudis CA (2003). Risk models for neutropenia in patients with breast cancer. *Oncology (Huntingt)* **17**:14-20

49 Delnoij D, Van Merode G, Paulus A, Groenewegen P (2000). Does general practitioner gatekeeping curb health care expenditure? *J Health Serv Res Policy* **5**:22-6

50 Denham JW, Hamilton CS, Christie D et al (1995). Simultaneous adjuvant radiation therapy and chemotherapy in high-risk breast cancer – toxicity and dose modification: a Transtasman Radiation Oncology Group Multi-Institution study. *Int J Radiat Oncol Biol Phys* **31**:305-13

51 Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY (1999). Methods for analyzing health care utilization and costs. *Annu Rev Public Health* **20**:125-44

52 Diez Roux AV (2002). A glossary for multilevel analysis. *J Epidemiol Community Health* **56**:588-94

53 Diez-Roux AV, Nieto FJ, Muntaner C et al (1997). Neighborhood environments and coronary heart disease: a multilevel analysis. *Am J Epidemiol* **146**:48-63

54 Diez-Roux AV (2000). Multilevel analysis in public health research. *Annu Rev Public Health* **21**:171-92

55 Dolan CM, Fraher KE, Bleecker ER et al (2004). Design and baseline characteristics of the epidemiology and natural history of asthma: Outcomes and Treatment Regimens (TENOR) study: a large cohort of patients with severe or difficult-to-treat asthma. *Ann Allergy Asthma Immunol* **92**:32-9

56 Duan N (1983). Smearing estimate: a non-parametric re-transformation method. *Journal of the American Statistical Association* **78**:605-10

57 Duan N, Manning WGJ, Morris CN, Newhouse JP (1983). A Comparison of Alternative Models for the Deman for Medical Care. *Journal of Business & Economic Statistics* **1**:115-26

58 Duncan C, Jones K, Moon G (1998). Context, composition and heterogeneity: using multilevel models in health research. *Soc Sci Med* **46**:97-117

59 Early Breast Cancer Trialists' Collaborative Group (1998). Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* **352**:930-42

60 Efron B (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**:316-33

61 Efron B, Tibshirani R (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* **92**:548-60

62 Ellis GK, Livingston RB, Gralow JR, Green SJ, Thompson T (2002). Dose-dense anthracycline-based chemotherapy for node-positive breast cancer. *J Clin Oncol* **20**:3637-43

63 Engels EA, Lau J, Barza M (1998). Efficacy of quinolone prophylaxis in neutropenic cancer patients: a meta-analysis. *J Clin Oncol* **16**:1179-87

64 Engelsman E, Klijn JC, Rubens RD et al (1991). "Classical" CMF versus a 3-weekly intravenous CMF schedule in postmenopausal patients with advanced breast cancer. An EORTC Breast Cancer Co-operative Group Phase III Trial (10808). *Eur J Cancer* **27**:966-70

65 Escarce JJ, Kapur K, Joyce GF, Van Vorst KA (2001). Medical care expenditures under gatekeeper and point-of-service arrangements. *Health Serv Res* **36**:1037-57

66 ESMO Guidelines Task Force (2001). ESMO recommendations for the application of haematopoietic growth factors (hGFs). *Ann Oncol* **12**:1219-20

67 Etter JF, Perneger TV (1998). Health care expenditures after introduction of a gatekeeper and a global budget in a Swiss health insurance plan. *J Epidemiol Community Health* **52**:370-6

68 Ettner SL (1999). The relationship between continuity of care and the health behaviors of patients: does having a usual physician make a difference? *Med Care* **37**:547-55

69 Ferris TG, Chang Y, Blumenthal D, Pearson SD (2001). Leaving gatekeeping behind – effects of opening access to specialists for adults in a health maintenance organization. *N Engl J Med* **345**:1312-7

70 Forrest CB, Glade GB, Baker AE et al (1999). The pediatric primary-specialty care interface: how pediatricians refer children and adolescents to specialty care. *Arch Pediatr Adolesc Med* **153**:705-14

71 Forrest CB, Nutting P, Werner JJ et al (2003). Managed health plan effects on the specialty referral process: results from the Ambulatory Sentinel Practice Network referral study. *Med Care* **41**:242-53

72 Franco SM, Mitchell CK, Buzon RM (1997). Primary care physician access and gatekeeping: a key to reducing emergency department use. *Clin Pediatr (Phila)* **36**:63-8

73 Franks P, Clancy CM, Nutting PA (1992). Gatekeeping revisited – protecting patients from overtreatment. *N Engl J Med* **327**:424-9

74 Frey W (1996). HMO- und Hausarztmodelle in der Schweiz. *KSK aktuell. Konkordat der Schweizerischen Krankenversicherer* **4**:54-5

75 Fuhlbrigge A, Carey VJ, Adams RJ et al (2004). Evaluation of asthma prescription measures and health system performance based on emergency department utilization. *Med Care* **42**:465-71

76 Furno P, Bucaneve G, Del Favero A (2002). Monotherapy or aminoglycoside-containing combinations for empirical antibiotic treatment of febrile neutropenic patients: a meta-analysis. *Lancet Infect Dis* **2**:231-42

77 Galt KA, Rich EC, Kralewski JE et al (2001). Group practice strategies to manage pharmaceutical cost in an HMO network. *Am J Manag Care* **7**:1081-90

78 Gendo K, Sullivan SD, Lozano P et al (2003). Resource costs for asthma-related care among pediatric patients in managed care. *Ann Allergy Asthma Immunol* **91**:251-7

79 Gendo K, Lodewick MJ (2005). Asthma economics: focusing on therapies that improve costly outcomes. *Curr Opin Pulm Med* **11**:43-50

80 Global Initiative for Asthma (GINA) (2002). Global Strategy for Asthma Management and Prevention. (Updated from: NHLBI/WHO Workshop Report: Global Strategy for Asthma Management and Prevention issued January, 1995). NIH Publication No. 02-3659. National Institutes of Health. National Heart, Lung, and Blood Institute, Bethesda

81 Godard P, Chanez P, Siraudin L, Nicoloyannis N, Duru G (2002). Costs of asthma are correlated with severity: a 1-yr prospective study. *Eur Respir J* **19**:61-7

82 Goldstein H, Spiegelhalter DJ (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society A* **159**:385-409

83 Goldstein H (1999). Multilevel Statistical Models. 1st Internet Edition. Edward Arnold, London

84 Goldstein H, Leyland A (2001). Further Topics in Multilevel Modelling. In: Leyland AH, Goldstein H (eds) Multilevel Modelling of Health Statistics. Wiley, Chichester, pp 175-86

85 Goldstein H, Browne W, Rasbash J (2002). Multilevel modelling of medical data. *Stat Med* **21**:3291-315

86 Graf von der Schulenburg JM, Greiner W, Molitor S, Kielhorn A (1996). Kosten der Asthmatherapie nach Schweregrad. Eine empirische Untersuchung. *Med Klin* **91**:670-6

87 Grant EN, Wagner R, Weiss KB (1999). Observations on emerging patterns of asthma in our society. *J Allergy Clin Immunol* **104**:S1-9

88 Green MD, Koelbl H, Baselga J et al (2003). A randomized double-blind multicenter phase III study of fixed-dose single-administration pegfilgrastim versus daily filgrastim in patients receiving myelosuppressive chemotherapy. *Ann Oncol* **14**:29-35

89 Greenfield S, Apolone G, McNeil BJ, Cleary PD (1993). The importance of co-existent disease in the occurrence of postoperative complications and one-

year recovery in patients undergoing total hip replacement. Comorbidity and outcomes after hip replacement. *Med Care* **31**:141-54

90 Greenland S (1989). Modeling and variable selection in epidemiologic analysis. *Am J Public Health* **79**:340-9

91 Greenland S (2000). Principles of multilevel modelling. *Int J Epidemiol* **29**:158-67

92 Grembowski DE, Martin D, Diehr P et al (2003). Managed care, access to specialists, and outcomes among primary care patients with pain. *Health Serv Res* **38**:1-19

93 Halpern MT, Khan ZM, Stanford RH, Spayde KM, Golubiewski M (2003). Asthma: resource use and costs for inhaled corticosteroid vs leukotriene modifier treatment – a meta-analysis. *J Fam Pract* **52**:382-9

94 Halpern MT, Schmier JK, Richner R, Guo C, Togias A (2004). Allergic rhinitis: a potential cause of increased asthma medication use, costs, and morbidity. *J Asthma* **41**:117-26

95 Hannan EL, Wu C, DeLong ER, Raudenbush SW (2005). Predicting risk-adjusted mortality for CABG surgery: logistic versus hierarchical logistic models. *Med Care* **43**:726-35

96 Hardin J, Hilbe J (2001). Generalized Linear Models and Extensions. Stata Press, College Station

97 Harrell FE, Jr., Lee KL, Mark DB (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15**:361-87

98 Hartert TV, Peebles RS, Jr. (2000). Epidemiology of asthma: the year in review. *Curr Opin Pulm Med* **6**:4-9

99 Healy MJR (2001). Multilevel Data and Their Analysis. In: Leyland AH, Goldstein H (eds) Multilevel Modelling of Health Statistics. Wiley, Chichester, pp 1-12

100 Hennekens CH, Buring JE (1987). Epidemiology in Medicine. Little, Brown and Company, Boston

101 Herjavecz I, Nagy GB, Gyurkovits K et al (2003). Cost, morbidity, and control of asthma in Hungary: The Hunair Study. *J Asthma* **40**:673-81

102 Holmes FA, O'Shaughnessy JA, Vukelja S et al (2002). Blinded, randomized, multicenter study to evaluate single administration pegfilgrastim once per cycle versus daily filgrastim as an adjunct to chemotherapy in patients with high-risk stage II or stage III/IV breast cancer. *J Clin Oncol* **20**:727-31

103 Hoskins G, McCowan C, Neville RG et al (2000). Risk factors and costs associated with an asthma attack. *Thorax* **55**:19-24

104 Hosmer DW, Lemeshow S (2000). Applied logistic regression. Wiley, New York

105 Hox J (2002). Multilevel Analysis. Techniques and Applications. Lawrence Erlbaum Associates, Mahwah

106 Hryniuk W, Frei E, 3rd, Wright FA (1998). A single scale for comparing dose-intensity of all chemotherapy regimens in breast cancer: summation dose-intensity. *J Clin Oncol* **16**:3137-47

107 Huber-Stemich F, Hees K, Baumann P, Berger D (1996). Sechs Jahre HMO Zürich-Wiedikon. Ein Erfahrungsbericht. *Ars Medici* **18**:1079-82

108 Humphreys K, Carr-Hill R (1991). Area variations in health outcomes: artefact or ecology. *Int J Epidemiol* **20**:251-8

109 Hurley RE, Paul JE, Freund DA (1989). Going into gatekeeping: an empirical assessment. *QRB. Quality Review Bulletin* **15**:306-14

110 Hurley RE, Freund DA, Gage BJ (1991). Gatekeeper effects on patterns of physician use. *J Fam Pract* **32**:167-74

111 Jackisch C, Jaber M, Burkamp U et al (2003). Maintenance of dose intensity in adjuvant chemotherapy of breast cancer in patients treated outside a clinical trial. Results of a retrospective study. *Geburtsh u Frauenheilk* **63**:333-43

112 Joyce GF, Kapur K, Van Vorst KA, Escarce JJ (2000). Visits to primary care physicians and to specialists under gatekeeper and point-of-service arrangements. *Am J Manag Care* **6**:1189-96

113 Kapur K, Joyce GF, Van Vorst KA, Escarce JJ (2000). Expenditures for physician services under alternative models of managed care. *Med Care Res Rev* **57**:161-81

114 Kerger JN, Bormans V, Dauwe M (2002). Adjuvant (adj) chemotherapy (CT) delivery in patients (pts) with breast cancer (BC): Results from the Chemodose Working Party Belgium-Luxembourg [abstract]. *Ann Oncol* **13 Suppl 5**:38

115 Kiivet RA, Kaur I, Lang A, Aaviksoo A, Nirk L (2001). Costs of asthma treatment in Estonia. *Eur J Public Health* **11**:89-92

116 Kleinbaum DG, Kupper LL, Muller KE, Nizam A (1998). Applied Regression Analysis and Other Multivariable Methods. Duxbury Press, Pacific Grove

117 Konkordat der Schweizerischen Krankenversicherer (1999). Tagestaxen in Heilanstalten (1999). Konkordat der Schweizerischen Krankenversicherer, Bern

118 Kuttner R (1998). Must good HMOs go bad? First of two parts: the commercialization of prepaid group health care [see comments]. *N Engl J Med* **338**:1558-63

119 Kwak LW, Halpern J, Olshen RA, Horning SJ (1990). Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: results of a tree-structured survival analysis. *J Clin Oncol* **8**:963-77

120 Laditka SB, Laditka JN (2001). Utilization, costs, and access to primary care in fee-for-service and managed care plans. *J Health Soc Policy* **13**:21-39

121 Langford IH, Lewis T (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, Series A* **161**:121-60

122 Langford IH, Day RJ (2001). Poisson Regression. In: Leyland AH, Goldstein H (eds) Mulitlevel Modelling of Health Statistics. Wiley, Chichester, pp 45-58

123 Leonard RC, Miles D, Thomas R, Nussey F (2003). Impact of neutropenia on delivering planned adjuvant chemotherapy: UK audit of primary breast cancer patients. *Br J Cancer* **89**:2062-8

124 Leuenberger P (1995). Pollution de l'air en Suisse et maladies respiratoires chez l'adulte. Resultats preliminaires de la partie transversale de l'etude Sapaldia. *Schweiz Rundsch Med Prax* **84**:1096-100

125 Leyland AH, Groenewegen PP (2003). Multilevel modelling and public health policy. *Scand J Public Health* **31**:267-74

126 Link BK, Budd GT, Scott S et al (2001). Delivering adjuvant chemotherapy to women with early-stage breast carcinoma: current patterns of care. *Cancer* **92**:1354-67

127 Lipscomb J, Ancukiewicz M, Parmigiani G et al (1998). Predicting the cost of illness: a comparison of alternative models applied to stroke. *Med Decis Making* **18**:S39-56

128 Long SH, Settle RF (1988). An evaluation of Utah's primary care case management program for Medicaid recipients. *Med Care* **26**:1021-32

129 Lyman GH (2003). Risk assessment in oncology clinical practice. From risk factors to risk models. *Oncology (Huntingt)* **17**:8-13

130 Lyman GH, Dale DC, Crawford J (2003). Incidence and predictors of low dose-intensity in adjuvant breast cancer chemotherapy: a nationwide study of community practices. *J Clin Oncol* **21**:4524-31

131 Lyman GH, Morrison VA, Dale DC et al (2003). Risk of febrile neutropenia among patients with intermediate-grade non-Hodgkin's lymphoma receiving CHOP chemotherapy. *Leuk Lymphoma* **44**:2069-76

132 Lyman GH, Lyman CH, Agboola O (2005). Risk models for predicting chemotherapy-induced neutropenia. *Oncologist* **10**:427-37

133 Lyons KS, Sayer AG (2005). Using multilevel modeling in caregiving research. *Aging Ment Health* **9**:189-95

134 Machin D (2004). On the evolution of statistical methods as applied to clinical trials. *J Intern Med* **255**:521-8

135 Malone DC, Lawson KA, Smith DH (2000). Asthma: an analysis of high cost patients. *Pharm Pract Manage* Q **20**:12-20

136 Manning WG, Leibowitz A, Goldberg GA, Rogers WH, Newhouse JP (1984). A controlled trial of the effect of a prepaid group practice on use of services. *N Engl J Med* **310**:1505-10

137 Martin DP, Diehr P, Price KF, Richardson WC (1989). Effect of a gatekeeper plan on health services use and charges: a randomized trial. *Am J Public Health* **79**:1628-32

138 Masoli M, Fabian D, Holt S, Beasley R (2004). The global burden of asthma: executive summary of the GINA Dissemination Committee report. *Allergy* **59**:469-78

139 Matteson DW, Burr JA, Marshall JR (1998). Infant mortality: a multi-level analysis of individual and community risk factors. *Soc Sci Med* **47**:1841-54

140 Maziak W, von Mutius E, Keil U et al (2004). Predictors of health care utilization of children with asthma in the community. *Pediatr Allergy Immunol* **15**:166-71

141 McCullagh P, Nelder JA (1989). Generalized Linear Models. Chapman & Hall/CRC, Boca Raton

142 Mickey RM, Greenland S (1989). The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* **129**:125-37

143 Mittlbock M, Schemper M (1999). Computing measures of explained variation for logistic regression models. *Comput Methods Programs Biomed* **58**:17-24

144 Morgan RO, Virnig BA, DeVito CA, Persily NA (1997). The Medicare-HMO revolving door – the healthy go in and the sick go out. *N Engl J Med* **337**:169-75

145 Mullahy J (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ* **17**:247-81

146 National Asthma Campaign (1992). Report on the cost of asthma in Australia. National Asthma Campaign, Melbourne

147 National Comprehensive Cancer Network (NCCN) (2005). Clinical Practice Guidelines in Oncology – v.2.2005. Myeloid Growth Factors in Cancer Treatment.

http://www.nccn.org/professionals/physician_gls/PDF/myeloid_growth.pdf. Accessed May 30, 2005

148 National Heart, Lung, and Blood Institute (2002). Morbidity and Mortality: 2002 Chartbook on Cardiovascular, Lung, and Blood Diseases. US Department of Health and Human Services, Bethesda

149 Neri M, Spanevello A (2000). Chronic bronchial asthma from challenge to treatment: epidemiology and social impact. *Thorax* **55**:S57-8

150 Neuhaus JM, Hauck WW, Kalbfleisch JD (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**:755-62

151 O'Connell AA, McCoach DB (2004). Applications of hierarchical linear models for evaluations of health interventions: demystifying the methods and interpretations of multilevel models. *Eval Health Prof* **27**:119-51

152 OECD (2005). OECD Health Data 2005. Total expenditure on health, % of gross domestic product. http://www.oecd.org/dataoecd/60/28/35529791.xls. Accessed May 20, 2006

153 Ozer H, Armitage JO, Bennett CL et al (2000). 2000 update of recommendations for the use of hematopoietic colony-stimulating factors: evidence-based, clinical practice guidelines. American Society of Clinical Oncology Growth Factors Expert Panel. *J Clin Oncol* **18**:3558-85

154 Paterson L, Goldstein H (1992). New statistical methods for analyzing social structures: an introduction to multilevel models. *British Educational Research Journal* **17**:387-93

155 Pati S, Shea S, Rabinowitz D, Carrasquillo O (2005). Health expenditures for privately insured adults enrolled in managed care gatekeeping vs indemnity plans. *Am J Public Health* **95**:286-91

156 Paul M, Borok S, Fraser A et al (2005). Additional anti-Gram-positive antibiotic treatment for febrile neutropenic cancer patients. *Cochrane Database Syst Rev*:CD003914

157 Pettengell R, Gurney H, Radford JA et al (1992). Granulocyte colony-stimulating factor to prevent dose-limiting neutropenia in non-Hodgkin's lymphoma: a randomized controlled trial. *Blood* **80**:1430-6

158 Piccart MJ, Biganzoli L, Di Leo A (2000). The impact of chemotherapy dose density and dose intensity on breast cancer outcome: what have we learned? *Eur J Cancer* **36 Suppl 1**:S4-10

159 Plaza V, Serra-Batlles J, Ferrer M, Morejon E (2000). Quality of life and economic features in elderly asthmatics. *Respiration* **67**:65-70

160 PROGNOS AG (1998). Evaluation neuer Formen der Krankenversicherung. Synthesebericht. Bundesamt für Sozialversicherung, Bern

161 Rabe-Hesketh S, Skrondal A, Pickles A (2004). GLLAMM Manual. The Berkeley Electronic Press, Berkeley

162 Rabe-Hesketh S (2006). Personal communication

163 Rasbash J, Browne W, Goldstein H, et al. (2002). A User's Guide to MLwiN. Institute of Education, University of London, London

164 Rasbash J, Browne W, Goldstein H, et al. (2004). A User's Guide to MLwiN. Version 2.0. Institute of Education, University of London, London

165 Raudenbush SW, Bryk AS (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics* **10**:75-98

166 Ray-Coquard I, Borg C, Bachelot T et al (2003). Baseline and early lymphopenia predict for the risk of febrile neutropenia after chemotherapy. *Br J Cancer* **88**:181-6

167 Rice N, Leyland A (1996). Multilevel models: applications to health data. *J Health Serv Res Policy* **1**:154-64

168 Rice N, Jones A (1997). Multilevel models and health economics. *Health Econ* **6**:561-75

169 Rice N (2001). Binomial Regression. In: Leyland AH, Goldstein H (eds) Mulitlevel Modelling of Health Statistics. Wiley, Chichester, pp 27-44

170 Ripley BD (1996). Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge

171 Rivera E, Erder MH, Moore TD et al (2003). Targeted filgrastim support in patients with early-stage breast carcinoma: toward the implementation of a risk model. *Cancer* **98**:222-8

172 Rivera E, Haim Erder M, Fridman M, Frye D, Hortobagyi GN (2003). First-cycle absolute neutrophil count can be used to improve chemotherapy-dose delivery and reduce the risk of febrile neutropenia in patients receiving adjuvant therapy: a validation study. *Breast Cancer Res* **5**:R114-20

173 Robinson WS (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review* **15**:351-7

174 Rodwin MA (1995). Conflicts in managed care. *N Engl J Med* **332**:604-7

175 Rogers WH (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin* **13**:19-23

176 Rothman KJ, Greenland S (1998). Causation and causal inference. In: Rothman KJ, Greenland S (eds) Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia, pp 7-28

177 Rutten-van Molken MP, Van Doorslaer EK, Jansen MC, Kerstjens HA, Rutten FF (1995). Costs and effects of inhaled corticosteroids and bronchodilators in asthma and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* **151**:975-82

178 Schillinger D, Bibbins-Domingo K, Vranizan K et al (2000). Effects of primary care coordination on public hospital patients. *J Gen Intern Med* **15**:329-36

179 Schneeweiss S, Sangha O (2001). [Provider profiling: needs, methodologic requirements and means to increase acceptance]. *Dtsch Med Wochenschr* **126**:918-24

180 Schramm B, Ehlken B, Smala A et al (2003). Cost of illness of atopic asthma and seasonal allergic rhinitis in Germany: 1-yr retrospective study. *Eur Respir J* **21**:116-22

181 Sears MR (1997). Descriptive epidemiology of asthma. *Lancet* **350**:SII1-4

182 Serra-Batlles J, Plaza V, Morejon E, Comella A, Brugues J (1998). Costs of asthma according to the degree of severity. *Eur Respir J* **12**:1322-6

183 Shenkman E, Wu SS, Nackashi J, Sherman J (2003). Managed care organizational characteristics and health care use among children with special health care needs. *Health Serv Res* **38**:1599-624

184 Shortell SM, Waters TM, Clarke KW, Budetti PP (1998). Physicians as double agents: maintaining trust in an era of multiple accountabilities. *JAMA* **280**:1102-8

185 Silber JH, Fridman M, DiPaola RS et al (1998). First-cycle blood counts and subsequent neutropenia, dose reduction, or delay in early-stage breast cancer therapy. *J Clin Oncol* **16**:2392-400

186 Skrondal A, Rabe-Hesketh S (2004). Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC, Boca Raton

187 Sly RM (1999). Changing prevalence of allergic rhinitis and asthma. *Ann Allergy Asthma Immunol* **82**:233-48; quiz 48-52

188 Smith DH, Malone DC, Lawson KA et al (1997). A national estimate of the economic costs of asthma. *Am J Respir Crit Care Med* **156**:787-93

189 Smith MJ, Rascati KL, McWilliams BC (2004). Inhaled anti-inflammatory pharmacotherapy and subsequent hospitalizations and emergency department visits among patients with asthma in the Texas Medicaid program. *Ann Allergy Asthma Immunol* **92**:40-6

190 Smith RB (2001). Gatekeepers and sentinels. Their consolidated effects on inpatient medical care. *Eval Rev* **25**:288-330

191 Smith TJ, Khatcheressian J, Lyman CG (2006). 2006 Update of Recommendations for the Use of White Blood Cell Growth Factors: An Evidence-Based Clinical Practice Guideline. *J Clin Oncol* **24**:published ahead of print on May 8, 2006 as 10.1200/JCO.2006.06.4451

192 Snijders TAB, Bosker RJ (1999). Multilevel analysis. Sage, Newbury Park

193 Snijders TAB (2001). Sampling. In: Leyland AH, Goldstein H (eds) Multilevel Modelling of Health Statistics. Wiley, Chichester, pp 159-74

194 Sommers AR, Wholey DR (2003). The effect of HMO competition on gatekeeping, usual source of care, and evaluations of physician thoroughness. *Am J Manag Care* **9**:618-27

195 Soondergaard B, Davidsen F, Kirkeby B, Rasmussen M, Hey H (1992). The economics of an intensive education programme for asthmatic patients: a prospective controlled trial. *Pharmacoeconomics* **1**:207-12

196 Stangl D, Huerta G (2000). Assessing the impact of managed-care on the distribution of length-of-stay using Bayesian hierarchical models. *Lifetime Data Anal* **6**:123-39

197 Stanworth SJ, Massey E, Hyde C et al (2005). Granulocyte transfusions for treating infections in patients with neutropenia or neutrophil dysfunction. *Cochrane Database Syst Rev*:CD005339

198 Stevens CA, Turner D, Kuehni CE, Couriel JM, Silverman M (2003). The economic impact of preschool asthma and wheeze. *Eur Respir J* **21**:1000-6

199 Stock S, Redaelli M, Luengen M et al (2005). Asthma: prevalence and cost of illness. *Eur Respir J* **25**:47-53

200 Sullivan SD (2003). Asthma in the United States: recent trends and current status. *J Manag Care Pharm* **9**:3-7

201 Sullivan SD, Buxton M, Andersson LF et al (2003). Cost-effectiveness analysis of early intervention with budesonide in mild persistent asthma. *J Allergy Clin Immunol* **112**:1229-36

202 Sung L, Nathan PC, Lange B, Beyene J, Buchanan GR (2004). Prophylactic granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor decrease febrile neutropenia after chemotherapy in children with cancer: a meta-analysis of randomized controlled trials. *J Clin Oncol* **22**:3350-6

203 Szucs TD, Anderhub H, Rutishauser M (1999). The economic burden of asthma: direct and indirect costs in Switzerland. *Eur Respir J* **13**:281-6

204 Szucs TD, Anderhub HP, Rutishauser M (2000). Determinants of health care costs and patterns of care of asthmatic patients in Switzerland. *Schweiz Med Wochenschr* **130**:305-13

205 Takkouche B, Cadarso-Suarez C, Spiegelman D (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *Am J Epidemiol* **150**:206-15

206 Talcott JA, Siegel RD, Finberg R, Goldman L (1992). Risk assessment in cancer patients with fever and neutropenia: a prospective, two-center validation of a prediction rule. *J Clin Oncol* **10**:316-22

207 Thatcher N, Girling DJ, Hopwood P et al (2000). Improving survival without reducing quality of life in small-cell lung cancer patients by increasing the dose-intensity of chemotherapy with granulocyte colony-stimulating factor support: results of a British Medical Research Council Multicenter Randomized Trial. Medical Research Council Lung Cancer Working Party. *J Clin Oncol* **18**:395-404

208 Tolpin HG, Bentkover JD (1983). Economic cost of illness: decision-making applications and practical considerations. *Adv Health Econ Health Serv Res* **4**:165-98

209 Ungar WJ, Coyte PC (2001). Prospective study of the patient-level cost of asthma care in children. *Pediatr Pulmonol* **32**:101-8

210 Valdivia G (2000). Asma bronquial y enfermedades atopicas como problema emergente de Salud Publica: nuevas hipotesis etiologicas. La experiencia de sociedades desarrolladas. *Rev Med Chil* **128**:339-46

211 Van Ganse E, Antonicelli L, Zhang Q et al (2006). Asthma-related resource use and cost by GINA classification of severity in three European countries. *Respir Med* **100**:140-7

212 Vertrees JC, Manton KG, Mitchell KC (1989). Case-mix adjusted analyses of service utilization for a Medicaid health insuring organization in Philadelphia. *Med Care* **27**:397-411

213 Vogel CL, Wojtukiewicz MZ, Carroll RR et al (2005). First and subsequent cycle use of pegfilgrastim prevents febrile neutropenia in patients with breast cancer: a multicenter, double-blind, placebo-controlled phase III study. *J Clin Oncol* **23**:1178-84

214 Von Korff M, Koepsell T, Curry S, Diehr P (1992). Multi-level analysis in epidemiologic research on health behaviors and outcomes. *Am J Epidemiol* **135**:1077-82

215 Weinmann S, Kamtsiuris P, Henke KD et al (2003). The costs of atopy and asthma in children: assessment of direct costs and their determinants in a birth cohort. *Pediatr Allergy Immunol* **14**:18-26

216 Weiss KB, Sullivan SD (2001). The health economics of asthma and rhinitis. I. Assessing the economic impact. *J Allergy Clin Immunol* **107**:3-8

217 Weissflog D, Matthys H, Virchow JC, Jr. (2001). [Epidemiology and costs of bronchial asthma and chronic bronchitis in Germany]. *Dtsch Med Wochenschr* **126**:803-8

218 Welte K, Gabrilove J, Bronchud MH, Platzer E, Morstyn G (1996). Filgrastim (r-metHuG-CSF): the first 10 years. *Blood* **88**:1907-29

219 Wickizer TM, Lessler D (2002). UTILIZATION MANAGEMENT: Issues, Effects, and Future Prospects. *Ann Rev Public Health* **23**:233-54

220 Woodhouse G, Goldstein H (1989). Educational performance indicators and LEA league tables. *Oxford Review of Education* **25**:469-83

221 Yang M, Goldstein H (2003). Modelling Survival Data in MLwiN 1.20. Institute of Education, University of London, London

222 Yelin E, Trupin L, Cisternas M et al (2002). A national study of medical care expenditures for respiratory conditions. *Eur Respir J* **19**:414-21

223 Zheng B, Agresti A (2000). Summarizing the predictive power of a generalized linear model. *Stat Med* **19**:1771-81

Appendix I

Multilevel re-analysis details

Asthma dataset

Structure of dataset

The asthma dataset by Szucs et al. is characterised by a two-level hierarchical data structure, with patients being the level 1 units und physicians being the level 2 units [203].

Table 1 shows potential predictor variables by level. Details regarding data collection and variable definitions are contained in chapter 4.

Table 1. Potential predictors of direct medical costs (log scale) induced by asthma

Predictor variable	**Remarks regarding level**	**Used in main conventional model?**
	Level 1 (patients)	
Age and age squared	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Duration of asthma	Impact on target variable may be influenced by unmeasured level 2 characteristic	No[a]
Height		No due to missing values
Weight		No due to missing values
BMI		No due to missing values
FEV_1		No due to missing values
FVC		No due to missing values
Degree of asthma severity (based on 1995 Global Initiative for Asthma (GINA) recommendations on medication use) [80]	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Presence of asthma exacerbations	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes

Table 1 ctd.

Interaction of degree of severity and presence of asthma exacerbations	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Presence of asthma-related comorbidities	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Quick reliever vs controller therapy	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Involvement of a pulomonologist in diagnosis or treatment	Level unclear, to be split up in "Diagnosis of asthma by a pulmologist" and "Specialty of treating physician" (see level 2)	Yes
Employment status in those aged 65 years or younger	Impact on target variable may be influenced by unmeasured level 2 characteristic	No[b]
Absences from work	Impact on target variable may be influenced by unmeasured level 2 characteristic	No[b]
Insurance status	Impact on target variable may be influenced by unmeasured level 2 characteristic	No[b]
Level 2 (physicians)		
Specialty of treating physician		No, non-significant
Location in a rural or urban area		No, non-significant
Language region		No, non-significant

a Reason for non-inclusion in main conventional model: Due to many missing values, the number of usable observations would have been reduced to N = 268. Also, there was no indication of a substantial contribution to the model, not even if duration was replaced by age at diagnosis to reduce the correlation with age.

b Reason for non-inclusion in main conventional model: Regression coefficients for these predictors were (near) significant but their inclusion did not lead to a substantial improvement of the model. (Erroneously, insurance status is not mentioned in chapter 4.)

The majority of potential predictors were level 1 variables. Height, weight, BMI, FEV_1 and FVC were excluded from the multilevel analyses due to a massive amount of missing values. (Tentative use of these variables in conventional models hinted at an impact of weight, BMI and FEV_1, but this could not be pursued further due to data quality issues.) The impact of some other level 1 variables describing disease characteristics (e.g., degree of asthma control) and treatment characteristics (e.g., choice of quick reliever vs. controller therapy) on the target variable could be assumed to be influenced by the behaviour of the treating physician, an unmeasured level 2 characteristic. Group level predictors were constructed by aggregating the values of these variables within their level 2 units (i.e., physicians), where they had an intuitive meaning. For example, the approach to the treatment of employed asthma patients and thus, the costs induced, might have differed between physicians with a low vs. high proportion of employed patients. Physicians frequently dealing with employed patients might be particularly aware of the specific requirements of this group, and behave accordingly.

Multilevel modelling process and intermediate results

Step 1. Random intercept model (variance components model).

A random intercept model was estimated (Table 2) and showed a highly significant random intercept term ($p < 0.001$). The fixed parameter estimates remained essentially stable.

Table 2. Two-level random intercept model of direct medical costs (log scale) induced by asthma[a]

Number of level 1 units = 420					**Log likelihood = -539.880**	
Number of level 2 units = 107					**AIC 1109.76**	
Variable	**Esti-mates**	**Std. Err.**	**Test stat.[b]**	**p value**	**95% Confidence Interval**	
Fixed part						
Degree of severity:						
Mild persistent[c]	0.747	0.187	3.99	< 0.001	0.379	1.114
Moderate persistent[c]	1.026	0.181	5.66	< 0.001	0.671	1.381
Severe persistent[c]	0.907	0.197	4.60	< 0.001	0.521	1.294
Exacerbations present	0.326	0.289	1.13	0.260	-0.241	0.892
Interaction of degree of severity and presence of exacerbations, ordinal:						
Level 1[d]	-0.422	0.346	-1.22	0.224	-1.100	0.257
Level 2[d]	0.111	0.331	0.34	0.737	-0.537	0.759
Level 3[d]	0.504	0.330	1.53	0.127	-0.142	1.149
Age (centered)	0.006	0.002	2.66	0.008	0.002	0.011
Age squared (centered)	-0.0002	0.0001	-2.11	0.035	-0.0004	-0.0000
Asthma-related comorb. present	0.321	0.123	2.62	0.009	0.080	0.561
Involvement of pulmonologist	0.524	0.115	4.56	< 0.001	0.299	0.750
Controller therapy[e]	0.249	0.094	2.65	0.008	0.065	0.433
Intercept	4.249	0.310	13.71	< 0.001	3.641	4.856
Random part - level 1						
Residual variance	0.685	0.053			0.554	0.762
Random part - level 2						
Intercept variance	0.140	0.047	14.69	< 0.001	0.047	0.232

a Conventional regression model for comparison: chapter 4, Table 4.

b Fixed parameters, Wald test based on z statistic; random parameters, likelihood ratio test, p-value divided by 2 (see Methods, p. 35). Standard errors and CIs based on the Wald statistic throughout.

c Compared to mild intermittent. Wald test for this set of variables, p < 0.001.

d Compared to level 0. Wald test for this set of variables, p = 0.002.

e Compared to quick reliever therapy.

Step 2. Analysis of the variance structure.

Based on the covariates with a potential for level 2 variation, the variance structure was analysed as described in the methods section. The resulting model (Table 3) contained additional random effects for the variable distinguishing quick reliever from controller therapy and for the variable describing employment status in those aged 65 years or younger. No significant covariance terms were found and all covariance terms were set to zero. The random intercept term turned non-significant.

Table 3. Two-level model of direct medical costs (log scale) induced by asthma, including random effects for the intercept, quick reliever versus controller therapy, and employment status in those aged 65 or younger[a]

Number of level 1 units = 408 **Number of level 2 units = 106**					**Log likelihood = -521.200** **AIC 1080.40**	
Variable	**Esti-mates**	**Std. Err.**	**Test stat.[b]**	**p value**	**95% Confidence Interval**	
Fixed part						
Degree of severity:						
Mild persistent[c]	0.799	0.190	17.646	< 0.001	0.427	1.171
Moderate persistent[c]	1.085	0.186	34.172	< 0.001	0.720	1.450
Severe persistent[c]	0.879	0.202	18.934	< 0.001	0.483	1.275
Exacerbations present	0.333	0.282	1.390	0.238	-0.220	0.886
Interaction of degree of severity and presence of exacerbations, ordinal:						
Level 1[d]	-0.414	0.341	1.475	0.225	-1.082	0.254
Level 2[d]	0.072	0.326	0.049	0.825	-0.567	0.712
Level 3[d]	0.539	0.326	2.723	0.099	-0.010	1.118
Age (centered)	0.009	0.003	11.269	0.001	0.004	0.013
Age squared (centered)	-0.0002	0.0001	4.948	0.026	-0.0004	-0.0000
Asthma-related comorb. present	0.293	0.124	5.539	0.019	0.050	0.536
Diagnosis by a pulmonologist:						
No[e]	-0.231	0.107	4.608	0.032	-0.441	-0.021
Unknown[e]	-0.351	0.149	5.519	0.019	-0.643	-0.059
Controller therapy[f]	0.512	0.119	18.595	< 0.001	0.279	0.745
No employment despite under age 65 (employment status)	0.219	0.127	2.966	0.085	-0.030	0.468
Intercept	5.431	0.190	811.768	< 0.001	5.059	5.803
Random part - level 1						
Residual variance	0.621	0.054			0.515	0.727
Random part - level 2						
Intercept variance	0.049	0.054	0.823	0.182	-0.057	0.155
Controller therapy variance	0.152	0.078	4.376	0.018	-0.001	0.305
Employment status variance	0.226	0.143	3.605	0.029	-0.054	0.506

[a] Conventional regression model for comparison: chapter 4, Table 4.

[b] Fixed parameters, chi squared-based Wald test; random parameters, likelihood ratio test, p-value divided by 2 (see Methods, p. 35). Standard errors and CIs based on the Wald statistic throughout.

[c] Compared to mild intermittent. Wald test for this set of variables, p < 0.001.

[d] Compared to level 0. Wald test for this set of variables, p = 0.002.

[e] Compared to yes. Wald test for this set of variables, p = 0.016.

[f] Compared to quick reliever therapy.

NOTE: The AIC shown in Table 3 is not comparable with the AIC shown in Table 2, as the number of observations differed due to some missing values in the employment status variable. The correct comparator AIC, based on the restricted set of 408 observations, was determined to be 1083.61. The AICs given in the text below were based on this same set of 408 observations, to allow for comparisons.

Step 3. Goodness-of-fit.

For the model shown in Table 3, the inverse normal plot of the studentised level 1 residuals showed acceptable properties, with some deviation from the normal distribution in the tail areas (Figure 1). Plotting the studentised level 1 residuals against the fixed part predicted values (Figure 2) showed no particularities apart from a slight tendency towards heteroskedasticity, which was not unexpected in a regression model of health care costs.

The three sets of studentised level 2 residuals, representing random deviations from the average intercept, the average coefficient of the variable distinguishing quick reliever from controller therapy, and the average coefficient of the working status variable, showed no serious deviations from the normal distribution (Figure 3).

Step 4: Influential level 2 units.

Caterpillar plots of the ranked level 2 residuals and diagnostics plots showing influence values [163:169-71] were used to identify influential level 2 units (marked in Figures 4, 5) and their corresponding observations at level 1 (marked in Figure 6). Four out of the seven observations marked had high residuals at level 1. Re-estimating the model using dummy variables for these four observations led to a significantly improved model, the AIC being reduced from 1080.40 to 1040.30. Figure 7 shows a repetition of the caterpillar plots, now based on the re-estimated model, with the residuals representing the "red" level 2 unit moved towards the middle of the distribution, but the residuals for the "green" level 2 unit still being located at its lower end. A possible interpretation would be that in the former case, the extreme position of the affected level 2 unit was mostly due to the level 1 characteristics of the outlying observations contained therein, while it was due to "true" level 2 characteristics (i.e., special characteristics of the affected physician) in the latter case. No such distinction could be made for the "blue" level 2 unit as it contained only one observation at level 1.

Figure 1. Inverse normal plot of studentised level 1 residuals

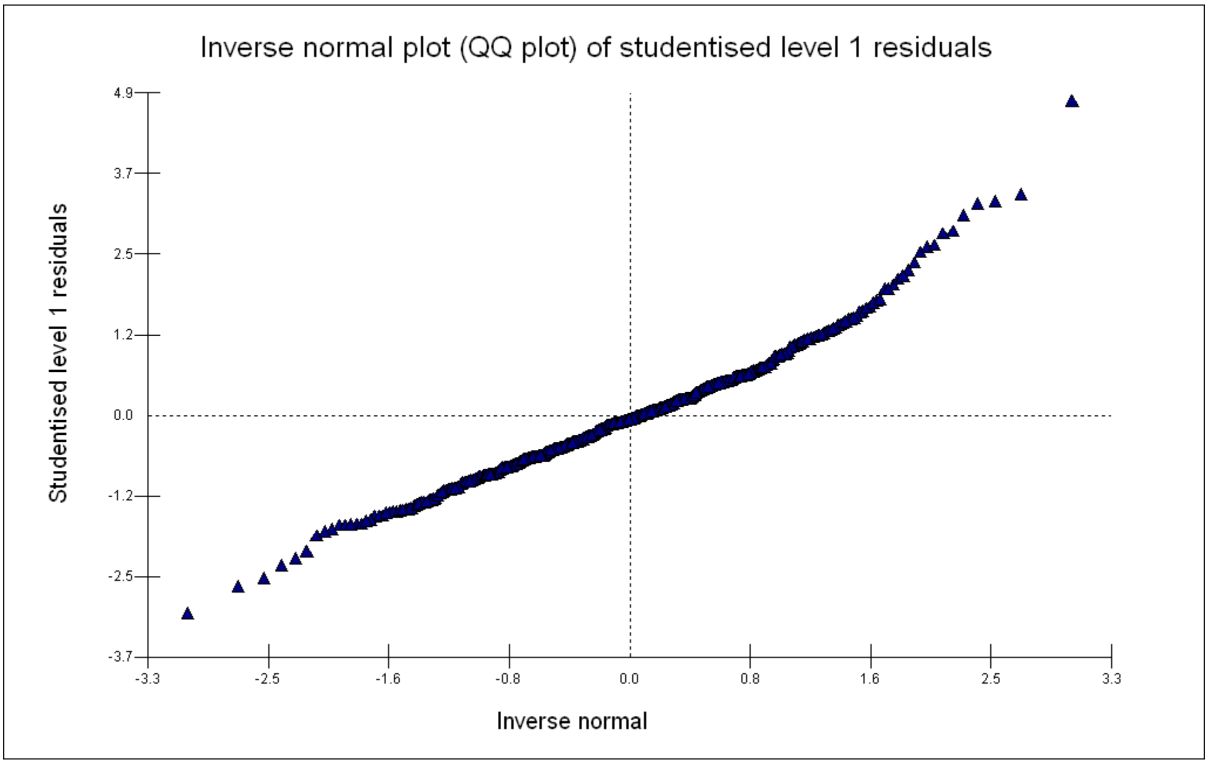


Figure 2. Plot of studentised level 1 residuals against fixed part predicted values

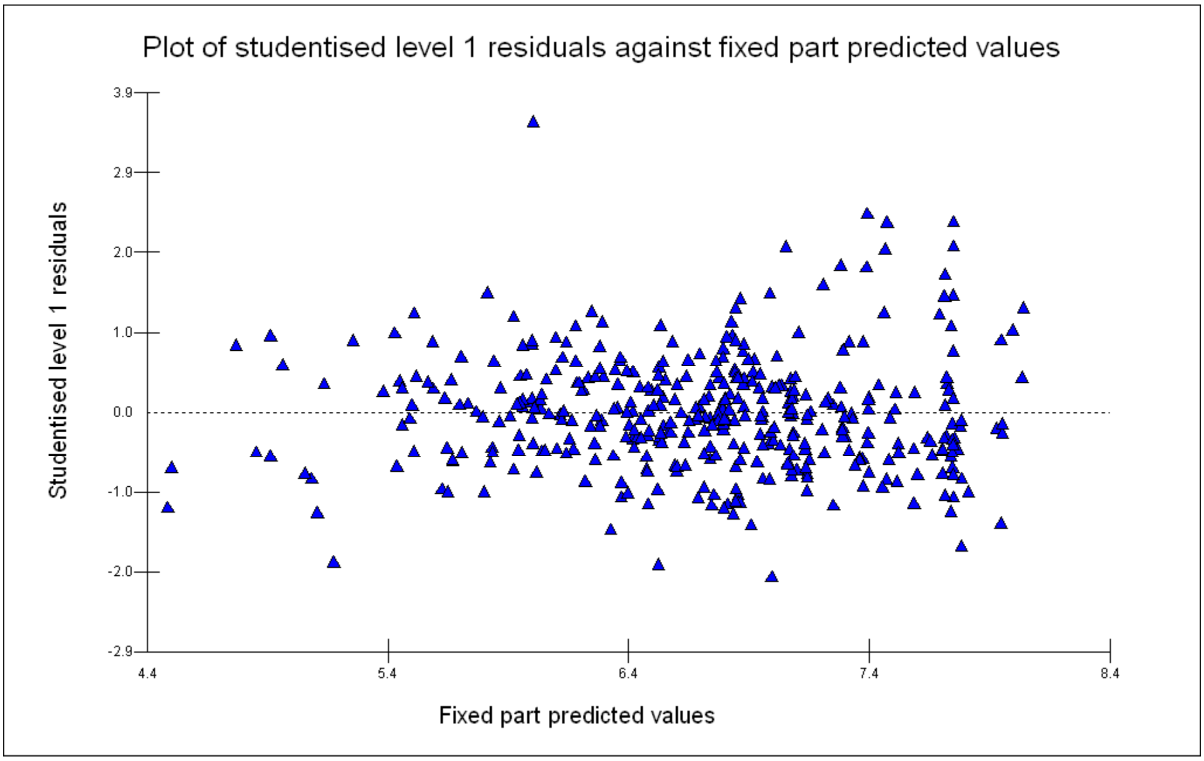


Figure 3. Inverse normal plots of studentised level 2 residuals

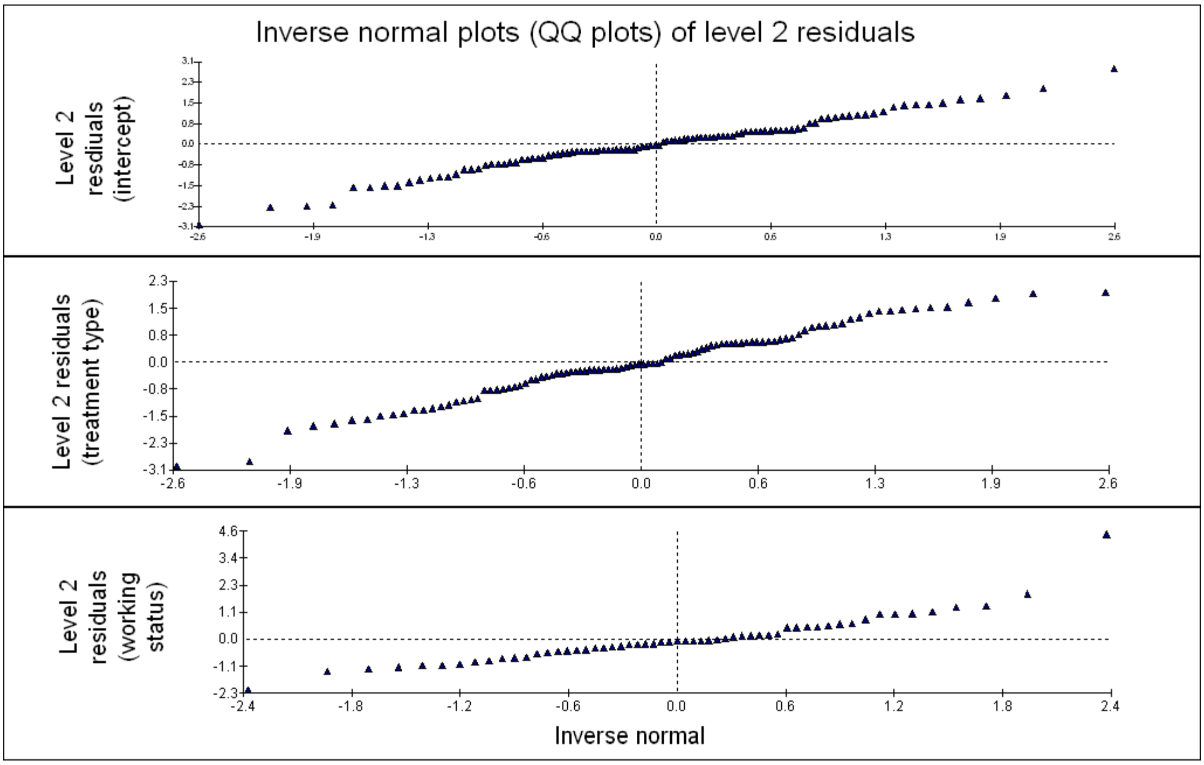


Figure 4. Caterpillar plots of ranked level 2 residuals with 95% CIs

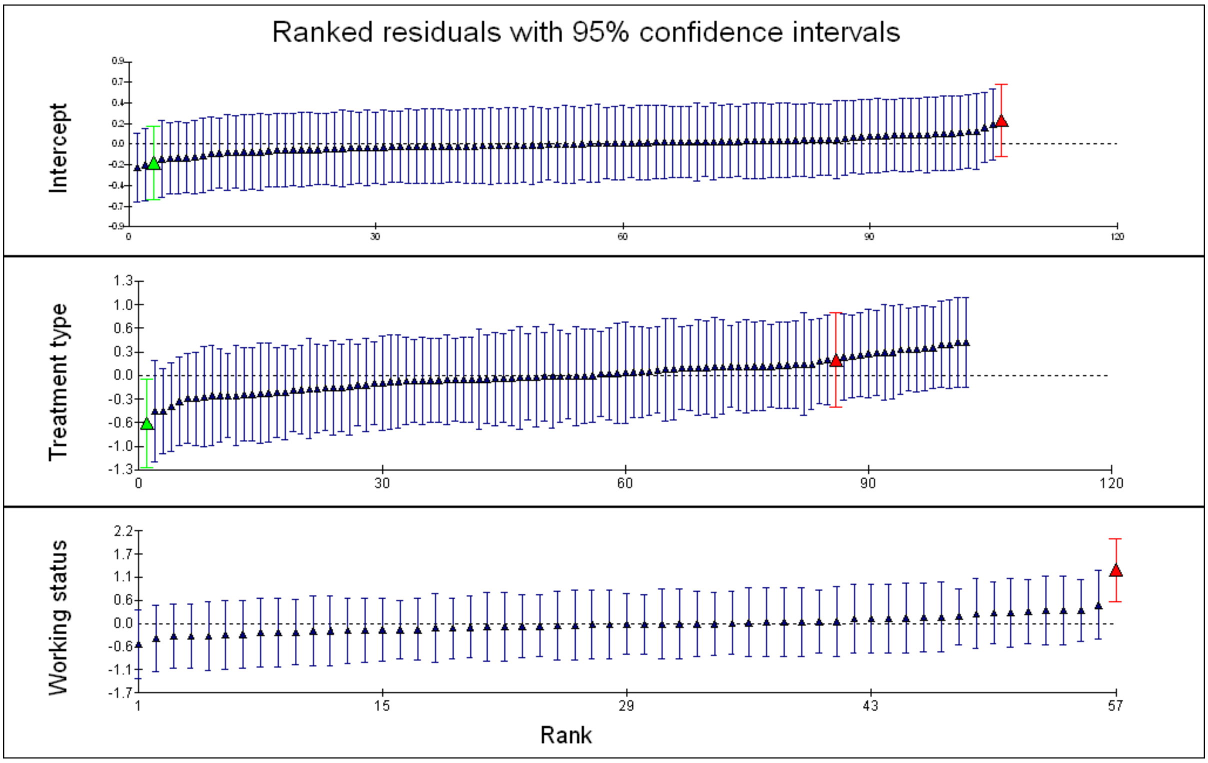


Figure 5. Diagnostics plots as available from MLwiN statistical package, example of level 2 intercept residuals

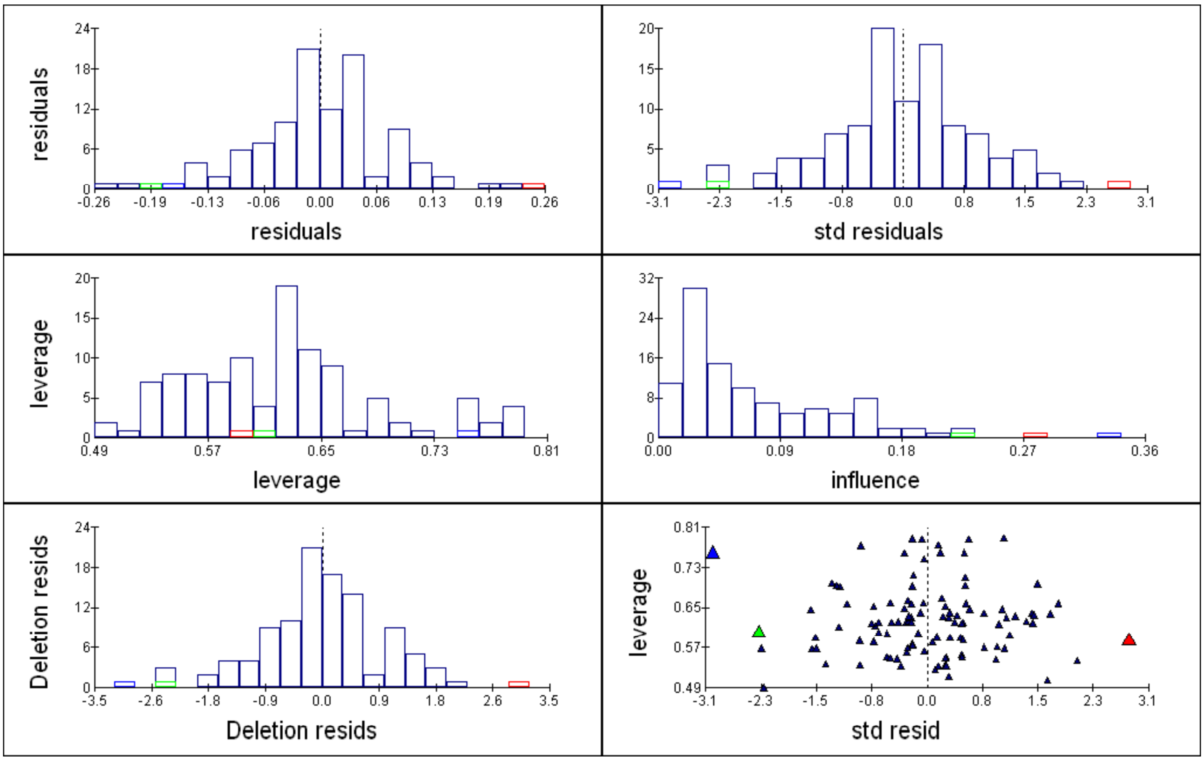


Figure 6. Inverse normal plot of studentised level 1 residuals, observations representing influential level 2 units marked

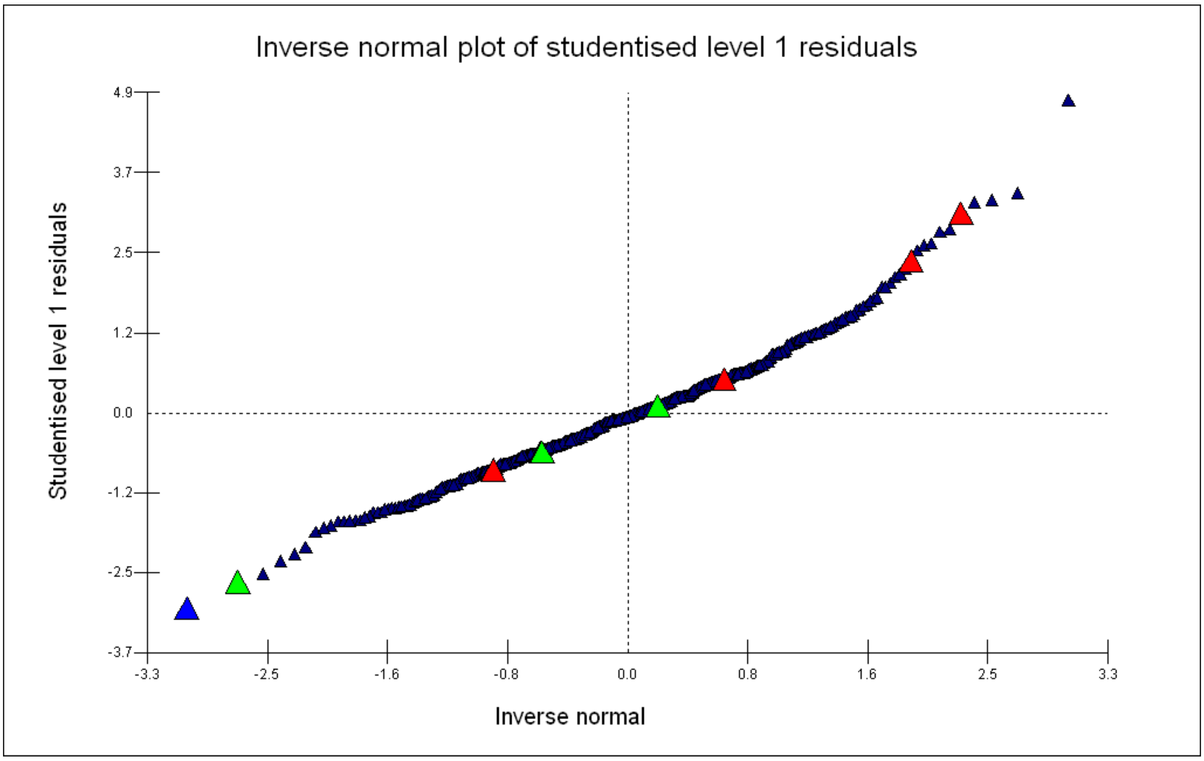
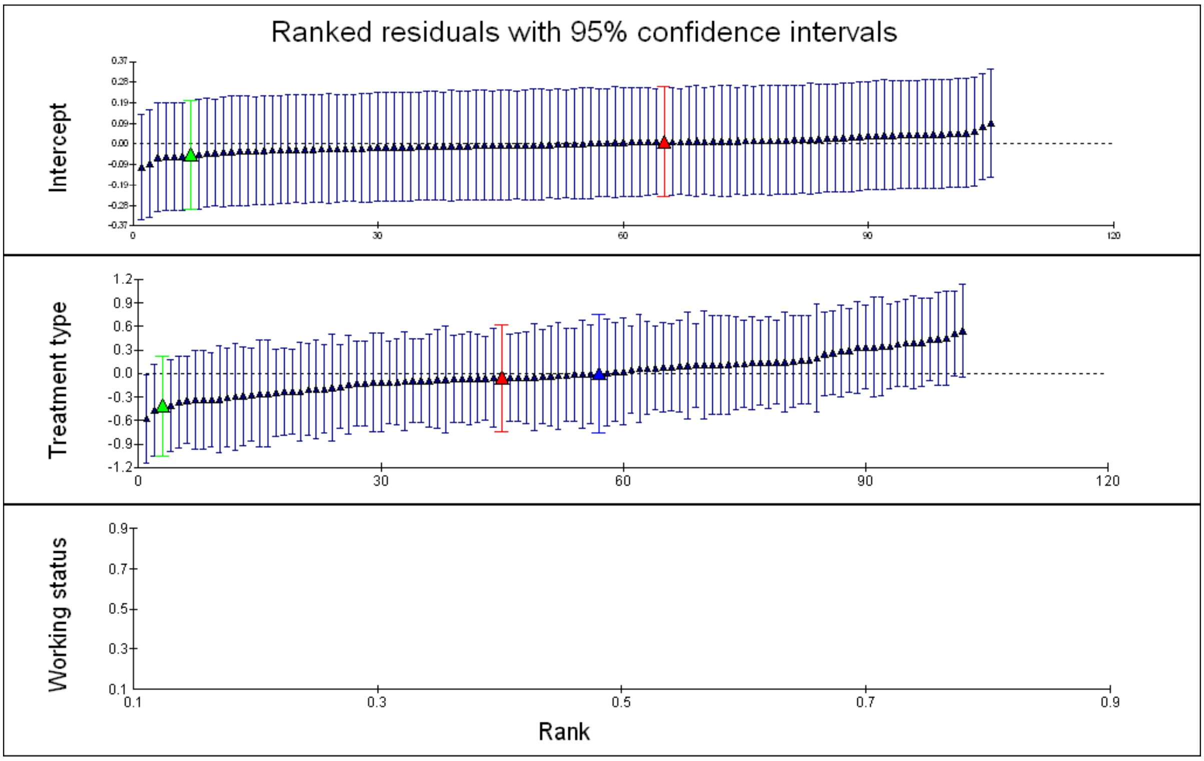


Figure 7. Caterpillar plots of ranked level 2 residuals with 95% CIs, based on a re-estimation of the multilevel model using dummy variables for four influential



The logical next step was to absorb all observations nested within the "green" level 2 unit into a single dummy variable, while still representing the outlying "red" and "blue" observations by individual dummy variables. The explanatory value of the model would thus be optimised by neutralising, in its entirety, the impact of a level 2 unit confirmed to have abnormal and influential characteristics. The AIC of the resulting model was reduced further to 1038.63.

Step 5: Final multilevel model.

The last mentioned model was finalised by removing the working status variable, which became non-significant at both levels when the influential level 2 units identified were taken into account as described above (in the last mentioned model with AIC 1038.63, p = 0.500 for the variance term and p = 0.251 for the fixed term). The random intercept term was also removed as it remained non-significant in all models containing a random effect for the variable distinguishing quick reliever from controller therapy (in the last mentioned model with AIC 1038.63, p = 0.182).

Details of the final model are shown in chapter 4, Table 3. Residual-based re-diagnosis showed acceptable properties for this model. Re-estimating the final model from a set from 408 instead of 420 observations, to allow for comparisons with the above shown AICs, led to an AIC of 1038.95.

Step 6: Model re-calculation using alternative algorithms.

Estimating the final model using Adaptive Quadrature (as implemented in Stata) or the Restricted Iterative Generalized Least Squares approach (RIGLS; as implemented in MLwiN) resulted in essentially identical parameter estimates. A final re-estimation used the non-parametric bootstrap method implemented in MLwiN. Five sets of 500 replicates were run, and the bias-corrected results were again fully confirmatory.

Remark on centring the "quick reliever vs. controller therapy" covariate

It has been recommended to use centred random slope covariates in multilevel modelling, in order to make the intercept and its variance interpretable (see Methods, pp. 34-5) [105:57-8,70-1]. Leaving numerical problems aside, all other model parameters are expected to remain unaffected by this procedure, and the resulting models are expected to be equivalent.

However, in the multilevel analysis of the asthma dataset, using the non-centred "quick reliever vs. controller therapy" covariate turned the random intercept term small and non-significant, which led to its removal from the model. When the "quick reliever vs. controller therapy" covariate was subsequently centred, the model changed considerably and was no longer equivalent to the main model. (The log-likelihood differed substantially.) In line with expectations, this problem was not observed when the random intercept term was re-added. It may be a point for further discussion if removing non-significant random intercept terms, as happened here, should be avoided to rule out such problems.

Independently of this issue, the centring did not facilitate interpretation in this case where the random slope covariate was binary and coded 0/1. In the non-centred situation, and disregarding the impact of the scaling of the other covariates for a moment, the intercept represented the patients receiving quick-reliever therapy, with no substantial random variation. After centring the random slope covariate, the intercept represented a difficult to conceive theoretical patient with a theoretical average treatment [105:56].

Gatekeeping dataset

Preliminary remark

The gatekeeping dataset described in chapter 5 allowed to assess covariates of health plan membership as well as covariates of health care costs. Conventional analysis focused on health care costs, and this same response of interest was used in the multilevel re-analysis described here. A distinction between reduced and extended cost models was made in chapter 5, and multilevel assessments were based on the latter, more complete models.

Structure of dataset

The gatekeeping dataset formally has a cross-classification structure, with health plan beneficiaries being the level 1 units and belonging to two groupings each at level 2 (insurance companies and physicians). In the primary analysis of cost to the Swiss statutory sick funds, no significant influence of insurance company was identified and a brief multilevel re-assessment confirmed this finding. Therefore, multilevel re-analysis used a two-level model with beneficiaries at level 1 and physicians at level 2.

List of potential predictor variables

Collection of potential covariates in the gatekeeping study aimed at achieving an appropriate casemix adjustment in a general population and was thus very comprehensive. Table 4 lists the potential level 1 predictor variables which showed to be of importance in the primary analysis. In addition, all potential level 2 predictor variables are shown.

Table 4. Potential predictors of total and outpatient costs to the Swiss statutory sick funds in 2000

Predictor variable	**Remarks regarding level**	**Used in extended two-part, non-multilevel cost models?**[a]
	Level 1 (patients)	
Health plan membership (gatekeeping vs. fee-for-service)	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Age	(Impact on target variable unlikely to be strongly influenced by unmeasured level 2 characteristic)	Yes
Gender	(Impact on target variable unlikely to be strongly influenced by unmeasured level 2 characteristic)	Yes
SF-36 scales	(Impact on target variable unlikely to be strongly influenced by unmeasured level 2 characteristic)	Yes (SF-2 scale and general health scale)
Morbidity, measured by ICED [89], in the year before the gatekeeping plan started	(Impact on target variable unlikely to be strongly influenced by unmeasured level 2 characteristic)	Yes
Increase in morbidity, measured by ICED [89], between the year before the gatekeeping plan started, and 2000	(Impact on target variable unlikely to be strongly influenced by unmeasured level 2 characteristic)	Yes
History of mental illness		No, non-significant
Smoking behaviour		No, non-significant
Alcohol consumption		No, non-significant
BMI		No, non-significant
Physical activity		No, non-significant
Importance assigned to healthy nutrition		No, non-significant
Self-reported avoidance of seeing a doctor		Yes
Outpatient costs in the year before the gatekeeping plan started	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes
Educational level		No, non-significant due to presence of covariates with partially overlapping content
Professional status		Yes (being retired)
Living in a partnership		Yes
Marital status		Yes
Household size		Yes

Table 4 ctd.

Household income		No, non-significant due to presence of a covariate with overlapping content
Integration score (composite measure of being Swiss born and / or a Swiss citizen		Yes
Residency in the Aarau area in 1996		Yes
Nursing home residency	Impact on target variable may be influenced by unmeasured level 2 characteristic	No, non-significant (after controlling for morbidity!)
Importance assigned to low insurance premiums (highly correlated with household income)		Yes
Duration of health insurance with the same company		No, non-significant
Number and type of complementary insurance contracts	Impact on target variable may be influenced by unmeasured level 2 characteristic	Yes (complementary semi-private insurance, complementary dental insurance)
	Level 2 (physicians)	
Year of medical degree of physician		No, non-significant
Specialty of physician		No, non-significant
Physician treating gatekeeping and fee-for-service beneficiaries, or fee-for-service beneficiaries only		No, non-significant
Physician working in an individual or group practice		No, non-significant
Time since entry in current practice		No, non-significant
Number of consultations per year		No, non-significant

a See chapter 5 for details of the total cost models, particularly Tables 3, 4. Details of outpatient cost models not shown.

Assessment of higher level variation

Part 1 of two-part models: logistic models of any health care costs occurring

The main logistic regression model of any costs to the Swiss statutory sick funds occurring (chapter 5, Table 3) was used as a basis and the intercept was allowed to vary at random. No significant level 2 variation was detected (intercept 5.62; level 2 variance 0.26 with standard error 0.39; log likelihood -99.05 compared to -98.70 for the original fixed effects model; likelihood ratio statistic 0.70 (1dgf); p value 0.201).

This Adaptive Quadrature-based result was re-evaluated using RIGLS in combination with Penalized Quasi Likelihood (PQL) estimation. The finding of no significant level 2 variation being present was confirmed, although the results differed in detail (intercept 7.14; level 2 variance 0.48 with standard error 0.41).

Tentatively allowing the covariates with a potential for higher level variation to vary at random did not produce any substantial or significant results. No further related assessments were performed.

Part 2 of two-part models: GLMs of total and outpatient health care costs in those with non-zero costs

The GLMs describing total costs and outpatient costs to the Swiss statutory sick funds in those beneficiaries with non-zero costs were used as a basis and their intercepts were allowed to vary at random. Virtually no level 2 variance was detected. For example, in the total cost model, the intercept was 7.36 and the level 2 variance was < 0.001 with standard error < 0.001.

This Adaptive Quadrature-based result was fully confirmed by RIGLS and PQL-based estimation. To match with the requirements of the software used for this, the cost variables were now treated as discrete count variables and the original GLMs (assuming a gamma distribution of errors) were reconstructed under a negative binomial distributional assumption [83: 7/10].

Tentatively allowing the covariates with a potential for higher level variation to vary at random did neither produce any substantial or significant results.

In order to rule out a situation where over-modelling at level 1 might hide any level 2 variation, the number of covariates in the fixed part of the model was gradually reduced, but even after removal of all covariate terms apart from the intercept, the level 2 random variation of the intercept remained near-zero and non-significant.

In order to make incorrect estimation results an even less likely explanation, an additional multilevel model assuming normality was tentatively fitted to the logarithm of the response variable. Again, no substantial or significant level 2 variation was found. This is supportive of the above-described findings, non-regarding that even on the log scale, the normality assumption was inadequate in this case, given a coefficient of variation of the untransformed response variable of around 0.73 [141: 292-3, 296-297]. No further related assessments were performed.

Neutropenia dataset

Structure of dataset

The neutropenia dataset is a combination of six audits of breast cancer chemotherapy conducted in different European countries (see chapter 6). It is characterised by a three-level hierarchical data structure, with patients being the level 1 units, study centres being the level 2 units, and audits conducted in different countries being the level 3 units.

Table 5 shows potential predictor variables by level. For details regarding data collection and variable definitions, see chapter 6.

There were no level 2 predictors in the strict sense, as type of centre (e.g., academic vs. non-academic) and centre size were unavailable, and no relevant level 3 predictors in the strict sense. However, many of the level 1 predictors could be assumed to be "intertwined" with the higher levels of the hierarchy in different ways. In some cases, their impact on the target variable was potentially influenced by level 2 characteristics (e.g., the aggressiveness of treatment of old cancer patients could be expected to vary across centres and treating physicians). The values of some other level 1 variables (e.g., chemotherapy regimen chosen) could be assumed to be directly influenced by their corresponding level 2 units, i.e. by centre-level differences in clinical practice and physician-level differences in treatment behaviours and clinical experience. The values of some level 1 variables could even have been influenced by their corresponding level 3 units (audits). For example, different national drug coverage policies could have influenced the chemotherapy regimens chosen. A strong impact of the audits (i.e., the study designs) themselves appeared unlikely, as they were very similar and purely observational in nature.

Table 5. Potential predictors of any neutropenic event occurrence

Predictor variable	Remarks regarding level	Used in main conventional model?
	Level 1 (patients)	
Age and age squared	Impact on target variable may be influenced by level 2 characteristics (clinical practice)	Yes
Menopausal status	Impact on target variable may be influenced by level 2 characteristics (clinical practice)	No, highly correlated with age
Body surface area	Impact on target variable may be influenced by level 2 characteristics (clinical practice)	Yes
BMI	Impact on target variable may be influenced by level 2 characteristics (clinical practice)	Yes
Oestrogen receptor status	Impact on target variable may be influenced by level 2 characteristics (clinical practice)	No, non-significant
Disease stage	Impact on target variable may be influenced by level 2 characteristics (clinical practice)	No, non-significant
Chemotherapy regimen	Influenced by level 2 characteristics (clinical practice)	Yes
Summation dose intensity (SDI)	Influenced by level 2 characteristics (clinical practice)	Yes
Planned chemotherapy cycles	Influenced by level 2 characteristics (clinical practice)	Yes
Concomitant radiotherapy administration	Influenced by level 2 characteristics (clinical practice)	Yes
Is treatment part of a clinical trial protocol?	Influenced by level 2 characteristics (clinical practice)	No, not available
	Level 2 (centres)	
Academic vs. non-academic centre		No, not available
Centre size (patients with disease of interest treated per year		No, not available
	Level 3 (audit)	
Country		No, highly correlated with audit itself
	Level difficult to assign	
Year of treatment and year of treatment squared	If interpreted as level 1, impact on target variable influenced by level 2 characteristics (clinical practice) and level 3 characteristics (timing of study)	Yes

Alternative presentation of conventional model

Presentation of the main conventional logistic regression model of any neutropenic event occurrence in chapter 6, Table 3 used simple effects to describe two-way interactions. However, the more familiar approach of showing main effects plus interaction terms was adopted to present the multilevel models. Moreover, linear coefficients are shown in the tables displaying multilevel models, instead of odds ratios. For reference purposes, Table 6 shows the main conventional logistic regression model of any neutropenic event occurrence re-written in this form. The model shown is fully mathematically equivalent to the model shown in chapter 6, Table 3.

Table 6. Influences on any neutropenic event occurrence (logistic regression allowing for clustering by audit). Two-way interactions re-written as main effects plus interaction terms[a]

N = 2'358				**Pseudo R squared 0.070**		
Std. err. adjusted for 6 clusters (audits)				**Log pseudolikelihood -1147.969 AIC 2355.94**		
Variable	**Esti-mates**	**Std. Err.**	**Test stat.**[b]	**p value**[b]	**95% Confidence Interval**	
Age[c]	0.017	0.005	3.53	< 0.001	0.008	0.026
Body surface area[c]	1.349	0.377	3.57	< 0.001	0.609	2.088
BMI[c]	-0.035	0.014	-2.47	0.013	-0.062	-0.007
Chemotherapy regimen:[d]						
Three weekly CMF	1.010	0.149	6.79	< 0.001	0.719	1.301
Anthracycline-based	0.407	0.217	1.87	0.061	-0.019	0.833
Taxane-based	0.520	0.080	6.48	< 0.001	0.363	0.677
Other	-0.144	0.477	-0.30	0.764	-1.080	0.793
Normal to high SDI[e]	0.529	0.088	5.99	< 0.001	0.356	0.702
Planned chemotherapy cycles[c]	0.359	0.100	3.59	< 0.001	0.163	0.555
Concomitant radiotherapy administration (Rx):[f]						
Rx yes	3.288	1.205	2.73	0.006	0.927	5.649
Rx unknown	0.015	2.327	0.01	0.995	-4.545	4.575
Year of treatment (centered)[c]	-0.006	0.003	-1.82	0.068	-0.013	0.001
Year of treatment squared (centered)[c]	-0.0008	0.0004	-1.78	0.075	-0.0016	0.0001
Interaction of chemotherapy regimen and year of treatment:						
Three weekly CMF	0.085	0.047	1.81	0.071	-0.007	0.176
Anthracycline-based	-0.106	0.045	-2.37	0.018	-0.194	-0.018
Taxane-based	0.707	0.078	9.11	< 0.001	0.555	0.859
Other	0.468	0.036	13.05	< 0.001	0.398	0.538

Table 6 ctd.

Interaction of chemotherapy regimen and Rx:						
Three weekly CMF, Rx yes	-0.553	0.469	-1.18	0.238	-1.472	0.366
Anthracycline-based, Rx yes	-0.902	0.212	-4.25	< 0.001	-1.319	-0.486
Taxane-based, Rx yes	0.432	0.219	1.97	0.049	0.002	0.862
Other, Rx yes	--	--	--	--	--	--
Three weekly CMF, Rx unkn.	0.206	0.393	0.52	0.601	-0.566	0.977
Anthracycline-based, Rx unkn.	0.805	0.399	2.02	0.044	0.023	1.587
Taxane-based, Rx unknown	--	--	--	--	--	--
Other, Rx unknown[g]	4.320	0.873	4.95	< 0.001	2.608	6.031
Interaction of body surface area and Rx:						
Rx yes	-1.397	0.449	-3.11	0.002	-2.278	-0.516
Rx unknown	1.231	0.850	1.45	0.148	-0.435	2.897
Interaction of planned chemotherapy cycles and Rx:						
Rx yes	-0.197	0.096	-2.06	0.039	-0.385	-0.010
Rx unknown	-0.446	0.082	-5.46	< 0.001	-0.607	-0.286
Interaction of normal to high SDI and Rx:						
Rx yes	1.136	0.314	3.61	< 0.001	0.520	1.752
Rx unknown	0.171	1.095	0.16	0.876	-1.974	2.317
Intercept	-6.494	0.640	-10.15	< 0.001	-7.748	-5.240

a Mathematically equivalent to chapter 6, Table 3.

b Combined Wald tests (z tests) for all sets of categorical or ordinal variables and for all sets of interaction terms, $p < 0.05$.

c Per one unit increase.

d Compared to four weekly CMF.

e Second to 4th quartiles compared to 1st quartile.

f Compared to no concomitant radiotherapy administration.

g Parameter estimate based on 4 observations, assumed to be an artefact

Abbreviations: C, cyclophosphamide; F, 5-fluorouracil; M, methotrexate; SDI, summation dose intensity.

Multilevel modelling process and intermediate results

Step 1. Random intercept model (variance components model).

The intercept was allowed to vary at random at levels 2 (centres) and 3 (audits), revealing substantial and significant variation at level 2, but no significant variation at level 3. Dropping the level 3 random intercept term resulted in virtually identical parameter estimates otherwise. The resulting two-level random intercept model (Table 7) showed coefficient and standard error estimates which were mostly similar to those seen in the main conventional model. However, the significance of the year of treatment and year of treatment squared terms was lost, and the same was true for the term representing interaction of year of treatment and use of a three weekly CMF regimen. The set of terms representing interaction of year of treatment and chemotherapy regimen type remained highly significant as a whole. Other substantial changes were limited to coefficients which were based on very few observations, and thus doubtful anyway (terms representing taxane-based and "other" chemotherapy regimens).

Table 7. Two-level random intercept model of any neutropenic event occurrence (logistic regression allowing for clustering by audit)[a]

Number of level 1 units = 2'358 **Number of level 2 units = 93**	**Log likelihood -1114.342**
Std. err. adjusted for 6 clusters (audits)	**AIC 2290.68**

Variable	**Esti-mates**	**Std. Err.**	**Test stat.[b]**	**p value[b]**	**95% Confidence Interval**	
Fixed part						
Age[c]	0.022	0.006	3.54	< 0.001	0.010	0.034
Body surface area[c]	1.431	0.289	4.96	< 0.001	0.865	2.000
BMI[c]	-0.045	0.012	-3.85	< 0.001	-0.068	-0.022
Chemotherapy regimen:[d]						
Three weekly CMF	0.777	0.130	5.96	< 0.001	0.521	1.302
Anthracycline-based	0.609	0.232	2.62	0.009	0.154	1.064
Taxane-based	1.182	0.201	5.88	< 0.001	0.788	1.576
Other	0.264	0.449	0.59	0.557	-0.616	1.143
Normal to high SDI[e]	0.478	0.176	2.72	0.007	0.133	0.822
Planned chemotherapy cycles[c]	0.351	0.110	3.21	0.001	0.137	0.566
Concomitant radiotherapy administration (Rx):[f]						
Rx yes	3.237	0.862	3.75	< 0.001	1.547	4.927
Rx unknown	-0.504	2.338	-0.22	0.829	-5.086	4.077
Year of treatment (centered)[c]	-0.004	0.005	-0.81	0.418	-0.014	0.006
Year of treatment squared (centered)[c]	-0.0006	0.0004	-1.63	0.104	-0.0013	0.0001
Interaction of chemotherapy regimen and year of treatment:						
Three weekly CMF	0.021	0.031	0.66	0.506	-0.041	0.083
Anthracycline-based	-0.095	0.047	-2.03	0.043	-0.186	-0.003
Taxane-based	0.679	0.097	7.01	< 0.001	0.489	0.869
Other	0.419	0.035	11.85	< 0.001	0.350	0.489

Table 7 ctd.

Interaction of chemotherapy regimen and Rx:						
Three weekly CMF, Rx yes	-0.376	0.479	-0.79	0.432	-1.315	0.563
Anthracycline-based, Rx yes	-0.915	0.186	-4.93	< 0.001	-1.279	-0.551
Taxane-based, Rx yes	0.645	0.170	3.97	< 0.001	0.311	0.979
Other, Rx yes	--	--	--	--	--	--
Three weekly CMF, Rx unkn.	0.845	0.478	1.77	0.077	-0.091	1.782
Anthracycline-based, Rx unkn.	0.872	0.459	1.90	0.057	-0.027	1.771
Taxane-based, Rx unknown	--	--	--	--	--	--
Other, Rx unknown[g]	4.644	0.982	4.73	< 0.001	2.719	6.570
Interaction of body surface area and Rx:						
Rx yes	-1.409	0.294	-4.79	< 0.001	-1.985	-0.833
Rx unknown	1.358	0.797	1.70	0.089	-0.205	2.920
Interaction of planned chemotherapy cycles and Rx:						
Rx yes	-0.211	0.089	-2.36	0.018	-0.386	-0.036
Rx unknown	-0.425	0.083	-5.11	< 0.001	-0.588	-0.262
Interaction of normal to high SDI and Rx:						
Rx yes	1.186	0.352	3.37	0.001	0.497	1.876
Rx unknown	0.008	1.033	0.01	0.993	-2.015	2.032
Intercept	-6.733	0.727	-9.26	< 0.001	-8.158	-5.307
Random part - level 1						
Binomial variance	1[h]	0				
Random part - level 2						
Intercept variance	0.587	0.065	67.25	< 0.001	0.460	0.714

a Conventional regression model for comparison: chapter 6, Table 3, re-written to facilitate comparison in Appendix I, Table 6.

b Fixed parameters, Wald test based on z statistic; random parameters, likelihood ratio test, p-value divided by 2 (see Methods, p. 35). Standard errors and CIs based on the Wald statistic throughout. Wald tests for all sets of categorical or ordinal variables and for all sets of interaction terms, p < 0.03.

c Per one unit increase.

d Compared to four weekly CMF.

e Second to 4th quartiles compared to 1st quartile.

f Compared to no concomitant radiotherapy administration.

g Parameter estimate based on 4 observations, assumed to be an artefact

h Constrained to 1 [169: 35].

Abbreviations: C, cyclophosphamide; F, 5-fluorouracil; M, methotrexate; SDI, summation dose intensity.

Step 2. Analysis of the variance structure.

Using the covariates with a potential for level 2 variation, the variance structure was analysed as described in the methods section. Significant level 2 variation was observed for the intercept and for the variable indicating the use of an anthracycline-based chemotherapy regimen. The corresponding covariance term was also significant. In contrast, the fixed effects terms representing year of treatment, year of treatment squared, and interaction between year of treatment and chemotherapy regimen type became now clearly non-significant and were removed from the model. The resulting final multilevel model is shown in chapter 7, Table 4.

Some level 1 predictors with a potential for level 2 variation were aggregated at the centre level, divided into quartiles and tentatively added to the model. None of them were significant at level 2 or interacted significantly with other covariates, but the following were significant at level 1.

- Proportion of patients aged 65 or older: the centres representing the highest quartile of this variable showed a reduced neutropenic event risk for their patients. In order to limit the number of additional model terms, a binary variable was constructed on this basis.
- Proportion of patients with a BMI ≥ 30 kg/m^2: the centres representing the second to fourth quartiles of this variable showed an increased neutropenic event risk for their patients. In order to limit the number of additional model terms, a binary variable was constructed on this basis.
- Proportion of anthracycline-based regimens used: the quartiles of this variable showed a continuously decreasing neutropenic event risk with increasing anthracyclines use. In order to limit the number of additional model terms, the continuous equivalent of this variable was used.
- Proportion of three-weekly CMF regimens used: this variable showed a complex relationship with neutropenic event occurrence. The risk was highest in the centres which did not use any three weekly CMF regimens (first and second quartiles combined). It was lower in the centres which used three-weekly CMF rarely (third quartile; three-weekly CMF used in 6.5% of patients on average) than in those who used it frequently (fourth quartile; three-weekly CMF used in 47.0% of patients on average).

As these covariates did not contribute to the analysis of higher level variation and in order to reduce the risk of over-modelling, it was decided not to use them in the final multilevel model of any neutropenic event occurrence. However, the extended model containing these covariates is shown in Table 8 and was fully assessed. Including the above-described terms reduced the AIC from 2284.63 to 2267.05. There were no massive changes of the other fixed effects coefficients or their standard errors, but increased coefficients and smaller standard errors were observed in some cases. Predictive ability was not increased. The apparent prediction error was, MSE 0.138 and classification error 19.55% in the extended model vs. MSE 0.137 and classification error 19.38% in the final model. Ten-fold cross-validation restricting the test set to observations from the level 2 units contributing to model estimation showed a similar result, with MSE 0.152 and classification error 21.17% in the extended model vs. MSE 0.152 and classification error 21.04% in the final model. If the test set was restricted to observations from the level 2 units not used for model estimation, the MSE was 0.167 and classification error 23.05% in the extended model vs. MSE 0.170 and classification error 22.44% in the final model.

Table 8. Two-level model of any neutropenic event occurrence (logistic regression); including variance and covariance terms for the intercept and use of an anthracycline-based chemotherapy regimen, and additional predictors aggregated at the centre level[a]

Number of level 1 units = 2'358 **Number of level 2 units = 93**					**Log likelihood -1101.527**	
Std. err. adjusted for 6 clusters (audits)					**AIC 2267.05**	
Variable	**Esti-mates**	**Std. Err.**	**Test stat.[b]**	**p value[b]**	**95% Confidence Interval**	
Fixed part - level 1						
Age[c]	0.025	0.007	3.49	< 0.001	0.011	0.039
Body surface area[c]	1.418	0.315	4.50	< 0.001	0.800	2.036
BMI[c]	-0.050	0.011	-4.63	< 0.001	-0.071	-0.029
Chemotherapy regimen:[d]						
Three weekly CMF	0.708	0.087	8.09	< 0.001	0.536	0.879
Anthracycline-based	0.570	0.177	3.22	0.001	0.223	0.918
Taxane-based	1.841	0.144	12.75	< 0.001	1.559	2.124
Other	1.036	0.395	2.62	0.009	0.261	1.810
Normal to high SDI[e]	0.522	0.146	3.58	< 0.001	0.236	0.808
Planned chemotherapy cycles[c]	0.353	0.098	3.62	< 0.001	0.162	0.544
Concomitant radiotherapy administration (Rx):[f]						
Rx yes	3.042	0.665	4.58	< 0.001	1.739	4.346
Rx unknown	-1.082	1.841	-0.59	0.557	-4.690	2.526
Interaction of chemotherapy regimen and Rx:						
Three weekly CMF, Rx yes	-0.399	0.433	-0.92	0.357	-1.247	0.450
Anthracycline-based, Rx yes	-0.950	0.114	-8.32	< 0.001	-1.174	-0.726
Taxane-based, Rx yes	1.524	0.246	6.18	< 0.001	1.041	2.007
Other, Rx yes	--	--	--	--	--	--
Three weekly CMF, Rx unkn.	0.830	0.440	1.89	0.059	-0.032	1.693
Anthracycline-based, Rx unkn.	0.789	0.400	1.97	0.048	0.005	1.572
Taxane-based, Rx unknown	--	--	--	--	--	--
Other, Rx unknown[g]	1.992	1.005	1.98	0.047	0.022	3.962
Interaction of body surface area and Rx:						
Rx yes	-1.248	0.207	-6.03	< 0.001	-1.654	-0.842
Rx unknown	1.659	0.687	2.41	0.016	0.312	3.006
Interaction of planned chemotherapy cycles and Rx:						
Rx yes	-0.196	0.089	-2.20	0.028	-0.371	-0.021
Rx unknown	-0.386	0.075	-5.17	< 0.001	-0.532	-0.239
Interaction of normal to high SDI and Rx:						
Rx yes	1.037	0.367	2.83	0.005	0.318	1.756
Rx unknown	-0.076	0.969	-0.08	0.938	-1.974	1.823
Intercept	-6.561	0.640	-10.25	< 0.001	-7.815	-5.307

Table 8 ctd.

Fixed part - level 2						
High proportion of patients above 65 years of age[h]	-0.598	0.195	-3.07	0.002	-0.980	-0.217
Higher proportion of patients with BMI ≥ 30 kg/m2 (lowest quartile)[e]	0.519	0.186	2.80	0.005	0.155	0.882
Proportion of patients receiving anthracycline-based chemotherapy regimens	-1.173	0.176	-6.68	0.008	-1.517	-0.829
Small proportion of patients receiving three weekly CMF regimens[i]	-0.422	0.146	-2.89	0.004	-0.709	-0.136
High proportion of patients receiving three weekly CMF regimens[i]	0.244	0.176	1.11	0.267	-0.187	0.675
Random part - level 1						
Binomial variance	1[j]	0				
Random part - level 2						
Intercept variance	0.647	0.139	50.44 (2dgf)[k]	< 0.001	0.315	0.931
Anthracycline-based chemotherapy regimen variance	0.838	0.248	20.35 (2dgf)[k]	< 0.001	0.134	1.252
Covariance	-0.627	0.120	13.28	< 0.001	-0.942	-0.206

a Conventional regression model for comparison: chapter 6, Table 3, re-written to facilitate comparison in Appendix I, Table 6.

b Fixed parameters, Wald test based on z statistic; random parameters, likelihood ratio test, p-value divided by 2 (see Methods, p. 35). Standard errors and CIs based on the Wald statistic throughout. Wald tests for all sets of categorical or ordinal variables and for all sets of interaction terms, $p < 0.05$.

c Per one unit increase.

d Compared to four weekly CMF.

e Second to 4th quartiles compared to 1st quartile.

f Compared to no concomitant radiotherapy administration.

g Parameter estimate based on 4 observations, assumed to be an artefact.

h Fourth quartile compared to 1st to 3rd quartiles.

i Third and 4th quartiles separately compared to 1st and 2nd quartiles combined. (Mean proportion of three-weekly CMF use by quartile: 1st and 2nd quartiles combined, 0.0%; 3rd quartile, 6.5%; 4th quartile, 47.0%.)

j Constrained to 1 [169: 35].

k Removing any of the level 2 random variance terms also removes the covariance term.

Abbreviations: C, cyclophosphamide; F, 5-fluorouracil; M, methotrexate; SDI, summation dose intensity.

Step 3. Goodness-of-fit.

For the final multilevel model of any neutropenic event occurrence, the plot of mean observed against mean predicted event probabilities, by deciles of the linear predictor, showed acceptable albeit not ideal properties (chapter 7, Figure 4). The two sets of studentised level 2 residuals representing random deviations from the average intercept and from the average coefficient of the variable indicating use of an anthracycline-based chemotherapy regimen showed no serious deviations from the normal distribution (chapter 7, Figure 5).

The corresponding plots for the extended model described in Table 8 are shown below in Figures 8 and 9. The plot of mean observed against mean predicted event probabilities might be interpreted as showing slightly superior properties for this model.

Figure 8. Extended multilevel model of neutropenic event occurrence - mean observed against mean predicted event probabilities, by deciles of the linear predictor

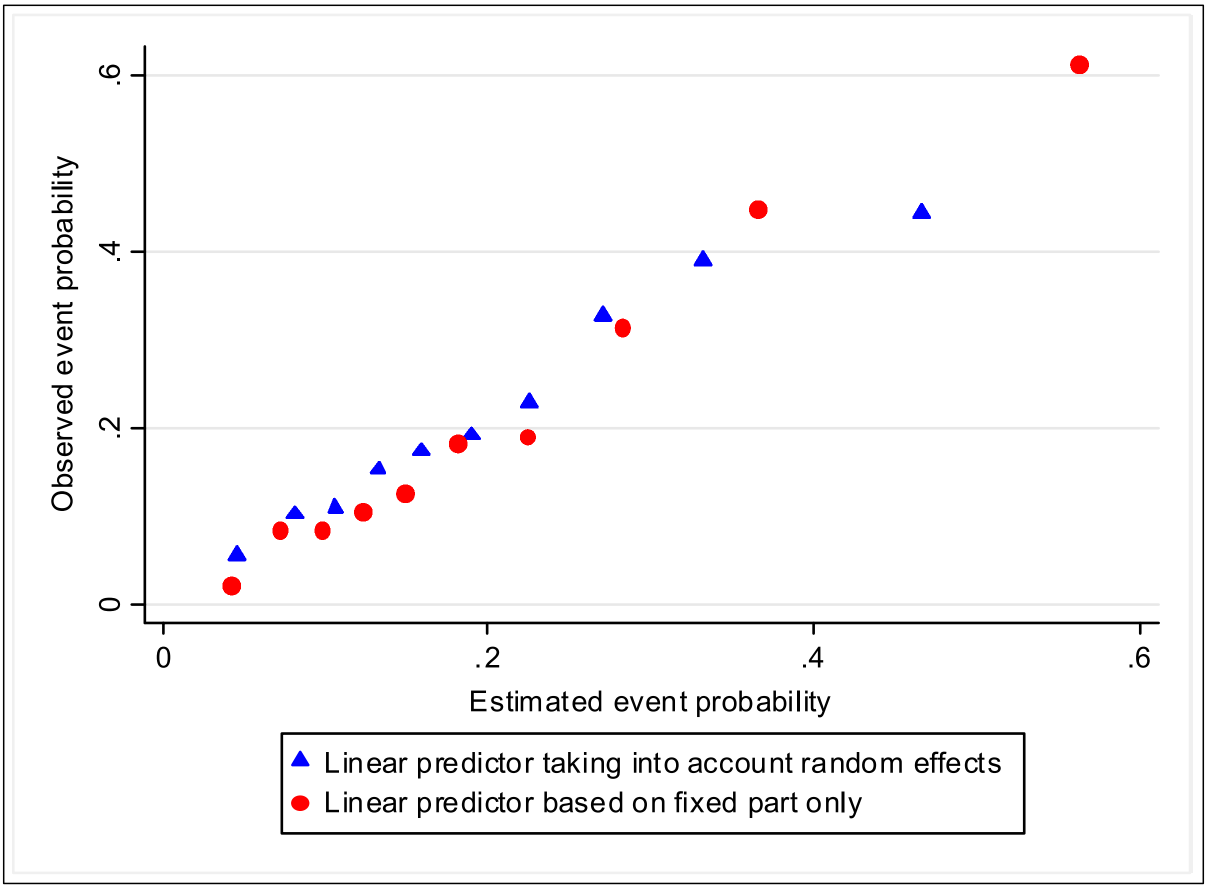
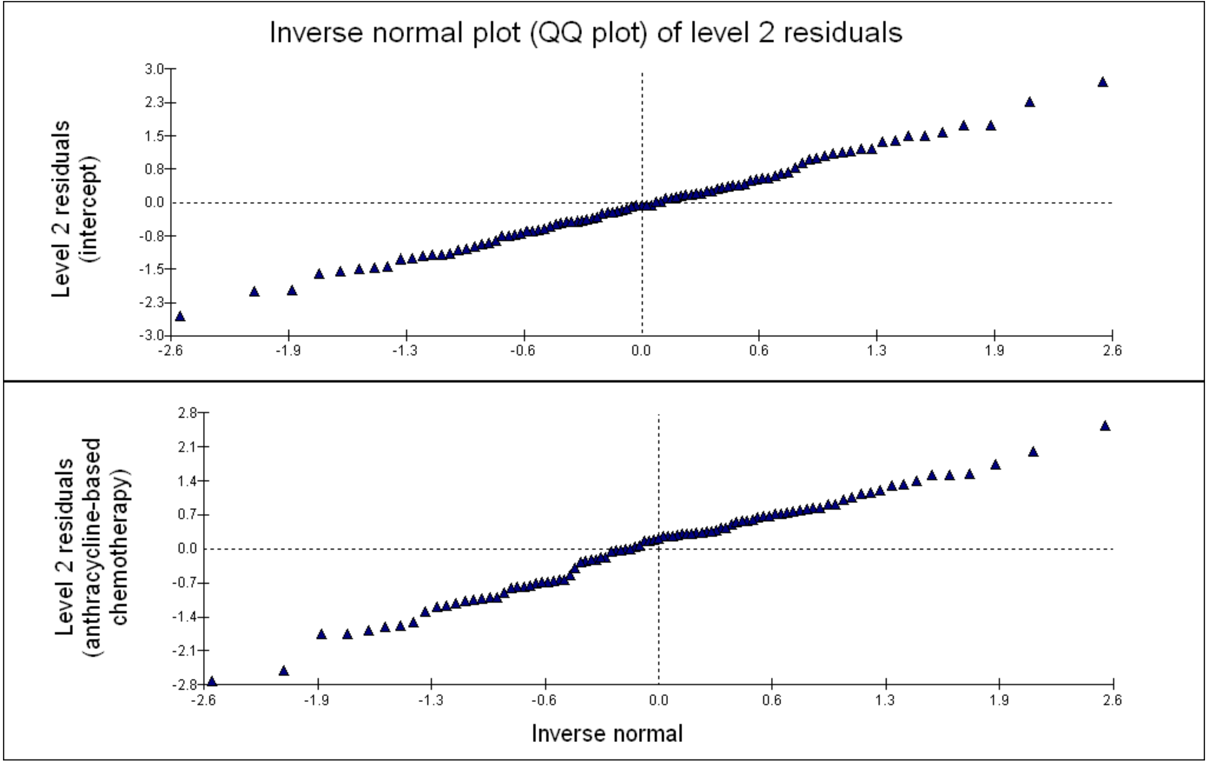


Figure 9. Extended multilevel model of neutropenic event occurrence - inverse normal plots of studentised level 2 residuals



Lifting the constraint on the residual level 1 variance and allowing it to vary at random led to an estimate of 0.939 (standard error 0.028) for the final multilevel model and to an estimate of 0.960 (standard error 0.029) for the extended multilevel model. These minor deviations from the one did not indicate the presence of any substantial extra-binomial variation. There was no hint of a misspecification of the link function or of an omission of relevant covariates or interaction terms according to this criterion [169: 35].

Step 4: Influential level 2 units.

Plots of ranked level 2 residuals and influence values hinted at two influential level 2 units in both models. Tentatively absorbing these into dummy variables did not change the other model parameters substantially or decrease the AIC in a relevant way. No changes to the final and extended multilevel models of any neutropenic event occurrence were induced. (Details not shown.)

Step 5: Final multilevel model.

The multilevel model shown in chapter 7, Table 4 is perceived as the final multilevel model of any neutropenic event occurrence resulting from this dataset.

Step 6: Model re-calculation using alternative algorithms.

Estimating the final model using Adaptive Quadrature (as implemented in Stata) or RIGLS followed 2nd order PQL (as implemented in MLwin) resulted in similar, but not identical parameter estimates. (In the tables, adaptive quadrature-based results are shown throughout.) A final re-estimation using the non-parametric bootstrap method implemented in MLwiN was planned but could not be performed due to numerical problems.

Remark on the use of robust standard error estimates

In the conventional analysis of any neutropenic event occurrence, the possibility of non-independence within the level 3 units (audits), or within the level 2 units (centres), was taken into account by estimating GEE-based robust standard errors allowing for clustering of observations. Alternatively, dummy variables were used to represent the level 3 units in the model. The changes induced by these different options were small. In most instances, using the robust method led to slightly reduced standard errors in this dataset. More substantial changes were only observed in some variables representing taxane-based or "other" chemotherapies, or interactions of these. These chemotherapy regimen types were represented by very few observations only and the validity of any related results was considered questionable from the beginning. Moreover, the dummy option did not decrease the AIC substantially.

In the multilevel models, similar observations were made. Robust standard error estimates allowing for clustering by audit tended again to be smaller, but the differences were not such that they would have affected any decisions regarding the inclusion or exclusion of model parameters. (Decisions regarding the inclusion or exclusion of variance and covariance terms were based on likelihood ratio tests, which were not influenced by the choice to use or not to use robust standard errors.)

Appendix II

Cross-validation details

This appendix addresses the implementation of ten-fold and bootstrap-based cross-validation to assess the prediction error of conventional and multilevel models. Three different situations were regarded:

(A) The assessment of predictive ability used observations from any higher level units, irrespective of whether these did or did not contribute to model estimation. This was considered to be equivalent to ignoring the hierarchical structure of the data.

(B) The assessment of predictive ability used observations whose corresponding higher level units did (through other observations) contribute to model estimation.

(C) The assessment of predictive ability used observations whose corresponding higher level units did not contribute to model estimation.

Conventional models

(A) Hierarchical structure of data ignored

In order to perform ten-fold cross-validation, the dataset was randomly split into ten equal subsets.

The parameters of the given model of interest were then estimated ten times, from training sets consisting of 90% of the observations each (always leaving out one of the ten subsets).

Each time, indicators of prediction error were calculated from the remaining 10% observations (tenth subset). The results were averaged.

The bootstrap approach was implemented analogously. e0 and .632 bootstrap estimators were used [60·61].

(B) Test set restricted to those observations whose corresponding higher level units, through other observations, contributed to model estimation

The same approach as above was used, but observations whose corresponding higher level units did not contribute to model estimation were excluded from the test sets.

Alternatively, in order to loose fewer observations from the test sets and make better use of the available information, the splitting or resampling (in the booststrap case) of the dataset was done within each higher level unit (stratified approach). (The results of this alternative approach turned out to be near-identical to those of the main approach and are not reported in detail.)

Only the e0 bootstrap estimator was used, as the correct way of deriving an .632 estimator in this situation was not clear.

(C) Test set restricted to those observations whose corresponding higher level units did not contribute to model estimation

Here, the splitting or resampling was performed on the higher level units, not on the individual observations. The training sets consisted of all observations from the selected higher level units, and the test sets consisted of all other observations.

Only the e0 bootstrap estimator was used, as the correct way of deriving an .632 estimator in this situation was not clear.

Multilevel models

(A) Hierarchical structure of data ignored

Approach not used, as the interpretation of the results would have been unclear.

(B) Test set restricted to those observations whose corresponding higher level units, through other observations, contributed to model estimation

The same approach as for the conventional models was used, but only ten-fold cross-validation was performed. Due to computation time issues, no bootstrap-based cross-validation was performed on the multilevel models (see Appendix III).

Prediction on the test sets (out of sample prediction) was based on the covariate values of the test set observations and on the random effects estimates (empirical Bayes estimates) for their corresponding higher level units as generated by the training set-based multilevel modelling process (i.e. information from the test set observations was not used to update the random effects estimates; also see Appendix III).

(C) Test set restricted to those observations whose corresponding higher level units did not contribute to model estimation

The same approach as for the conventional models was used, but again, only ten-fold cross-validation was performed.

Prediction on the test sets (out of sample prediction) was based on the covariate values of the test set observations and on values of zero for the higher level unit-specific random effects, as no informative empirical Bayes estimates were available for these higher level units not contributing to model estimation (also see Appendix III) [162].

Appendix III

Statistical software and estimation methods

All descriptive and univariate analyses, and conventional regression analyses, were performed using standard methods and algorithms as implemented in Stata® and Stata/SE®, versions 6.0-9.0 (Stata Corporation, College Station, USA). Multilevel re-analyses were performed in Stata/SE® version 9.0, and in MLwiN® version 2.02 (Multilevel Models Project, Institute of Education, London, UK).

Multilevel modelling is implemented in Stata by way of a user-written procedure named *gllamm* (abbreviating "Generalized Linear Latent and Mixed Models") and an accompanying prediction procedure named *gllapred* [161]. *gllamm* uses Gaussian Quadrature or Adaptive Quadrature and can be used with various types of response variables. (An additional Stata procedure, *xtmixed*, also allows to estimate multilevel linear regression models but cannot be used with non-continuous responses. In order to maintain consistency, only *gllamm* was used to estimate multilevel models in Stata.)

MLwiN is a highly specialized multilevel modelling software and can be used with continuous, count, binary/binomial, and survival data. Parameter estimation is based on the Iterative Generalized Least Squares (IGLS) and Restricted Iterative Generalized Least Squares (RIGLS) methods. These are combined with first and second order Marginal Quasi Likelihood (MQL), or optionally with Penalized Quasi Likelihood (PQL), in the case of non-linear responses [164: 206]. Bayesian methods (Markov Chain Monte Carlo - MCMC simulation) and bootstrap-based methods are offered as alternatives. The bootstrap solution implemented in MLwiN uses the resampling of residuals approach and parametric as well as non-parametric versions are available [34].

Both Stata's *gllamm* procedure and MLwiN have some strenghts and weaknesses, which are listed below.

Stata (*gllamm / gllapred* procedure)

Positives:

- Extremely flexible with respect to types of response variables, related distributional assumptions (distributional families) and available link functions.
- Allows to estimate GEE-based robust standard errors allowing for clustering (non-independence of observations) within higher level units. (For example, in a situation with many level 2 units and few level 3 units, it may make sense to estimate a two-level model allowing for clustering at level 3. This approach has been used in the neutropenia study.)
- Provides the log likelihood statistic for all types of responses, i.e. allows to perform likelihood ratio tests on all model parameters and in all situations.

Negatives:

- Extremely slow, particularly in the case of non-normal responses.
- No possibility to restrict individual covariance terms to zero (all or none approach). Therefore, not all MLwiN models can be re-estimated in Stata.

Remark regarding out of sample prediction:

- An "fsample" option for out of sample prediction is integrated in *gllapred*, the prediction procedure for *gllamm* [161:27-9]. However, if "fsample" is used as is, the likelihood, and subsequently the random effects estimates for the higher level units involved, are updated using the observed responses of the observations for which the prediction is to be made [162]. This behaviour makes it more difficult to assess the out of sample predictive ability of multilevel models, non-regarding if the prediction aims at observations whose corresponding higher level units did, or did not, contribute to estimating the multilevel model. For such assessments, the original random effects estimates as derived from the multilevel modelling process should be used, or values of zero for higher level units which did not contribute to model estimation [162]. The technical solution used by the author to achieve this involved use of the "us()" option with *gllapred*, which opens up a possibility to force predictions to be based on the original, or on zero, random effects estimates.

MLwiN

Positives:

- Very fast and flexible when using IGLS/RIGLS, reasonably fast when using resampling-based estimation methods (MCMC or bootstrap).
- Very flexible and efficient graphical tools for residuals and outlier assessments.

Negatives:

- No possibility to directly make a gamma distributional assumption for continuous or quasi-continuous responses. (As an alternative, quasi-continuous responses can be interpreted as discrete counts and a negative binomial distributional assumption can be made [83:7/10·141:373].
- No possibility to estimate GEE-based robust standard errors allowing for clustering by higher level units.
- No likelihood ratio tests can be performed for discrete response models (i.e., logistic and negative binomial models). The log likelihood statistic is unavailable in these cases, as the quasi-likelihood methods used for estimation are thought to produce unreliable likelihood estimates [164:113].

Taking these strengths and weaknesses into account, the following approach to multilevel modelling was chosen. For each study, the main conventional regression model(s) were reconstructed using Stata's *gllamm* procedure. The intercept terms were then allowed to vary at random, in order to assess the presence of substantial and significant higher level variation. Where such higher level variation was present, the corresponding random intercept model was reconstructed in MLwiN. The efficient IGLS/RIGLS algorithms (and, eventually, MQL/PQL algorithms) implemented in MLwiN were used to further analyse the variance structure and to perform goodness-of-fit assessments. Influential higher level units were also identified in MLwiN and related action was taken as applicable. The final IGLS/RIGLS-based multilevel models were then re-estimated using the non-parametric bootstrap method implemented in MLwiN, and using Stata's *gllamm* procedure (switching back to GEE-based standard error estimates in the case of the neutropenia study). This approach was chosen as a validity check, because the multilevel algorithms used have not yet reached the degree of maturity and reliability which one is now used to when standard non-multilevel regression techniques are applied. The *gllamm*-based results were defined as the final multilevel models.

The author used the estimation methods and settings recommended for the various types of multilevel models assessed, according to the documentations provided with the statistical packages. All *gllamm* models used Adaptive Quadrature with eight to twelve quadrature points.

In MLwiN, RIGLS and MQL were used to estimate normal response models. Logistic and negative binomial models were based on RIGLS and second order PQL, after deriving MQL-based starting values [164: 111·169: 39]. In these cases, the likelihood statistic was only available and the AIC could only be calculated for the *gllamm*-based models. Likelihood ratio tests for the variance parameters were thus postponed until the final *gllamm* models were available. In MLwiN, Wald tests were alternatively used. The results of both approaches were compared. Assessments of the ability of groups of fixed parameters (i.e., sets of predictor variables or interaction terms) to significantly improve the model were based on Wald tests only, although known to be sub-optimal [186:261], in order to keep computation time requirements within realistic limits. (NOTE: Stata uses the variant of the Wald test which is based on the standard normal distribution and z statistic, while MLwiN uses the chi squared-based equivalent [116: 647]. This lead to differences in test statistics, but not in the resulting p values.)

Non-parametric bootstrap-based parameter re-estimation for each model was based on five sets of 500 replicates per set, and bias-corrected estimates were derived (if no numerical problems occurred).

Cross-validation was programmed in Stata and as Stata's *gllamm* procedure is very slow, bootstrap-based cross-validation could not realistically be used with the multilevel models. Therefore, only ten-fold cross-validation was used with the multilevel models.

Curriculum vitae

Personal data

Name: Matthias Michael Schwenkglenks

Born on June 26, 1965 in Geislingen an der Steige, Federal Republic of Germany.

Married to Sabine Schmid. Two children, Jonathan Malte, born on May 8, 2000 and Judith Amélie, born on December 18, 2004.

German citizenship.

Education

2006: PhD in Epidemiology. Swiss Tropical Institute, University of Basel, Switzerland.

2002: Master of Public Health. Postgraduate program in Public Health, Universities of Basel, Bern and Zürich, Switzerland. Major area of study: Quantitative methodology. Thesis: Design, conduct and multivariate analysis of an observational study comparing health care costs in a gatekeeping vs. a fee-for-service plan.

1997: Magister Artium (Master of Social Sciences) in sociology and political sciences, University of Tübingen, Germany.

1997: Completed job training in intensive care nursing, University Hospital Tübingen, Germany.

1991: Completed job training in nursing, University Hospital Tübingen, Germany.

1984: 50th International Summer School, University of Exeter, UK. (British language, society, literature, and arts.)

1984: High school diploma, Michelberg-Gymnasium Geislingen an der Steige, Germany.

Additional training courses

2005: Key Issues in Drug Discovery & Development, Modules 6 (Clinical Phases I-III) and 7 (Registration and Clinical Phase IV) (Pharmacenter Basel-Zürich, University of Basel, Switzerland)

2005: Spatial Epidemiology (Erasmus Summer Programme, Erasmus Medical Centre, Rotterdam, The Netherlands; instructors: Brad Carlin, Alan Gelfand).

2005: Bayesian Analysis (Erasmus Summer Programme, Erasmus Medical Centre, Rotterdam, The Netherlands; instructors: Brad Carlin, Alan Gelfand).

2004: Survival analysis (Erasmus Summer Programme, Erasmus Medical Centre, Rotterdam, The Netherlands; instructors: David Kleinbaum, Jeff Klein).

2004: Advanced Methods in Epidemiology: Multilevel Modelling in Epidemiology and Prevention Research (Institute of Social and Preventive Medicine, University of Berne, Switzerland).

2003: Multilevel Modelling using MLwiN (University of Bristol, Bristol, UK; instructors: Harvey Goldstein; John Rasbash).

2001: Project Management; Good Clinical Practice (Pharmapart, Thalwil, Switzerland).

2000: Decision Analysis using DATA® (Division of Medical Economics, University of Zürich, Zürich, Switzerland).

Professional experience

Since August 2003: Head of Research, European Center of Pharmaceutical Medicine (ECPM), University Hospital, Basel, Switzerland (Heads of ECPM: Prof F. Bühler, Prof Th. D. Szucs).

July 2001 to July 2003: Head, Division of Medical Economics, Hirslanden Holding AG, Zürich, Switzerland (Head of Medical Department: Prof Th. D. Szucs).

April 2000 to June 2001: Research fellow, Division of Medical Economics, University Hospital Zürich, Switzerland (Head of Division: Prof. Th. D. Szucs).

April 1991 to March 2000: Nurse, internal intensive care unit, University Hospital Tübingen, Germany.

Teaching assignments

Biology programme, University of Zürich. (Topics: epidemiologic study designs; decision analysis; fundamentals of health economics; health economic evaluation of diagnostic tools.

Postgraduate programme in Public Health, Universities of Basel, Bern and Zürich, Switzerland. (Topics: decision analysis and Markov modelling; cost of illness analysis; health economic evaluation of diagnostic tools.)

Languages

German (native), English, French