Borg, Sixten et al.

Working Paper

Cost-Effectiveness and Heterogeneity: Using Finite Mixtures of Disease Activity Models to Identify and Analyze Phenotypes

Working Paper, No. 2015:5

Provided in Cooperation with:
Department of Economics, School of Economics and Management, Lund University

Suggested Citation: Borg, Sixten et al. (2015) : Cost-Effectiveness and Heterogeneity: Using Finite Mixtures of Disease Activity Models to Identify and Analyze Phenotypes, Working Paper, No. 2015:5, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/260143

Working Paper 2015:5

Department of Economics
School of Economics and Management

Cost-Effectiveness and Heterogeneity: Using Finite Mixtures of Disease Activity Models to Identify and Analyze Phenotypes

Sixten Borg
Ulf-G. Gerdtham
Tobias Rydén
Pia Munkholm
Selwyn Odes
Bjørn Moum
Reinhold Stockbrügger
Stefan Lindgren

February 2015

LUND UNIVERSITY

Cost-effectiveness and heterogeneity: Using finite mixtures of disease activity models to identify and analyze phenotypes.

Borg, Sixten* MSc (Health Economics Unit, Department of Clinical Science in Malmö, Lund University, Sweden)

Gerdtham, Ulf-G Professor, PhD (Department of Economics, Lund University, Lund, Sweden; Health Economics Unit, Faculty of Medicine, Lund University, Sweden)

Rydén, Tobias Adjunct Professor, PhD (Department of Information Technology, Uppsala University, Sweden)

Munkholm Pia, Professor, MD, DMSci (Department of Gastroenterology, Medical Section, Capital Region of Copenhagen, North Zealand University Hospital, Denmark)

Odes, Selwyn, Professor, MD, FCPSA, AGAF (Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel)

Moum, Bjørn, Professor, MD, PhD (Department of Gastroenterology, Oslo University Hospital and University of Oslo, Institute of Clinical Medicine, Oslo, Norway)

Stockbrügger, Reinhold, Professor, MD, PhD (Department of Gastroenterology and Hepatology, University Hospital Maastricht, Maastricht, Netherlands)

Lindgren, Stefan Professor, MD, PhD (Lund University, University Hospital Skåne, Malmö, Sweden, and Department of Clinical Sciences Malmö, Lund University, Sweden)

* Corresponding author. Sixten Borg. Lund University, Department of Clinical Sciences in Malmö, Health Economics Unit, Medicon Village, SE-223 81 Lund, Sweden. Phone +46 46 222 0000, Fax: +46 46-222 4720, Email: Sixten.Borg@med.lu.se

Keywords: Cost-effectiveness; Disease heterogeneity; Phenotypes; Latent classes; Disease activity model; Crohn's disease.

JEL classification: C18; D70.

Statement of funding: Unrestricted grant from Abbott Scandinavia, Sweden (Sixten Borg). The Health
Economics Program (HEP) at Lund University receives core funding from FAS (ref. 2006-1660), the Government Grant for Clinical Research (ALF), and Region Skåne (Ulf-G Gerdtham).

Disclosure of conflicts of interest: The authors declare that there are no conflicts of interest. Some authors have received honoraria for lecturing and/or advisory board participation.

Abstract

Heterogeneity in patient populations is an important issue in health economic evaluations, as the cost-effectiveness of an intervention can vary between patient subgroups, and an intervention which is not cost-effective in the overall population may be cost-effective in particular subgroups. Identifying such subgroups is of interest in the allocation of healthcare resources.

Our aim was to develop a method for cost-effectiveness analysis in heterogeneous chronic diseases, by identifying subgroups (phenotypes) directly relevant to the cost-effectiveness of an intervention, and by enabling cost-effectiveness analyses of the intervention in each of these phenotypes.

We identified phenotypes based on healthcare resource utilization, using finite mixtures of underlying disease activity models: first, an explicit disease activity model, and secondly, a model of aggregated disease activity. They differed with regards to time-dependence, level of detail, and what interventions they could evaluate. We used them for cost-effectiveness analyses of two hypothetical interventions.

Allowing for different phenotypes improved model fit, and was a key step towards dealing with heterogeneity. The cost-effectiveness of the interventions varied substantially between phenotypes. Using underlying disease activity models for identifying phenotypes as well as cost-effectiveness analysis appears both feasible and useful in that they guide the decision to introduce an intervention.

1. Introduction

Heterogeneity in patient populations is an important general issue in health economic evaluations of health interventions. The level of cost-effectiveness is often not constant across the population, but varies between different patient subgroups. An intervention which is not cost-effective in the patient population as a whole may well be cost-effective in particular subgroups, and vice versa. Thus, it is of great interest to identify such subgroups in order to optimize allocation of health care resources.

Typically, the cost-effectiveness of an intervention is evaluated in pre-specified subgroups of patients, divided for example by age group, gender, or disease severity. Such subgroups may influence the level of cost-effectiveness, and they fit well into a situation of decision-making over whether an intervention should be given to a specific patient subgroup. However, it could be that important aspects of heterogeneity are not discovered in such pre-specified subgroups. More relevant subgroups might, for instance, be described by combinations of different factors. We hypothesize that an approach to modeling disease activity may provide the means to identify such subgroups. Interventions can affect disease activity in many different ways, including avoidance of undesired events, slowing down irreversible disease progression, improvement of some function, shortening undesirable episodes, or postponing future episodes. Estimates of the relative value of interventions that differ in terms of how they affect disease activity can depend on model structure and design (Hoogendoorn, Feenstra et al. 2014). Depending on the nature of the intervention under study, different aspects may determine the level of cost-effectiveness, and so different aspects may be central for the subdivision of patients into subgroups. We will use a disease activity model designed to evaluate the intervention under study. Our subgroups will be defined on the basis of parameters to this model, and the subgroups will therefore automatically be related to the aspect of disease activity on which the intervention has an effect.

The aim of the current work was to develop a method for studying cost-effectiveness of interventions in heterogeneous chronic diseases, by subdividing the patient population into subgroups directly relevant for the cost-effectiveness of an intervention, and by enabling cost-effectiveness analysis of the intervention in each of the subgroups

In order to address the issue of heterogeneity, this article considers latent classes of patients, here denoted phenotypes. While the term “phenotype” is often associated with genetics or with localization and clinical behavior of the disease, here we identify phenotypes straightforwardly using

disease activity defined in terms of health care resource utilization which may be a relevant basis for cost-effectiveness analysis. We use a finite mixture model to identify the phenotypes. This technique is used to study populations composed of groups of patients with different behavior, by modeling the population as a mixture of components. We let a disease activity model describe each phenotype, and thus we work with a finite mixture of disease activity models. Finite mixture and similar models have been used previously. The following three examples are concerned with healthcare resource utilization in the general population. Gerdtham and Trivedi studied equity in Swedish healthcare resource utilization (Gerdtham and Trivedi 2001), distinguishing between frequent and infrequent users with a finite mixture model. They assumed that a patient's resource utilization came from a mixture of components, each with a negative binomial distribution, and then fitted a regression model with a negative binomial error term to each component on the basis of cross-sectional data. The components of Gerdtham and Trivedi correspond to our phenotypes. Deb and Holmes used a finite mixture model to distinguish between different classes of healthcare users (Deb and Holmes 2000). Like Gerdtham and Trivedi, they used cross-sectional data and assumed their observations were from a finite mixture of components with negative binomial distributions. Deb and Trivedi used a latent class model to distinguish between infrequent and frequent users (Deb and Trivedi 2002). They also used a finite mixture model with negative binomial components for annual total counts of utilization, using 3-year or 5-year periods. None of these studies attempted to develop any explicit disease activity models. Two studies quite similar to our work are those of Stull and Houghton in cardiovascular disease (Stull and Houghton 2013) and chronic obstructive pulmonary disease (Stull, Wiklund et al. 2011). These authors wanted to identify subgroups with different response to treatment. They used a linear growth model of individual longitudinal data on creatinine levels in cardiovascular disease and questionnaire scores in chronic obstructive pulmonary disease. They identified subgroups of patients which each had a common intercept and slope that was different from those in other subgroups. These subgroups correspond to our phenotypes, but neither of these studies examined the influence of phenotypes on the cost-effectiveness of an intervention.

We examine our method as applied to Crohn's disease, which is a chronic relapsing-remitting inflammatory bowel disease with heterogeneous disease course, usually requiring life-long follow-up and often also surgical interventions and/or life-long medical treatment. An intervention in Crohn's disease typically shortens relapses or prolongs remission. These ways to express effectiveness work well in a disease activity model where the transitions between relapse and remission are explicitly modeled, and where a quicker return to remission or a longer time before the next relapse can easily be parameterized. The heterogeneity in Crohn's disease has been studied previously. For instance,

Munkholm et al. described the disease in terms of its disease course by level of activity (Munkholm, Langholz et al. 1995). Solberg et al. pre-defined four different disease patterns, of which two had changing intensity over time, and asked Norwegian patients to classify themselves into one of these patterns (Solberg, Vatn et al. 2007; Solberg, Lygren et al. 2009). Odes et al. classified patients using Montreal criteria in a study of resource consumption (Odes, Vardi et al. 2007). The Montreal classification is based on disease localization, clinical behavior, and age at diagnosis (Silverberg, Satsangi et al. 2005). Montreal classification in inflammatory bowel disease was described at diagnosis in a recent inception cohort from Europe (Burisch, Pedersen et al. 2013). These previous studies did not estimate any disease models for their phenotypes, nor did they study the influence of phenotypes on the cost-effectiveness of an intervention. In summary, there is heterogeneity in disease behavior between individual patients (Sachar and Walfish 2013) and also over time (Louis, Reenaers et al. 2008), such as increasing or decreasing level of activity (Munkholm, Langholz et al. 1995; Solberg, Vatn et al. 2007). Thus it is relevant to use disease activity models that explicitly parameterize the transitions between relapse and remission, and models that can parameterize change over time.

The remainder of this article is organized as follows. We first describe the data, the aggregated nature of which presents challenges when estimating models of the relapsing-remitting disease. We then describe our method for identifying phenotypes and the two disease activity models that we use. We define two hypothetical interventions that we will use to study cost-effectiveness by phenotype. We then present results on the number of phenotypes identified and the model fit, describe the phenotypes, and examine the cost-effectiveness by phenotype. The article ends with a discussion of the results, the methodology, and the potential for wider application.

2. Data

We use data from the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD) (Shivananda, Hordijk et al. 1989; Shivananda, Lennard-Jones et al. 1996; Wolters, Russel et al. 2006). Patients in a number of countries diagnosed with Crohn's disease or ulcerative colitis in 1991-1993 were enrolled at diagnosis and followed for ten years until 2001-2003. Data were recorded as individual data on events such as relapse and surgery, aggregated over three-month periods. The relapse data comprise the number of relapses per period, with a relapse defined as a contact with a physician due to worsening of gastrointestinal symptoms that led to a new drug, an increased drug dose, or surgery. Surgery was recorded as present or absent in each relapse, and further by type of surgery, including resection, stricturoplasty, and fistulectomy (Wolters, van Zeijl et al. 2005; Hoie, Wolters et al. 2007).

Data were extracted from Crohn's disease patients in Denmark, Greece, Ireland, Israel, Italy, Netherlands, Norway, and Spain, with at least two years of follow-up. The number of relapses and the number of surgical operations were extracted and aggregated into annual data. The resulting dataset contained 3776 patient-years of observation from 380 patients (Denmark: 54 patients, 546 patient-years; Greece: 16, 126; Ireland: 29, 307; Israel: 16, 132; Italy: 54, 557; Netherlands: 75, 705; Norway: 104, 1067; and Spain: 32, 336).

3. Methods

We develop a phenotype estimator to identify the number of phenotypes that best describes the heterogeneity in the data. It is a finite mixture model based on the likelihood of an underlying disease activity model. At each moment, we consider a fixed number of different phenotypes, each with its own set of parameters to the underlying disease activity model. The likelihoods of the individual patients' data are weighted into an overall likelihood which is optimized to obtain an estimate of the global parameter vector, consisting of the disease model parameters for each phenotype and the probability of each patient belonging to each phenotype. The estimator alternately (1) determines the probabilities of belonging to the phenotypes based on the likelihoods and the model parameters of each phenotype, and (2) determines the parameter vector that best describes each of the phenotypes, as judged by likelihoods weighed by the probabilities. These two steps are iterated until a convergence criterion is reached. We identify increasing numbers of phenotypes until the number of different phenotypes that best explains the heterogeneity in the data is found (i.e. the model order). This is judged using the Bayesian Information Criterion (Hjorth

1994). The face validity of the estimated disease model parameters is judged by comparing observed disease activity to the disease activity predicted by the model. Further details on the estimator are given in the appendix.

The phenotypes are defined by their model parameters; that is, these parameter estimates provide a description of the disease activity in the phenotype. The proportion of patients belonging to each phenotype is derived from the probabilities of each patient belonging to each phenotype. The probabilities are also used as weights in descriptive statistics of the phenotypes. Thus we classify patients into phenotypes using their probabilities of belonging there.

We use two different underlying disease activity models. First, we use a previously developed Markov model of the relapsing and remitting disease (Borg, Persson et al. 2010), modified to use a separate parameter set for each phenotype. In the Markov model, the parameters are assumed to be fixed over time and it is extremely complicated to allow time-dependent parameters. Since disease activity patterns sometimes change over time, our second approach allows time-dependent parameters. Here we work directly with the aggregated data, assuming they come from a given distribution. We call this the count data model. We primarily consider the binomial distribution, but we also try the Poisson and negative binomial distributions. Our two underlying disease activity models are described below.

3.1 The Markov model

The Markov model is a simple disease activity model with four states: two for relapse and two for remission (Figure 1). The pair of relapse states represent the first month in relapse and the subsequent months in relapse, respectively. Each has a probability of remission (p_1 and p_2, respectively). The two remission states work correspondingly, with probabilities of relapse p_3 and p_4. In the relapse states, there is a probability of a surgical operation (p_5), which is handled as an event within the states. In the original model, the model parameters $\theta=(p_1, p_2, ..., p_5)$ were used to describe the entire patient population, and we estimated the parameters using an exact maximum likelihood estimator (Borg, Persson et al. 2010). Using the Markov model as the underlying disease activity model, we now estimate the model parameters for each phenotype. The mean number of relapses and surgical operations, the mean and median duration of a period of remission and the mean and median duration of a relapse were derived from the model parameters to describe the phenotypes. The parameters in this model are assumed to be constant over time, which motivates our second model with its time-dependent parameters.

3.2 The count data model

Our count data model assumes that the number of relapses is binomially distributed as Bin(6, ϕ_1) in year 1 and Bin(6, $\phi_2 + t^*\phi_3$) in the following years; that is, the probabilities depend linearly on time t. A binomial distribution with six tries was chosen because we consider an underlying process that alternates between relapse and remission, with up to six relapses per year (by analogy with the Markov model). Similarly, the model assumes that the numbers of surgical operations are distributed as Bin(6, ψ_1) and Bin(6, $\psi_2 + t^*\psi_3$), respectively. The expected annual numbers of relapses and surgical operations are derived from the model parameters to describe the phenotypes, henceforth denoted binomial phenotypes. We also try alternative distributions: first the Poisson distribution with mean number of relapses and surgical operations parameterized as μ_1 and γ_1, respectively, in year 1, and $\mu_2+t\mu_3$ and $\gamma_2+t\gamma_3$, respectively, in the following years, by analogy with the parameterization above. Then, the negative binomial distribution is parameterized with the corresponding means μ'_1 and γ'_1, respectively, in year 1, and $\mu'_2+t\mu'_3$ and $\gamma'_2+t\gamma'_3$, respectively, in the following years, and a size common for all years, σ for the number of relapses and σ' for the number of surgical operations. The results using the binomial distribution are presented explicitly. The other two distributions are used to examine the influence of the choice of distribution.

3.3 Cost-effectiveness

To illustrate the influence of phenotypes on the cost-effectiveness of interventions, we consider two hypothetical interventions: one which shortens the duration of a relapse by 25%, and one which reduces the annual number of relapses by 25%. Both have an annual cost of 500 Euros, a hypothetical value chosen to not resemble any particular drug. We use the phenotypes identified in Denmark together with Swedish cost and quality-adjusted life year (QALY) weight estimates (Mesterton, Jonsson et al. 2009). Both interventions can be evaluated using the Markov model, but only the second intervention can be evaluated using the count data model. To evaluate the first intervention, we use the Markov model with relapse as starting point, and state costs and QALY weights of 277 Euros and 0.92 in remission, and 969 Euros and 0.82 in relapse (year 2013 Euros). A five-year time frame is used. For the second intervention, we derive the cost and QALY loss of a relapse lasting two months as 1 383 Euros and 0.0167, and use these to compute total costs and QALYs of the reduced number of relapses. The incremental cost-effectiveness ratios (ICERs) in each of the phenotypes are compared to the ICER corresponding to the situation where the patient population is not subdivided into phenotypes, to explore the influence of the phenotypes. An intervention with an ICER of around 50 000 Euros or less per QALY is judged cost-effective.

4. Results

Below we identify phenotypes using different disease activity model specifications and compare the results across specifications in the area of Crohn's disease. We then examine the influence of phenotypes on cost-effectiveness.

4.1 Model order

The best model fit using the Markov model is with four phenotypes in Denmark and two in each of the other countries (Table I). Using the count data model, the best fit is with five phenotypes in Denmark[1]; four in Ireland; three in Norway, Netherlands, Israel, Spain, and Italy; and two in Greece (Table II). The count data model's ability to allow time-dependent parameters appears to result in more phenotypes in e.g. Denmark, Netherlands and Italy than with time-independent parameters (Figure 2). Especially in Denmark and Italy, phenotypes differ in how the relapse intensity changes over time, i. e. whether it increases or decreases.

4.2 Model fit

Model predictions of the mean annual number of relapses and surgical operations over ten years are presented together with weighted averages in Table III. The predictions of mean are fairly similar, but there are larger discrepancies in the number of relapses in one Markov phenotype in Denmark and in one phenotype in Italy. The predicted disease activity over time also appears similar to the weighted averages (Figure 2), with larger discrepancies for Markov phenotypes where activity varies more over time (Denmark, Italy). All these discrepancies are seen in the same two phenotypes.

4.3 Disease activity

The general pattern in the phenotypes is at least one relapse in the year of onset (year 1), followed by either zero/nearly zero relapses each subsequent year, or remaining on an active level (Figure 2). A visible time-dependence is seen in phenotypes in some countries, for instance Denmark and Netherlands.

The mean duration of a period of remission ranges from 11 to 88 months (median 1-56 months), with a shorter duration seen in more active phenotypes (Table II). The mean duration of a relapse ranges between 1 and 14 months (median 1-8), and more active phenotypes often have longer duration (Table II).

[1] The same model order and parameter estimates resulting in very similar predicted annual counts were obtained when the Poisson and negative binomial distributions were used instead of the binomial distribution.

The distributions of the annual number of relapses and surgical operations in the populations of Denmark, Norway, Netherlands, and Italy according to the Markov and count data models, respectively, are presented in Figure 3. The Markov phenotypes appear to agree well with the binomial phenotypes in Denmark and Norway, but there are larger discrepancies in Netherlands and Italy.

Without subdivision into phenotypes, the Markov model estimates the duration of a relapse as 1.1 months, the duration of a remission as 28.5 months, and the mean rate of relapses per year in Denmark as 0.56. The binomial model estimates this rate as 0.54 relapses per year.

4.4 Influence of phenotypes on cost-effectiveness

We consider two hypothetical interventions, one which shortens the duration of relapses by 25% and one which reduces the annual number of relapses by 25%. The first of these can only be evaluated in the Markov phenotypes, since relapse duration is not modeled in the count data model.

4.4.1 Shortening the duration of relapses

In the population as a whole, the ICER is 292 000 Euros per QALY gained, and so the intervention is not cost-effective. However, when analyzing by phenotype, the ICER ranges from 17 000 to 604 000 Euros per QALY gained, and the intervention is cost-effective in Markov phenotypes 2 and 4 (37% of the patients).

4.4.2 Reducing the number of relapses

In the population as a whole, the ICER is 132 000 Euros per QALY gained in the Markov model and 140 000 Euros per QALY gained in the annual counts model. Thus, both indicate that the intervention is not cost-effective. However, the ICER ranges between 32 000 and 361 000 Euros in the Markov phenotypes, and between 27 000 and 517 000 Euros in the binomial phenotypes. This makes it cost-effective in Markov phenotypes 3 and 4 (27% of the patients) and binomial phenotypes 2 and 5 (32% of the patients).

5. Discussion

In order to study cost-effectiveness in a heterogeneous disease, we separated the patients into different phenotypes. We then used two different disease activity models to determine how many phenotypes would best explain the heterogeneity of the patient population, and to estimate model parameters that describe the disease activity in these phenotypes. This was carried out in eight different countries. Under both models, two or more phenotypes improved the model fit in all

countries compared to just one phenotype. Thus the introduction of phenotypes improved our modeling of the population, regardless of our choice of model.

The data we used were a patient's number of relapses and number of surgical operations aggregated over one-year periods. We used a count data model to model these aggregated data directly. We also used a Markov model of the explicit transitions between relapse and remission, and fitting this model to these aggregated data required an indirect estimation approach. The Markov phenotypes showed higher discrepancies than their binomial counterparts. This may be due to the challenge of aggregation over time that faces the Markov estimator, leading to less precision in the estimates. Another alternative explanation may be that the Markov model is not a very good model of the disease activity, whereas the annual counts model is a fair model of the aggregated data. Moreover, the two models suggest different phenotypes. The disease phenotypes that we can identify result from the discrimination ability of the underlying disease activity model. The count data model allows probabilities that change over time, and in Denmark, for example, it identified phenotypes that differ from those identified by the Markov model (Figure 2). Still, the distribution of annual relapse and surgery rates agree well, at least in Denmark and Norway (Figure 3). The Markov model gave some rather high estimates of the mean duration of remission. These may have been inflated by ongoing remission periods at the point of censoring, and in some cases because the estimates were based on small amounts of data. Still there is mostly a fair agreement between model predictions and observations (Figure 2, Table III).

We identified different number of phenotypes in the different countries, with the highest number in Denmark. This is likely due to the different practices at the various centers, patient populations, and amounts of data available in the different countries.

We used the model parameters to evaluate whether two hypothetical interventions were cost-effective. Both interventions turned out to be cost-effective in part of the population if phenotypes were considered, whereas neither intervention would be cost-effective in the patient population as a whole. The introduction of these interventions is therefore dependent on phenotypes being considered. The identification of the phenotypes and evaluation of interventions in each phenotype can therefore guide how to optimize the use of healthcare resources.

While the work presented here has been fairly successful and potentially useful, several aspects deserve attention. The clinical relevance of our work to date is very limited, as we classify patients according to ten years of disease history. To become clinically useful, our method must provide results much sooner, for example by indicating the likely future path of the patient's disease. Ideally, this would occur close to diagnosis, but observations over some period of time would probably be needed to identify a patient's phenotype. Given the classification, one could decide on treatment according to the prognosis or the predicted future disease activity. The longer this period of time is, the more data there is for classifying the phenotype. The best length is a trade-off between the uncertainty of the classification, and the waiting time before a better-informed treatment choice can be made. Another aspect of the usefulness of our work is whether we can make good treatment choices for the phenotypes that we have identified. One could imagine that an aggressive treatment strategy would be cost-effective in patients with increasing disease activity, whereas a more modest treatment strategy would be a better choice for patients with decreasing or inactive disease activity. We found phenotypes with increasing as well as decreasing disease activity over time (Figure 2).

Our work could be transferred to other applications or other types of data; the only requirement is a likelihood function of an underlying model of the data and a reasonable amount of data. We have used longitudinal data with only two variables over ten years, but one might consider data organized differently, for example many variables at fewer points in time. If that turns out well, it could provide a phenotype classification much sooner than after ten years of follow-up, which would obviously improve the clinical usefulness. It is uncertain whether this is feasible with resource utilization data, but it may be more likely with other types of variables.

The phenotypes that we can identify are functions of the data we have and choose to use. Using data on resource consumption gives us a resource-oriented view of phenotypes which appears relevant for a health economic analysis. Others who have studied phenotypes have used data such as disease markers, genetic data, or data on clinical behavior, which give them different views of phenotypes.

Would it be just as good to pre-define subgroups of patients according to disease activity, for instance using frequency of relapse? This would require choosing suitable thresholds to discriminate between subgroups. For example, in Denmark, three Markov phenotypes were relatively active, but neither of the examined interventions were cost-effective in all three; one intervention was cost-effective in the two phenotypes with the highest relapse frequencies, and the other in the two

phenotypes with the longest relapses. Thus, it does not appear to be completely self-evident how to pre-define subgroups. Such pre-definition may have to rely on an amount of subjective judgment, whereas our method uses an objective identification algorithm (provided that the input data are objective, which in this case they are).

The patients have been given different treatments at different times. Our disease behavior observations inevitably depend on each patient's underlying disease and the patient's response to whatever treatment was given at the moment. Therefore, one might not be able to identify a treatment effect using our approach. For example, it may not be possible to discriminate between a truly inactive phenotype and a phenotype which is well controlled by an ongoing treatment but which would be active otherwise. When considering our cost-effectiveness results, it must be asked whether a new treatment would actually accomplish the effect we assume on top of inherently existing treatment in the data.

We could potentially extend the way in which our framework deals with time dependence. First, we could allow time-dependent probabilities in the Markov model. However, more parameters would then have to be estimated, so more data would be needed or the uncertainty in the estimates would increase. Moreover, the computational burden of the Markov estimator would increase. Second, we could allow patients to change phenotype over time. This might improve model fit to data from patients that do change behavior. On the other hand, fitting the additional parameters required for phenotypes to change over time would require more data and increase the computational burden. Importantly, these two aspects of time dependence may interfere. It may be hard to distinguish patients with a phenotype that changes over time from patients who move between phenotypes. Furthermore, it could make it even harder to discriminate a treatment effect from any of the above changes over time.

We have not examined the uncertainty in our results. The most straightforward approach would be a bootstrap analysis, but our estimator has very long execution times and this makes bootstrap infeasible. The lack of uncertainty analysis is nonetheless a weakness of our study.

Conclusions

Allowing for different phenotypes improves model fit, and is a key step towards dealing with heterogeneity. The cost-effectiveness of an intervention can vary between phenotypes, and identifying phenotypes and evaluating the intervention in each phenotype can guide the decision to introduce the intervention, though this is not without pitfalls.

Acknowledgements

This study was sponsored by Abbott Scandinavia AB, Solna, Sweden through an unrestricted grant. The Health Economics Program (HEP) at Lund University also receives core funding from FAS (ref. 2006-1660), the Government Grant for Clinical Research (ALF), and Region Skåne (Gerdtham). The authors gratefully acknowledge support from *the EC-IBD study group* consisting of Professor Vito Annese, Florence, Italy; Dr Marina Beltrami, Reggio Emilia, Italy; Dr Juan Clofent, Vigo, Spain; Professor Michael Friger, Beer Sheva, Israel; Professor Giovanni Fornaciari, Reggio Emilia, Italy; Dr Ebbe Langholz, Copenhagen, Denmark; Dr Monica Milla, Florence, Italy; Professor Iannis Mouzas, Heraklion, Greece; Professor Colm O'Morain, Dublin, Ireland; Dr Patricia Politi, Cremona, Italy; Dr Lene Riis, Herlev, Denmark; Professor Reinhold Stockbrugger, Maastricht, Netherlands; Professor Epameinondas Tsianos, Ioannina, Greece; Dr Hilel Vardi, Beer Sheva, Israel; and Professor Morten H Vatn, University of Oslo, Norway. The authors further wish to thank Dr Ulf Persson at the Swedish Institute for Health Economics (IHE) and Mr Thomas Lundqvist at Abbott, for valuable discussions.

References

Borg, S., U. Persson, et al. (2010). "A maximum likelihood estimator of a Markov model for disease activity in Crohn's disease and ulcerative colitis for annually aggregated partial observations." Medical Decision Making **30**(1): 132-142.

Burisch, J., N. Pedersen, et al. (2013). "East-West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom inception cohort." Gut.

Deb, P. and A. M. Holmes (2000). "Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models." Health Economics **9**(6): 475-489.

Deb, P. and P. K. Trivedi (2002). "The structure of demand for health care: latent class versus two-part models." Journal of Health Economics **21**(4): 601-625.

Gerdtham, U. G. and P. K. Trivedi (2001). "Equity in Swedish health care reconsidered: new results based on the finite mixture model." Health Economics **10**(6): 565-572.

Hjorth, J. S. U. (1994). Computer intensive statistical methods : validation model selection and bootstrap. London ; New York, Chapman & Hall.

Hoie, O., F. Wolters, et al. (2007). "Ulcerative colitis: patient characteristics may predict 10-yr disease recurrence in a European-wide population-based cohort." The American journal of gastroenterology **102**(8): 1692-1701.

Hoogendoorn, M., T. L. Feenstra, et al. (2014). "Cost-Effectiveness Models for Chronic Obstructive Pulmonary Disease: Cross-Model Comparison of Hypothetical Treatment Scenarios." Value in Health.

Louis, E., C. Reenaers, et al. (2008). "Does the behavior of Crohn's disease change over time?" Inflammatory Bowel Diseases **14 Suppl 2**: S54-55.

Mesterton, J., L. Jonsson, et al. (2009). "Resource use and societal costs for Crohn's disease in Sweden." Inflamm Bowel Dis **15**(12): 1882-1890.

Munkholm, P., E. Langholz, et al. (1995). "Disease activity courses in a regional cohort of Crohn's disease patients." Scand J Gastroenterol **30**(7): 699-706.

Odes, S., H. Vardi, et al. (2007). "Effect of phenotype on health care costs in Crohn's disease: A European study using the Montreal classification." Journal of Crohn's & colitis **1**(2): 87-96.

Sachar, D. B. and A. Walfish (2013). "Inflammatory bowel disease: one or two diseases?" Current gastroenterology reports **15**(1): 298.

Shivananda, S., M. L. Hordijk, et al. (1989). "Proceedings of the first European community workshop on inflammatory bowel disease " Scandinavian Journal of Gastroenterology **24**(Suppl 170): 1-101.

Shivananda, S., J. Lennard-Jones, et al. (1996). "Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD)." Gut **39**(5): 690-697.

Silverberg, M. S., J. Satsangi, et al. (2005). "Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology." Canadian journal of gastroenterology = Journal canadien de gastroenterologie **19 Suppl A**: 5A-36A.

Solberg, I. C., I. Lygren, et al. (2009). "Clinical course during the first 10 years of ulcerative colitis: results from a population-based inception cohort (IBSEN Study)." Scandinavian Journal of Gastroenterology **44**(4): 431-440.

Solberg, I. C., M. H. Vatn, et al. (2007). "Clinical course in Crohn's disease: results of a Norwegian population-based ten-year follow-up study." Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association **5**(12): 1430-1438.

Stull, D. E. and K. Houghton (2013). "Identifying differential responders and their characteristics in clinical trials: innovative methods for analyzing longitudinal data." Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research **16**(1): 164-176.

Stull, D. E., I. Wiklund, et al. (2011). "Application of latent growth and growth mixture modeling to identify and characterize differential responders to treatment for COPD." Contemporary clinical trials **32**(6): 818-828.

Wolters, F. L., M. G. Russel, et al. (2006). "Disease outcome of inflammatory bowel disease patients: general outline of a Europe-wide population-based 10-year clinical follow-up study." Scand J Gastroenterol Suppl(243): 46-54.

Wolters, F. L., G. van Zeijl, et al. (2005). "Internet-based data inclusion in a population-based European collaborative follow-up study of inflammatory bowel disease patients: description of methods used and analysis of factors influencing response rates." World journal of gastroenterology : WJG **11**(45): 7152-7158.

Tables and Figures

Table I: Model parameter estimates of the identified Markov phenotypes, and duration of remission and relapse, by country.

Country	Phenotype (proportion of patients)	p_1	p_2	p_3	p_4	p_5	Duration of remission, mean (median)	Duration of relapse, mean (median)
Denmark N=54	M1 (49%)	0.9900	0.9900	0.2473	0.0100	0.3796	76.3 (44.5)	1.0 (1.0)
	M2 (24%)	0.3223	0.1840	0.5307	0.0205	0.0272	23.9 (1.0)	4.7 (1.0)
	M3 (14%)	0.9625	0.9900	0.7228	0.0161	0.5767	18.2 (1.0)	1.0 (2.0)
	M4 (13%)	0.2224	0.8075	0.4713	0.0521	0.1804	11.1 (1.0)	2.0 (1.0)
Greece N=16	M1 (82%)	0.9900	0.9900	0.1316	0.0100	0.0813	87.8 (56.0)	1.0 (1.0)
	M2 (18%)	0.9900	0.7875	0.0100	0.0393	0.0100	26.2 (16.0)	1.0 (1.0)
Ireland N=29	M1 (59%)	0.0100	0.9653	0.2671	0.0100	0.0374	74.3 (40.0)	2.0 (1.0)
	M2 (41%)	0.5771	0.9304	0.4484	0.0453	0.2033	13.2 (2.0)	1.5 (1.0)
Israel N=16	M1 (50%)	0.9900	0.9900	0.2091	0.0436	0.0100	19.1 (11.0)	1.0 (1.0)
	M2 (50%)	0.0100	0.1959	0.4340	0.0100	0.0506	57.6 (12.0)	6.1 (4.0)
Italy N=54	M1 (92%)	0.0618	0.9152	0.1702	0.0142	0.0983	59.6 (33.0)	2.0 (1.0)
	M2 (8%)	0.1859	0.0626	0.7562	0.0100	0.0199	25.4 (1.0)	14.0 (8.0)
Netherlands N=75	M1 (55%)	0.9900	0.9900	0.2357	0.0106	0.4508	73.1 (42.0)	1.0 (1.0)
	M2 (45%)	0.6129	0.4236	0.1699	0.0325	0.0334	26.6 (15.0)	1.9 (1.0)
Norway N=104	M1 (86%)	0.0100	0.9900	0.1365	0.0100	0.1123	87.4 (55.5)	2.0 (1.0)
	M2 (14%)	0.0100	0.8678	0.1735	0.0393	0.0438	22.1 (13.0)	2.1 (1.0)
Spain N=32	M1 (92%)	0.0100	0.4236	0.2347	0.0123	0.0344	62.8 (37.0)	3.3 (2.0)
	M2 (8%)	0.9900	0.9900	0.0695	0.0708	0.1311	14.1 (9.0)	1.0 (1.0)

Note: The parameters p_1, p_2, ..., p_5 are transition probabilities (see Figure 1).

Table II: Model parameter estimates of the identified binomial phenotypes, by country.

Country	Phenotype (proportion of patients)	Probability *p* of relapse in annual count, Bin(6,*p*)			Probability *p* of surgery in annual count, Bin(6,*p*)		
		Year 1	Year 2	Year 10	Year 1	Year 2	Year 10
Denmark N=54	B1 (39%)	0.2628	0.0190	0.0052	0.1097	0.0000	0.0000
	B2 (24%)	0.4169	0.2665	0.0727	0.1628	0.1060	0.0289
	B3 (16%)	0.1914	0.1278	0.0349	0.0000	0.0293	0.0080
	B4 (12%)	0.5611	0.0005	0.0391	0.4275	0.0000	0.0133
	B5 (8%)	0.2405	0.0000	0.2724	0.0000	0.0000	0.0149
Greece N=16	B1 (52%)	0.2190	0.0000	0.0000	0.0201	0.0000	0.0000
	B2 (48%)	0.1970	0.0525	0.0468	0.0000	0.0000	0.0050
Ireland N=29	B1 (54%)	0.2718	0.0110	0.0231	0.0312	0.0000	0.0000
	B2 (24%)	0.3571	0.2200	0.0600	0.0234	0.0387	0.0106
	B3 (16%)	0.5470	0.0450	0.1079	0.1152	0.0000	0.0228
	B4 (7%)	0.6561	0.3069	0.1240	0.3228	0.2087	0.0569
Israel N=16	B1 (54%)	0.3560	0.1133	0.0599	0.0000	0.0000	0.0000
	B2 (39%)	0.2220	0.0085	0.0023	0.0526	0.0000	0.0000
	B3 (7%)	0.5021	0.1982	0.1338	0.0000	0.0968	0.0464
Italy N=54	B1 (52%)	0.2463	0.0941	0.0466	0.0386	0.0179	0.0097
	B2 (38%)	0.2159	0.0120	0.0033	0.0287	0.0000	0.0000
	B3 (10%)	0.2113	0.0000	0.0738	0.0929	0.0000	0.0000
Netherlands N=75	B1 (50%)	0.2667	0.0762	0.0266	0.0320	0.0097	0.0035
	B2 (26%)	0.2602	0.1345	0.0367	0.0918	0.0373	0.0102
	B3 (24%)	0.1808	0.0000	0.0007	0.0767	0.0000	0.0007
Norway N=104	B1 (45%)	0.2023	0.0309	0.0084	0.0140	0.0040	0.0011
	B2 (28%)	0.2465	0.0693	0.0811	0.0118	0.0118	0.0212
	B3 (27%)	0.2166	0.0000	0.0092	0.0674	0.0000	0.0000
Spain N=32	B1 (43%)	0.3149	0.0686	0.0187	0.0125	0.0130	0.0035
	B2 (41%)	0.1974	0.0000	0.0245	0.0122	0.0000	0.0000
	B3 (16%)	0.2083	0.1930	0.0658	0.0000	0.0000	0.0309

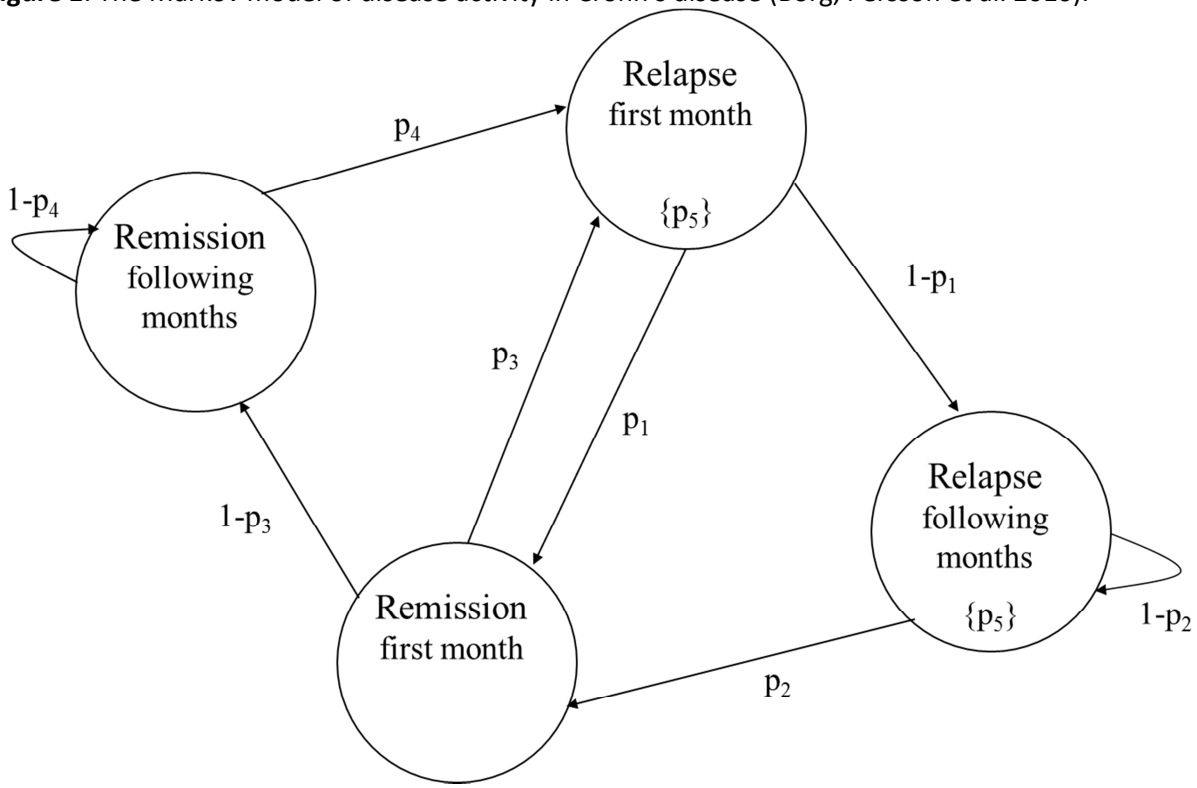
Note: Our model assumes a linear relationship between probabilities in years 2 through 10.

Table III: Average annual number of relapses and surgical operations in ten years, predicted by the model (Model), and weighted observations (Observed)* (Markov, M, and binomial, B).

Country			Markov phenotypes				Binomial phenotypes				
			M1	M2	M3	M4	B1	B2	B3	B4	B5
Denmark	Relapses	Model	0.27	0.56	0.89	1.04	0.20	1.09	0.52	0.43	0.97
		Observed	0.27	0.70	0.90	1.10	0.21	1.09	0.51	0.47	0.99
	Surgery	Model	0.07	0.07	0.48	0.35	0.06	0.43	0.09	0.27	0.05
		Observed	0.10	0.06	0.52	0.36	0.06	0.43	0.09	0.29	0.02
Greece	Relapses	Model	0.24	0.53	---	---	0.12	0.38	---	---	---
		Observed	0.20	0.49	---	---	0.12	0.38	---	---	---
	Surgery	Model	0.01	0.00	---	---	0.01	0.02	---	---	---
		Observed	0.02	0.00	---	---	0.01	0.02	---	---	---
Ireland	Relapses	Model	0.28	0.94	---	---	0.25	0.90	0.74	1.47	---
		Observed	0.30	0.98	---	---	0.26	0.88	0.73	1.54	---
	Surgery	Model	0.02	0.26	---	---	0.02	0.14	0.13	0.85	---
		Observed	0.02	0.28	---	---	0.02	0.14	0.13	0.85	---
Israel	Relapses	Model	0.70	0.32	---	---	0.65	0.15	1.16	---	---
		Observed	0.62	0.34	---	---	0.62	0.14	1.09	---	---
	Surgery	Model	0.01	0.09	---	---	0.00	0.03	0.37	---	---
		Observed	0.00	0.08	---	---	0.00	0.03	0.35	---	---
Italy	Relapses	Model	0.30	0.48	---	---	0.50	0.16	0.34	---	---
		Observed	0.32	0.85	---	---	0.50	0.16	0.36	---	---
	Surgery	Model	0.05	0.12	---	---	0.09	0.02	0.05	---	---
		Observed	0.05	0.13	---	---	0.09	0.02	0.05	---	---
Nether-lands	Relapses	Model	0.27	0.52	---	---	0.41	0.58	0.10	---	---
		Observed	0.27	0.52	---	---	0.41	0.57	0.10	---	---
	Surgery	Model	0.09	0.03	---	---	0.05	0.17	0.04	---	---
		Observed	0.12	0.03	---	---	0.05	0.17	0.04	---	---
Norway	Relapses	Model	0.24	0.59	---	---	0.21	0.55	0.15	---	---
		Observed	0.23	0.63	---	---	0.21	0.55	0.15	---	---
	Surgery	Model	0.04	0.05	---	---	0.02	0.10	0.04	---	---
		Observed	0.04	0.06	---	---	0.02	0.10	0.04	---	---
Spain	Relapses	Model	0.29	0.87	---	---	0.39	0.18	0.78	---	---
		Observed	0.32	0.87	---	---	0.39	0.18	0.78	---	---
	Surgery	Model	0.03	0.11	---	---	0.05	0.01	0.09	---	---
		Observed	0.03	0.12	---	---	0.05	0.01	0.10	---	---

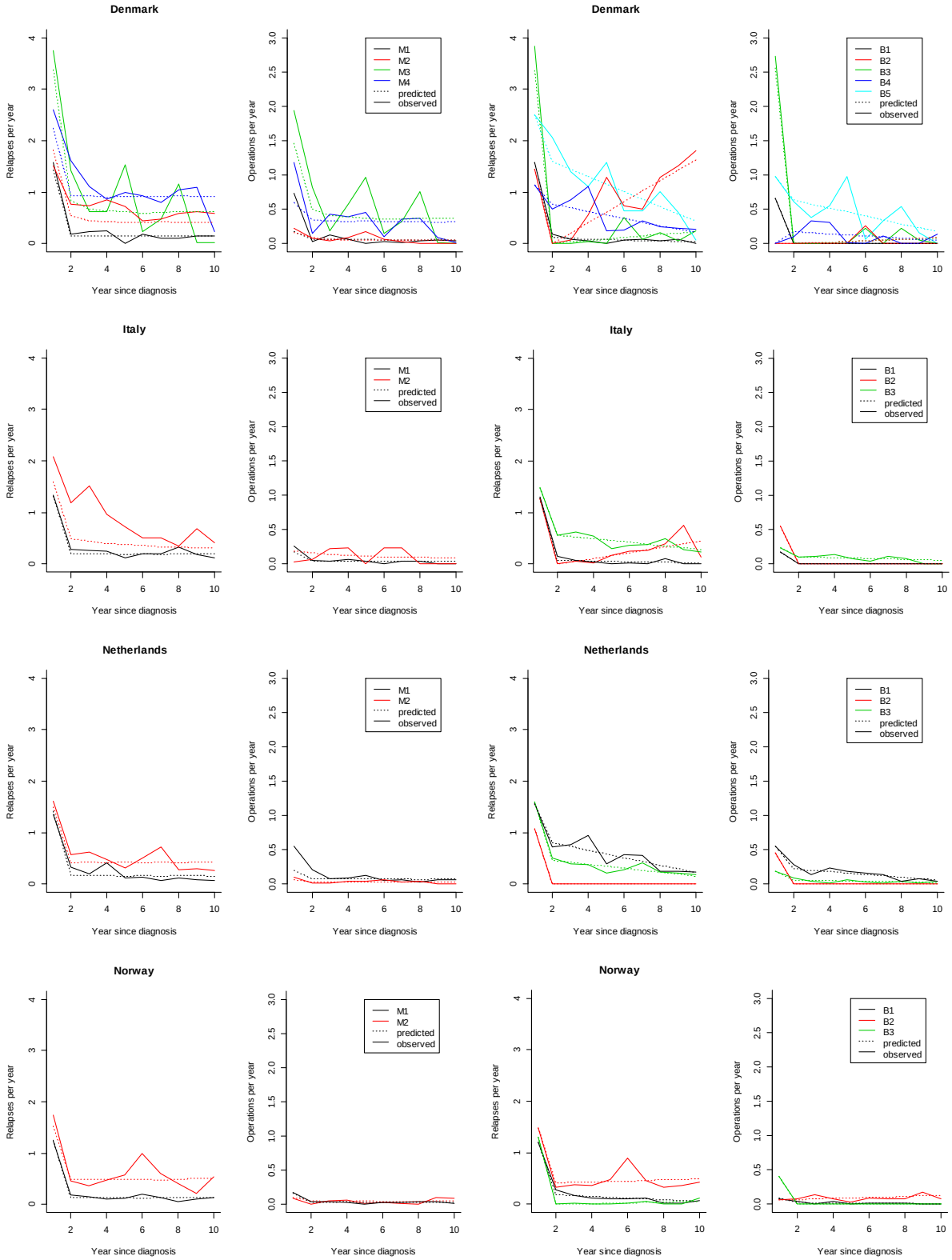
Notes: --- = phenotype not identified. * Weighted according to individual probabilities of belonging to each phenotype

Figure 1: The Markov model of disease activity in Crohn's disease (Borg, Persson et al. 2010).



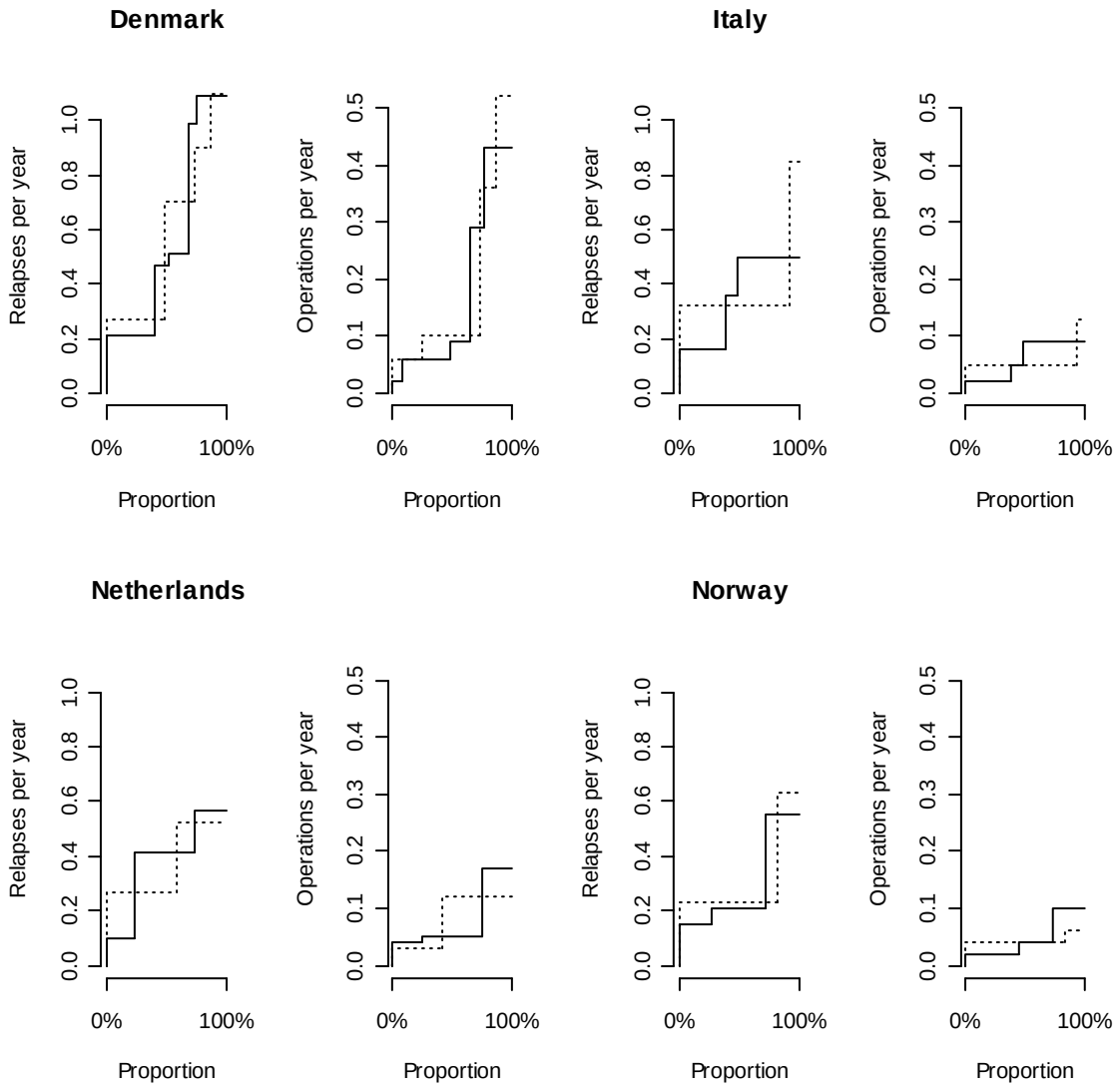
There is a probability of surgery (p_5) in the two states representing relapse.

Figure 2: Observed and predicted annual number of relapses and surgical operations, by country and Markov (M) and binomial (B) phenotype.



Notes: The observed quantities are shown as weighted averages, with individual phenotype probabilities used as weights. Therefore the observed quantities shown here depend on which kind of phenotype is used for weighting (Markov/binomial).

Figure 3: Distribution of the annual number of relapses and surgical operations in the populations of four countries, estimated using Markov phenotypes (dotted line) and binomial phenotypes (solid line).



Notes: Cumulative proportion of patients.

Appendix

3 November 2014

1 Data

We observe an individual i for n_i years and obtain the annual number of surgical operations $Z_i = (Z_{i_1}, Z_{i_2}, \ldots, Z_{in_i})$, and the annual number of relapses $V_i = (V_{i_1}, V_{i_2}, \ldots, V_{in_i})$, during each observed year $j = 1, 2, \ldots, n_i$. We denote the set of data from an individual as $y_i = (Z_i, V_i)$.

2 The Phenotype estimator

We consider r different phenotypes. For each phenotype $u = 1, 2, \ldots, r$, we have the probability α_u of a patient belonging to the phenotype, and the underlying disease activity model parameters θ_u. We use $\kappa = (\alpha_1, \alpha_2, \ldots, \alpha_r, \theta_1, \theta_2, \ldots, \theta_r)$ to denote the entire parameter vector, and y_i to denote the data from patient i. The likelihood for an individual i is $L_i(\theta) = \Pr\{Z_i = z_i, V_i = v_i|\theta\}$ The likelihood for phenotype u is $L_F(\theta_u) = \prod_i L_i(\theta_u)^{z_{iu}}$. The likelihood $L_i(y_i; \theta_u)$ of each individual i is weighted according to its probability z_{iu} of belonging to phenotype u.

We use an Expectation-Maximization algorithm with an initial starting point (κ) generated randomly. The algorithm then iterates a number of steps for a prespecified number of times (default 20).

In the first step, the probabilities w_{ij} of patient i belonging to phenotype u are determined as

$$w_{iu} = \alpha_u * L_i(y_i; \theta_u) / [\sum_{k=1}^{r} \alpha_k * L_i(y_i; \theta_k)]$$

In the second step, for each phenotype u with $z_{iu} = w_{iu}$, $L_F(\theta_u)$ is maximized with regards to θ_u. We denote the updated parameter vector that results in a new maximum, with θ'_u.

Finally, new probabilities α'_j are derived as $\alpha'_j = W_j / \sum_{k=1}^{r} W_k$, where $W_j = \sum_{i=1}^{n} w_{ij}$. The updated overall parameter set, $\kappa' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_r, \theta'_1, \theta'_2, \ldots, \theta'_r)$ is obtained and we carry on the iteration by setting $\kappa = \kappa'$, unless the difference is small, i.e. if $|\kappa' - \kappa| < \epsilon$ (default $\epsilon = 10^{-8}$), in which case the iteration is terminated. A pre-defined number of starting points (default 50) are used, and the resulting estimate with the highest overall likelihood $L_r = \prod_{u=1}^{r} L_F(\theta_u)$ is taken as the Maximum Likelihood estimate.

2.1 The Markov model

We used a previously developed Markov model, where the number of relapses and the number of surgical operations are determined from an individual patient's pathway through the model. The likelihood $L_i(\theta)$ is a function of the disease activity model parameters $\theta = (p_1, p_2, \ldots, p_5)$, and it is computed using an exact likelihood estimator [1].

2.2 The Count data model

The aggregated number of relapses, and the number of surgical operations, in any given year is modelled as a stochastic variable of a given distribution. Our main choice was the binomial distribution, $Bin(6, p)$. For the number of relapses: $p = \phi_1$ in year 1, $\phi_2 + t\phi_3$ in years $t = 2, 3, \ldots$ This imposes obvious boundaries on $\phi_1 \in [0, 1]$ and on ϕ_2, ϕ_3 so that $\phi_2 + t\phi_3 \in [0, 1]$. For the number of surgeries, $p = \psi_1$ in year 1, $\psi_2 + t\psi_3$ in years $t = 2, 3, \ldots$

We also considered the Poisson and negative binomial distribution. For the Poisson distribution the mean number of relapses was μ_1 and surgical operations γ_1, year 1; and corresponding, relapses, $\mu_2 + \mu_3 t$ and surgical operations $\gamma_2 + \gamma_3 t$, year $t = 2, 3, \ldots$ For the negative binomial distribution the mean number of relapses was μ'_1 and surgical operations was γ'_1, year 1; and corresponding, relapses, $\mu'_2 + \mu'_3 t$ and surgical operations $\gamma'_2 + \gamma'_3 t$, year $t = 2, 3, \ldots$ and a size common for all years, σ for relapses and σ' for surgical operations.

With this approach, we denote the parameter of phenotype u with $\Phi_u = (\phi_1, \phi_2, \phi_3, \psi_1, \psi_2, \psi_3)$ so that we have the likelihood of each individual's data, $L'_i(y_i; \Phi_u)$ based on the binomial probability function according to above. The likelihood for the phenotype estimator was obtained by substituting $L_i(y_i; \theta_u)$ with $L'_i(y_i; \Phi_u)$ and θ with Φ in $L_F(\theta_u)$ wherever relevant. The estimator using the Poisson or negative binomial distributions operates in analogy, except for bounds for parameters being different in the numerical searches.

3 Model order

The model order is judged using the Bayesian Information Criterion, $BIC_r = -2 \ln L_r + m \ln n$, where L_r is the maximum likelihood (among the ones we have obtained), m is the number of parameters with r phenotypes ($m = r - 1 + 5r$ with the Markov model, $m = r - 1 + 6r$ with the Count data model), and n is the number of observed patient-years. We prefer a model with $r + 1$ phenotypes over a model with r phenotypes, if $BIC_{r+1} < BIC_r$.

References

[1] Borg S, Persson U, et al, A Maximum Likelihood Estimator of a Markov Model for Disease Activity in Crohn's Disease and Ulcerative Colitis for Annually Aggregated Partial Observations. Med Decis Making 2010;30(1) 132-142.