4-21-2017 12:00 AM

Development of Clinical Prediction Models for Surgery and Complications in Crohn's Disease

Leonard M. Guizzetti, *The University of Western Ontario*

Supervisor: Dr. Guangyong Zou, *The University of Western Ontario*
Joint Supervisor: Dr. Brian Feagan, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

Background and Objective: Crohn's disease (CD)-related complications account for a substantial proportion of IBD-related healthcare expenditure. Identifying patients at risk for complications may allow for targeted use of early therapeutic interventions to alter this natural course. The objective of this project was to develop risk prediction models of CD-related surgery and complications.

Methods: Using data from the REACT cluster-randomized clinical trial (N=1898 from 41 community practices), prediction models were developed and internally validated for CD-related surgery and CD-related complications, defined as the first CD-related surgery, hospitalization or complication within 24 months. Performance of each model was assessed in terms of discrimination and calibration, as well as decision curves and net benefit analyses.

Results: There were 130 (6.8%) CD-related surgeries and 504 (26.6%) CD-related complications during the 24-month follow-up period. Selected baseline predictors for predicting surgery included age, gender, disease location, HBI score, stool frequency, antimetabolite or 5-aminosalicylates use, and the presence of a fistula, abscess or abdominal mass. Selected predictors of complications included those same factors for surgery, corticosteroid and TNF-antagonist use and excluded 5-aminosalicylate use. The discrimination ability, as measured by optimism-corrected c-statistic, was 0.70 for the surgery model, and 0.62 for the complication model. Score charts and nomograms were developed to facilitate future risk score calculation.

Conclusions: Risk prediction models for CD-related surgery and CD-related complications were developed using clinical trial data involving community gastroenterology practices. These models need to be externally validated before being used to guide management of CD.

Keywords: prognostic model, Crohn's disease, surgery, complications, risk estimate

Dedication

To my wife, Heather, and our beloved Jack Russell ~~Terrier~~ Terrorist, Sally.

Acknowledgments

I am especially grateful to my advisors, Dr. Guangyong (GY) Zou and Dr. Brian Feagan, for giving me their guidance and the opportunity to work with so many wonderful people at Robarts Clinical Trials Inc. I thank Dr. Reena Khanna for her guidance and assistance. I thank Dr. Vipul Jairath for his generosity and guidance. I also thank Mr. Larry Stitt and Ms. Claire Parker for their assistance.

I have met many exceptional and talented colleagues while in the Epidemiology and Biostatistics program. Thank you for your friendship and making our time at Western a fun and productive environment.

Lastly, I thank Mr. Digby Isnor. He constantly reminded me to take delight in the small things in life and to spend more time outdoors.

Statement of Co-Author Contributions

Chapters 1 and 2 are wholly the work of **Dr. Guizzetti**. In Chapter 3, **Dr. Guizzetti** developed and validated the prediction models, performed the data analysis, prepared the figures and wrote the thesis. **Dr. Zou, Dr. Feagan, Dr. Khanna and Dr. Jairath** edited the manuscript form of the results presented here.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC, Akaike information criterion
ASCA, anti-*Saccharomyces cerivisiae* mannan antibodies
CD, Crohn's disease
CRP, C-reactive protein
ECI, early-combined immunosuppression
EPV, events per variable
FC, Fecal calprotectin
FN, false negative
FP, false positive
FPR, false-positive rate
GEE, generalized estimating equations
HBI, Harvey-Bradshaw index
IBD, inflammatory bowel disease
ICC, intra-class cluster correlation
IL23R, interleukin-23 receptor
NB, net benefit
NPV, negative predictive value
NOD2, nucleotide-binding oligomerization domain-containing protein 2
PPV, positive predictive value
REACT, randomized evaluation of an algorithm for Crohn's disease treatment
SL, stool lactorferrin
TN, true negative
TP, true positive
TPR, true positive rate
UC, ulcerative colitis

Chapter 1

Introduction to prediction models

1.1 Introduction

This thesis concerns the development of prediction models for the risk of Crohn's disease (CD)-related surgery and CD-related complications. Crohn's disease is an idiopathic disorder of the gastrointestinal tract characterized by chronic, segmental inflammation. While the symptoms of CD typically flare up between periods of remission, many patients will have persistent sub-clinical inflammation of the bowel despite the use of maintenance drug therapy [1]. This inflammatory process predisposes patients to complications of stricture and fistula formation. Eventually, surgery will be required for up to 80% of patients [2]. These complications may also recur. Surgery is also a risk factor for disease recurrence [3, 4] and future disability [5, 6]. Thus, surgery is a clinically relevant outcome to patients.

The objective of this thesis is to develop separate models for predicting the risk that a patient will experience either a CD-related surgery and CD-related complications within the proceeding two years of follow up.

This chapter provides a review, within the context of CD, of the general methodology to develop and validate clinical prediction models.

1.2 Crohn's disease

Crohn's disease is one of the two major subtypes of inflammatory bowel disease (IBD), along with ulcerative colitis. Crohn's disease is characterized by periods of active inflammation and symptom flare ups that are separated by periods of quiescence or low disease activity. Typical symptoms experienced in CD include abdominal pain and cramping, (frequent) diarrhea, blood in the stool, fever, fatigue, reduced appetite and weight loss. The inflammation in Crohn's disease is transmural, affecting the entire thickness of the intestinal wall, and even in asymptomatic periods, subclinical transmural inflammation may persist.

1.3 What are prediction models?

Over the past decade, approximately 3,000 to 4,000 clinical prediction models (sometimes called clinical prediction rules, clinical decision rules or prediction models) have been described in the biomedical literature based on a search of PubMed for the MeSH heading "clinical prediction model." A prediction model can be used to assign a patient an expected risk of receiving a particular diagnosis or prognosis or suggest a therapeutic or diagnostic course of action. One of the most well-known prediction models is the Framingham Heart Study coronary heart disease risk model, which predicts the ten-year risk of developing heart disease [7]. An example of a decision prediction rule is the Ottawa Angle Rule which determines whether a patient should undergo radiography for ankle injuries [8]. The ultimate goal of the prediction model is to inform decisions about patient care. When existing data are not available from which to develop a prediction

model, they must be collected from a study specific to this task, such a retrospective chart review, prospective cohort or even a clinical trial. If data are to be collected from a proposed new study, this collection of patient data (especially from tests and procedures) can be expensive. Therefore, it is essential that prediction models be well developed and fully validated.

Development of prediction models is especially useful to guide patient care towards the goal of personalized medicine. With personalized medicine, an algorithm (or model) has the ability to risk stratify a patient and appropriately tailor their subsequent therapy. Personalized medicine in IBD currently plays a limited role [9]. A clinician may identify those patients at high risk of rapid progression to a complication and then must select, monitor and adapt the most appropriate therapeutic strategies in order to achieve an appropriate endpoint for that patient. Numerous clinical, genetic and immunological items have already been identified as significant prognostic factors, to be discussed in the next chapter. However, these have generally demonstrated heterogeneous prognostic ability in prospective studies. Part of this variability is due to the choice of endpoint: a symptom-based endpoint, a target disease activity score, or mucosal healing and deep remission. The former is significantly limited by its poor association with objective markers of inflammation [10]. While the latter is still debated as being the optimal endpoint of healing and resolution of inflammation, clinical trials of anti-TNF and anti-integrin biologic agents have achieved mucosal healing and durable remission [11–17]. Mucosal healing is associated with lower rates of inflammation, requirement of lower cumulative dose of corticosteroids and lower rates of both intestinal surgery and hospitalization [18, 19]. Mucosal healing appears to be the more objective and reliable treatment target for IBD.

Using disease-based activity scores, those patients with low clinical activity may still experience sub-clinical inflammation. Thus, they may be under-treated and face increased risk of progression to more complicated disease behaviour or development of related complications. If mucosal healing is the target endpoint, then a subset of patients without clinical symptoms, but who have failed to reach endoscopic remission, would require es-

calation to more intensive or effective therapies. However, this strategy may impose unnecessary over-treatment for some. Whatever the choice of endpoint, a well-constructed and validated prediction model can be a key part of risk stratifying patients while aiding the clinician's judgment with a patient's predicted risk.

The development of a risk prediction model may proceed in four general steps: model development, internal validation, external validation and optional model updating. Multivariable logistic regression models are usually constructed for predicting binary outcomes while proportional hazard regression models are commonly used for predicting the time to an event. The remainder of this section is a summary of the consensus approach to developing and validating prediction models that has been advocated for in the biostatistics literature [20–25].

1.4 Model development

Before expending the effort to develop a model, the authors must ask themselves if this model will be used, and who are the potential users. While there are thousands of examples of prediction models in the biomedical literature, most of them have failed to be translated into clinical settings. Models fail for various reasons, such as those listed by Reilly *et al.* [26] and Laupacus *et al.* [27]. Important obstacles to application include: (1) the potential users did not trust the variables, relationships or weights used in the model; (2) the variables necessary to make predictions would not be routinely available; or (3) the setting envisioned by the author for prediction is not relevant. If the authors deem the modelling task to be worthwhile, the question arises of where and how to obtain the data.

1.4.1 Data collection and study design

A prospective cohort study can be designed for the specific purpose to develop a prediction model. This may even be considered ideal, since the study can be designed with an adequate sample size in order to estimate measures of discrimination or calibration with pre-specified precision. The study could also be designed to recruit patients with respect to a common temporal reference point, such as the onset of disease or time since diagnosis. For instance, one could recruit newly diagnosed patients in order to predict risks in early disease. Another advantage to the prospective design is the specificity to the distribution of disease severity. For example, the prediction model may be intended to predict risk in those with severe disease. Additionally, one could ensure enrollment for specific covariate patterns, such as stratifying on age or gender. Here then, the prediction model may be intended to have adequate power to make predictions in both elderly and young males and females.

When it is too expensive or prohibitive to design a purpose-driven prospective study, authors will naturally obtain a *convenience* sample, such as from patient chart reviews. This approach is likely to present several issues, such as: missing clinically important variables; lacking consensus (or operational) definitions of predictors or outcomes; poorly defined patient populations; biases of retrospective data collection; large degrees of missing variables (especially when data are not missing at random); lack of standardized data collection methods and observer/rater variability. Several of these limitations can be avoided by careful collection of data from prospective studies, including randomized trials, for the deliberate purpose of developing prediction models. Though the data used in this thesis were obtained from a controlled trial, this was not the primary purpose of the trial, though we used this data nevertheless.

When seeking a convenient dataset to use, randomized controlled trials offer many advantages over data obtained from cohort studies. Major strengths of such studies include: prospective design and data collection; standardized and stringent criteria for par-

ticipant selection; objective outcome and variable definitions; and close follow up of study participants. For all of its potential advantages, a trial may have limited generalizability, especially when restrictive inclusion and exclusion criteria are applied, and may also be limited when the trial is pragmatic in design [28, 29]. One simple example to see this is to consider that many trials automatically exclude children and pregnant women for ethical reasons. Once the data source has been chosen, the next concern is to select predictors.

1.4.2 Selection of potential predictors

Typical candidate predictors include: patient characteristics (age and sex are generally included at a minimum), presence of co-morbidities, disease severity and laboratory or diagnostic test results. Since the sample size and the number of events are (often) limited, the variables must be selected judiciously to include only the most relevant predictors. In contrast to etiologic study designs, the choice of predictors need not be causal, but rather only associated with the outcome of interest. Even non-causal variables can be highly predictive of an outcome [24]. Having chosen the predictors is itself claiming an association between the predictor and the target outcome, while the specific modelling procedure will assign a weight of "importance," or strength of association.

It has been argued that as much information as possible should be extracted from the variables [30]. Where one has the choice between categorizing a continuous variable or keeping it continuous, the continuous case is preferred. Categorization of continuous variables has been recognized as a practice that is relatively less statistically efficient and encourages potentially arbitrary, data-driven choice of a cut point and increasing the risk of attaining biased results [31–33]. If categorization is to be performed, then the loss of information resulting from categorization must be weighed against the bias introduced due to violating the assumptions of the relationship between the continuous predictor and the outcome (e.g., the linear relationship between the continuous predictor and the logit of risk). In the simple case, if the continuous predictor is dichotomized, it implies

that the risk is constant within categories and that there is a discontinuous jump in risk across categories. This may or may not be reasonable assumption to make.

It may be necessary to exclude predictors that are difficult to measure. The predictor may be too expensive or too infrequently measured. It may be unavailable to the target user. Or the variable may be more subjective to measure and have high inter-rater variability. A standardized and accurate measurement is necessary to enhance the predictive ability of the model and its applicability [30].

Sample size considerations are an important issue in any research project. In the framework of hypothesis testing and inferential statistics, a formal calculation or simulation of sample size is a fundamental aspect of such empirical studies. These calculations are done to meet a target amount of power for a statistical test at a pre-specified type I error rate (alpha). Sample size estimation may also be done to ensure a target variance or width of a confidence interval and ensure a desired level of precision. Formal sample size estimation in prediction modeling is complicated due primarily to the fact that one needs to consider multiple target statistics which each describe different aspects of the model performance, such as discrimination and calibration. The literature has not yet reached a consensus on which of these statistics is most important for calculation of sample size. As a result, most researchers have turned to simulations and have instead used the idea of events per variable (EPV) in order to compare the relative precision of a discrimination and calibration statistics. For logistic and Cox proportional hazards models, the widely accepted *rule-of-thumb* for the minimum number of events and non-events per variable can be as few as 10 EPV to more than 50 EPV [34–36], though there are some situations in which less than 10 EPV are acceptable [37]. The greater the EPV, the smaller the risk of bias in estimating the model coefficients and hence more accurate model performance.

There are several methods which may be used to select predictors. Predictor selection should be informed by prior domain knowledge. Selecting appropriate predictors will signal to the target users that the model could be sensible, *prima facie*. Solely relying on statistically significant predictor-outcome relationships from univariate analyses risks

potentially severe model bias and erroneously including predictors by chance [20, 21, 38, 39]. It has been recommended to use a backward selection procedure in combination with bootstrap resampling shrinking estimated regression coefficients to correct for over-fitting [21]. If there are multiple candidate predictors that the authors wish to include, but limited data, a commonly used strategy is to combine multiple related variables into a single (summary) variable.

1.4.3 Outcome of interest

Any outcome may be predicted in principle. However, in the clinical prediction context the authors should decide on an important outcome that is relevant to either the patient or clinician. The outcome should have an objective definition, or failing that, a standardized operational definition (e.g., presence or absence of a disease; a specific type of surgical procedure; 30-day mortality). When given a choice, using a standardized, objective outcome (e.g., 30-day mortality, surgery by one year) is preferable to a broad or subjective definition (e.g., need for treatment escalation or a composite of many different specific outcomes). For example, a subjective outcome may be the decision of whether there was inadequate response to treatment as judged by a clinician. Or an outcome that defines need for treatment escalation that actually reflects many implicit decisions, including physician experience, patient preference, hospital policies and treatment guidelines, all of which are likely to vary between centres, geographic regions and maybe even across time periods. Subjective outcomes increase the possibility of introducing many biases due to detection and assessment of the outcome, selection bias (e.g., the investigator may decide the outcome definition was not reached for a particular patient due to their own personal bias), and recall bias if the outcome cannot be evaluated prospectively [40]. Using standardized, objective definitions along with standardized measurements and blinded outcome assessment is more in keeping with the PICOTS approach to risk of bias assessment [41], in which every effort is made to assess the choice of outcome for risk of bias.

1.4.4 Modelling

Many modelling strategies exist, such as non-parametric (rank-based), data-driven approaches, Bayesian models, and other statistical learning approaches. This thesis will adopt and focus solely on a regression modeling approach. Details of regression modelling have been extensively discussed in the literature [21].

1.5 Characterizing model performance

Model performance can be quantified using multiple measures [22, 42, 43]. First, calibration measures estimate the agreement between observed and predicted outcome frequencies (or risks). Second, discrimination measures estimate the ability of the model to distinguish between patients with different outcomes (e.g., distinguishing the diseased from the non-diseased). Third, overall performance measures are those which incorporate some aspects of both calibration and discrimination. Each measure assesses a different aspect of performance. For instance, calibration measures are more sensitive to systematic deviations from predicted risks, whereas discrimination is better at detecting differences in case-mix. Lastly, clinical usefulness (or utility) of the model must be assessed to determine whether the model provides additional benefit to the clinical user. Both discrimination and calibration are critical to interpreting the potential usefulness of the model and they must be reported [25]. However, multiple measures of performance are recommended to fully describe prediction models [25] and the most important are summarized in **Table 1.1**.

Table 1.1. Measures of prediction model validity and performance.

Aspect	Measure	Characteristics
Calibration	Calibration plot	Visual representation of agreement between observed and predicted probabilities
	Calibration slope	quantifies the agreement between predicted and observed outcome
	Calibration intercept ("calibration in the large")	Degree of systematic bias of predicted probabilities (too high or too low)
	Hosmer-Lemeshow statistic	Degree of goodness-of-fit
Discrimination	Concordance (c-statistic); area under ROC curve	Overall quality of predicted outcomes across all possible risk thresholds; overall agreement in predicted risks when ranking a pair of patients, each with a different outcome.
	Boxplot of predicted probabilities	Visual representation of spread of predicted probabilities for each outcome value
	Discrimination slope	Mean squared difference in predicted probabilities between those with the outcome and those without the outcome
Overall	Brier score	A measure of the error of the prediction
	R^2	Proportion of outcome variation that can be explained by the model
Clinical utility	Sensitivity	Percentage of patients with the outcome correctly classified as having the outcome
(requires selection of risk threshold)	Specificity	Percentage of patients without the outcome correctly classified as not having the outcome
	Net benefit (NB)	Model and reference policy are compared using a weighting for the relative costs of false-positive and false-negative classification to determine the net number of true-positive decisions gained for a single risk threshold.
	Decision curve analysis	Same as NB, but for a range of risk thresholds.

1.5.1 Discrimination measures

One way to quantify discrimination is to use Harrell's concordance index (also called the c-index or c-statistic) [44]. It quantifies the ability of the model to differentiate (discriminate) between patients having different outcomes of interest. Among all pairs of patients with different outcomes, the c-statistic is concordance percentage when the higher predicted risk of the pair correctly corresponds to having the outcome and the lower risk corresponds to not having the outcome. Thus, the c-statistic is the probability across all patients that a model will correctly predict that one high-risk patient has a higher probability of the outcome than another low-risk patient. In logistic regression models, the c-statistic is equal to the area under the receiver operating characteristic curve [45]. A discrimination value of 0.5 corresponds to predictions that are equally as good as chance (i.e., flipping a fair coin). A useful observed value should therefore be judged in the range from 0.5 to 1.0, the higher the better. Since it is based on the rank of predicted probabilities, this measure cannot distinguish a pair of patients with probabilities of 0.1 and 0.11 with that of 0.1 and 0.99.

1.5.2 Calibration measures

Calibration is a complementary performance measure to discrimination, quantifying the agreement between the predicted and observed outcome. For example, if a predicted risk is 10%, then (approximately) 10% of patients should have the outcome. Calibration has commonly been assessed using the Hosmer-Lemeshow goodness-of-fit test. However, the value of the test statistic, and thus a test of significantly poor calibration, varies greatly when the number of groups are even slightly changed (say 8 or 9 groups compared to 10 groups) and has been reported to have low statistical power to detect meaningful miscalibration [46].

Calibration may be assessed graphically using a calibration plot, with the predictions

on the x-axis and the observations on the y-axis. If a model has perfect calibration, then the resulting points of the calibration curve should all lie in a diagonal line coinciding with the $45°$ line. Deviations from perfect calibration result in an imperfect curve. Most often when developing a model, over-fitting (optimism) will occur and yield a calibration curve that is skewed to extreme predictions (slope <1); the predicted risk for low-risk patients will be too low and for high-risk patients too high. The calibration curve is also useful to detect if the model predicts risks that are systematically biased (too high or too low) and this is captured by the calibration curve intercept and is sometimes called "calibration-in-the-large".

A model may be calibrated to varying degrees of robustness. Recently, Van Calster and colleagues have proposed a hierarchy of calibration [47]. In the preceding paragraph, a calibration curve approach corresponds to "moderate calibration" while the systematic bias ("calibration-in-the-large") corresponds to "weak calibration". The strongest form of calibration, according to their hierarchy, would be to check calibration for each and every possible covariate pattern implied by the model. While this may be possible with only a few binary predictors, they note that this is an impossible task when continuous predictors are present. Instead, model development and validation efforts should focus on moderate calibration [47].

1.5.3 Overall performance measures

It is also possible to compute measures which estimate both calibration and discriminative ability into a single summary measure. The Brier score was derived for binary outcomes [48] as a measure of the error of the prediction. For an example with actual outcome Y and predicted probability p, the Brier score is $[B = Y \times (1 - \bar{\hat{p}_1})^2 + (1 - Y) \times \bar{\hat{p}_0^2}) = \sum_{i=1}^{N} (Y_i - \hat{p}_i)^2]$. In other words, the Brier score computes the sum of the mean squared difference between each predicted and observed outcome. It may be decomposed into calibration and discrimination [49, 50]. Sensible models have a Brier score that ranges from 0

to 0.25; a perfectly calibrated and discriminating model has no prediction error and thus a Brier score of 0, whereas a model that performs no better than chance has a score of 0.25 with an outcome prevalence of 50%. Because the Brier score is difficult to interpret, a scaled Brier score can be calculated in which the scaled Brier ranges from 0% to 100%. This is done by estimating the maximum Brier score under the assumption that the model is non-informative $[B_{scaled} = 1 - (B/B_{max}); B_{max} = mean(p) \times (1 - mean(p))]$. Similar to the Brier score are the R^2 measures of explained variance. Nagelkerke's R^2 is commonly computed for generalized linear models [51]. For binary outcomes, Nagelkerke's R^2 is a logarithm score of predicted probabilities $[R^2 = Y \times \ln(\bar{p_1}) + (Y-1) \times (\ln(1 - \bar{p_0}))]$. Nagelkerke's R^2 is interpretable as the proportion of outcome variation that can be explained by the model's predictors.

1.5.4 Clinical utility and net benefit

A model is useful in a clinical setting only if it can help a clinician to make a decision, such as order a test or initiate a treatment. Even if calibration and discrimination ability are satisfactory for a group of patients, the decisions implied by the current guidelines (or policy) and by the model may be the same. Thus, the model does (not) perform better than current guidelines and the model has (no) additional usefulness to the clinician. This may be especially true when guidelines and policy documents already factor in clinical variables into decision algorithms. Traditional measures of classification based on contingency tables (i.e., sensitivity, specificity, positive predictive value, negative predictive value) are already well known to clinicians. However, the clinical relevance of these measures greatly depend on the relative benefit and costs of correctly classifying disease (true-positive prediction) versus false classifications (specifically, false-positive and false-negative predictions).

The clinical net benefit index is means of assessing the potential decisions made by using the model by explicitly accounting for the clinical consequences of false-positive

and false-negative diagnoses [52, 53]. The net benefit may be calculated by choosing risk threshold at which point we would be indifferent to delivering a treatment to a patient or not. For the calculation, this risk threshold is converted to odds, as in $(p_t/(1-p_t))$. At this specific threshold, the absolute number of true positives and false positives (TP and FP, respectively) are computed over all patients (N), and the net benefit is calculated as $NB = (TP/N) - (FP/N) \times (p_t/(1-p_t))$. For example, if we assume that each benefit gained by a true positive diagnosis is worth the harms realized by three false positives, then it implies an odds of 1:3 or a risk threshold of 25%. By definition then, the net benefit is exactly zero when the assumed costs (or harms) of false positive diagnoses are exactly matched to the benefits of true positive diagnoses. By repeating this process over a range of risk thresholds, a decision curve can be constructed to inspect and compare clinical utility over a range of operating risk thresholds. Since the unit for net benefit is the number of true cases found per patient, its maximum value is therefore defined at the prevalence. In other words, its maximum is when all true cases are found, with no false positives. The net benefit is directly interpretable as the rate of additional true-positive classifications made at a specific risk threshold when using the model over an existing policy. It can also be used to calculate the unnecessary interventions avoided.

When evaluating decision curves, they should at least be compared to two default strategies in which either all patients are classified as positive (the treat all strategy) or no patients are classified as positive (the treat none strategy). In the treat none strategy, since no positive diagnoses are ever made, the net benefit is always exactly zero for any risk threshold. In contrast, for the treat all strategy, all of the true positives are correctly classified, and everyone else is by definition a false positive. In other words, the true positive rate (TP/N) represents the event rate and the false positive rate (FP/N) represents its complement (1 - event rate). Thus, the net benefit under a treat all strategy is exactly zero when the threshold is equal to the event rate in order to satisfy this condition. Also under the treat all strategy, the net benefit is maximized at the event rate when the risk threshold is set to 0%.

1.6 Internal validation

Every model is prone to over-fitting (optimism) which inevitably results in predictions that are too extreme, especially when the number of events is small relative to the number of predictors [20, 21]. Once the candidate model has been decided, internal validation is conducted to ascertain the best-fitted, most stable version of the model [20, 21]. It is necessary that any prediction model study include some form of internal validation technique, which uses only the original study sample (development data), to quantify initial optimism.

Internal validation requires some form of resampling technique, but may also involve some (or all) of the model development steps. A popular method is split-sample testing in which the data are randomly divided into a development subset and test subset (e.g., a 70-30 split). Unless the dataset contains many (say 20,000) events of interest, this split-sample approach is relatively statistically inefficient because some data are withheld. In a large data set, this is not very concerning because the advantage is that the test performance in the hold out data gives a clean estimate of performance without having seen the data and the analyst being biased by multiple comparisons and exploratory analyses; however, in small data sets which are common in the biomedical literature, holding out data increases the risk of producing biased effect estimates [35, 54–56].

Cross-validation and bootstrap resampling are the two preferred methods of assessing internal validity by means of computing model performance [20, 21, 24, 57]. Cross-validation is a technique to validate predictive performance by randomly splitting the data into a training set to estimate (train) the model and a hold out set to evaluate the model [58]. In *k*-fold cross-validation, the dataset is randomly divided into equal subsamples of size k. One of the k subsamples is held out, and the model is estimated using the remaining $k - 1$ subsamples. The estimate of performance may be improved by repeating the entire *k*-fold cross-validation procedure several times (say 100 to 200 times) and averaging the desired performance result [21]. In contrast to data splitting, the principle underlying

the bootstrap is to use resampling. One bootstrap sample requires drawing random samples of observations with replacement in order to create a new dataset of equal size to the original dataset. Model performance can be estimated in this bootstrap sample, and the statistic saved. This process is repeated many times (say 500 to 2000 times), each time computing the same performance measure each bootstrap sample. At the end of the iterations, the bootstrap performance measures are averaged for an estimate of model performance [59]. The approach used in this thesis was to implement Efron's optimism bootstrap to characterize the degree of over-fitting [59] because it mimics the empirical distribution of samples as if one were drawing random samples from the theoretical source population. This bootstrap method naturally yields an estimate of the average amount of optimism in initial regression coefficients and also in measures of model performance, which may be adjusted accordingly [20, 21]. The initial (original) performance may be compared to the performance measured using either cross-validation or bootstrap resampling, to estimate the degree of optimism present in the original model building process.

It is also possible to use shrinkage techniques to reduce or remove this initial over-fitting. These methods diminish coefficients toward zero in order to reduce error in risk prediction for new individuals [60]. The shrinkage process implies a process of regression to the mean, in which future predictions lie closer to the overall mean than might be expected from the original predicted value [61]. Shrinkage methods reduce over-fitting and thereby improve model calibration [38, 62]. Several shrinkage techniques exist. The first and simplest is applied after a regression model has been fit by using a common shrinkage factor to reduce all coefficients by a fixed amount. This method is compatible with bootstrapping techniques to estimate the degree of shrinkage required [21]. Penalized regression is a more rigorous approach to estimating model coefficients, rather than applying shrinkage following model fitting. The least absolute shrinkage and selection operator (LASSO) and ridge regression are popular penalization methods and are also considered within the TRIPOD statement for clinical models [25]. LASSO applies a constraint that the sum of all p absolute model coefficients must remain smaller than some value, such that $|\sum_{j=1}^{p} |\beta_j| \ | < c$. In contrast, the constraint applied in penalized regression is to min-

imize the sum of squared model coefficients, $\sum_{j=1}^{p} \beta_j^2 < c$. The constraint value may be estimated by cross-validation. The LASSO also incorporates shrinkage into variable selection, and works within logistic and Cox proportional hazards regression models [63–65]. Penalized maximum likelihood, a generalization of ridge regression, may also be used with these models [66, 67]. Penalization methods are an active area of research in their application to clinical models. While penalization methods are outside the scope of this thesis, such methods are promising for model development using datasets with few observations or when there are many more potential predictors than available observations (e.g., bioinformatic predictions) [68, 69].

1.7 External validation

Before a model can be used for clinical decision making, it is crucial to perform external validation. This process involves using completely new dataset with information on the same predictors and outcome. Several forms of external validation may be considered, including: (1) *prospective validation* of a model developed retrospectively; (2) *temporal validation* using the same cohort definition at two different time periods; (3) *geographic validation* using cohorts from two independent locales; or (4) *multi-centre validation*. Some authors have arranged these types of external validation into a hierarchy which is characterized by increasing stringency and generalizability [70, 71].

Prospective validation requires a collecting data on a prospectively defined cohort. Ideally all variables will also be assessed prospectively and in an unbiased manner. Prospective validation is necessary for models developed using a retrospectively-defined cohort. Temporal validation is by definition a form of prospective validation. When this form of validation is chosen, it may be a useful alternative to a split-sample internal validation approach. If the rate of collection of new observations is reasonably fast, then it provides a reasonable way to externally validate the model. However, some view this approach as a disadvantage since it may be argued that if one could simply wait for more observations,

then one should do so in order to build the model using more information. One clear disadvantage is when outcomes are rare or concern time-to-event outcomes, where it will necessarily take a long time to acquire an adequate cohort for validation. Alternatively, prospective and temporal validation are useful to assess whether the same predictors are still useful and whether the estimate coefficients (such as their magnitude or direction) are still compatible with theory and clinical expertise. However, neither prospective or temporal validation address the wider generalizability of the model since they are derived from cohorts within the same centre(s) and setting. Geographic and multi-centre validation are more stringent forms of assessing the wider external validity of a model. The difference is whether data are collected from one or multiple independent centres and gives a more realistic perspective on whether accuracy is maintained in these different cohorts and different locales.

Formal sample size estimation is rarely done in practice for external validation studies. It is recommended that at least 100 events of interest (and ideally >200 events) be included in validation cohorts [72, 73]. These sample sizes are attainable from clinical trial datasets or from large, population registries (such as insurance databases or national data clearinghouses), though rare outcomes present greater costs to data collection.

The process of external validation is very similar to that of internal validation. The developed model and its coefficients are used on the new cohort to compute estimated risks. Validation cohort performance is especially emphasized as this directly relates to how the model can realistically perform. Ideally, the performance in the development and validation cohorts will be similar, suggesting that the model will generalize to the source populations of both. However, it is unsurprising to observe a (usually small) performance loss upon external validation. Currently there are no strict guidelines to denote poor or acceptable or very good performance [26, 70, 71, 74]. Authors will describe performance as “acceptable”, “good”, or “excellent” without theoretical basis for such qualifiers. Rather,these qualifiers are often chosen arbitrarily based on the context of other models in the same or a related research discipline. If validated performance is lacking,

the model may be updated by adjusting baseline risk or regression coefficients, or adding or removing predictors [21]. If model updating is performed, it should technically then be externally validated on the basis that this is now a new model and hence untested [28, 75].

An important step beyond external validation would be to determine the impact or effectiveness of using the model. This should be done in the context of randomized controlled trial, in which the patient is randomized or using a clustered design (e.g., randomizing the clinician or the practice) [75]. Such a trial would demonstrate, in a controlled way, that patient care is truly improved using the algorithm compared to usual care. Such a design can also be adapted to include economic impacts of treatment decisions providing an economic impact analysis, especially when associated treatments or diagnostic testing is expensive or resource intensive.

1.8 Thesis scope and organization

The primary purpose of this thesis is to develop and internally validate risk prediction models for CD-related complications, defined as hospitalization, surgery or serious disease complication. The data were obtained from the Randomized Evaluation of an Algorithm for Crohn's Treatment (REACT) multi-centre, open-label, cluster-randomized controlled trial [76]. REACT is one of the largest clinical trials in CD, and evaluated an early, combined immunomodulation therapy algorithm for patients with CD in the community clinics compared to conventional care.

The remaining thesis chapters are organized as follows. **Chapter 2** will review important clinical predictors associated with disease-related complications which have informed variable selection in the model development stage. **Chapter 3** will focus on development and validation of the prediction models using data from REACT. **Chapter 4** will present general conclusions and discuss further work.

Chapter 2

Predictors in Crohn's disease

2.1 Introduction

Crohn's disease (CD) is idiopathic in nature and there is an extensive literature in epidemiology to identify potential factors associated with its onset and its complications. This chapter reviews the most important prognostic factors associated with disease-related outcomes. The first major consensus on risk factors of the disease course arose from the Vienna classification in 1998 and its Montréal update [77, 78], in which a set of criteria were created to stratify patients by disease phenotype and classify risk of disease progression based on characteristics at diagnosis (**Table 2.1**).

In **Table 2.1**, age at diagnosis refers to when positive clinical diagnosis was established. Disease location refers to the largest anatomic extension of the disease in any period prior to the first surgery. There are four possible locations: terminal ileum (the inferior third of the small intestine with or without involvement of the cecum); colon (any location between the cecum and the rectum); ileocolic (terminal ileum disease and any place between the ascending colon and the rectum); and upper gastrointestinal tract (any location proximal to

the terminal ileum). In the Montréal classification, L4 may modify any of the L1-L3 classifications. The behaviour phenotype may be either inflammatory (non-stricturing and non-penetrating), stricturing or penetrating with an additional modifier of perianal disease. Stricturing disease is defined by the occurrence of constant stricture of the lumen demonstrated by either radiology, endoscopy or surgery without evidence of penetration. However, penetrating disease is defined by any occurrence of abdominal or perianal fistulas, inflammatory masses, and abscesses during any period of disease evolution. Finally, inflammatory disease is defined only by the absence of evidence of stricture or penetration.

This chapter will review predictors that pertain to development of Crohn's disease-related complications, including surgery, reoperation, hospitalization, or severe disease behaviour.

2.2 Conceptual model

The conceptual model used in this thesis is presented in **Figure 2.1**. Notably, the biologic activity is represented as an unmeasurable abstract concept. The true pathological state will subsequently influence measurable disease characteristics, including current stricturing or penetrating behaviour, visible signs and symptoms, extraintestinal manifestations of the disease (e.g. mouth and skin ulcers) and the disease location. These markers of disease severity are measured using clinical severity indices (in this case, the Harvey-Bradshaw Index) and the patient and clinician will use this measure to make treatment decisions which may involve medication or eventual surgery. CD-related complications represent serious worsening of disease or disease-related hospitalizations. A person may develop inflammatory bowel disease given the right combination of family history of IBD, genetic susceptibility an environmental exposures. Implicitly I assume that a person's socioeconomic status (SES) will influence their toxic habits (i.e., smoking and drinking) and thus may modify the severity of their pathology. Furthermore, a person's SES will also

impact their access to care and ability to make decisions about and afford medical treatment.

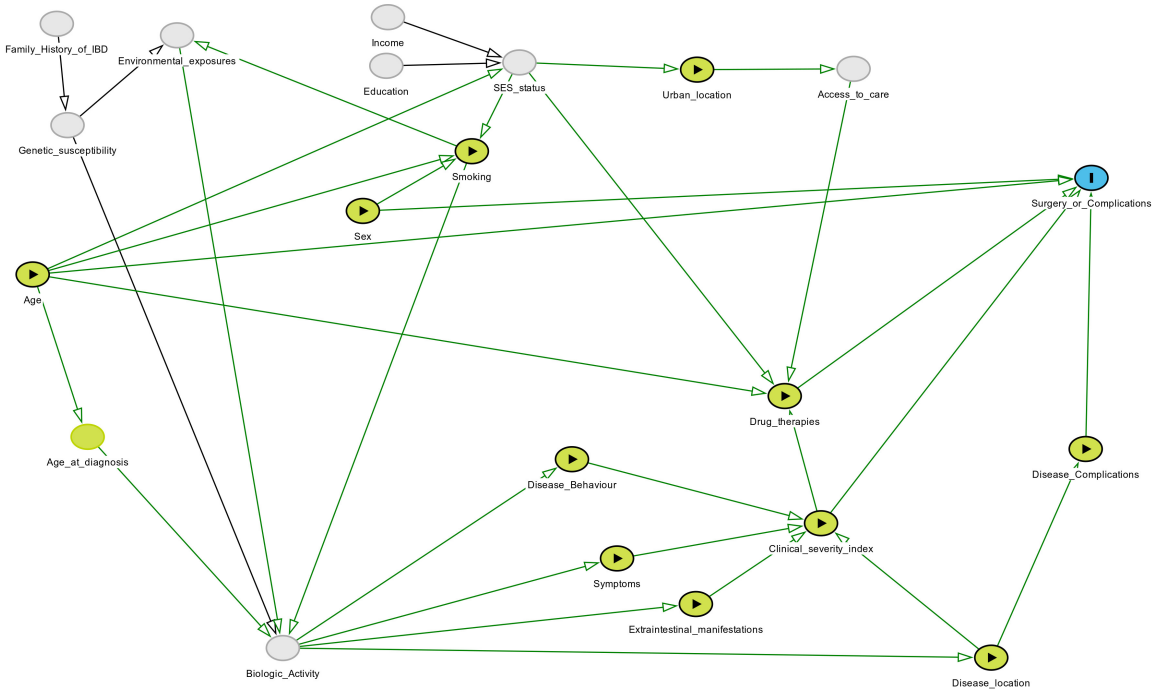

Figure 2.1. Directed acyclic graph conceptual model relating predictors to CD-related outcomes. Grey icons depict unmeasured or abstract variables. Green icons represent measured variables in the REACT trial. The blue icon represents the outcome of CD-related surgeries or CD-related complications. Abbreviations: socioeconomic status (SES), inflammatory bowel disease (IBD).

Table 2.1. Vienna and Montréal classification of Crohn's disease at diagnosis

	Vienna	**Montréal**
Age at diagnosis	A1 - <40 years A2 - ≥40 years	A1 - <17 years A2 - 17–40 years A3 - ≥40 years
Location of disease	L1 - ileal L2 - colonic L3 - ileocolonic L4 - upper GI	L1 - ileal L2 - colonic L3 - ileocolonic ±L4 - upper GI disease
Behaviour	B1 - inflammatory B2 - stricturing B3 - penetrating	B1 - inflammatory B2 - stricturing B3 - penetrating ±p - perianal disease

2.3 Sex

As a risk factor, the data regarding sex is equivocal at best. Most studies have not detected any strong or statistically significant association of sex (i.e., RR = 1) [79–89], yet others have found small yet significant increased risk for complex disease behaviour or increased severity of disease in males [90–94] or females [3, 95, 96]. It was recently reported that females experience a statistically significant absolute increase in the number of TNF-antagonist side-effects (81.3% versus 64.2%) and subsequent therapy withdrawal due to these side-effects (35.4% versus 18.4%) [97]. Thus, sex-specific approaches to therapy may be warranted.

2.4 Age at diagnosis

Age of diagnosis is categorized using the Vienna/Montŕeal criteria (**Table 2.1**). The age of onset is approximately 10% in adolescents and children (Montréal A1), 60% in adults 40 years old or younger (A2) and 30% in adults older than 40 years (A3) [2]. A positive diagnosis is usually made in the patient's twenties to early thirties, however CD may be diagnosed older than 90 years and even as young as 6 months old. Referral centre and population-based cohorts generally find that younger age at diagnosis (A1 vs. A2 and A2 vs. A3; **Table 2.1**) is generally associated with poorer prognosis compared to older patients at diagnosis, with hazard ratios and odds ratios typically ranging from 1.2 to 1.5. This difference may either be with respect to greater severity of disease [79, 89, 98–100], increased risk of surgery or time to recurrence [3, 79, 80, 98, 100–103] or a penetrating phenotype at diagnosis [89, 104]. In contrast, others have reported no associated with age, or the opposite finding that older age at diagnosis results in poorer prognosis compared to younger age at diagnosis [80, 81, 83, 99, 105]. One limitation noted from all of these studies is the insistence to categorize age rather than keep it continuous. Most studies, but not all, align these categories to the Vienna/Montréal age groups. As a result, this

categorization may exaggerate the true effect size and lose information from a statistical perspective. Nevertheless, younger age at diagnosis is generally associated with poorer prognosis.

2.5 Smoking

Smoking is the most consistent risk factor of developing CD and is associated with more severe and refractory disease [106–109]. The pathomechanism is not yet known, yet current smokers experience greater frequency of strictures, fistulas and a larger proportion of perianal involvement [109–112] and extraintestinal manifestations [113].

The risks of hospitalization and first and subsequent surgery are as much as doubled while smoking [110, 112, 114, 115] as well as post-operative complications and readmissions [116]. A large UK registry found that the negative effects of smoking most notably manifest after when diagnosed with CD after age 40 years, suggesting a disease modifying relationship between age and smoking [117]. Population-based studies consistently show the negative effects of smoking [84, 98, 101, 105, 113, 118–121], with the notable and unexplained exception from Olmsted County [81, 122]. Interestingly, immunosuppressive use was shown by Cosnes and colleagues to partially reverse the deleterious effects cigarette smoking [107].

Smoking can also affect the response to treatment. While receiving TNF-antagonists, smoking increases the risk of relapse [123]. The PRECISE-3 study found that smokers were more likely to lose response to certolizumab pegol [124].

Smoking cessation is beneficial in CD. As an intervention, the disease pattern of ex-smokers at one year were the same as never smokers [125]. Those ex-smokers also required less immunosuppressive and immunomodulatory medication [125]. These results have also been replicated after widespread use of TNF-antagonists [126]. Thus, former

smokers can reasonably expect to resume a similar prognosis to never smokers after one year [117]. Interestingly, many patients with IBD are unaware of the impact of quitting smoking [127, 128]. Smoking cessation is not only beneficial to the course of disease but is the only known modifiable environmental risk factor, and should be included in the design of future clinical trials.

2.6 Family history

A family history of IBD, that is having any relatives with IBD diagnoses, is a strong predictor of also developing IBD [129, 130]. CD-patients with first-degree relatives are enriched for risk variants of multiple IBD-associated genes compared to healthy controls with a family history of IBD, resulting in cumulatively greater genetic risk scores [131]. Having closer consanguineous relatives with IBD is associated with increased risk of developing IBD, and being diagnosed at a slightly earlier but statistically significant younger age, based on a 35-year study of the Danish population [132]. The incident rate ratio for having a twin with CD is 51.4 (95% CI: 29.1, 90.7) compared to a first-degree relative (IRR 7.77, 95% CI: 7.05, 8.06) [132]. Similar results were documented in a large Spanish cohort of 100,983 cases of IBD documented in hospitals [84], and confirming the younger age at diagnosis. Having a close relative affected with IBD is still the greatest unmodifiable risk factor for one to develop IBD [133].

In contrast, a family history of IBD is not generally considered a risk factor for experiencing more severe disease behaviour (stricturing or penetrating activity). A few small studies from single institutions have found risk of penetrating disease to be increased in familial CD pairs [1, 134]. Familial CD cases have a nominally increased risk of first major surgery by two years after diagnosis based on the Danish national cohort, and significantly increased risk of surgery for two to 15 years after diagnosis (HR 1.60, 95% CI: 1.21, 2.11) [135]. Familial cases of CD, identified from a large Spanish cohort, had small but significant increases in the rates of extraintestinal manifestations, penetrating disease behaviour

and perianal disease [84]. The risk of pouchitis, an inflammation of the lining of a pouch that is surgically created in the treatment of IBD, in CD was tripled with a family history of CD compared to no history of CD from a prospective study of a referral clinic [136]. Overall, a family history of CD increases the long-term risks for CD-related surgery and complications of disease.

2.7 Site of disease involvement

The site of disease involvement has long been recognized as an important clinical characteristic. Approximately one-third of patients will experience either ileal involvement, colonic involvement or ileocolonic involvement at diagnosis [2, 137]. Up to 33% of patients will also show evidence of stricturing or penetrating complications at diagnosis [2, 137]. Disease behaviour in approximately one-third of CD patients will evolve over time from an inflammatory disease to a stricturing and penetrating phenotype within 5 years and 50% by 20 years (e.g., [104, 122, 138]). Progression in disease behaviour was strongly associated with all disease locations (L1, L3, L4) other than strictly colonic disease [122]. However, the anatomic site of involvement generally remains stable over time [104, 122, 138–140].

The results from population-based cohorts are considered first. In Olmsted County, Minnesota, ileocolonic and small bowel localization are each significantly associated with intestinal surgery (HR 3.3 and 3.4, respectively) [141]. The IBSEN population-based cohort showed that small bowel involvement (L1 or L4) increases the risk of intestinal surgery by approximately 20% [142]. A modern Danish inception cohort further identified ileal involvement (L1) to be associated with increased surgery, surgical recurrence and hospitalization compared to all other anatomic locations [101], providing a valuable update on the historical inception cohort from Copenhagen County [143]. Ileal disease doubled the risk of progression to complicated disease behaviour compared to only colonic disease (HR, 2.1) in a Hungarian cohort [144], and doubled risk of CD-related surgery (OR, 2.3) in a

Singaporean cohort [145]. These results generally agree with an older Stockholm County, Sweden cohort [146]. Furthermore, ileal disease was also associated with the development of perianal disease in the Canterbury, New Zealand cohort, which in turn may lead to more serious complications [89]. Small bowel disease leads to more rapid progression towards complicated CD behaviour [104, 109].

While rare, jejunal disease leads to a significantly earlier first disease-related complication in Olmsted County [122] and increased risk of surgery from the IBSEN cohort [142].

The above results largely agree with large observational studies. The Inflammatory Bowel Disease Genetics Consortium cross-sectional study showed that ileal involvement and jejunal disease were independent risk factors for multiple surgeries and progression to stricturing behaviour [147]. In a retrospective cohort of 2,573 patients, ileal involvement (HR, 2.78) and the absence of rectal involvement (HR, 0.34) nearly tripled the risk of first surgical resection [148]. Upper GI involvement was strongly associated with surgical and non-surgical recurrence in a large European inception cohort [98]. Colonic disease was also found to be mildly protective for resective surgery [98]. Taken together, ileal involvement (especially isolated ileal disease) and jejunal involvement are strong predictors of serious complications and need for surgery.

2.8 Perianal disease

Perianal disease encompasses a multitude of perianal manifestations, including hemorrhoids, anal canal lesions (anal fissures, anal ulcers and anorectal strictures), fistulas and abscesses [149]. Perianal fistulas are especially difficult because they cause significant morbidity and reduce quality of life. Perianal fistulas tend to be multiple, recur, and predict greater disease severity, faster disease progression, and greater need medical and surgical intervention [79, 86, 104, 109, 146, 150]. Placebo-controlled trials also consistently show

that the use of TNF-antagonists for the induction and maintenance of response also improves healing of perianal fistulas [11, 151–156]. A meta-analysis of 12 placebo-controlled trials also showed fistulizing disease is associated with surgical recurrence [157].

The lifetime cumulative incidence of perianal fistulas is estimated at 25-35% [87, 149]. Those with colonic involvement and especially those develop proctitis are most likely to develop a perianal fistula [86, 93, 158]. Perianal disease is associated with younger age and complicated disease behaviour [89], poorer prognosis [89, 122, 159, 160] and increased risk of intestinal surgery [93, 158, 161, 162]. Perianal disease is also an established risk factor for postoperative recurrence [163].

2.9 Prior surgery

The most common operations in CD are stoma creation (colostomy or ileostomy), ileocolonic resection, small bowel resection with or without strictureplasty, colectomy, and ileo-rectal anastomosis [164]. Terminal ileal disease is a common indication for ileocolectomy.

While surgery is not curative of Crohn's disease, it can temporarily induce remission for some patients, albeit with a highly variable duration. However, it is almost inevitable that patients will eventually have post-operative recurrence. At one year following resection, endoscopic recurrence occurs in 70-80% of patients, while 10-20% will show clinical recurrence and rarely (<5%) will surgery be curative for the long-term [141, 143, 165, 166]. Surgical recurrence rates are estimated between 4-25% by one year after diagnosis [4, 167]. However, the trend from placebo-controlled trials in the last decade have indicated a decline in the annual intestinal surgery rate toward the lower end of this range. A meta-analysis of 12 placebo-controlled trials found that prior surgery independently also influenced endoscopic recurrence [157]. About 30-50% will require reoperation by 10 years, with a mean time to surgical recurrence of 15 years [4, 148, 167].

Prior surgery for CD is considered a marker for for postoperative disease recurrence and subsequent surgery [168]. For example, population-based cohorts consistently show that prior intestinal surgery more quickly leads to the next surgery [102, 141, 169], in agreement with other reports [4, 170, 171]. The post-operative period is still one of the most poorly understood clinical periods [170]. The authors remark that part of this poor understanding of postoperative recurrence of CD is that these studies have not applied a universal or consistent definition of recurrence and evaluation of the primary recurrence or re-operation and that there is inconsistent reference to the type of primary involvement [170]. Yet, the surgical procedure itself may cause subsequent disease activity as up to half of all Crohn's disease patients with an anastomosis, the disease will recur at this site [172, 173]. Overall, prior surgery is a reliable marker for disease recurrence, albeit with a variable time frame.

2.10 Genetic factors

Genetic risk factors are an important aspect of the etiology of inflammatory bowel disease. The known variants mostly cluster within genes related to innate immunity, mucosal homeostasis and autophagy. The most consistent and extensively studied gene region in association with CD is NOD2, the first risk locus identified for IBD [174]. The list of risk loci has since expanded to over 200 [175–177], with several common to both ulcerative colitis (UC) and CD. Together, these identified risk loci explain only approximately 13% of variance in disease susceptibility [177].

Genetic analysis is attractive to stratify patients by their genetic susceptibility of future disease phenotype or complications, and even surgery. However, there are two important limitations to consider. The first is that risk variants often have (very) low prevalence. The second is that most of the data concerning risk loci for IBD have originated from European cohorts of mostly Caucasian ethnicity and these results may therefore suffer

from considerable ethnic bias. As such, the major *NOD2* risk variants are not present in individuals of East Asian descent [177].

Genetic information is of great interest, though rarely used, in risk prediction models for CD-related complications. Nevertheless, genetic risk scores and the total number of risk variants are useful to predict disease phenotype and site of involvement [178]. Sometimes this is due to low risk variant prevalence (e.g., [142], or only using information about *NOD2* risk variants [118, 162, 179].

2.10.1 Nucleotide-binding oligomerization domain-containing protein 2

Nucleotide-binding oligomerization domain-containing protein 2 (NOD2) (also referred to as CARD15 or IBD1) is an intestinal bacterial peptidoglycan receptor. There are three risk variants associated with CD (R702W, G908R and L1007fsX) [180, 181]. The largest genetype-phenotype study of IBD patients (16,902 with CD and 12,597 with UC) found that having a NOD2 risk variant approximately doubles the risk of developing ileal disease compared to colonic disease, with ileocolonic disease presenting intermediate risk [178]. The greatest risk is from the rs2066847 (3020*ins*C/p.Leu1007fsX) mutation [178], which happens to be the most common among Caucasians.

After accounting for disease location, NOD2 risk variants do not further increase risk of stricturing or penetrating disease [178]. NOD2 risk variants only slightly increased risk by 10%-31% for index surgery [178], similar to a previous meta-anaylsis [182]. However, an different meta-analysis found that no association with surgical recurrence [183]. An Australian cohort found that the same frame-shift mutation was associated with rapid progression to more complicated disease and significantly earlier time to index surgery [184].

It's important to consider that NOD2 status on its own can have poor predictive power [182]. Minor allele frequencies in healthy Caucasians and Caucasians with CD are generally <5% and with large geographic heterogeneity [178, 182, 185]. Thus, genotyping *NOD2* status may only be useful for a small subset of patients.

2.11 Biomarkers

Biomarkers such as C-reactive protein (CRP), neutrophil-derived proteins, cytokines, and anti-microbial antibodies have been investigated for their associations with disease course and treatment outcomes.

2.11.1 Serum C-reactive protein

Serum C-reactive protein (CRP) is an acute-phase inflammation protein of systemic inflammation [10, 186]. Serum CRP concentrations is frequently have elevated in IBD, and is more pronounced in CD compared to UC [19, 187–190]. This only moderately correlates with the severity of inflammation and hence of disease [187]. However, a substantial portion of individuals are genetically predisposed to not mount a serum CRP response [10, 191–193]. Elevated serum CRP also correlates with endoscopic and severe histologic inflammation, yet not with radiographic activity [10, 19, 194, 195]. Elevated CRP levels also weakly correlate to relapses, possibly requiring hospitalization over the short-term (1 to 2 years) [188, 195, 196], including in asymptomatic patients to predict future relapse [160, 197]. These results led to the view that CRP is a sensitive marker of IBD-related inflammation.

The association between CRP concentration and either future treatment response or disease activity is relatively weak. Several placebo-controlled trials demonstrated that baseline CRP concentration *per se* was generally not predictive of future remission status

or substantial clinical improvement in disease-activity scores [198–203]. In contrast, an early CRP response was shown to be a better predictor of treatment response and durable remission [199, 204–206], in accord with findings from large prospective cohorts [18, 207–209].

Few studies have investigated CRP as a predictor for CD-related surgery. In the IBSEN cohort, a high CRP concentration (>53 mg/L) was strongly predictive of intestinal surgery by one year, but only for those 46 patients with terminal ileitis (L1) (OR 6.0) [190]. If CRP remained elevated for one year with refractory CD, this also tended to suggest increased risk of surgery for the subsequent four years [190]. Others still have found no statistically significant association between CRP and surgery [142, 179]. Despite its limitations, serum CRP tests are relatively inexpensive, widely available and may be readily included into prediction models.

2.11.2 Fecal calprotectin and lactoferrin

Two fecal biomarkers, calprotectin and lactoferrin, have garnered significant interest as surrogates of active inflammation. Calprotectin is a heterodimeric protein that binds zinc and manganese ions and its binding activity is calcium-dependent. It is mainly produced by neutrophils which is excreted in feces [210]. Fecal calprotectin (FC) concentrations are significantly elevated in IBD compared to healthy controls [211–213]. Lactoferrin is an iron-transporting globular protein, also secreted from granules of neutrophils and into the serum by mucosal secretion. Fecal lactoferrin (FL) concentrations are elevated in active IBD compared to inactive disease and healthy controls [214].

With respect to other biomarkers, calprotectin and lactoferrin strongly correlate (Pearson correlation, $r > 0.6$) with each other [10, 215–218], whereas fecal markers only moderately correlate with CRP (Pearson correlation, $0.3 < r < 0.6$) [10, 215, 218–220].

As surrogates of disease activity

FC concentrations are moderately to strongly correlated with the severity of endoscopic or histologic inflammation in active CD [10, 212, 215, 216, 220–225]. Additionally, Low FC concentrations are generally useful to discriminate between inactive and active disease [216, 220, 222, 223]. FC correlates moderately with the HBI [212, 215] and weakly with CDAI [212, 213, 216, 219, 225] to indicate clinically active disease.

FL concentrations are moderately to to strongly correlated with the severity of endoscopic or histologic inflammation in active CD, and may discriminate between inactive and active disease [10, 216, 218, 219, 221]. However, FL correlates weakly with clinically active disease [10, 215, 216, 218, 219].

Heterogeneity among disease location

Fecal calprotectin and lactoferrin concentrations are heterogeneous when grouped by site of disease involvement. Active and quiescent ileal disease (Montréal L1) tend to induce less secretion of these fecal markers compared to active colonic or ileocolonic disease [10, 216, 219, 226]. Endoscopic inflammation also correlated poorly with ileal CD [10, 216]. However, reports conflict as to whether there are significant differences among disease location in inactive disease [10, 218–220].

Heterogeneity as predictors of post-operative recurrence

Few studies have examined the relationship between fecal markers and the risk of post-operative recurrence. Fecal calprotectin and lactoferrin concentrations normalize within 1-2 months of intestinal surgery [215, 217]. One study showed that these fecal markers could predict postoperative recurrence in symptomatic post-operative patients, but recurrence

was not defined [215]. The FC concentration has been shown to prospectively predict endoscopic recurrence following ileocolonic resection [225, 227–229].

Fecal markers, especially calprotectin, show high sensitivity in prospectively evaluating surgical or endoscopic postoperative recurrence [225, 227–232]. A key limitation in postoperative recurrence is the high variability in the choice of cut-off values for each marker to demarcate remission from recurrence, making the clinical utility dependent on the choice of cut-off. Ileal disease especially may be especially hard to monitor using these fecal markers.

Diagnostic utility

Information on the comparative diagnostic performance of fecal calprotectin and lactoferrin tests in IBD is relatively sparse. Nevertheless, most authors suggest that FL is superior [10, 215, 218, 233, 234]. A recent meta-analysis of the pooled diagnostic characteristics of serum and fecal biomarkers reported that fecal biomarkers are overall more accurate for the diagnosis of endoscopically active IBD [235]. Compared to FC, FL had nominally lower pooled sensitivity (82% vs 88%) and nominally greater pooled specificity (79% vs 73%), while fecal markers were significantly more sensitive than CRP (sensitivity = 0.49, 95% CI: 0.34, 0.64) [235]. However, these stool biomarkers have yet to be evaluated in clinical prediction models.

These fecal biomarker tests are yet to be perfect surrogates for the gold standard of endoscopy for the evaluation of disease activity. However, in suspected disease, these tests may be an added tool to expedite management while waiting for the more expensive and invasive endoscopy [235].

2.12 Development of clinical prediction models for complications and surgery related to Crohn's disease

Subclinical transmural inflammation persists in many patients despite the use of immunosuppressive maintenance therapies [1]. This inflammatory process predisposes patients to complications such as strictures and fistulas [2]. Recent studies estimate the long-term risk of surgery to be approximately 60-80% [2, 80, 141, 165], with the greatest risk in the first few years following diagnosis. Typical symptoms experienced in CD include abdominal pain and cramping, (frequent) diarrhea, blood in the stool, fever, fatigue, reduced appetite and weight loss.

In the last two decades, the advent of biologic therapies and refinement of treatment paradigms have revolutionized the medical management of CD. Specific advances include the introduction of TNF-antagonists [11, 202, 203, 236–240] and integrin inhibitors [200, 201, 241–243], the use of combination therapy [76, 244–246], therapeutic drug monitoring [247, 248] and earlier initiation of effective therapies in high-risk patients [13, 15, 16, 76, 240, 244, 245, 249, 250]. However, one of the greatest challenges to implementing these strategies is determining which patients are most appropriate for intensive therapy. Usually this decision is based on clinical judgment and is heavily weighted by the patient's disease activity as assessed by symptoms. While this approach is clinically sensible it ignores prognostic factors that ultimately drive the risk for disease-related complications. Accordingly, identification of patients at highest risk of complications and disease progression who have the greatest chance of benefiting from early initiation of highly effective therapy is an aspirational goal.

In this regard our approach to therapy in CD has changed dramatically. Formerly "step care," in which drugs are used sequentially to attain symptomatic remission was the preferred paradigm. While this approach is attractive because it avoids over-treating low risk patients, step-care delays initiation of effective therapy in patients most at risk for

adverse outcomes. More recently, attention has turned to early introduction of combination immunosuppression therapy in high risk patients to promote mucosal healing and to minimize exposure to corticosteroids [244, 245]. This "top-down" approach requires accurate identification of high risk patients to minimize treatment-related adverse events and costs in low-risk patients. Conversely, mis-classification of high-risk patients delays administration of effective therapies and potentially results in increased risk of complications. Therefore, the ability to accurately risk stratify patients has garnered considerable interest [118, 251, 252].

Retrospective analyses of single centre and population-based cohorts have identified multiple prognostic factors in CD [1, 79, 80, 98–100, 102, 104, 118, 122, 141, 165, 252, 253] including younger age at diagnosis; ileal disease location; perianal disease, stricturing or penetrating phenotype (sub-clinical behaviour); current smoking; (clinically apparent) stricturing or penetrating disease; treatment with corticosteroids at diagnosis or corticosteroid dependence; and extensive disease involvement.

Although several prediction models have been developed, they are encumbered by several limitations, and none of which are in widespread use. Specifically, these models were developed using cohorts from before the widespread use of TNF-antagonists, cohorts from single centres, retrospective sample selection, small development samples or predict excessively long-term risk estimates of outcome (5-10 years) [79, 100, 103, 105, 142, 179] (**Table 2.2**). This study is uniquely able to develop a clinical prediction model using data from a large, controlled trial which is currently lacking in the literature. This design is additionally advantageous to limit the possible biases from the noted limitations.

Data arising from controlled randomized trials offer a unique opportunity to develop clinical prediction models given their scope, size, multi-centre participation and prospective nature, thus overcoming some of the aforementioned design limitations. This thesis presents the development of prediction models using many of the predictors just presented. Specifically, the model development is based on data from the REACT study, a

large cluster-randomized trial of two treatment algorithms for CD [76]. Multivariable logistic regression models are developed for each of the two binary outcomes: (i) a composite of CD-related surgery, disease-related complications or hospitalization, and (ii) CD-related surgery alone. Both endpoints are defined as 24 months since patients entered the study.

Table 2.2. Characteristics of related clinical risk prediction models.

Beaugerie 2006 (Logistic model) **Outcome: Severe disease**: more than two steroid courses or steroid dependence; hospitalization for a flare-up or complication; cumulatively one year or more of severe symptoms; need for immunosuppression; or intestinal surgery. Derivation: Retrospective cohort of 1,188 adults; French centre. Time frame: 5 years following diagnosis Predictors: Diagnosis below 40 years; perianal lesions at diagnosis; requirement of steroids to treat the first flare-up Validation: Partial validation using prospective cohort of 302 adults from same centre.
Loly 2008 (Cox PH model) **Outcome: Time to severe disease**: complex perianal disease; colonic resection; two or more small-bowel resections (or a single small-bowel resection measuring >50 cm in length) or the construction of a definite stoma. Derivation: Retrospective cohort of 361 people, mostly adults; Belgian centre. Time frame: 5 years following diagnosis Predictors: Stricturing behaviour; weight loss in excess of 5 kg Validation: None
Solberg 2014 (Logistic model) **Outcome: Severe disease**: intestinal resective surgery, stricturing or penetrating disease behaviour, or need for thiopurines. Derivation: Prospective population-based cohort of 132 adults and children; Norway (IBSEN) Time frame: 5 years following diagnosis Predictors: ASCA seropositive status (IgA or IgG), disease location (L2/L3 vs L1/L4), age at diagnosis, corticosteroids at diagnosis Validation: None.
Solberg 2014 (Logistic model) **Outcome: Intestinal surgery**: CD-related fistula surgery or intestinal surgery. Derivation: Population-based cohort of 190 adults and children Time frame: 10 years following diagnosis Predictors: ASCA seropositive status (IgA or IgG), disease location, age at diagnosis, stricturing or penetrating behaviour, corticosteroids at diagnosis Validation: None

Note: table continued on next page.

Characteristics of related clinical risk prediction models continued.

Lakatos 2015 (Logistic model)
Outcome: Severe disease: having intestinal resection or progression to stricturing/penetrating disease behaviour.
Derivation: Prospective cohort of 271 adults; Hungarian centre
Time frame: 3, 5 or 7 years following diagnosis, with >3 years disease duration
Predictors: ASCA seropositive status (IgA or IgG), disease location and need for early azathioprine (within 3 years of diagnosis)
Validation: None

Siegel 2016 (Cox PH model)
Outcome: Time to first complication: bowel stricture, internal penetrating disease or non-perianal surgery (bowel resection or stricturoplasty).
Derivation: Retrospective-prospective cohort of 695 adults; 2 North American centres
Time frame: 3 years following diagnosis
Predictors: Small bowel disease; left colonic disease; perianal disease; NOD2 frame-shift mutation; seropositive status of ASCA, Cbir, ANCA; log of ASCA total antibody concentration
Validation: Internal validation using bootstrap resampling. External validation using multi-centre registry including 109 adults and 392 children.

Dubinksy 2013 (Logistic model)
Outcome: Resective surgery: intestinal resection only for penetrating or stricturing CD, excluding perianal surgery or stricturoplasty
Derivation: Retrospective cohort of 1,115 adults and children; 3 American referral centres
Time frame: 5 years following diagnosis
Predictors: NOD2 risk variant; 5 genetic variants (IL23R, IL12B, C11orf30, RXRA, CACNA2D1)
Validation: None

Dubinksy 2013 (Cox PH model)
Outcome: Time to resective surgery: intestinal resection only for penetrating or stricturing CD, excluding perianal surgery or stricturoplasty
Derivation: Retrospective cohort of 1,115 adults and children; 3 American referral centres
Time frame: 5 years following diagnosis
Predictors: Multiple models were developed. The complete list of included predictors were: NOD2 risk variant; 5 genetic variants (IL12B, SLC22A4, 21q21, ZNRF1); sex, age at diagnosis <16 years, stricturing or penetrating behaviour, ANCA serology, disease duration, small bowel disease location.
Validation: Internal validation only using cross-validation.

Chapter 3

Methods

3.1 Data source

The dataset includes all participants from the Randomized Evaluation of an Algorithm for Crohn's Disease (REACT) trial, as previously reported [76] (NCT01030809). Briefly, REACT was a large cluster-randomized, controlled trial of two distinct algorithms for the management of CD. Forty Canadian and Belgian community-based gastroenterology clinics were randomized, in a 1:1 ratio, to either early combined immunosuppression (ECI) or step care. In each cluster, consecutive adult (18 years or older) patients with CD were enrolled from urban community clinics, regardless of disease activity or concurrent therapy, and followed up to 24 months. Two separate models were developed to predict the risk over 24 months of having (i) CD-related surgery, and (ii) a composite outcome consisting of the first occurrence of CD-related surgery, complications or hospitalization (the latter is henceforth referred to as a disease-related complication). The flow chart for patient recruitment throughout the trial is presented in **Figure 3.1**.

The reporting of this study conforms to the *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis* (TRIPOD) statement [25].

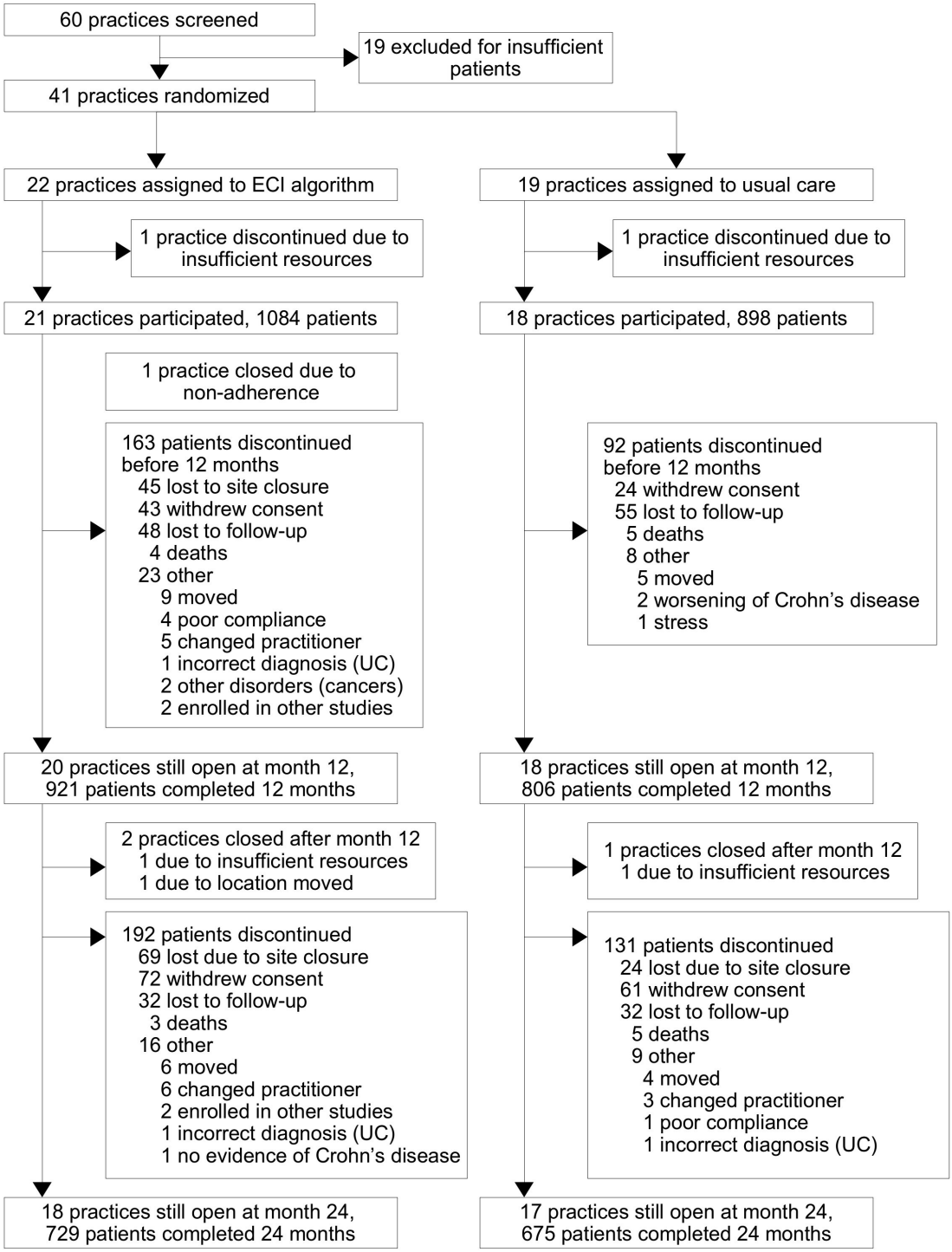

Figure 3.1. Trial profile flow diagram. Adapted from Khanna *et al.* [76].

3.2 Clinical outcomes and definitions

Two binary outcomes were defined within two years of follow up. The first was the occurrence of CD-related surgery. The second was a CD-related complication. Surgery was considered separately since it is a readily measured event that is important to patients. All of the events in the REACT studies were evaluated by an adjudication committee who were blinded to treatment assignment. Disease-related surgeries included resective bowel surgery (for example, ileal resection, ileocecal resection, proctocolectomy, colectomy, enterectomy, ostomy formation and repair), fistula repair (incision and drainage of abscess, seton placement, fistulotomy, fistulectomy), (re)anastomosis and strictureplasty. Disease-related complications were defined as serious worsening of disease (development of penetrating or stricturing disease, serious rectal bleeding, abdominal pain, increased bowel frequency), serious extra-intestinal manifestations, severe perianal disease, fistula or abscess [76].

3.3 Selection of predictors

Predictors were selected from the demographic and disease-related variables collected at baseline using standard clinical definitions. These are referred to here as baseline predictors. A search of the literature concerning potential markers and risk factors for Crohn's disease outcomes identified potential items of interest that were augmented through items identified by clinical judgment. These variables were age at enrollment, age at diagnosis, gender, smoking status, disease location, perianal disease, prior surgical resection for CD, use of each medication at baseline including 5-aminosalicylates, corticosteroids, immunosuppressives and TNF-antagonists, abdominal pain, abdominal mass, extra-intestinal manifestations, strictures, fistula status and stool frequency. No laboratory parameters, biomarkers or cross-sectional imaging items were included. Treatment allocation was intentionally excluded in our prediction models for two reasons. First, the primary interest was to

develop prediction models using only baseline predictive factors that are available in practice. Second, the effect of treatment was found to be relatively small as compared with the factors in our prediction models (OR, 1.4-1.5). This is consistent with prediction literature in other areas such as cardiovascular disease [254].

Age at diagnosis was considered first as a predictor since it forms part of the Vienna/Montréal classification system and it is commonly understood that age at diagnosis is associated with disease severity and associated complications, as discussed earlier. However, age at diagnosis did not have a statistically significant association with either outcome (OR = 1.00). Since age at baseline was highly correlated with age at diagnosis (Pearson's r = 0.74), this was substituted in place of age at diagnosis, and was found to have a more precise and stronger effect than age at diagnosis. Medication use at baseline is important to consider since these modulate disease activity.

A commonly used clinical index of disease severity is the Harvey-Bradshaw Index (HBI) [255]. The HBI score (**Table 3.1**) is a simple tally of symptom-related items (e.g., extraintestinal manifestations, complications of disease, general well-being) and frequency of liquid stools. By dividing the HBI into the components and stool frequency, more variation of the disease outcomes could be explained than when just using the HBI score. For example, a patient may be considered to be out of remission (HBI > 4) because they have frequent diarrhea, or because of much more serious symptoms (e.g., severe abdominal pain and a fistula). However, the clinical severity of the latter situation is much more severe than the former. Furthermore, many patients with inactive Crohn's disease still have frequent stools, which may be managed with anti-diarrheals or dietary modifications, yet the stool frequency can contribute the most out of any other item in the HBI score. Patients with objectively mild disease tend to have 2-4 stools per day, whereas patients with more active and more severe disease will also have substantially more frequent stools (often >8). Therefore, the HBI score was divided into these two components and an interaction between them was introduced in the modeling process to account for non-additive effects (on the log odds scale).

Table 3.1. Harvey-Bradshaw index

Item	**Item Value**
General well-being	Very well (+0) Slighty below average (+1) Poor (+2) Very poor (+3) Terrible (+4)
Abdominal pain (yesterday)	None (+0) Mild (+1) Moderate (+2) Severe (+3)
Number of liquid or soft stools yesterday	Add
Abdominal mass	None (+0) Dubious (+1) Definite (+2) Definite and tender (+3)
Complications (check all that apply, +1 point each)	Arthralgia Uveitis Erythyma nodosum Aphthous ulcers Pyoderma gangrenosum Anal fissure New fistula Abscess
Total HBI score = sum of points for each item.	

3.4 Missing data and loss to follow up

There were 4.2% (n=84) of participants that had at least one missing variable at baseline. In the original trial, there were 323 participants that (323/1982, 16.3%) were lost to follow up. These individuals were kept in the the study since they were followed for at least some duration and were considered to have not had the outcome during their follow up time. Because of a large sample size, only participants with complete baseline data were used for model development.

3.5 Model development

Exploratory univariate data analysis was initially conducted to assess adequate event frequency between each outcome and the candidate predictors. Univariable associations between candidate predictors and the outcomes were assessed by simple logistic regression. Each model was then constructed using multivariable logistic regression providing log-odds ratios and standard errors.

Candidate predictors were selected arbitrarily by a manual review of the literature augmented by expert clinical opinion. Unnecessary variables were removed by evaluation of performance in bootstrap replicates. Since the focus was to predict individual risk, the standard logistic regression approach could still be applied even though the data arose from a cluster-randomized trial, as the degree of clustering mainly inflates te estimated standard errors of the coefficients, and an intra-class correlation coefficient less than 0.05 does not much affect prediction performance [256, 257].

3.6 Predictive performance and model validation

Model performance was characterized by the discrimination ability and calibration. Discrimination refers to the capacity of a prediction model to distinguish between patients with the outcome and patients without the outcome. In the present context, discrimination is measured using the c-statistic, which is identical to the area under the receive operating characteristic curve [45]. A value of 0.50 for the c-statistic represents the prognostic ability of a coin flip, suggesting a model without discrimination ability. There are no accepted guidelines on assigning qualitative labels to the degree magnitude of discrimination. We arbitrarily considered discrimination values below 60% are unacceptable, 60-70% are acceptable and greater than 70% are considered good to excellent. Model discrimination may also be visualized by a discrimination plot, wherein the distributions of predicted probabilities are compared for those with and without the outcome. The slope of the discrimination plot is then the mean difference in predicted values. The greater the separation between groups, or slope, implies greater discrimination.

Calibration refers to the agreement between predicted and observed risks [21]. This can be assessed by plotting the predicted risks and the actual percentages of patients who have the outcome. The Brier score measures overall prediction errors. It is the original (apparent) accuracy, or average prediction error, and is calculated as the mean square of the difference between the observed outcome and predicted outcome probability. Calibration curves are useful to examine the model fit and validation process. Linear scores from a model are used to predict the observed outcomes. The curves show whether predictions are systematically too large or too small (called calibration-in-the-large) [45]. Since model development tends to over-fit the data, commonly referred to as optimism, this would result in predictions that are too extreme. The prediction model is said to have poor calibration when this type of extreme prediction occurs. Therefore, it is preferable to reduce optimism using internal validation or external validation [30].

Internal validation of the final model was performed using a bootstrap procedure that

was specifically designed for prediction models [258]. This procedure proceeds with the following steps. First, draw a sample (with replacement) of the same size as the original data set. Second, build a prediction model using the bootstrap sample and obtain the performance indices (termed bootstrap performance indices). Apply this model to the original data and obtain model performance indices (termed test performance indices). Third, obtain estimates of optimism by subtracting test performance indices from bootstrap performance indices. Next, repeat steps one to three several times and obtain the average of the optimism. This thesis project used 500 repetitions. Lastly, the optimism-corrected performance indices are computed by subtracting the average amount of optimism from the performance indices obtained using the original data. The optimism-corrected calibration slope can then be used to shrink regression coefficients in the prediction model [45]. Shrinking coefficients is analogous to regression toward the mean which provides more accurate risks for new individuals. Index performance characteristics were calculated for the initial and validated models following correction for optimism.

3.7 Score charts, nomograms and clinical utility

Score charts were derived from the regression model equations and simplified for ease of use [259]. The integer sum from all of the predictors in a model can then be converted to a risk estimate using the conversion chart at the bottom of the score chart. Each model is accompanied by a nomogram for risk estimation.

Clinical utility was evaluated using the net benefit index and is graphically represented using a decision curve [52, 53]. At a specified risk threshold, the net benefit (NB) summarizes how many additional true-positive classifications can be made when using a model compared to not, without sacrifice to the false-positive rate [52]. The NB considers the harm of being treated unnecessarily (i.e., a false-positive compared to a true-negative) to the benefit of avoiding treatment where none was necessary (i.e., a true-negative compared to a false-negative). A decision curve can be plotted from all possible NB values.

3.8 Sample size

Formal sample size calculation was not performed due to the lack of a formula. Nevertheless, there were 1982 participants in REACT trial, among them 504 had disease-related complications and 130 had surgeries over two years of follow up. According to the guideline of 10 events per variable [36], this dataset is large enough to construct prediction models with at most ten degrees of freedom.

3.9 General statistical methods

Statistical analysis was performed using Stata (version 14.1/IC; StataCorp, College Station, TX), R (version 3.3.1; Linux; R Core Team) and RStudio (version 0.99; RStudio Team) software packages. Summary statistics are presented as mean $\pm$ standard deviation (SD), median or frequencies and proportions as appropriate.

Chapter 4

Results

4.1 Patient characteristics and model specifications

The REACT trial enrolled 1,982 patients. From this, 84 patients had some degree of missing baseline predictor information and were excluded from the present analysis. Of the 1,898 (1,982 - 84) patients included, 1,097 (58%) were female. Other baseline characteristics are outlined in **Table 4.1**. Median disease duration was 148.8 months, and patients had, for the most part, predominantly low disease activity (mean HBI score = 4.1). Overall, 6.9% (n=130) underwent CD-related surgery, whereas 26.6% (n=504) of participants experienced a CD-related complication. Univariate associations for each outcome are shown in **Table 4.3**.

A table presenting pairwise correlations between potential predictors and each outcome is presented in **Table 4.2**. Most correlations among predictors are small (Pearson's $r < 0.15$). Note that age at baseline and age at diagnosis were very strongly positively correlated (r=0.74). As expected, total HBI score is also strongly correlated with the stool frequency component (r=0.88) and the remaining symptom-based components (r=0.71).

Table 4.1. Overall patient characteristics in REACT.

Patient Characteristics	**Overall N=1,898**	**Surgery group N=130 6.9%**	**No surgery group N=1,768 93.1%**	**Complication group N=504 26.6%**	**No complication group N=1,394 73.5%**
Demographics					
Age, years (mean ± SD)	44.0 ± 14.6	42.7 ± 14.0	44.1 ± 14.6	42.1 ± 14.6	44.7 ± 14.5
Gender, female	1,097 (57.8%)	68 (52.3%)	1,029 (58.2%)	311 (61.7%)	786 (56.4%)
Smoking history at baseline					
Non-smoker	939 (49.5%)	58 (44.6%)	881 (49.9%)	241 (47.9%)	698 (50.1%)
Ex-smoker	550 (29.0%)	36 (27.7%)	514 (29.1%)	405 (29.1%)	145 (28.8%)
Current smoker	407 (21.5%)	36 (27.7%)	371 (21.0%)	290 (20.8%)	23.3 (117%)
Disease characteristics					
Duration, months [mean (median)]	148.8 (119)	144.8 (111.1)	149.1 (119.2)	147.0 (114.0)	149.5 120.9
HBI score at baseline [median; mean ± SD]	3; 4.1 ± 1.1	5; 6.0 ± 3.5	3; 4.0 ± 1.1	4; 5.0 ± 4.9	3; 3.8 ± 3.8
Steroid-free remission (HBI 4) at baseline	1,065 (56.1%)	51 (39.2%)	1,014 (57.4%)	236 (46.8%)	829 (59.5%)
Involved intestinal areas					
Colon	417 (22.0%)	17 (13.1%)	400 (22.6%)	92 (18.3%)	325 (23.3%)
Small bowel	654 (34.5%)	52 (40.0%)	602 (34.1%)	169 (33.5%)	485 (34.8%)
Colon and small bowel	827 (43.6%)	61 (46.9%)	766 (43.3%)	243 (48.2%)	584 (41.9%)
Extra-intestinal manifestations at baseline	594 (31.3%)	55 (42.3%)	539 (30.5%)	185 (36.7%)	409 (29.3%)
Prior history for disease-related surgery	864 (45.5%)	56 (43.1%)	808 (45.7%)	244 (48.4%)	620 (44.5%)
Medications at baseline					
Aminosalicylates	539 (28.4%)	22 (16.9%)	517 (29.2%)	116 (23.0%)	423 (30.3%)
Corticosteroids	348 (18.3%)	31 (23.9%)	317 (17.9%)	113 (22.4%)	235 (16.9%)
Antimetabolites	826 (43.5%)	48 (36.9%)	778 (44.0%)	204 (40.5%)	622 (44.6%)
TNF-antagonists	622 (32.8%)	53 (40.8%)	569 (32.2%)	201 (39.9%)	421 (30.2%)

Note: Values are presented as [n (%)], unless otherwise specified. Figures vary slightly from those in Khanna et al. (2015) due to 84 patients with at least one missing baseline variable.

Table 4.2. Pair-wise Pearson correlation matrix among considered predictors for Crohn's disease-related complications or surgery.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	1													
B	0.74*	1												
C	-0.02	-0.01	1											
D	-0.12*	-0.21*	0.02	1										
E	0.13*	-0.18*	0.03	0.23*	1									
F	0.01	-0.08*	-0.1*	0.08*	0.14*	1								
G	0.04	-0.08*	-0.07*	0.08*	0.21*	0.88*	1							
H	-0.03	-0.05*	-0.13*	0.07*	0.02	0.71*	0.39*	1						
I	-0.02	-0.03	-0.01	0.05*	0.01	0.21*	0.06*	0.36*	1					
J	-0.02	0.02	-0.01	-0.03	-0.07*	0.21*	0.16*	0.24*	0.04	1				
K	-0.05*	-0.03	0.05*	0.01	0.01	0	-0.03	0.05*	0	-0.01	1			
L	-0.15*	-0.15*	0.02	0.09*	0.06*	0.04	0.04	0.03	-0.01	-0.09*	-0.08*	1		
M	-0.03	-0.02	0.03	0.04	-0.01	0.12*	0.06*	0.18*	0.25*	0.04	-0.04	0.05*	1	
N	-0.08*	-0.08*	-0.05*	0.06*	0.03	0.13*	0.09*	0.14*	0.14*	0.06*	-0.04	0.09*	0.45*	1

*Note: * denotes significance at 0.05 level.*

A = Age at baseline
B = Age at diagnosis
C = Sex
D = Disease location
E = Number of previous surgeries
F = Baseline total HBI score
G = Baseline stool frequency
H = Baseline HBI score (except stool component)
I = Fistula, abscess or abdominal mass
J = Baseline steroid use
K = Baseline antimetabolite use
L = Baseline TNF-antagonist use
M = CD-related Surgery outcome
N = CD-related complication outcome

Table 4.3. Univariate associations for Crohn's disease-related complications or surgery.

	Surgery model		Complication model*	
Baseline variable	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
Current age (year)	0.99 (0.98, 1.00)	0.27	0.99 (0.98, 1.00)	<0.0001
Gender (male vs. female)	1.27 (0.89, 1.81)	0.19	0.80 (0.65, 0.99)	0.04
HBI score (except stool freq.)	1.36 (1.25, 1.48)	<0.0001	1.18 (1.12, 1.24)	<0.0001
Stool frequency (total from yesterday)	1.08 (1.15, 1.02)	<0.01	1.07 (1.03, 1.11)	0.0001
Location of disease				
Colon only (ref.)	1	0.03	1	0.02
Small bowel and colon	1.87 (1.08, 3.25)		1.47 (1.12, 1.94)	
Small bowel only	2.03 (1.16, 3.57)		1.23 (0.92, 1.65)	
Presence of new fistula, abscess or definite abdominal mass (yes vs no)	10.29 (6.23, 16.99)	<0.0001	3.94 (2.48, 6.26)	<0.0001
Antimetabolite use (yes vs no)	0.75 (0.52, 1.08)	0.11	0.84 (0.69, 1.04)	0.1
5-Aminosalicylate use (yes vs no)	0.49 (0.31, 0.79)	<0.01	0.69 (0.54, 0.87)	<0.01
Corticosteroid use (yes vs no)	1.43 (0.94, 2.18)	0.1	1.43 (1.11, 1.84)	<0.01
TNF-antagonist use (yes vs no)	1.45 (1.01, 2.09)	0.05	1.53 (1.24, 1.89)	<0.0001
Smoking status				
Non-smoker (ref.)	1		1	
Ex-smoker	1.06 (0.69, 1.64)	0.28	1.04 (0.82, 1.32)	0.77
Current smoker	1.47 (0.96, 2.27)	0.08	1.17 (0.90, 1.52)	0.24

*Note: * CD-related complication is the first occurrence of a CD-related surgery, hospitalization or complication.*

4.2 Model performance

Separate logistic regression models were estimated for CD-related surgery (**Table 4.4**) and CD-related complication (**Table 4.5**) by 24 months. The original and validated log-odds coefficients are presented for both models. The baseline predictors included in the surgery model were age, gender, disease location, HBI score, stool frequency, immunosuppressive use, 5-aminosalicylate use and the presence of a fistula, abscess or abdominal mass. The baseline predictors for the disease-related outcome model also uniquely included the use of corticosteroids and TNF-antagonists, in addition to all the variables identified for surgery.

Table 4.4. Risk prediction model for Crohn's disease-related surgery.

Baseline predictor	**Original** β	**Optimism corrected** $\beta\dagger$	**SE**(β)	**P-value**
Intercept	-3.607	-3.511	0.327	< 0.0001
Age subtract 45 years	-0.0027	-0.0025	0.007	0.72
Male vs female	0.3946	0.3634	0.194	0.06
HBI score (except stool frequency)	0.2425	0.2234	0.065	< 0.001
Stool frequency*	0.0741	0.0683	0.06	0.25
Location of disease				
Colon only (*ref.*)	0	0	–	–
Small bowel and colon	0.3614	0.3328	0.293	0.26
Small bowel only	0.5167	0.4759	0.297	0.11
Antimetabolite use	-0.4519	-0.4162	0.2	0.04
5-aminosalicylate use	-0.6015	-0.554	0.253	0.03
Presence of new fistula, abscess or definite abdominal mass	1.7019	1.5674	0.306	< 0.0001
Interaction				
HBI score $\times$ Stool frequency	-0.0182	-0.0167	0.014	0.22

Notes: SE, standard error. $\dagger$ *the*β *coefficients are presented after shrinkage (shrinkage factor = 0.92).*
** Stool frequency has a maximum value of 12.*

4.3 Model validation and calibration

The original and validated performance statistics were computed for each model (**Table 4.6**). The validated discrimination ability for the CD-related surgery model was good,

with a c-statistic of 0.70, whereas the discrimination ability of disease-related complication model was acceptable at 0.62 (**Table 4.6**).

Table 4.5. Risk prediction model for a Crohn's disease-related complication.

Baseline predictor	**Original** β	**Optimism corrected** $\beta\dagger$	**SE**(β)	**P-value**
Intercept	-1.6134	-1.552	0.162	< 0.0001
Age subtract 45 years	-0.0102	-0.0092	0.004	0.02
Male vs female	-0.1848	-0.1669	0.11	0.13
HBI score (except stool frequency)	0.1139	0.1028	0.042	0.01
Stool frequency*	0.0752	0.0679	0.034	0.04
Location of disease				
Colon only (*ref.*)	0	0	–	–
Small bowel and colon	0.2525	0.228	0.145	0.12
Small bowel only	0.1739	0.157	0.153	0.31
Steroid use	0.2465	0.2226	0.138	0.11
Antimetabolite use	-0.1598	-0.1443	0.11	0.19
TNF-antagonist use	0.3887	0.351	0.114	< 0.01
Presence of new fistula, abscess or definite abdominal mass	1.084	0.9789	0.261	< 0.001
Interaction				
HBI score $\times$ Stool frequency	-0.0136	-0.0123	0.009	0.18

Notes: SE, standard error. † $the\beta$ *coefficients are presented after shrinkage (shrinkage factor = 0.90).*
** Stool frequency has a maximum value of 12.*

Both models show low degrees of optimism (shrinkage coefficients were 0.92 for the CD-related surgery model and 0.90 for the CD-related complication model) (**Table 4.6**). Each model had overall moderate prediction error as measured by the Brier score. Both models have good calibration (**Figures 4.1 and 4.2**) and goodness-of-fit statistics. The models for CD-related surgery (H-L $\chi^2(8) = 3.65; P = 0.89$) and CD-related complication (H-L $\chi^2(8) = 10.30; P = 0.24$) following optimism correction. The average difference in predicted risk was 8% greater for having CD-related surgery compared to not having surgery, and 5% greater for having the CD-related complications compared to not having the complications (**Figures 4.3 and 4.4**). While surgery was less prevalent than the disease-related complications, the surgical predictive model was more accurate. The calibration curves trend upward since a few individuals in the high predicted risk of surgery group had experienced the outcome.

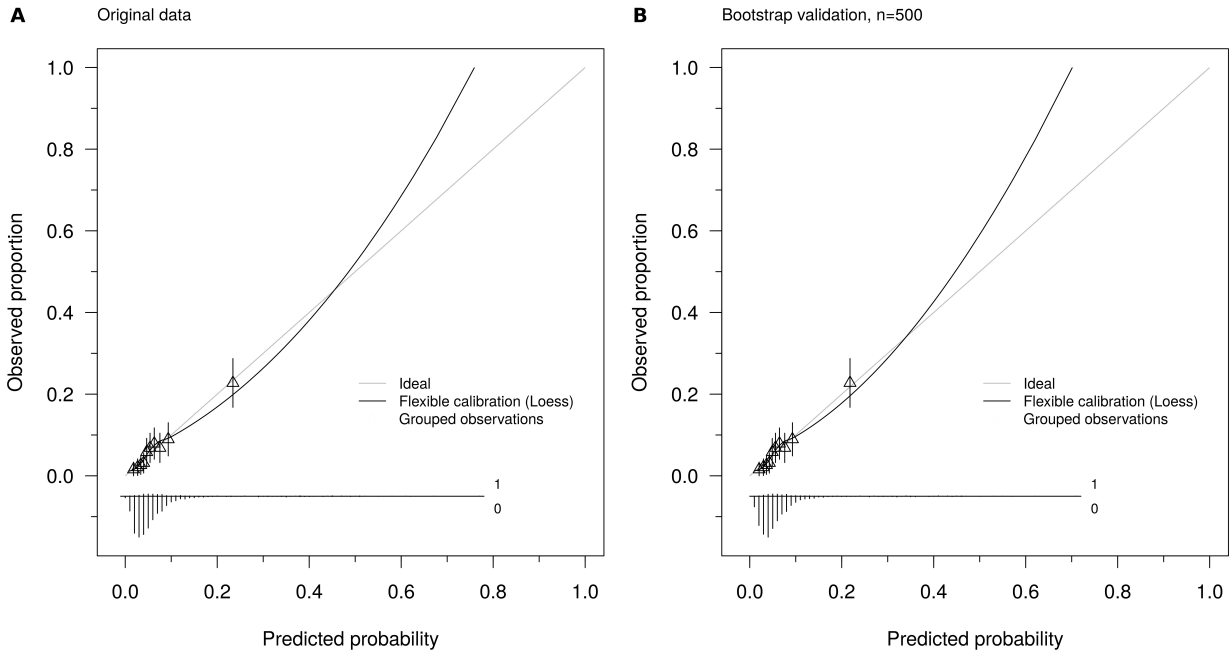

Figure 4.1. Original (*A*) and bootstrap validated calibration curves (*B*) for the CD-related surgery model. The proportion of patients that experienced a CD-related surgery was 6.9%. The ideal line (*gray line*) represents predicted and observed risks perfectly match. A flexible calibration curve is drawn for observed calibration (*black line*). Risk predictions were grouped into deciles and plotted (*triangles*) with vertical lines representing 95% confidence intervals. The bottom of each plot shows the distribution events and non-events over the range of predicted risks.

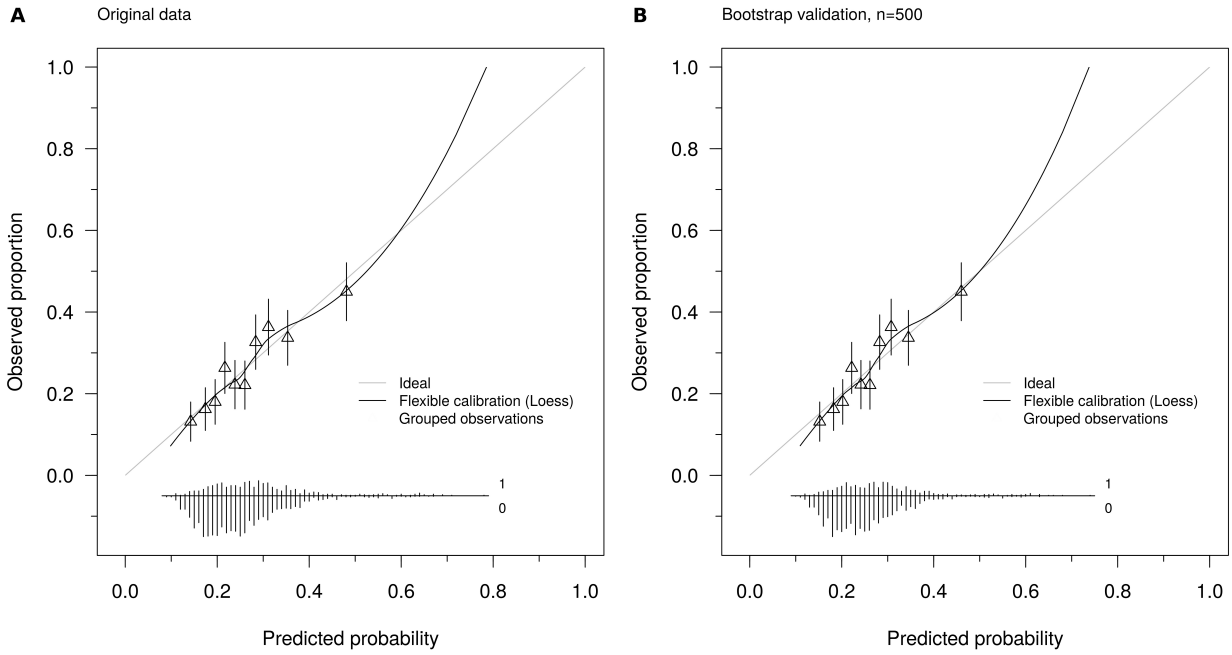

Figure 4.2. Original (*A*) and bootstrap validated calibration curves (*B*) for the CD-related complication model. The proportion of patients that experienced a CD-related complication was 26.6%. The ideal line (*gray line*) represents predicted and observed risks perfectly match. A flexible calibration curve is drawn for observed calibration (*black line*). Risk predictions were grouped into deciles and plotted (*triangles*) with vertical lines representing 95% confidence intervals. The bottom of each plot shows the distribution events and non-events over the range of predicted risks.

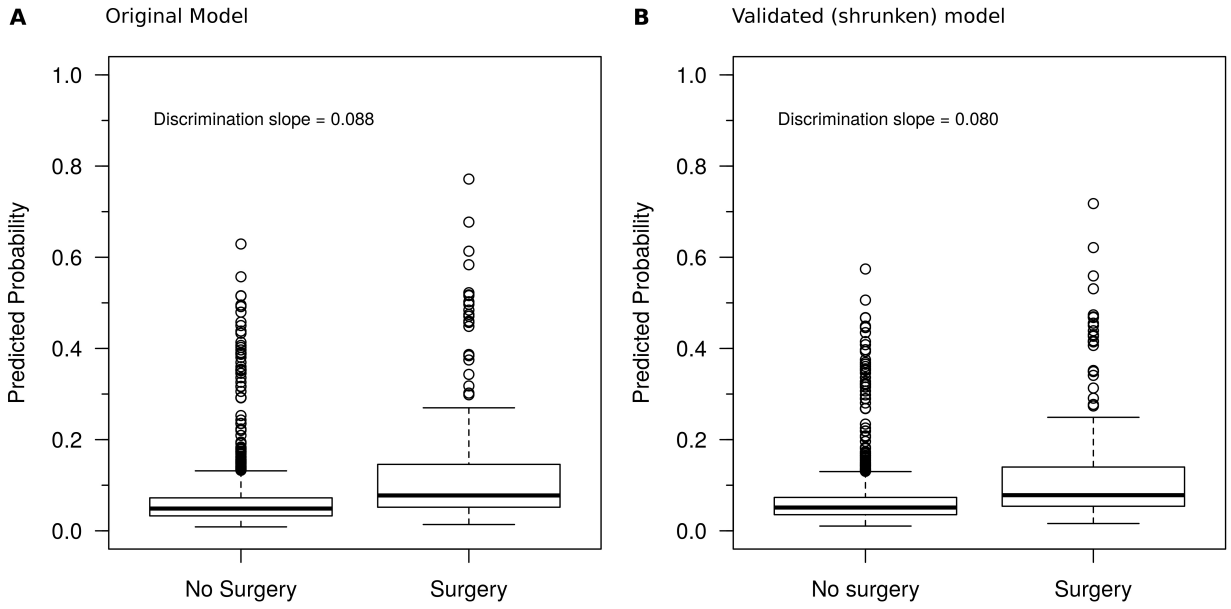

Figure 4.3. Original (*A*) and bootstrap validated discrimination plots (*B*) for the CD-related surgery model. The discrimination slope is the difference in average predicted risk between those who did or did not experience the event.

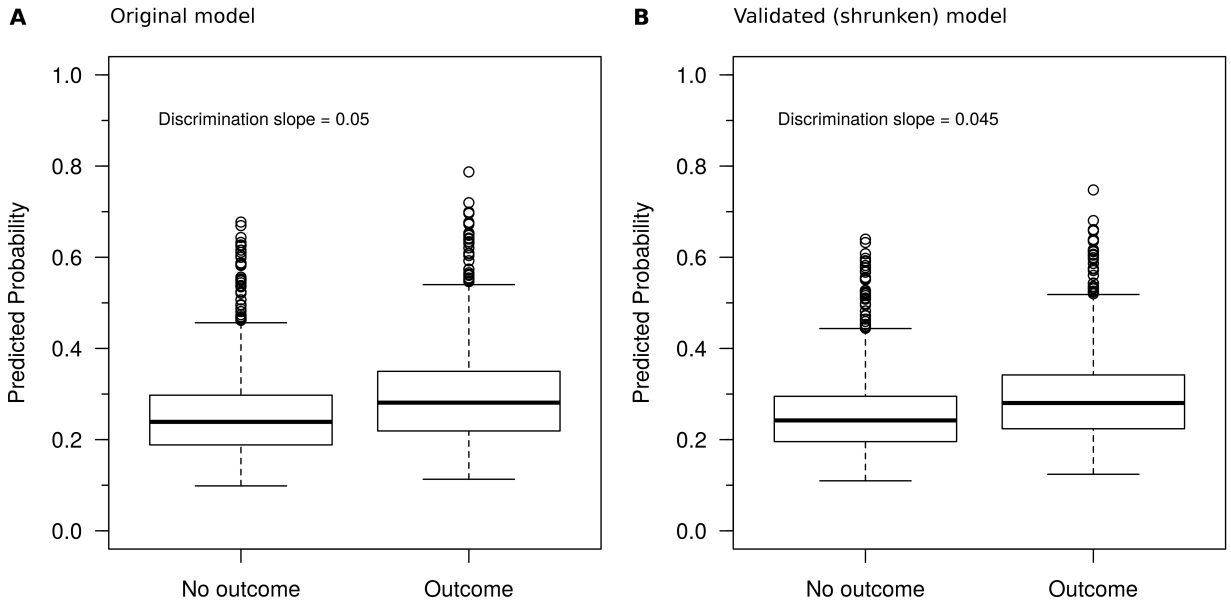

Figure 4.4. Original (*A*) and bootstrap validated discrimination plots (*B*) for the CD-related complication model. The discrimination slope is the difference in average predicted risk between those who did or did not experience the event.

Table 4.6. Summary of original and bootstrap optimism-corrected performance measures (500 bootstrap replicates).

CD-related surgery model					
Statistic	**Index Performance**	**Optimism**	**Corrected Statistic**	**Range**	**Ideal**
Nagelkerke's R^2	0.133	0.0252	0.107	[0, 1]	1
Calibration Intercept	0	0.18	-0.18	(-Inf, +Inf)	0
Calibration Slope	1	0.079	0.921	(-Inf, +Inf)	1
C-statistic	0.719	0.0243	0.695	[0, 1]	1
Dxy	0.438	0.0486	0.39	[0.5, 1]	1
Brier score	0.058	-0.0013	0.059	[0, 0.5]	0
Brier (max)	0.064				
Brier (scaled)	0.088	–	0.072	[0,1]	1
CD-related complication model					
Statistic	**Index Performance**	**Optimism**	**Corrected Statistic**	**Range**	**Ideal**
Nagelkerke's R^2	0.0683	0.0145	0.0538	[0, 1]	1
Calibration Intercept	0	0.0941	-0.0941	(-Inf, +Inf)	0
Calibration Slope	1	0.097	0.903	(-Inf, +Inf)	1
C-statistic	0.636	0.0139	0.622	[0, 1]	1
Dxy	0.271	0.0278	0.243	[0.5, 1]	1
Brier score	0.185	-0.0024	0.0188	[0, 0.5]	0
Brier (max)	0.195				
Brier (scaled)	0.049	–	0.036	[0,1]	1

Notes: Optimism is estimated as the difference of the training and test estimates.

Dxy = (c + 1)/2. Nagelkerke's R^2 and Dxy are provided as additional measures of model fit, see e.g., Harrell (2015).

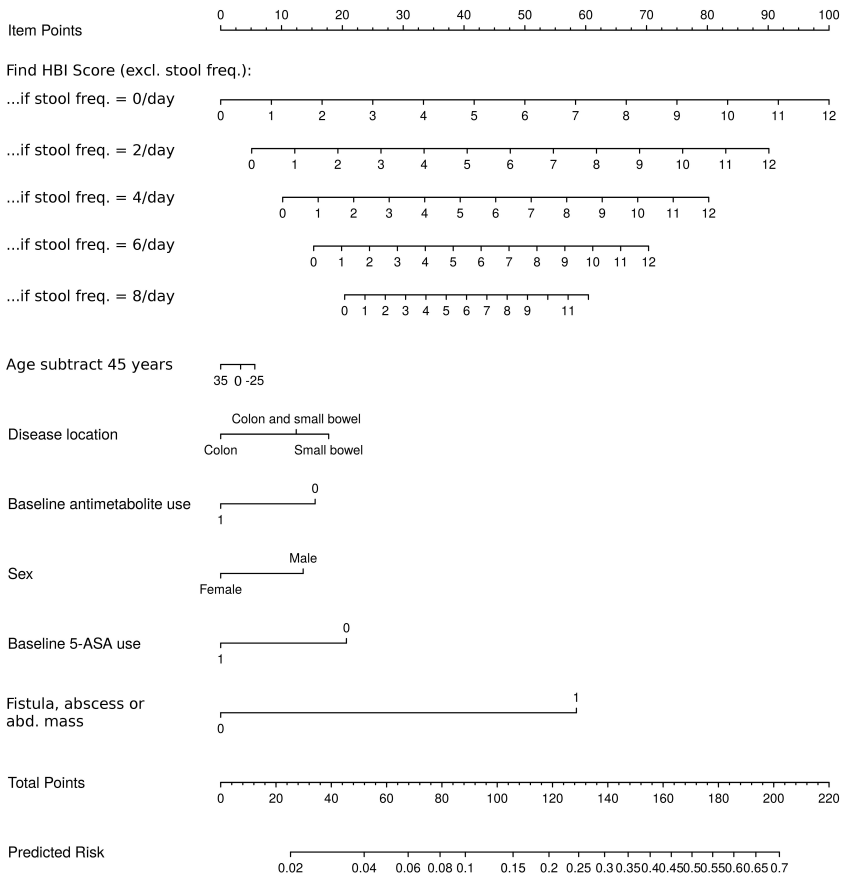

Figure 4.5. Nomogram for the computation of CD-related surgery risk.

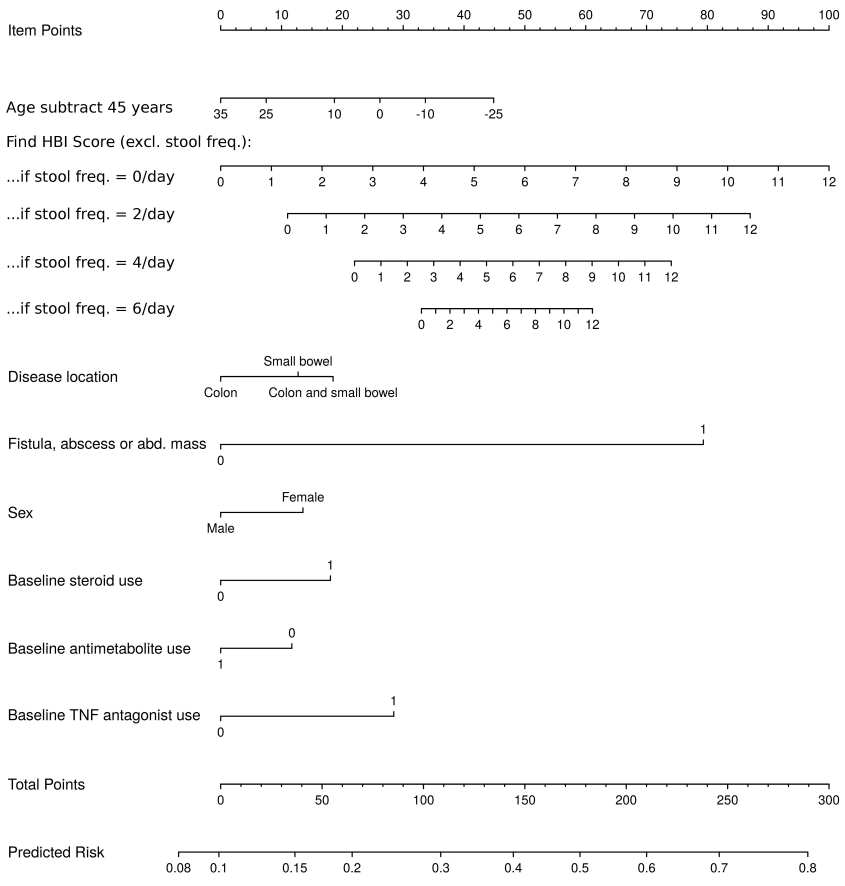

Figure 4.6. Nomogram for the computation of CD-related complication risk.

4.4 Computing a risk estimate, the score chart and the nomogram

For an individual patient, the predicted risk estimate is calculated as demonstrated as follows. The procedure for computing a risk estimate for a patient involves multiplying each β coefficient with the value of the associated variable. These are summed together with the intercept to produce a linear predictor, lp, such that lp = Intercept + $[X_1 \times \beta_1 + X_2 \times \beta_2 + ... + X_n \times \beta_n]$. The lp is converted to a risk estimate using the inverse log-odds formula, risk (%) = $100\% \times \frac{1}{1+\exp(-lp)}$. The presence of a condition takes a value of 1, while its absence takes a value of 0.

As an example, suppose a 50 year old male patient presents to the clinic with disease confined to the small bowel. His total HBI score is 3 with a stool frequency of 2 per day, he is currently taking 5-aminosalicylates and does not have a fistula, stricture or definite abdominal mass. The clinician would now like to predict their risk of surgery within the next two years.

Start by plugging these values into the equation for the linear predictor.

$$\begin{aligned} lp = & -3.511 + (-0.0025) \times (50 - 45; \text{age, years}) + \\ & (0.3634) \times (1; \text{male}) + (0.2234) \times (1\text{point; HBI subtract stools/day}) + \\ & (0.0683) \times (2 \text{ stools/day}) + (0.4759) \times (1; \text{ small bowel only}) + \\ & (-0.4162) \times (0; \text{ antimetabolite use}) + (-0.5540) \times (1; 5 - \text{ASA use}) + \\ & (1.5674) \times (0; \text{fistula, abscess, or abdominal mass}) + \\ & (-0.0167) \times (1 \text{ point; HBI subtract stools/day}) \times (2 \text{ stools/day}) + \\ = & -2.912 \end{aligned}$$

Now convert the linear predictor into percent risk.

Risk = $100\% \times \frac{1}{1+\exp(-lp)} = 5.2\%$ (expected risk is 5.2%)

Alternatively, a risk score may be computed using score charts (**Tables 4.7 and 4.8**) and nomograms (**Figures 4.5 and 4.6**) which may then be converted into a risk estimate.

For prediction of CD-related surgery, the score ranges from 0 to approximately 203 (**Tables 4.7**). A score of 130 predicts a 25% chance of requiring surgery within 24 months follow up, a score of 171 predicts a 50% chance of surgery within 24 months of follow up and a score of 203 predicts greater than 70% chance of surgery within 24 months of follow up.

For prediction CD-related complications, the score ranges from 0 to approximately 266 (**Tables 4.8**). A score of 88, 177 or 266 respectively predict a 25%, 50% or 75% chance of experiencing a disease-related complication within 24 months of follow up.

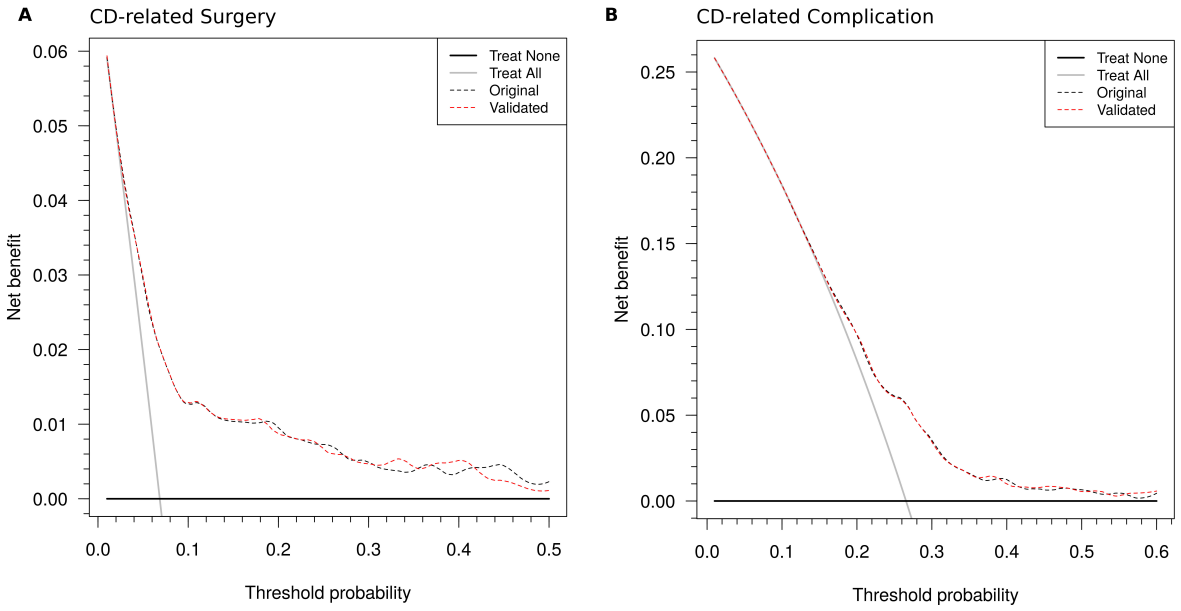

Figure 4.7. Decision curves of clinical net benefit for the models predicting Crohn's disease-related (*A*) surgery and (*B*) complications. The preferred model is the model with the highest net benefit at a given threshold. The choice of prevalence is the value at which net benefit is maximized, while the additional risk thresholds were chosen to be arbitrarily larger for the purposes of illustration.

4.5 Decision curve analysis and clinical net benefit

Positive clinical net benefit indicates potential improvement to the clinician's ability to risk stratify patients. Both models show positive net benefit over a wide range of risk thresholds (**Figure 4.7**; **Table 4.9**). Additionally, the decision curves are superior to the default "treat all" or "treat none" strategies implying improved prediction compared to not using such models. Multiple classification statistics were computed for range of risk thresholds for illustration, including: true and false positives and negatives, sensitivity, specificity, true and false positive classification rate, net benefit and interventions avoided (**Table 4.9**).

Table 4.7. Score chart for risk of CD-related surgery.

Choose HBI score (without stool frequency) when stool frequency is

HBI	0 times/day	2 times/day	4 times/day	6 times/day
0	0	5	10	15
1	8	12	16	20
2	17	19	22	24
3	25	26	28	29
4	33	33	34	34
5	42	41	39	38
6	50	48	45	43
7	58	55	51	47
8	67	62	57	52
9	75	69	63	57
10	83	76	69	61
11	92	83	74	66
12	100	90	80	70

Patient's Age (years)		**Gender**	
20 to 29	6	Male	14
30 to 39	5	Female	0
40 to 49	4		
50 to 59	3	**Disease location**	
60 to 69	2	Colon	0
70 to 79	1	Colon and small bowel	12
80 years or older	0	Small bowel	18
Baseline antimetabolite use		**Baseline 5-aminosalicylate use**	
No	16	No	21
Yes	0	Yes	28

New fistula, abscess or definite abdominal mass

No	21
Yes	0

Total the points, then convert them to predicted risk using the below scale.

Total Points	**Predicted Risk**	**Points**	**Predicted Risk**
0	<1%	130	25%
26	2%	139	30%
41	3%	148	35%
52	4%	156	40%
61	5%	163	45%
68	6%	171	50%
74	7%	178	55%
80	8%	186	60%
89	10%	194	65%
106	15%	203	>70%
119	20%		

Table 4.8. Score chart for risk of a CD-related complication.

Choose HBI score (without stool frequency) when stool frequency is

HBI	0 times/day	2 times/day	4 times/day	6 times/day
0	0	11	22	33
1	8	17	26	35
2	17	24	31	38
3	25	30	35	40
4	33	36	39	42
5	42	43	44	45
6	50	49	48	47
7	58	55	52	49
8	67	62	57	52
9	75	68	61	54
10	83	74	65	56
11	92	81	70	59
12	100	87	74	61

Patient's Age (years)		**Gender**	
20 to 29	45	Male	0
30 to 39	37	Female	14
40 to 49	30		
50 to 59	22	**Disease location**	
60 to 69	15	Colon	0
70 to 79	7	Colon and small bowel	18
80 years or older	0	Small bowel	13
Baseline antimetabolite use		**Baseline TNF-antagonist use**	
No	12	No	0
Yes	0	Yes	28
Baseline steroid use		**New fistula, abscess or definite abdominal mass**	
No	0	No	0
Yes	18	Yes	79

Total the points, then convert them to predicted risk using the below scale.

Total Points	**Predicted Risk**	**Total Points**	**Predicted Risk**
0	<10%	161	45%
37	15%	177	50%
65	20%	193	55%
88	25%	210	60%
109	30%	227	65%
127	35%	247	70%
144	40%	>266	>75%

Table 4.9. Original and optimism-corrected classification table of the CD-related models for selected risk thresholds.

Surgical Model, Original										
Threshold	NB	Interventions avoided*	TP	FP	FN	TN	TPR	FPR	PPV	NPV
6.8% (Prev.)	0.02	27.1	73	506	57	1262	56.2%	28.6%	12.6%	95.7%
7.0%	0.02	28.3	73	472	57	1296	56.2%	26.7%	13.4%	95.8%
8.0%	0.017	33.4	61	354	69	1414	46.9%	20.0%	14.7%	95.3%
9.0%	0.014	37.9	49	251	81	1517	37.7%	14.2%	16.3%	94.9%
10.0%	0.013	42.9	45	187	85	1581	34.6%	10.6%	19.4%	94.9%
Surgery Model, Validated										
Threshold	NB	Interventions avoided*	TP	FP	FN	TN	TPR	FPR	PPV	NPV
6.8% (Prev.)	0.02	27.2	78	524	52	1244	60.0%	29.6%	13.0%	96.0%
7.0%	0.02	28.4	73	489	57	1279	56.2%	27.7%	13.0%	95.7%
8.0%	0.017	33.4	63	359	67	1409	48.5%	20.3%	14.9%	95.5%
9.0%	0.014	37.9	49	248	81	1520	37.7%	14.0%	16.5%	94.9%
10.0%	0.013	43.1	45	181	85	1587	34.6%	10.2%	19.9%	94.9%
Complication Model, Original										
Threshold	NB	Interventions avoided*	TP	FP	FN	TN	TPR	FPR	PPV	NPV
26.6% (Prev.)	0.056	15.8	295	507	209	887	58.5%	36.4%	36.8%	80.9%
27.0%	0.054	16.3	281	482	223	912	55.8%	34.6%	36.8%	80.4%
28.0%	0.046	17.1	254	436	250	958	50.4%	31.3%	36.8%	79.3%
29.0%	0.041	18.4	238	381	266	1013	47.2%	27.3%	38.4%	79.2%
30.0%	0.036	19.8	210	337	294	1057	41.7%	24.2%	38.4%	78.2%
31.0%	0.029	20.8	187	298	317	1096	37.1%	21.4%	38.6%	77.6%
32.0%	0.024	22.1	163	253	341	1141	32.3%	18.1%	39.2%	77.0%
Complication Model, Validated										
Threshold	NB	Interventions avoided*	TP	FP	FN	TN	TPR	FPR	PPV	NPV
26.6% (Prev.)	0.056	15.6	296	512	208	882	58.7%	36.7%	36.6%	80.9%
27.0%	0.054	16.2	282	485	222	909	56.0%	34.8%	36.8%	80.4%
28.0%	0.046	17.1	252	432	252	962	50.0%	31.0%	36.8%	79.2%
29.0%	0.04	18.3	234	372	270	1022	46.4%	26.7%	38.6%	79.1%
30.0%	0.035	19.5	204	324	300	1070	40.5%	23.2%	38.6%	78.1%
31.0%	0.028	20.6	175	278	329	1116	34.7%	19.9%	38.6%	77.2%
32.0%	0.023	21.9	154	236	350	1158	30.6%	16.9%	39.5%	76.8%

Notes: NB, net benefit, TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; TPR, true positive rate (sensitivity); FPR, false positive rate (1 - specificity).

** Interventions avoided per 100 people.*

Chapter 5

Discussion and conclusions

5.1 Discussion of results

Two clinical models were developed and internally developed for predicting risk of experiencing CD-related complications and CD-related surgery within 24 months of follow up, using data from one of the largest clinical trials of CD patients to date [76]. Although both models demonstrated high specificity and negative predictive values, the surgery model had greater predictive ability, likely because surgery is an easily defined and more objective clinical outcome, in distinction to disease-related complications. In addition, the development of a score chart using basic clinical variables which are readily calculable facilitates implementation of this score in the out-patient clinic setting, allowing enhanced decision making between patients and their physicians.

The baseline predictors identified in the CD-related prediction models, especially disease activity, presence of fistula, abscess or abdominal mass, and medication use, were all prospectively defined and measured. This is a distinct advantage to retrospective cohort studies in which a greater potential exists for bias during the data collection process.

Use of this model may help to discriminate those at higher risk of disease-related complications and surgery and thus may benefit most from more intensive treatment and or combination therapy, or closer monitoring for development of disease related symptoms that require planned surgery (e.g. obstructive symptoms in the setting of fibrostenotic disease).

Prediction of CD-related complications is limited by the use of non-standardized or broadly inclusive outcome definitions (**Table 2.2**). This may explain the lower performance of the model for prediction of CD-related complications. The time frame for prediction must also be considered. Since the greatest risk of complications occurs within the first few years after diagnosis, the early reports of long-term risk prediction models that were developed in this area have become confounded or outdated with evolving treatment strategies, most notably the introduction and widespread use of biologic drugs. For example, some models were developed in the pre-biologic era, while others define the need for thiopurines as advanced disease, but currently thiopurine monotherapy represents a treatment strategy generally reserved for less severe disease rather than severe disease [260, 261].

Existing prediction models have been previously described for CD-related outcomes (summarized in Table 7) which have been limited by selection or referral centre bias, small sample size, evolving drug treatments and management strategies, long-term time horizons, or broadly inclusive outcome definitions. Nevertheless, it is notable that many of the items identified as independent predictors have previously been shown to have prognostic value in population-based cohorts with longer-term follow up [1, 3, 81, 141, 144, 262]. Some alternative prediction models have incorporated one or more genetic factors (e.g, *NOD2* risk allele; **Table 2.2**). In principle, the inclusion of genetic predictors can increase the prognostic ability of the model [103, 178, 179].

An important clinical concern is how best to risk stratify patients according to disease severity. This ability may be critical to determine the optimal choice and timing of drug therapies [263]. These models demonstrate the ability to acceptably discriminate those at

higher risk of disease-related complications and surgery, and thus who may benefit most from more intensive management such as early combined immunosuppression, and reduce over-treatment and the associated risk of adverse outcomes [260]. Prediction models can ideally help clinicians to more rapidly identify those at high risk of progression to severe disease while managing those at low risk with conventional treatment.

A major strength of the study is use of data from large RCT conducted across community centres, meaning that the endpoints were well-defined and data complete to 24 months of follow up. The pragmatic nature of the trial design meant that consecutive patients were enrolled, regardless of disease activity, phenotype or treatment and an algorithm of care applied which mirrors clinical practice. Thus, we believe that these models are generalizable to routine clinical care.

Several limitations should be acknowledged. First, participants in the REACT study had a longer average duration of disease and the operating characteristics of these models in newly diagnosed patients remains to be determined. Second, the participants of REACT were recruited from community clinics which could realistically recruit 60 patients. Thus, the source population is more reflective of adults living within the catchment areas of urban community clinics, such as those associated with teaching hospitals and universities. Third, the performance of the CD-related surgery model performed better than the complication model, which may be related to using a more well-defined and objective outcome definition. Fourth, the REACT study did not include biomarkers, serological or genetic factors and thus these factors could not be incorporated into the current models. Fifth, the model was validated in the same cohort in which is was developed and thus independent validation in an external dataset is required. We further opted to use the age at baseline rather than other models in CD risk prediction which use age at diagnosis. The reason for this was practical, since age at diagnosis did not show any relationship on risk prediction and was highly correlated with age at baseline ($p = 0.74$). While the patient population tended to be recruited more urban centres, these centres were balanced on CD caseload (>100 and <100 patients in the practice). It was not possible to investigate the

distribution of these patients in terms of CD behaviour according to Montréal classification, though it is expected that this study population would resemble the average urban community practice due to continuous enrollment from may of these centres. However, it would be unlikely that this study population match either remote and rural populations or those with severe disease who are referred to specialized clinics.

What follows next are some general remarks and consideration of developing and using prediction models.

5.2 Additional considerations concerning clustered data

Clustered data frequently arise in risk modelling, such as observing individuals within families, or patients within clinics. Examples when clustering designs may be used are when interventions are most appropriately or most feasibly assigned to groups of people, especially when the investigators are concerned about contamination of the treatment, or when it is not possible to acquire an individual's informed consent, such as a trial involving treatment for acute cardiac arrest or stroke. Patients in the same centre are expected to be more alike than patients from different centres and patients can no longer be assumed to be independent. Therefore, it is most appropriate to acknowledge this clustering in subsequent analyses, including prognostic modelling. This added complexity will also have direct consequences on model performance.

5.2.1 Modelling choices for cluster randomization designs

Clinical prediction models have traditionally been based on non-hierarchical or flat designs, thereby ignoring clustering by using standard regression models. Even if clinical prediction models are developed from clustered data sets, they are scarcely developed using techniques that explicitly or implicitly take this structure into account.

Regressions methods for clustered data are needed when one wishes to account for clustered designs, such as cluster-randomized or multi-centre trials [264]. The two most popular regression methods for analyzing clustered data are the mixed effect (also called hierarchical or multilevel) models (more specifically, random intercept models) and generalized estimating equations (GEEs) (e.g., [265–270]). These models allow for cluster-specific (conditional) or population-averaged (marginal) interpretations [271]. Another means of adjusting for clustering could be to fit a GEE or standard logistic model, and apply a robust variance estimator to account for the clustering effect [272]. Although the literature on estimation of multi-level methods is quite extensive, issues specific to prediction have been less extensively characterized. This perhaps contributes to the less rapid adoption of clinical prediction models accounting for clustering.

Accounting for clustering is most important when one is interested in making inferences about the effect size of a variable, since clustering designs can significantly inflate the standard error. However, the coefficients of each variable will be quite similar when using standard or mixed-effect models. Thus, for the purposes of prediction of risk at the individual level, the standard error is not needed to estimate the individual's risk. The issue of clustering in the REACT trial was initially explored by constructing random-intercept logistic regression models. However, the cluster-specific models agreed closely with the standard logistic regression models. Therefore, the standard logistic regression models were preferred.

5.2.2 Options for accounting for clustering

It is not popular to make predictions for individuals using mixed effects models. A simple approach is to ignore clustering and use standard regression methods. These are marginalized (population-averaged) predictions. An appeal of this method is that model performance is primarily based on regression coefficient estimates which are less affected by clustering effects than are the standard errors. As a result, prediction performance

should not suffer greatly [257, 273–275]. Another option is to assume an average cluster effect. This can be done in a mixed effects model by setting the random intercept to its expected value (zero) prior to making predictions, as was done previously [36, 256]. A related option is that the centre-specific (conditional) effect can be estimated from the outcome prevalence at the new centre [257] or to assume the intercept of a related centre [276]. These choices may make sense for predictions about patients from very similar or even the same centre. Lastly, a mixed effects model may be converted to a marginal model by integrating over the cluster-specific effects to produce marginal predictions analogous to those from standard regression models [277]. A rebuttal to the approach of using simple regression is that mixed effects models may be used to derive any other related model, thus preserving information fidelity and maximizing the potential uses for the model.

5.2.3 When should clustering be taken into account?

Consideration of the level at which predictions are desired and the purpose of the model are beginning to be emphasized in the literature where clustered observations are specifically concerned [257, 278]. The investigator and analyst should be forced to consider whether the scientific question concerns prediction about new individuals or new centres. If the goal is to implement a prediction model at a provincial or national level, then performance at that level should be considered. In contrast, when risk models should be used for decision-making within specific centres, then centre-level performance is of interest. If the prediction model will be used at a level higher than the cluster, such as implementing national screening or policy guidelines, then the population-level performance may be of greater interest. The choice of which type of model to build and which type of predictions to use was recently organized into a proposed decision framework by Wynants and colleagues [257]. In this work, they consider model performance using logistic regression on clustered data [257]. They showed that calibration and discrimination performance can vary not only as a consequence of over-fitting or under-fitting in the developed model, but also because of the choice of standard or mixed effects logistic model, and the level at

which discrimination and calibration are considered (i.e., individual or cluster level) [257]. They find that when the intraclass cluster correlation is low ($p < 0.05$), a standard logistic model will strictly be mis-calibrated and have lower discrimination compared to a mixed effects model, but this difference is negligible. However, the deviations will become much more serious as the clustering effect grows large (i.e., $p > 0.20$).

In the context where patients are clustered within hospitals, the ICC is typically below 0.15 [256, 279]. This is certainly the case for the REACT trial [76]. In those cases, the marginalized predictions will strictly be mis-calibrated but this error will be small with sufficient events per variable. However, the mis-calibration will become much more serious as the ICC increases [256, 257, 277, 279]. Marginal prediction models tend to be well calibrated at the population level but mis-calibrated at the centre level, while mixed effects models show the opposite trend, particularly for datasets with relatively few observations [257]. On the other hand, conditional predictions from mixed effects models can be well calibrated at both levels [257]. When using marginal predictions derived from mixed effects logistic regression model [277], it is possible to obtain calibrated results [279]. Predictive performance in clustered data contexts have been assessed at the population level [279] and also distinguishing between performance at both the population and centre levels [36, 270, 280].

In the specific case of the REACT trial, the focus was on the point estimate for risk, rather than its variance. Certainly, the degree of clustering was not severe for either CD-related surgery or complication outcomes (ICC $p < 0.05$). Simulations under these scenarios showed that the different types of predictions that can be obtained from mixed effects models yield quite similar results about individual risk compared to standard regression models [257]. Under these conditions, when only marginal or average centre-specific effects are assumed, standard and mixed effects models produced similar discrimination ability at the levels of the population or centre [36, 257]. In other words, the multi-level model performs reasonably similarly to standard regression models. However, conditional predictions tended to produce superior discrimination at the population-level when

clustering was greater than 5% and especially when it was more severe [257].

5.3 Final remarks on model building

The model building and predictor selection process is central to the construction of prediction models as much as any other regression analysis. What follows is a general overview of methods and concerns as they relate to these issues.

The choice of model is to be informed by the type of outcome available. As previously mentioned, the two most common outcomes and their corresponding regression models used in clinical prediction models are to predict binary outcomes using logistic regression and time-to-event outcomes using Cox proportional hazards regression. In the REACT dataset, time-to-event data was recorded in days, with a slight bias to experience events in year one compared to year two. The main interest from these models was only to predict whether the outcome would occur in the two year period, as opposed to the time to the particular event, which was deemed sufficient for informing potential therapeutic decisions.

5.3.1 Comments on predictor selection and model simplification

Variable selection is just as important a concern for specifying a reasonably correct model. In the case of the REACT models, predictor selection was informed by a manual literature search and mainly guided by the judgment of expert gastroenterologists. This agrees with the approach advocated by Steyerberg [20] and Harrell Jr. [21], for example. However, much work has been done to develop semi-automated methods of variable selection that try to select the most important variables while limiting bias from trial-and-error approaches. It is important to remember that selection predictors should still integrate domain knowledge or expert opinion.

There is less agreement on whether treatment allocation should be included in the final predictive model. Partly related to this choice is whether the prediction question to be answered is interested in specific effects of treatment, or a marginal effect. Steyerberg noted that a commonly observed phenomena from clinical trials is that treatment allocation does have a relatively small effect on outcome, even when statistically significant, in which case the relevance of other predictors are more important [20]. It was observed in the REACT models that inclusion of treatment allocation group made a negligible impact on the prediction and performance of the CD-related surgery model. We also grouped both treatment arms together to produce a marginal prediction that can more broadly be applied in different health care settings where intensive therapy may not be economically feasible, and the specific choice of treatment algorithm may vary, since these are not yet standardized approaches in the field of gastroenterology.

Harrell [21] favours a backward elimination bootstrap variable selection procedure as superior to forward selection (of the most significant predictor), backwards elimination (of the least significant predictor) or step-wise selection methods. In work by Derksen and Keselmen [281], there was little practical significance on the size of the sample in determining the number of selected variables. Instead, the number of candidate predictors, and especially the correlation between independent predictors, had a strong influence on the number of noise variables that enter the model, specifically, the greater the degree of correlation among true predictors led to increasingly greater fractions of noise variables being selected as predictors [281]. Overall, they found that noise variables were selected between 20-74% of the time using these methods.

The bootstrap selection method should in theory correct for bias in variable selection [282]. The procedure selects any step-wise variable selection method and applies it to a large number of bootstrap samples (often 500 or 1,000). The frequency with which each variable was selected in the resulting models is tabulated and a model composed of the most frequently selected variables by the selection procedure is built. The correlation among candidate predictors is then examined and the redundant variables may be re-

moved. The final model is then selected based on the conditional relative frequency of the remaining variables. The bootstrap method is advantageous where there are large numbers of noise variables [282] yet can fail when there is a high degree of correlation among candidate predictors (i.e., in situations of multicollinearity). The latter problem of collinearity is a limitation of stepwise variable selection in general, and is not limited to just bootstrap methods. When there are no clear patterns in the resulting set of chosen predictors, the final choice of the model still requires external criteria such as expert knowledge, model simplification or shrinkage.

There are several methods of model simplification by removing or shrinking irrelevant predictors. Linear shrinkage of predictors has already been discussed in the context of the REACT models. Harrell has suggested using a step-down approach [21] in which the predictors that have the weakest relationship with the linear predictor are removed and the smaller, remaining model reasonably approximates the full model. One advantage to this approach is that reportedly avoids over-fitting since any penalization that has been applied remains in the linear predictor [21]. Similarly, the backwards elimination approach is guided by a rule that drops variables if they fail to reach a significance threshold (often α = 0.05 or 0.10). This process occurs iteratively until there are no more non-significant variables to remove [283]. A third method would be to focus on optimizing the Akaike information criterion (AIC). The step-down and backwards elimination approaches tended to produce similarly performing models (based on both the c-statistic and Brier score) when the model included some irrelevant predictors [283]. However, model simplification based on the AIC was a successful strategy, regardless of the above simplification methods [283]. This was desirable because it sometimes produced models that performed even better than the full model and its ease of application [283, 284]. Some conditions of the AIC stopping rule are that the model and predictors should be fully pre-specified and that the method works best when the models considered are "nested" (subsets of a larger model) [285]. However, when the number of observations is small, the risk of over-fitting and extreme bias in variable selection is high. Here, step-wise methods are to be avoided in favour of using expert knowledge or developing full models and employing some method

of shrinkage of the coefficients by penalized regression [30, 38, 43, 68, 284].

5.3.2 When should a prediction model be used?

Presume for this discussion that the hypothetical clinical prediction model has been validated and shows favourable performance characteristics. When such a model is available, what are the barriers to their use? The most common reasons that clinical prediction models are not used are because clinicians are either unaware of their existence or that they are unsure of how to take advantage of such models [286, 287]. The former is remedied by education, but the latter requires implicit trust in the modelling process and preferably to understand the workings of the model. For these reasons, traditional regression models are still preferred to data-driven approaches such as artificial neural networks and classification and regression trees, despite no clear superiority of any one modelling technique [286, 287]. Several other barriers to use have been discussed in the literature, including model performance that is not superior to clinical judgment or care-as-usual, the clinician did not fully adhere to the (changes in) risk-dependent decisions, or increased attention to the risk prediction may lead to increased treatment rather than as a result of increased need for treatment (i.e., a Hawthorne effect) [288].

The advantage of using a predictive model is to produce a risk estimate or suggest a decision that is tailored to the individual and thereby offering better risk management and avoiding unnecessary investigations or diagnostics [27]. However, it is not enough just to validate a clinical prediction model. Ideally, two conditions must be met for successful adoption of prediction rules. First, the clinician behaviour must be modified to use the model as a decision aid. Second, the model must accurately differentiate patients with and without the disease or outcome so that there is measurable benefit to patient care. In this way, the model complements the clinician's skill and abilities, the patient benefits from better risk management, naturally leading to better cost-effectiveness.

When a candidate model is produced, if the clinician is the intended user, it is worth considering whether the model can augment their own diagnostic or prognostic abilities. There is some evidence from systematic reviews and meta-analyses to suggest that clinical judgment can be just as sensitive as the model to diagnose the disease state, yet the models may have superior specificity [289, 290]. In other words, no more cases would have been identified by using a prediction model, yet using the model could reduce the need for potentially time-consuming, expensive or invasive diagnostic testing in order to rule out disease. Some clinical prediction models have been tested in formal impact assessments in which aided clinical management had superior performance to management by clinical judgment alone. For example, the bacterial pneumonia score was tested against the unaided clinician's judgment [291]. The group managed according to the risk prediction model used significantly less antibiotics and did not experience increased treatment failure [291]. The use of the Padua prediction score, assessing the risk of venous thromboembolism in hospitalized patients, was also superior to clinical judgment alone [292]. In contrast, there are some outcomes in which a prediction model can inform risk-dependent treatment strategies to achieve better patient care. One example is the use of model for postoperative nausea and vomiting in which low-risk patients can reasonably be managed by usual care, whereas high-risk patients benefit from management according to the model [288]. There is still much room to provide clinicians with decision support tools. However, the decision of whether or not to implement such models are best decided in a head-to-head controlled trial to determine if the aided clinical judgment performs better than clinical expertise alone and should also account for the relative harms or costs of misclassification errors.

5.4 Conclusion

In summary, we have developed and validated clinical prediction models for CD-related surgery and complications using data from a large, pragmatic RCT, with good overall performance for predicting the outcome of surgery within 24 months. We have

transformed these models into scoring tools facilitating their use in the clinic and these now require external validation.

References

1. Henriksen, M., Jahnsen, J., Lygren, I., Aadland, E., *et al.* Clinical course in Crohn's disease: results of a five-year population-based follow-up study (the ibsen study). *Scand J Gastroenterol* **42,** 602–10 (2007).
2. Peyrin-Biroulet, L., Loftus, E. V., Colombel, J.-F. & Sandborn, W. J. The natural history of adult Crohn's disease in population-based cohorts. *Am J Gastroenterol* **105,** 289–97 (2010).
3. Bernell, O., Lapidus, A. & Hellers, G. Risk factors for surgery and postoperative recurrence in Crohn's disease. *Ann Surg* **231,** 38–45 (2000).
4. Riss, S., Schuster, I., Papay, P., Mittlböck, M., *et al.* Repeat intestinal resections increase the risk of recurrence of Crohn's disease. *Dis Colon Rectum* **56,** 881–7 (2013).
5. Feagan, B. G., Bala, M., Yan, S., Olson, A., *et al.* Unemployment and disability in patients with moderately to severely active Crohn's disease. *J Clin Gastroenterol* **39,** 390–5 (2005).
6. Van der Have, M., Fidder, H. H., Leenders, M., Kaptein, A. A., *et al.* Self-reported disability in patients with inflammatory bowel disease largely determined by disease activity and illness perceptions. *Inflamm Bowel Dis* **21,** 369–77 (2015).
7. Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97,** 1837–47 (1998).
8. Stiell, I. G., McKnight, R. D., Greenberg, G. H., McDowell, I., *et al.* Implementation of the ottawa ankle rules. *JAMA* **271,** 827–32 (1994).
9. Viennois, E., Zhao, Y. & Merlin, D. Biomarkers of inflammatory bowel disease: from classical laboratory tools to personalized medicine. *Inflamm Bowel Dis* **21,** 2467–74 (2015).
10. Jones, J., Loftus, E. V., Panaccione, R., Chen, L.-S., *et al.* Relationships between disease activity and serum and fecal biomarkers in patients with Crohn's disease. *Clin Gastroenterol Hepatol* **6,** 1218–24 (2008).
11. Hanauer, S. B., Feagan, B. G., Lichtenstein, G. R., Mayer, L. F., *et al.* Maintenance infliximab for Crohn's disease: the accent i randomised trial. *Lancet* **359,** 1541–9 (2002).
12. Colombel, J.-F., Sandborn, W. J., Ghosh, S., Wolf, D. C., *et al.* Four-year maintenance treatment with adalimumab in patients with moderately to severely active ulcerative colitis: data from ultra 1, 2, and 3. *Am J Gastroenterol* **109,** 1771–80 (2014).
13. Colombel, J. F., Rutgeerts, P. J., Sandborn, W. J., Yang, M., *et al.* Adalimumab induces deep remission in patients with Crohn's disease. *Clin Gastroenterol Hepatol* **12,** 414–422.e5 (2014).
14. Pineton de Chambrun, G., Peyrin-Biroulet, L., Lémann, M. & Colombel, J.-F. Clinical implications of mucosal healing for the management of IBD. *Nat Rev Gastroenterol Hepatol* **7,** 15–29 (2010).
15. Rutgeerts, P., Van Assche, G., Sandborn, W. J., Wolf, D. C., *et al.* Adalimumab induces and maintains mucosal healing in patients with Crohn's disease: data from the extend trial. *Gastroenterology* **142,** 1102–1111.e2 (2012).

16. Colombel, J. F., Reinisch, W., Mantzaris, G. J., Kornbluth, A., *et al.* Randomised clinical trial: deep remission in biologic and immunomodulator naïve patients with Crohn's disease - a sonic post hoc analysis. *Aliment Pharmacol Ther* **41,** 734–746 (2015).

17. Peyrin-Biroulet, L., Reinisch, W., Colombel, J.-F., Mantzaris, G. J., *et al.* Clinical disease activity, C-reactive protein normalisation and mucosal healing in Crohn's disease in the sonic trial. *Gut* **63,** 88–95 (2014).

18. Schnitzler, F., Fidder, H., Ferrante, M., Noman, M., *et al.* Long-term outcome of treatment with infliximab in 614 patients with Crohn's disease: results from a single-centre cohort. *Gut* **58,** 492–500 (2009).

19. Frøslie, K. F., Jahnsen, J., Moum, B. A., Vatn, M. H., *et al.* Mucosal healing in inflammatory bowel disease: results from a Norwegian population-based cohort. *Gastroenterology* **133,** 412–22 (2007).

20. Steyerberg, E. *Clinical prediction models* (Springer International Publishing, Switzerland, 2009).

21. Harrell Jr., F. E. *Regression modeling strategies* 2nd ed., (Springer International Publishing, Switzerland, 2015).

22. Hendriksen, J. M. T., Geersing, G. J., Moons, K. G. M. & de Groot, J. A. H. Diagnostic and prognostic prediction models. *J Thromb Haemost* **11,** 129–141 (2013).

23. Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., *et al.* Prognosis research strategy (progress) 3: prognostic model research. *PLoS Med* **10,** e1001381 (2013).

24. Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E., *et al.* Prognosis and prognostic research: what, why, and how? *BMJ* **338,** b375 (2009).

25. Collins, G. S., Reitsma, J. B., Altman, D. G., Moons, K. G. M., *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Circulation* **131,** W1–W73 (2015).

26. Reilly, B. M. & Evans, A. T. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* **144,** 201–9 (2006).

27. Laupacis, A., Sekar, N. & Stiell, I. G. Clinical prediction rules. a review and suggested modifications of methodological standards. *JAMA* **277,** 488–94 (1997).

28. Moons, K. G. M., Altman, D. G., Vergouwe, Y. & Royston, P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* **338,** b606 (2009).

29. Post, P. N., de Beer, H. & Guyatt, G. H. How to generalize efficacy results of randomized trials: recommendations based on a systematic review of possible approaches. *J Eval Clin Pract* **19,** 638–43 (2013).

30. Harrell Jr., F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15,** 361–87 (1996).

31. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *BMJ* **332,** 1080 (2006).

32. Royston, P., Altman, D. G. & Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25,** 127–41 (2006).

33. Altman, D. G. Problems in dichotomizing continuous variables. *Am J Epidemiol* **139,** 442–5 (1994).

34. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., *et al.* A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* **49,** 1373–9 (1996).

35. Austin, P. C. & Steyerberg, E. W. Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res,* 1–13 (2014).

36. Wynants, L., Bouwmeester, W., Moons, K. G. M., Moerbeek, M., *et al.* A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol* **68,** 1406–14 (2015).

37. Vittinghoff, E. & McCulloch, C. E. Relaxing the rule of ten events per variable in logistic and cox regression. *Am J Epidemiol* **165,** 710–8 (2007).

38. Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E. & Habbema, J. D. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* **19,** 1059–79 (2000).

39. Sun, G. W., Shook, T. L. & Kay, G. L. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* **49,** 907–16 (1996).

40. Paradis, C. Bias in surgical research. *Ann Surg* **248,** 180–8 (2008).

41. Debray, T. P. A., Damen, J. A. A. G., Snell, K. I. E., Ensor, J., *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356,** i6460 (2016).

42. Miller, M. E., Hui, S. L. & Tierney, W. M. Validation techniques for logistic regression models. *Stat Med* **10,** 1213–26 (1991).

43. Spiegelhalter, D. J. Probabilistic prediction in patient management and clinical trials. *Stat Med* **5,** 421–33 (1986).

44. Harrell Jr., F. E., Lee, K. L., Califf, R. M., Pryor, D. B., *et al.* Regression modelling strategies for improved prognostic prediction. *Stat Med* **3,** 143–52 (1984).

45. Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., *et al.* Validation of probabilistic predictions. *Med Decis Making* **13,** 49–58 (1993).

46. Hosmer, D. W., Hosmer, T., Le Cessie, S. & Lemeshow, S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* **16,** 965–80 (1997).

47. Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., *et al.* A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* **74,** 167–76 (2016).

48. Brier, G. W. Verifications of forecasts expressed in terms of probability. *Mon Weather Rev* **78,** 1–4 (1950).

49. Murphy, A. A new vector partition of the probability score. *J Appl Meteorol* **12,** 595–600 (1973).

50. Arkes, H. R., Dawson, N. V., Speroff, T., Harrell, F. E., *et al.* The covariance decomposition of the probability score and its use in evaluating prognostic estimates. support investigators. *Med Decis Making* **15,** 120–31 (1995).

51. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78,** 691–692 (1991).

52. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* **26,** 565–74 (2006).

53. Vickers, A. J. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *Am Stat* **62,** 314–320 (2008).

54. Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M. J., *et al.* Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* **54,** 774–81 (2001).

55. Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., *et al.* Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* **56,** 441–7 (2003).

56. Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E. & Habbema, J. D. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* **21,** 45–56 (2001).

57. Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., *et al.* Risk prediction models: i. development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* **98,** 683–90 (2012).

58. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Series B Stat Methodol* **36,** 111–47 (1974).

59. Efron, B. How biased is the apparent error rate of a prediction rule? *J Am Stat Assoc* **81,** 461–470 (1986).

60. Copas, J. B. Regression, prediction and shrinkage. *J R Stat Soc Series B Stat Methodol* **45,** 311–354 (1983).

61. Copas, J. B. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* **6,** 167–83 (1997).

62. Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Application of shrinkage techniques in logistic regression analysis: a case study. *Stat Neerl* **55,** 76–88 (2001).

63. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodol* **58,** 267–288 (1996).

64. Tibshirani, R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* **58,** 267–88 (1996).

65. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat Med* **16,** 385–95 (1997).

66. Moons, K. G., Donders, A., Rogier, T., Steyerberg, E. W., *et al.* Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* **57,** 1262–70 (2004).

67. Verweij, P. J. & Van Houwelingen, H. C. Penalized likelihood in cox regression. *Stat Med* **13,** 2427–36 (1994).

68. Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., *et al.* How to develop a more accurate risk prediction model when there are few events. *BMJ* **351,** h3868 (2015).

69. Austin, E., Pan, W. & Shen, X. Penalized regression and risk prediction in genome-wide association studies. *Stat Anal Data Min* **6,** 315–328 (2013).

70. Altman, D. G. & Royston, P. What do we mean by validating a prognostic model? *Stat Med* **19,** 453–73 (2000).

71. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann Intern Med* **130,** 515–24 (1999).

72. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* **58,** 475–83 (2005).

73. Collins, G. S., Ogundimu, E. O. & Altman, D. G. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* **35,** 214–226 (2016).

74. Vergouwe, Y., Royston, P., Moons, K. G. M. & Altman, D. G. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol* **63,** 205–14 (2010).

75. Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., *et al.* Risk prediction models: ii. external validation, model updating, and impact assessment. *Heart* **98,** 691–8 (2012).

76. Khanna, R., Bressler, B., Levesque, B. G., Zou, G., *et al.* Early combined immunosuppression for the management of Crohn's disease (react): a cluster randomised controlled trial. *Lancet* **386,** 1825–34 (2015).

77. Gasche, C., Scholmerich, J., Brynskov, J., D'Haens, G., *et al.* A simple classification of Crohn's disease: report of the working party for the world congresses of gastroenterology, Vienna 1998. *Inflamm Bowel Dis* **6,** 8–15 (2000).

78. Silverberg, M. S., Satsangi, J., Ahmad, T., Arnott, I. D. R., *et al.* 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. *Can J Gastroenterol Hepatol* **19 Suppl A,** 5A–36A (2005).

79. Beaugerie, L., Seksik, P., Nion-Larmurier, I., Gendre, J.-P., *et al.* Predictors of Crohn's disease. *Gastroenterology* **130,** 650–6 (2006).

80. Solberg, I. C., Vatn, M. H., Høie, O., Stray, N., *et al.* Clinical course in Crohn's disease: results of a Norwegian population-based ten-year follow-up study. *Clin Gastroenterol Hepatol* **5,** 1430–8 (2007).

81. Golovics, P. A., Lakatos, L., Mandel, M. D., Lovasz, B. D., *et al.* Prevalence and predictors of hospitalization in Crohn's disease in a prospective population-based inception cohort from 2000-2012. *World J Gastroenterol* **21,** 7272–80 (2015).

82. Caprilli, R., Corrao, G., Taddei, G., Tonelli, F., *et al.* Prognostic factors for postoperative recurrence of Crohn's disease. Gruppo Italiano per lo Studio del Colon e del retto (GISC). *Dis Colon Rectum* **39,** 335–41 (1996).

83. Moum, B., Ekbom, A., Vatn, M. H., Aadland, E., *et al.* Clinical course during the 1st year after diagnosis in ulcerative colitis and Crohn's disease. results of a large, prospective population-based study in southeastern norway, 1990-93. *Scand J Gastroenterol* **32,** 1005–12 (1997).

84. Andreu, M., Márquez, L., Domènech, E., Gisbert, J. P., *et al.* Disease severity in familial cases of IBD. *J Crohns Colitis* **8,** 234–9 (2014).

85. Henriksen, M., Jahnsen, J., Lygren, I., Vatn, M. H., *et al.* Are there any differences in phenotype or disease course between familial and sporadic cases of inflammatory bowel disease? results of a population-based follow-up study. *Am J Gastroenterol* **102,** 1955–63 (2007).

86. Kanaan, Z., Ahmad, S., Bilchuk, N., Vahrenhold, C., *et al.* Perianal Crohn's disease: predictive factors and genotype-phenotype correlations. *Dig Surg* **29,** 107–114 (2012).

87. Schwartz, D. A., Loftus, E. V., Tremaine, W. J., Panaccione, R., *et al.* The natural history of fistulizing Crohn's disease in Olmsted County, Minnesota. *Gastroenterology* **122,** 875–80 (2002).

88. Ardizzone, S., MacOni, G., Sampietro, G. M., Russo, A., *et al.* Azathioprine and mesalamine for prevention of relapse after conservative surgery for Crohn's disease. *Gastroenterology* **127,** 730–740 (2004).

89. Eglinton, T. W., Roberts, R., Pearson, J., Barclay, M., *et al.* Clinical and genetic risk factors for perianal Crohn's disease in a population-based cohort. *Am J Gastroenterol* **107,** 589–96 (2012).

90. Nordenvall, C., Ekbom, A., Bottai, M., Smedby, K. E., *et al.* Mortality after total colectomy in 3084 patients with inflammatory bowel disease: a population-based cohort study. *Aliment Pharmacol Ther* **40,** 280–7 (2014).

91. Nemetz, A., Molnar, T., Zagoni, T., Kovacs, A., *et al.* Phenotypes defined by the "Vienna classification" in 100 Hungarian patients with Crohn's disease. *Rev Esp Enferm Dig* **95,** 533–8, 527–33 (2003).

92. Hofer, B., Böttger, T., Hernandez-Richter, T., Seifert, J. K., *et al.* The impact of clinical types of disease manifestation on the risk of early postoperative recurrence in Crohn's disease. *Hepatogastroenterology* **48,** 152–5 (2001).

93. Hellers, G., Bergstrand, O., Ewerth, S. & Holmström, B. Occurrence and outcome after primary treatment of anal fistulae in Crohn's disease. *Gut* **21,** 525–7 (1980).

94. Solberg, I. C., Lygren, I., Cvancarova, M., Jahnsen, J., *et al.* Predictive value of serologic markers in a population-based Norwegian cohort with inflammatory bowel disease. *Inflamm Bowel Dis* **15,** 406–14 (2009).

95. Vegh, Z., Kurti, Z., Gonczi, L., Golovics, P. A., *et al.* Association of extraintestinal manifestations and anaemia with disease outcomes in patients with inflammatory bowel disease. *Scand J Gastroenterol* **5521,** 1–7 (2016).

96. Jess, T., Winther, K. V., Munkholm, P., Langholz, E., *et al.* Mortality and causes of death in Crohn's disease: follow-up of a population-based cohort in copenhagen county, denmark. *Gastroenterology* **122,** 1808–14 (2002).

97. Lie, M. R. K. L., Kreijne, J. E. & van der Woude, C. J. Sex is associated with adalimumab side effects and drug survival in patients with Crohn's disease. *Inflamm Bowel Dis* **23,** 75–81 (2017).

98. Wolters, F. L., Russel, M. G., Sijbrandij, J., Ambergen, T., *et al.* Phenotype at diagnosis predicts recurrence rates in Crohn's disease. *Gut* **55,** 1124–30 (2006).

99. Yang, C. H., Ding, J., Gao, Y., Chen, X., *et al.* Risk factors that predict the requirement of aggressive therapy among chinese patients with Crohn's disease. *J Dig Dis* **12,** 99–104 (2011).

100. Loly, C., Belaiche, J. & Louis, E. Predictors of severe Crohn's disease. *Scand J Gastroenterol* **43,** 948–54 (2008).

101. Vester-Andersen, M. K., Vind, I., Prosberg, M. V., Bengtsson, B. G., *et al.* Hospitalisation, surgical and medical recurrence rates in inflammatory bowel disease 2003-2011 — a Danish population-based cohort study. *J Crohns Colitis* **8,** 1675–83 (2014).

102. Romberg-Camps, M. J. L., Dagnelie, P. C., Kester, a. D. M., Hesselink-van de Kruijs, M. a. M., *et al.* Influence of phenotype at diagnosis and of other potential prognostic factors on the course of inflammatory bowel disease. *Am J Gastroenterol* **104,** 371–83 (2009).

103. Dubinsky, M. C., Kugathasan, S., Kwon, S., Haritunians, T., *et al.* Multidimensional prognostic risk assessment identifies association between il12b variation and surgery in Crohn's disease. *Inflamm Bowel Dis* **19,** 1662–70 (2013).

104. Tarrant, K. M., Barclay, M. L., Frampton, C. M. A. & Gearry, R. B. Perianal disease predicts changes in Crohn's disease phenotype-results of a population-based study of inflammatory bowel disease phenotype. *Am J Gastroenterol* **103,** 3082–93 (2008).

105. Lakatos, P. L., Sipeki, N., Kovacs, G., Palyu, E., *et al.* Risk matrix for prediction of disease progression in a referral cohort of patients with Crohn's disease. *J Crohns Colitis* **9,** 891–8 (2015).

106. Holdstock, G., Savage, D., Harman, M. & Wright, R. Should patients with inflammatory bowel disease smoke? *Br Med J (Clin Res Ed)* **288,** 362 (1984).

107. Cosnes, J., Carbonnel, F., Beaugerie, L., Le Quintrec, Y., *et al.* Effects of cigarette smoking on the long-term course of Crohn's disease. *Gastroenterology* **110,** 424–31 (1996).

108. Cosnes, J., Carbonnel, F., Carrat, F., Beaugerie, L., *et al.* Effects of current and former cigarette smoking on the clinical course of Crohn's disease. *Aliment Pharmacol Ther* **13,** 1403–11 (1999).

109. Lakatos, P. L., Czegledi, Z., Szamosi, T., Banai, J., *et al.* Perianal disease, small bowel disease, smoking, prior steroid or early azathioprine/biological therapy are predictors of disease behavior change in patients with Crohn's disease. *World J Gastroenterol* **15,** 3504–10 (2009).

110. To, N., Gracie, D. J. & Ford, A. C. Systematic review with meta-analysis: the adverse effects of tobacco smoking on the natural history of Crohn's disease. *Aliment Pharmacol Ther* **43,** 549–61 (2016).

111. Parkes, G. C., Whelan, K. & Lindsay, J. O. Smoking in inflammatory bowel disease: impact on disease course and insights into the aetiology of its effect. *J Crohns Colitis* **8,** 717–25 (2014).

112. Lunney, P. C., Kariyawasam, V. C., Wang, R. R., Middleton, K. L., *et al.* Smoking prevalence and its influence on disease course and surgery in Crohn's disease and ulcerative colitis. *Aliment Pharmacol Ther* **42,** 61–70 (2015).

113. Ott, C., Takses, A., Obermeier, F., Schnoy, E., *et al.* Smoking increases the risk of extraintestinal manifestations in Crohn's disease. *World J Gastroenterol* **20,** 12269–76 (2014).

114. Reese, G. E., Nanidis, T., Borysiewicz, C., Yamamoto, T., *et al.* The effect of smoking after surgery for Crohn's disease: a meta-analysis of observational studies. *Int J Colorectal Dis* **23,** 1213–21 (2008).

115. Lawrance, I. C., Murray, K., Batman, B., Gearry, R. B., *et al.* Crohn's disease and smoking: is it ever too late to quit? *J Crohns Colitis* **7,** e665–71 (2013).

116. Kulaylat, A. N., Hollenbeak, C. S., Sangster, W. & Stewart, D. B. Impact of smoking on the surgical outcome of Crohn's disease: a propensity-score matched national surgical quality improvement program analysis. *Colorectal Dis* **17,** 891–902 (2015).

117. Frolkis, A. D., de Bruyn, J., Jette, N., Lowerison, M., *et al.* The association of smoking and surgery in inflammatory bowel disease is modified by age at diagnosis. *Clin Transl Gastroenterol* **7,** e165 (2016).

118. Nasir, B. F., Griffiths, L. R., Nasir, A., Roberts, R., *et al.* An envirogenomic signature is associated with risk of IBD-related surgery in a population-based Crohn's disease cohort. *J Gastrointest Surg* **17,** 1643–50 (2013).

119. Burisch, J., Pedersen, N., Cukovic-Cavka, S., Turk, N., *et al.* Environmental factors in a population-based inception cohort of inflammatory bowel disease patients in europe–an ecco-epicom study. *J Crohns Colitis* **8,** 607–16 (2014).

120. Lakatos, P. L., Vegh, Z., Lovasz, B. D., David, G., *et al.* Is current smoking still an important environmental factor in inflammatory bowel diseases? results from a population-based incident cohort. *Inflamm Bowel Dis* **19,** 1010–7 (2013).

121. Gearry, R. B., Richardson, A. K., Frampton, C. M., Dodgshun, A. J., *et al.* Population-based cases control study of inflammatory bowel disease risk factors. *J Gastroenterol Hepatol* **25,** 325–33 (2010).

122. Thia, K. T., Sandborn, W. J., Harmsen, W. S., Zinsmeister, A. R., *et al.* Risk factors associated with progression to intestinal complications of Crohn's disease in a population-based cohort. *Gastroenterology* **139,** 1147–55 (2010).

123. Gisbert, J. P., Marín, A. C. & Chaparro, M. Systematic review: factors associated with relapse of inflammatory bowel disease after discontinuation of anti-tnf therapy. *Aliment Pharmacol Ther* **42,** 391–405 (2015).

124. Sandborn, W. J., Melmed, G. Y., McGovern, D. P. B., Loftus, E. V., *et al.* Clinical and demographic characteristics predictive of treatment outcomes for certolizumab pegol in moderate to severe Crohn's disease: analyses from the 7-year precise 3 study. *Aliment Pharmacol Ther* **42,** 330–42 (2015).

125. Cosnes, J., Beaugerie, L., Carbonnel, F. & Gendre, J. P. Smoking cessation and the course of Crohn's disease: an intervention study. *Gastroenterology* **120,** 1093–9 (2001).

126. Nunes, T., Etchevers, M. J., Merino, O., Gallego, S., *et al.* High smoking cessation rate in Crohn's disease patients after physician advice–the tabaCrohn study. *J Crohns Colitis* **7,** 202–7 (2013).

127. De Bie, C., Ballet, V., Hendriks, N., Coenen, S., *et al.* Smoking behaviour and knowledge of the health effects of smoking in patients with inflammatory bowel disease. *Aliment Pharmacol Ther* **42,** 1294–302 (2015).

128. Saadoune, N., Peyrin-Biroulet, L., Baumann, C., Bigard, M.-A., *et al.* Beliefs and behaviour about smoking among inflammatory bowel disease patients. *Eur J Gastroenterol Hepatol* **27,** 797–803 (2015).

129. Franchimont, D., Belaiche, J., Louis, E., Simon, S., *et al.* Familial Crohn's disease: a study of 18 families. *Acta Gastroenterol Belg* **60,** 134–7 (1997).

130. Loftus, E. V. Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. *Gastroenterology* **126,** 1504–17 (2004).

131. Kevans, D., Silverberg, M. S., Borowski, K., Griffiths, A., *et al.* Ibd genetic risk profile in healthy first-degree relatives of Crohn's disease patients. *J Crohns Colitis* **10,** 209–15 (2016).

132. Moller, F. T., Andersen, V., Wohlfahrt, J. & Jess, T. Familial risk of inflammatory bowel disease: a population-based cohort study 1977-2011. *Am J Gastroenterol* **110,** 564–71 (2015).

133. Gabbani, T., Deiana, S., Annese, A. L., Lunardi, S., *et al.* The genetic burden of inflammatory bowel diseases: implications for the clinic? *Expert Rev Gastroenterol Hepatol* **4124,** 1–9 (2016).

134. Carbonnel, F., Macaigne, G., Beaugerie, L., Gendre, J. P., *et al.* Crohn's disease severity in familial and sporadic cases. *Gut* **44,** 91–5 (1999).

135. Trier Moller, F., Andersen, V., Andersson, M. & Jess, T. Hospital admissions, biological therapy, and surgery in familial and sporadic cases of inflammatory bowel disease: a population-based cohort study 1977-2011. *Inflamm Bowel Dis* **21,** 2825–32 (2015).

136. Shen, B., Remzi, F. H., Hammel, J. P., Lashner, B. A., *et al.* Family history of Crohn's disease is associated with an increased risk for Crohn's disease of the pouch. *Inflamm Bowel Dis* **15,** 163–70 (2009).

137. Burisch, J. & Munkholm, P. The epidemiology of inflammatory bowel disease. *Scand J Gastroenterol* **50,** 942–51 (2015).

138. Chow, D. K. L., Leong, R. W. L., Lai, L. H., Wong, G. L. H., *et al.* Changes in Crohn's disease phenotype over time in the chinese population: validation of the Montreal classification system. *Inflamm Bowel Dis* **14,** 536–41 (2008).

139. Louis, E., Collard, A., Oger, A. F., Degroote, E., *et al.* Behaviour of Crohn's disease according to the Vienna classification: changing pattern over the course of the disease. *Gut* **49,** 777–82 (2001).

140. Kalaria, R., Desai, D., Abraham, P., Joshi, A., *et al.* Temporal change in phenotypic behaviour in patients with Crohn's disease: do indian patients behave differently from western and other asian patients? *J Crohns Colitis* **10,** 255–61 (2016).

141. Peyrin-Biroulet, L., Harmsen, W. S., Tremaine, W. J., Zinsmeister, A. R., *et al.* Surgery in a population-based cohort of Crohn's disease from Olmsted County, Minnesota (1970-2004). *Am J Gastroenterol* **107,** 1693–701 (2012).

142. Solberg, I. C., Cvancarova, M., Vatn, M. H., Moum, B., *et al.* Risk matrix for prediction of advanced disease in a population-based study of patients with Crohn's disease (the ibsen study). *Inflamm Bowel Dis* **20,** 60–8 (2014).

143. Munkholm, P., Langholz, E., Davidsen, M. & Binder, V. Disease activity courses in a regional cohort of Crohn's disease patients. *Scand J Gastroenterol* **30,** 699–706 (1995).

144. Lovasz, B. D., Lakatos, L., Horvath, A., Szita, I., *et al.* Evolution of disease phenotype in adult and pediatric onset Crohn's disease in a population-based cohort. *World J Gastroenterol* **19,** 2217–26 (2013).

145. Pandey, A., Salazar, E., Kong, C. S. C., Lim, W. C., *et al.* Risk of major abdominal surgery in an asian population-based Crohn's disease cohort. *Inflamm Bowel Dis* **21,** 2625–33 (2015).

146. Lapidus, A., Bernell, O., Hellers, G. & Löfberg, R. Clinical course of colorectal Crohn's disease: a 35-year follow-up study of 507 patients. *Gastroenterology* **114,** 1151–60 (1998).

147. Lazarev, M., Huang, C., Bitton, A., Cho, J. H., *et al.* Relationship between proximal Crohn's disease location and disease behavior and surgery: a cross-sectional study of the IBD genetics consortium. *Am J Gastroenterol* **108,** 106–12 (2013).

148. Cosnes, J., Nion-Larmurier, I., Beaugerie, L., Afchain, P., *et al.* Impact of the increasing use of immunosuppressants in Crohn's disease on the need for intestinal surgery. *Gut* **54,** 237–41 (2005).

149. Ardizzone, S. & Porro, G. B. Perianal Crohn's disease: overview. *Dig Liver Dis* **39,** 957–8 (2007).

150. Ng, S. C., Zeng, Z., Niewiadomski, O., Tang, W., *et al.* Early course of inflammatory bowel disease in a population-based inception cohort study from 8 countries in asia and Australia. *Gastroenterology* **150,** 86–95.e3, quiz e13–4 (2016).

151. Present, D. H., Rutgeerts, P., Targan, S., Hanauer, S. B., *et al.* Infliximab for the treatment of fistulas in patients with Crohn's disease. *N Engl J Med* **340,** 1398–405 (1999).

152. Sands, B. E., Arsenault, J. E., Rosen, M. J., Alsahli, M., *et al.* Risk of early surgery for Crohn's disease: implications for early treatment strategies. *Am J Gastroenterol* **98,** 2712–8 (2003).

153. Sands, B. E., Anderson, F. H., Bernstein, C. N., Chey, W. Y., *et al.* Infliximab maintenance therapy for fistulizing Crohn's disease. *N Engl J Med* **350,** 876–85 (2004).

154. Targan, S. R., Hanauer, S. B., van Deventer, S. J., Mayer, L., *et al.* A short-term study of chimeric monoclonal antibody ca2 to tumor necrosis factor alpha for Crohn's disease. Crohn's disease ca2 study group. *N Engl J Med* **337,** 1029–35 (1997).

155. Ford, A. C., Sandborn, W. J., Khan, K. J., Hanauer, S. B., *et al.* Efficacy of biological therapies in inflammatory bowel disease: systematic review and meta-analysis. *Am J Gastroenterol* **106,** 644–59, quiz 660 (2011).

156. Behm, B. W. & Bickston, S. J. Tumor necrosis factor-alpha antibody for maintenance of remission in Crohn's disease. *Cochrane Database Syst Rev* (ed Bickston, S. J.) CD006893 (2008).

157. Pascua, M., Su, C., Lewis, J. D., Brensinger, C., *et al.* Meta-analysis: factors predicting post-operative recurrence with placebo therapy in patients with Crohn's disease. *Aliment Pharmacol Ther* **28,** 545–56 (2008).

158. Bell, S. J., Williams, A. B., Wiesel, P., Wilkinson, K., *et al.* The clinical course of fistulating Crohn's disease. *Aliment Pharmacol Ther* **17,** 1145–51 (2003).

159. Ng, S. C., Plamondon, S., Gupta, A., Burling, D., *et al.* Prospective evaluation of anti-tumor necrosis factor therapy guided by magnetic resonance imaging for Crohn's perineal fistulas. *Am J Gastroenterol* **104,** 2973–86 (2009).

160. Bitton, A., Dobkin, P. L., Edwardes, M. D., Sewitch, M. J., *et al.* Predicting relapse in Crohn's disease: a biopsychosocial model. *Gut* **57,** 1386–92 (2008).

161. Tougeron, D., Savoye, G., Savoye-Collet, C., Koning, E., *et al.* Predicting factors of fistula healing and clinical remission after infliximab-based combined therapy for perianal fistulizing Crohn's disease. *Dig Dis Sci* **54,** 1746–52 (2009).

162. Nasir, B. F., Griffiths, L., Nasir, A., Roberts, R., *et al.* Perianal disease combined with NOD2 genotype predicts need for IBD-related surgery in Crohn's disease patients from a population-based cohort. *J Clin Gastroenterol* **47,** 242–5 (2013).

163. Buisson, A., Chevaux, J.-B., Allen, P. B., Bommelaer, G., *et al.* Review article: the natural history of postoperative Crohn's disease recurrence. *Aliment Pharmacol Ther* **35,** 625–33 (2012).

164. Casillas, S. & Delaney, C. P. Laparoscopic surgery for inflammatory bowel disease. *Dig Surg* **22,** 135–42 (2005).

165. Cosnes, J., Gower-Rousseau, C., Seksik, P. & Cortot, A. Epidemiology and natural history of inflammatory bowel diseases. *Gastroenterology* **140,** 1785–94 (2011).

166. Bernell, O., Lapidus, A. & Hellers, G. Risk factors for surgery and recurrence in 907 patients with primary ileocaecal Crohn's disease. *Br J Surg* **87,** 1697–701 (2000).

167. Connelly, T. M. & Messaris, E. Predictors of recurrence of Crohn's disease after ileocolectomy: a review. *World J Gastroenterol* **20,** 14393–406 (2014).

168. Van Assche, G., Dignass, A., Reinisch, W., van der Woude, C. J., *et al.* The second european evidence-based consensus on the diagnosis and management of Crohn's disease: special situations. *J Crohns Colitis* **4,** 63–101 (2010).

169. Hellers, G. Crohn's disease in stockholm county 1955-1974. a study of epidemiology, results of surgical treatment and long-term prognosis. *Acta Chir Scand Suppl* **490,** 1–84 (1979).

170. Ng, S. C., Lied, G. A., Arebi, N., Phillips, R. K., *et al.* Clinical and surgical recurrence of Crohn's disease after ileocolonic resection in a specialist unit. *Eur J Gastroenterol Hepatol* **21,** 551–7 (2009).

171. Heimann, T. M., Greenstein, a. J., Lewis, B., Kaufman, D., *et al.* Comparison of primary and reoperative surgery in patients with Crohns disease. *Ann Surg* **227,** 492–5 (1998).

172. Olaison, G., Smedh, K. & Sjödahl, R. Natural course of Crohn's disease after ileocolic resection: endoscopically visualised ileal ulcers preceding symptoms. *Gut* **33,** 331–5 (1992).

173. Rutgeerts, P., Geboes, K., Vantrappen, G., Kerremans, R., *et al.* Natural history of recurrent Crohn's disease at the ileocolonic anastomosis after curative surgery. *Gut* **25,** 665–72 (1984).

174. Hugot, J. P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J. M., *et al.* Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* **379,** 821–3 (1996).

175. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–24 (2012).

176. Elding, H., Lau, W., Swallow, D. M. & Maniatis, N. Refinement in localization and identification of gene regions associated with Crohn disease. *Am J Hum Genet* **92,** 107–13 (2013).

177. Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47,** 979–86 (2015).

178. Cleynen, I., Boucher, G., Jostins, L., Schumm, L. P., *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387,** 156–67 (2016).

179. Siegel, C. A., Horton, H., Siegel, L. S., Thompson, K. D., *et al.* A validated web-based tool to display individualised Crohn's disease predicted outcomes based on clinical, serologic and genetic variables. *Aliment Pharmacol Ther* **43,** 262–71 (2016).

180. Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411,** 599–603 (2001).

181. Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411,** 603–6 (2001).

182. Adler, J., Rangwalla, S. C., Dwamena, B. A. & Higgins, P. D. R. The prognostic power of the NOD2 genotype for complicated Crohn's disease: a meta-analysis. *Am J Gastroenterol* **106,** 699–712 (2011).

183. Solon, J. G., Burke, J. P., Walsh, S. R. & Coffey, J. C. The effect of NOD2 polymorphism on postsurgical recurrence in Crohn's disease: a systematic review and meta-analysis of available literature. *Inflamm Bowel Dis* **19,** 1099–105 (2013).

184. Bhullar, M., Macrae, F., Brown, G., Smith, M., *et al.* Prediction of Crohn's disease aggression through NOD2/CARD15 gene sequencing in an Australian cohort. *World J Gastroenterol* **20,** 5008–5016 (2014).

185. Hugot, J.-P., Zaccaria, I., Cavanaugh, J., Yang, H., *et al.* Prevalence of CARD15/NOD2 mutations in caucasian healthy people. *Am J Gastroenterol* **102,** 1259–67 (2007).

186. Kathiresan, S., Larson, M. G., Vasan, R. S., Guo, C.-Y., *et al.* Contribution of clinical correlates and 13 C-reactive protein gene polymorphisms to interindividual variability in serum C-reactive protein level. *Circulation* **113,** 1415–23 (2006).

187. Fagan, E. A., Dyck, R. F., Maton, P. N., Hodgson, H. J., *et al.* Serum levels of C-reactive protein in Crohn's disease and ulcerative colitis. *Eur J Clin Invest* **12,** 351–9 (1982).

188. Boirivant, M., Leoni, M., Tariciotti, D., Fais, S., *et al.* The clinical significance of serum c reactive protein levels in Crohn's disease. results of a prospective longitudinal study. *J Clin Gastroenterol* **10,** 401–5 (1988).

189. Cellier, C., Sahmoud, T., Froguel, E., Adenis, A., *et al.* Correlations between clinical activity, endoscopic severity, and biological parameters in colonic or ileocolonic Crohn's disease. a prospective multicentre study of 121 cases. The Groupe d'etudes thérapeutiques des affections inflammatoires digestives. *Gut* **35,** 231–5 (1994).

190. Henriksen, M., Jahnsen, J., Lygren, I., Stray, N., *et al.* C-reactive protein: a predictive factor and marker of inflammation in inflammatory bowel disease. results from a prospective population-based study. *Gut* **57,** 1518–23 (2008).

191. Vermeire, S., Van Assche, G. & Rutgeerts, P. C-reactive protein as a marker for inflammatory bowel disease. *Inflamm Bowel Dis* **10,** 661–5 (2004).

192. Henderson, P., Kennedy, N. A., Van Limbergen, J. E., Cameron, F. L., *et al.* Serum C-reactive protein and CRP genotype in pediatric inflammatory bowel disease: influence on phenotype, natural history, and response to therapy. *Inflamm Bowel Dis* **21,** 596–605 (2015).

193. Suk Danik, J., Chasman, D. I., Cannon, C. P., Miller, D. T., *et al.* Influence of genetic variation in the C-reactive protein gene on the inflammatory response during and after acute coronary ischemia. *Ann Hum Genet* **70,** 705–16 (2006).

194. Solem, C. A., Loftus, E. V., Tremaine, W. J., Harmsen, W. S., *et al.* Correlation of C-reactive protein with clinical, endoscopic, histologic, and radiographic activity in inflammatory bowel disease. *Inflamm Bowel Dis* **11,** 707–12 (2005).

195. Koelewijn, C. L., Schwartz, M. P., Samsom, M. & Oldenburg, B. C-reactive protein levels during a relapse of Crohn's disease are associated with the clinical course of the disease. *World J Gastroenterol* **14,** 85–89 (2008).

196. Lémann, M., Mary, J.-Y., Colombel, J.-F., Duclos, B., *et al.* A randomized, double-blind, controlled withdrawal trial in Crohn's disease patients in long-term remission on azathioprine. *Gastroenterology* **128,** 1812–8 (2005).

197. Click, B., Vargas, E. J., Anderson, A. M., Proksell, S., *et al.* Silent Crohn's disease: asymptomatic patients with elevated C-reactive protein are at risk for subsequent hospitalization. *Inflamm Bowel Dis* **21,** 2254–61 (2015).

198. Hanauer, S. B., Sandborn, W. J., Rutgeerts, P., Fedorak, R. N., *et al.* Human anti-tumor necrosis factor monoclonal antibody (adalimumab) in Crohn's disease: the classic-i trial. *Gastroenterology* **130,** 323–33, quiz 591 (2006).

199. Melmed, G. Y., McGovern, D., Schreiber, S., Kosutic, G., *et al.* Early remission status predicts long-term outcomes in patients with Crohn's disease treated with certolizumab pegol. *Curr Med Res Opin,* 1–18 (2016).

200. Targan, S. R., Feagan, B. G., Fedorak, R. N., Lashner, B. A., *et al.* Natalizumab for the treatment of active Crohn's disease: results of the encore trial. *Gastroenterology* **132,** 1672–1683 (2007).

201. Sandborn, W. J., Colombel, J. F., Enns, R., Feagan, B. G., *et al.* Natalizumab induction and maintenance therapy for Crohn's disease. *N Engl J Med* **353,** 1912–25 (2005).

202. Schreiber, S., Khaliq-Kareemi, M., Lawrance, I. C., Thomsen, O. Ø., *et al.* Maintenance therapy with certolizumab pegol for Crohn's disease. *N Engl J Med* **357,** 239–50 (2007).

203. Sandborn, W. J., Feagan, B. G., Stoinov, S., Honiball, P. J., *et al.* Certolizumab pegol for the treatment of Crohn's disease. *N Engl J Med* **357,** 228–38 (2007).

204. Schreiber, S., Rutgeerts, P., Fedorak, R. N., Khaliq-Kareemi, M., *et al.* A randomized, placebo-controlled trial of certolizumab pegol (cdp870) for treatment of Crohn's disease. *Gastroenterology* **129,** 807–18 (2005).

205. Cornillie, F., Hanauer, S. B., Diamond, R. H., Wang, J., *et al.* Postinduction serum infliximab trough level and decrease of C-reactive protein level are associated with durable sustained response to infliximab: a retrospective analysis of the accent i trial. *Gut* **63,** 1721–7 (2014).

206. Reinisch, W., Wang, Y., Oddens, B. J. & Link, R. C-reactive protein, an indicator for maintained response or remission to infliximab in patients with Crohn's disease: a post-hoc analysis from accent i. *Aliment Pharmacol Ther* **35,** 568–76 (2012).

207. Jürgens, M., Mahachie John, J. M., Cleynen, I., Schnitzler, F., *et al.* Levels of C-reactive protein are associated with response to infliximab therapy in patients with Crohn's disease. *Clin Gastroenterol Hepatol* **9,** 421–7.e1 (2011).

208. Karmiris, K., Paintaud, G., Noman, M., Magdelaine-Beuzelin, C., *et al.* Influence of trough serum levels and immunogenicity on long-term outcome of adalimumab therapy in Crohn's disease. *Gastroenterology* **137,** 1628–40 (2009).

209. Kiss, L. S., Szamosi, T., Molnar, T., Miheller, P., *et al.* Early clinical remission and normalisation of CRP are the strongest predictors of efficacy, mucosal healing and dose escalation during the first year of adalimumab therapy in Crohn's disease. *Aliment Pharmacol Ther* **34,** 911–22 (2011).

210. Røseth, A. G., Schmidt, P. N. & Fagerhol, M. K. Correlation between faecal excretion of indium-111-labelled granulocytes and calprotectin, a granulocyte marker protein, in patients with inflammatory bowel disease. *Scand J Gastroenterol* **34,** 50–4 (1999).

211. Røseth, a. G., Fagerhol, M. K., Aadland, E. & Schjønsby, H. Assessment of the neutrophil dominating protein calprotectin in feces. a methodologic study. *Scand J Gastroenterol* **27,** 793–8 (1992).

212. Tibble, J., Teahon, K., Thjodleifsson, B., Roseth, A., *et al.* A simple method for assessing intestinal inflammation in Crohn's disease. *Gut* **47,** 506–13 (2000).

213. Costa, F., Mumolo, M. G., Bellini, M., Romano, M. R., *et al.* Role of faecal calprotectin as non-invasive marker of intestinal inflammation. *Dig Liver Dis* **35,** 642–7 (2003).

214. Kane, S. V., Sandborn, W. J., Rufo, P. A., Zholudev, A., *et al.* Fecal lactoferrin is a sensitive and specific marker in identifying intestinal inflammation. *Am J Gastroenterol* **98,** 1309–14 (2003).

215. Lamb, C. A., Mohiuddin, M. K., Gicquel, J., Neely, D., *et al.* Faecal calprotectin or lactoferrin can identify postoperative recurrence in Crohn's disease. *Br J Surg* **96,** 663–74 (2009).

216. Sipponen, T., Kärkkäinen, P., Savilahti, E., Kolho, K.-L., *et al.* Correlation of faecal calprotectin and lactoferrin with an endoscopic score for Crohn's disease and histological findings. *Aliment Pharmacol Ther* **28,** 1221–9 (2008).

217. Scarpa, M., D'Incà, R., Basso, D., Ruffolo, C., *et al.* Fecal lactoferrin and calprotectin after ileocolonic resection for Crohn's disease. *Dis Colon Rectum* **50,** 861–9 (2007).

218. Karczewski, J., Swora-Cwynar, E., Rzymski, P., Poniedziałek, B., *et al.* Selected biologic markers of inflammation and activity of Crohn's disease. *Autoimmunity* **48,** 318–27 (2015).

219. Sipponen, T., Savilahti, E., Kolho, K.-L., Nuutinen, H., *et al.* Crohn's disease activity assessed by fecal calprotectin and lactoferrin: correlation with Crohn's disease activity index and endoscopic findings. *Inflamm Bowel Dis* **14,** 40–6 (2008).

220. Schoepfer, A. M., Beglinger, C., Straumann, A., Trummler, M., *et al.* Fecal calprotectin correlates more closely with the simple endoscopic score for Crohn's disease (SES-CD) than CRP, blood leukocytes, and the CDAI. *Am J Gastroenterol* **105,** 162–9 (2010).

221. Sipponen, T., Björkesten, C.-G. a. F., Färkkilä, M., Nuutinen, H., *et al.* Faecal calprotectin and lactoferrin are reliable surrogate markers of endoscopic response during Crohn's disease treatment. *Scand J Gastroenterol* **45,** 325–31 (2010).

222. Falvey, J. D., Hoskin, T., Meijer, B., Ashcroft, A., *et al.* Disease activity assessment in IBD: clinical indices and biomarkers fail to predict endoscopic remission. *Inflamm Bowel Dis* **21,** 824–31 (2015).

223. Zittan, E., Kelly, O. B., Kirsch, R., Milgrom, R., *et al.* Low fecal calprotectin correlates with histological remission and mucosal healing in ulcerative colitis and colonic Crohn's disease. *Inflamm Bowel Dis* **22,** 623–30 (2016).

224. D'Haens, G., Ferrante, M., Vermeire, S., Baert, F., *et al.* Fecal calprotectin is a surrogate marker for endoscopic lesions in inflammatory bowel disease. *Inflamm Bowel Dis* **18,** 2218–24 (2012).

225. Lobatón, T., López-García, A., Rodríguez-Moranta, F., Ruiz, A., *et al.* A new rapid test for fecal calprotectin predicts endoscopic remission and postoperative recurrence in Crohn's disease. *J Crohns Colitis* **7,** e641–51 (2013).

226. Naismith, G. D., Smith, L. A., Barry, S. J. E., Munro, J. I., *et al.* A prospective single-centre evaluation of the intra-individual variability of faecal calprotectin in quiescent Crohn's disease. *Aliment Pharmacol Ther* **37,** 613–21 (2013).

227. Yamamoto, T., Shiraki, M., Bamba, T., Umegae, S., *et al.* Faecal calprotectin and lactoferrin as markers for monitoring disease activity and predicting clinical recurrence in patients with Crohn's disease after ileocolonic resection: a prospective pilot study. *United European Gastroenterol J* **1,** 368–74 (2013).

228. Herranz Bachiller, M. T., Barrio Andres, J., Fernandez Salazar, L., Ruiz-Zorrilla, R., *et al.* The utility of faecal calprotectin to predict post-operative recurrence in Crohn's disease. *Scand J Gastroenterol* **51,** 720–6 (2016).

229. Boschetti, G., Laidet, M., Moussata, D., Stefanescu, C., *et al.* Levels of fecal calprotectin are associated with the severity of postoperative endoscopic recurrence in asymptomatic patients with Crohn's disease. *Am J Gastroenterol* **110,** 865–72 (2015).

230. Gecse, K. B., Brandse, J. F., van Wilpe, S., Löwenberg, M., *et al.* Impact of disease location on fecal calprotectin levels in Crohn's disease. *Scand J Gastroenterol* **50,** 841–7 (2015).

231. Tibble, J. A., Sigthorsson, G., Bridger, S., Fagerhol, M. K., *et al.* Surrogate markers of intestinal inflammation are predictive of relapse in patients with inflammatory bowel disease. *Gastroenterology* **119,** 15–22 (2000).

232. Kallel, L., Ayadi, I., Matri, S., Fekih, M., *et al.* Fecal calprotectin is a predictive marker of relapse in Crohn's disease involving the colon: a prospective study. *Eur J Gastroenterol Hepatol* **22,** 340–5 (2010).

233. Silberer, H., Küppers, B., Mickisch, O., Baniewicz, W., *et al.* Fecal leukocyte proteins in inflammatory bowel disease and irritable bowel syndrome. *Clin Lab* **51,** 117–26 (2005).

234. Schröder, O., Naumann, M., Shastri, Y., Povse, N., *et al.* Prospective evaluation of faecal neutrophil-derived proteins in identifying intestinal inflammation: combination of parameters does not improve diagnostic accuracy of calprotectin. *Aliment Pharmacol Ther* **26,** 1035–42 (2007).

235. Mosli, M. H., Zou, G., Garg, S. K., Feagan, S. G., *et al.* C-reactive protein, fecal calprotectin, and stool lactoferrin for detection of endoscopic activity in symptomatic inflammatory bowel disease patients: a systematic review and meta-analysis. *Am J Gastroenterol* **110,** 802–19, quiz 820 (2015).

236. Rutgeerts, P., Feagan, B. G., Lichtenstein, G. R., Mayer, L. F., *et al.* Comparison of scheduled and episodic treatment strategies of infliximab in Crohn's disease. *Gastroenterology* **126,** 402–413 (2004).

237. Sandborn, W. J., Hanauer, S. B., Rutgeerts, P., Fedorak, R. N., *et al.* Adalimumab for maintenance treatment of Crohn's disease: results of the classic ii trial. *Gut* **56,** 1232–9 (2007).

238. Colombel, J.-F., Sandborn, W. J., Rutgeerts, P., Enns, R., *et al.* Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the charm trial. *Gastroenterology* **132,** 52–65 (2007).

239. Sandborn, W. J., Lee, S. D., Randall, C., Gutierrez, A., *et al.* Long-term safety and efficacy of certolizumab pegol in the treatment of Crohn's disease: 7-year results from the precise 3 study. *Aliment Pharmacol Ther* **40,** 903–16 (2014).

240. Schreiber, S., Colombel, J.-F., Bloomfield, R., Nikolaus, S., *et al.* Increased response and remission rates in short-duration Crohn's disease with subcutaneous certolizumab pegol: an analysis of precise 2 randomized maintenance trial data. *Am J Gastroenterol* **105,** 1574–82 (2010).

241. Sandborn, W. J., Feagan, B. G., Rutgeerts, P., Hanauer, S., *et al.* Vedolizumab as induction and maintenance therapy for Crohn's disease. *N Engl J Med* **369,** 711–21 (2013).

242. Sands, B. E., Feagan, B. G., Rutgeerts, P., Colombel, J.-F., *et al.* Effects of vedolizumab induction therapy for patients with Crohn's disease in whom tumor necrosis factor antagonist treatment failed. *Gastroenterology* **147,** 618–627.e3 (2014).

243. Vermeire, S., Loftus, E. V., Colombel, J.-F., Feagan, B. G., *et al.* Long-term efficacy of vedolizumab for Crohn's disease. *J Crohns Colitis,* 1–26 (2016).

244. D'Haens, G., Baert, F., van Assche, G., Caenepeel, P., *et al.* Early combined immunosuppression or conventional management in patients with newly diagnosed Crohn's disease: an open randomised trial. *Lancet* **371,** 660–667 (2008).

245. Colombel, J. F., Sandborn, W. J., Reinisch, W., Mantzaris, G. J., *et al.* Infliximab, azathioprine, or combination therapy for Crohn's disease. *N Engl J Med* **362,** 1383–95 (2010).

246. Vermeire, S., Noman, M., Van Assche, G., Baert, F., *et al.* Effectiveness of concomitant immunosuppressive therapy in suppressing the formation of antibodies to infliximab in Crohn's disease. *Gut* **56,** 1226–31 (2007).

247. Scott, F. I. & Lichtenstein, G. R. Advances in therapeutic drug monitoring of biologic therapies in inflammatory bowel disease: 2015 in review. *Curr Treat Options Gastroenterol* **14,** 91–102 (2016).

248. Vande Casteele, N., Ferrante, M., Van Assche, G., Ballet, V., *et al.* Trough concentrations of infliximab guide dosing for patients with inflammatory bowel disease. *Gastroenterology* **148,** 1320–9.e3 (2015).

249. Lémann, M., Mary, J.-Y., Duclos, B., Veyrac, M., *et al.* Infliximab plus azathioprine for steroid-dependent Crohn's disease patients: a randomized placebo-controlled trial. *Gastroenterology* **130,** 1054–61 (2006).

250. Feagan, B. G., Sandborn, W. J., Gasink, C., Jacobstein, D., *et al.* Ustekinumab as induction and maintenance therapy for Crohn's disease. *N Engl J Med* **375,** 1946–1960 (2016).

251. Benitez, J.-M. & Louis, E. Can we predict the high-risk patient? *Dig Dis* **32,** 328–36 (2014).

252. Ryan, J. D., Silverberg, M. S., Xu, W., Graff, L. A., *et al.* Predicting complicated Crohn's disease and surgery: phenotypes, genetics, serology and psychological characteristics of a population-based cohort. *Aliment Pharmacol Ther* **38,** 274–83 (2013).

253. Ramadas, A. V., Gunesh, S., Thomas, G. A. O., Williams, G. T., *et al.* Natural history of Crohn's disease in a population-based cohort from Cardiff (1986-2003): a study of changes in medical treatment and surgical resection rates. *Gut* **59,** 1200–6 (2010).

254. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an abcd for validation. *Eur Heart J* **35,** 1925–31 (2014).

255. Harvey, R. F. & Bradshaw, J. M. A simple index of Crohn's disease activity. *Lancet* **1,** 514 (1980).

256. Bouwmeester, W., Twisk, J. W. R., Kappen, T. H., van Klei, W. A., *et al.* Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med Res Methodol* **13,** 19 (2013).

257. Wynants, L., Vergouwe, Y., Van Huffel, S., Timmerman, D., *et al.* Does ignoring clustering in multicenter data influence the performance of prediction models? a simulation study. *Stat Methods Med Res* **In Press,** 1–14 (2016).

258. Efron, B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* **78,** 316–31 (1983).

259. Sullivan, L. M., Massaro, J. M. & D'Agostino, R. B. Presentation of multivariate data for clinical use: the framingham study risk score functions. *Stat Med* **23,** 1631–60 (2004).

260. Dignass, A., Van Assche, G., Lindsay, J. O., Lémann, M., *et al.* The second european evidence-based consensus on the diagnosis and management of Crohn's disease: current management. *J Crohns Colitis* **4,** 28–62 (2010).

261. Gomollón, F., Dignass, A., Annese, V., Tilg, H., *et al.* 3. european evidence-based consensus on the diagnosis and management of Crohn's disease 2016: part 1: diagnosis and medical management. *J Crohns Colitis,* jjw168 (2016).

262. Nguyen, G. C., Nugent, Z., Shaw, S. & Bernstein, C. N. Outcomes of patients with Crohn's disease improved from 1988 to 2008 and were associated with increased specialist care. *Gastroenterology* **141,** 90–7 (2011).

263. Peyrin-Biroulet, L., Panés, J., Sandborn, W. J., Vermeire, S., *et al.* Defining disease severity in inflammatory bowel diseases: current and future directions. *Clin Gastroenterol Hepatol* **14,** 348–354.e17 (2016).

264. Sullivan, L. M., Dukes, K. A. & Losina, E. Tutorial in biostatistics. an introduction to hierarchical linear modelling. *Stat Med* **18,** 855–88 (1999).

265. Moreno, R. P., Metnitz, P. G. H., Metnitz, B., Bauer, P., *et al.* Modeling in-hospital patient survival during the first 28 days after intensive care unit admission: a prognostic model for clinical trials in general critically ill patients. *J Crit Care* **23,** 339–48 (2008).

266. Connelly, C. R., Laird, A., Barton, J. S., Fischer, P. E., *et al.* A clinical tool for the prediction of venous thromboembolism in pediatric trauma patients. *JAMA Surg* **151,** 50–7 (2016).

267. Wimmer, N. J., Spertus, J. A., Kennedy, K. F., Anderson, H. V., *et al.* Clinical prediction model suitable for assessing hospital quality for patients undergoing carotid endarterectomy. *J Am Heart Assoc* **3,** 1–8 (2014).

268. Lagu, T., Rothberg, M. B., Nathanson, B. H., Steingrub, J. S., *et al.* Incorporating initial treatments improves performance of a mortality prediction model for patients with sepsis. *Pharmacoepidemiol Drug Saf* **21 Suppl 2,** 44–52 (2012).

269. Kappen, T. H., Moons, K. G. M., van Wolfswinkel, L., Kalkman, C. J., *et al.* Impact of risk assessments on prophylactic antiemetic prescription and the incidence of postoperative nausea and vomiting: a cluster-randomized trial. *Anesthesiology* **120,** 343–54 (2014).

270. Van Klaveren, D., Steyerberg, E. W., Perel, P. & Vergouwe, Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* **14,** 5 (2014).

271. Hanley, J. A., Negassa, A., deB Edwardes, M. D. & Forrester, J. E. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* **157,** 364–75 (2003).

272. Williams, R. L. A note on robust variance estimation for cluster-correlated data. *Biometrics* **56,** 645–6 (2000).

273. Paccagnella, O. Sample size and accuracy of estimates in multilevel models new simulation results. *Methodology* **7,** 111–120 (2011).

274. Moineddin, R., Matheson, F. I. & Glazier, R. H. A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol* **7,** 34 (2007).

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

276. Snell, K. I. E., Hua, H., Debray, T. P. A., Ensor, J., *et al.* Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* **69,** 40–50 (2016).

277. Skrondal, A. & Rabe-Hesketh, S. Prediction in multilevel generalized linear models. *J R Stat Soc Ser A Stat Soc* **172,** 659–687 (2009).

278. Debray, T. P. A., Moons, K. G. M., Ahmed, I., Koffijberg, H., *et al.* A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* **32,** 3158–80 (2013).

279. Pavlou, M., Ambler, G., Seaman, S. & Omar, R. Z. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Med Res Methodol* **15,** 59 (2015).

280. Van Oirbeek, R. & Lesaffre, E. Assessing the predictive ability of a multilevel binary regression model. *Comput Stat Data Anal* **56,** 1966–1980 (2012).

281. Derksen, S. & Keselman, H. J. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol* **45,** 265–282 (1992).

282. Sauerbrei, W. & Schumacher, M. A bootstrap resampling procedure for model building: application to the cox regression model. *Stat Med* **11,** 2093–109 (1992).

283. Ambler, G., Brady, A. R. & Royston, P. Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med* **21,** 3803–22 (2002).

284. Van Houwelingen, J. C. & Le Cessie, S. Predictive value of statistical models. *Stat Med* **9,** 1303–25 (1990).

285. Breiman, L. The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. *J Am Stat Assoc* **87,** 738–754 (1992).

286. Adams, S. T. & Leveson, S. H. Clinical prediction rules. *BMJ* **344,** d8312 (2012).

287. Liao, L. & Mark, D. B. Clinical prediction models: are we building better mousetraps? *J Am Coll Cardiol* **42,** 851–3 (2003).

288. Kappen, T. H., Vergouwe, Y., van Wolfswinkel, L., Kalkman, C. J., *et al.* Impact of adding therapeutic recommendations to risk assessments from a prediction model for postoperative nausea and vomiting. *Br J Anaesth* **114,** 252–60 (2015).

289. Sanders, S., Doust, J. & Glasziou, P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. *PLoS One* **10,** e0128233 (2015).

290. Lucassen, W., Geersing, G.-j., Erkens, P. M. G., Reitsma, J. B., *et al.* Clinical decision rules for excluding pulmonary embolism: a meta-analysis. *Ann Intern Med* **155,** 448–60 (2011).

291. Torres, F. A., Pasarelli, I., Cutri, A., Ossorio, M. F., *et al.* Impact assessment of a decision rule for using antibiotics in pneumonia: a randomized trial. *Pediatr Pulmonol* **49,** 701–706 (2014).

292. Germini, F., Agnelli, G., Fedele, M., Galli, M. G., *et al.* Padua prediction score or clinical judgment for decision making on antithrombotic prophylaxis: a quasi-randomized controlled trial. *J Thromb Thrombolysis* **42,** 336–9 (2016).

Curriculum Vitae

Name: Leonardo Guizzetti

Education

M.Sc. Epidemiology and Biostatistics
Western University
London, Ontario, Canada
2015-2017

Ph.D. Medical Biophysics and Molecular Imaging
Western University
2009-2015

B.Sc. Medical Biophysics, Hon. Spec. (Physical Sciences)
Western University
2005-2009

Peer-Reviewed Publications

1. McGirr, R*, **Guizzetti, L***, Dhanvantari, S. The sorting of proglucagon to secretory granules is mediated by CPE and intrinsic sorting signals. Journal of Endocrinology, 17(2):229-40. (2013) *** denotes co-first author.**

2. **Guizzetti, L**, McGirr, R, Dhanvantari, S. Two Dipolar -helices within hormone-encoding regions of proglucagon are sorting signals to the regulated secretory pathway. Journal of Biological Chemistry, 23;289(21):14968-80. (2014)

3. **Guizzetti, L**. Total versus partial splenectomy in pediatric hereditary spherocytosis: A systematic review and meta-analysis. Pediatric Blood Cancer, 63(10):1713-22. (2016)

4. **Guizzetti, L**, Zou, GY, Khanna, R, Dulai, PS, Sandborn, WJ, Jairath, V, Feagan, BG. Development of Clinical Prediction Models for Surgery and Complications in Crohn's Disease. (*(2017) Under review at the time of this writing*)