DEVELOPMENT, EVALUATION AND APPLICATION OF
A PEDIATRIC ULCERATIVE COLITIS ACTIVITY INDEX
(PUCAI)

By
Dan Turner

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy in Clinical Epidemiology
Graduate Department of Health Policy, Management, and Evaluation
University of Toronto

© Copyright by Dan Turner 2008
ABSTRACT

Dan Turner
Development, evaluation and application of a Pediatric Ulcerative Colitis Activity Index (PUCAI)
Doctor of Philosophy, 2008
Department of Health Policy, Management, and Evaluation (HPME)
University of Toronto

This thesis uses the methods of psychometrics and clinimetrics to develop and evaluate a Pediatric Ulcerative Colitis Activity Index (PUCAI). The initial phases of item generation and reduction were performed previously. This thesis comprises five main studies.

Study one: the weighting and formatting of an initial draft PUCAI using a cohort of 157 children with ulcerative colitis, enrolled prospectively in five pediatric IBD centers.

Study two: the validation of the final draft on a separate prospective cohort of 48 children undergoing complete colonoscopy. The PUCAI was highly correlated with physician global assessment (PGA) (r=0.91), Mayo score (r=0.95) and colonoscopic appearance (r=0.77). The PUCAI was able to differentiate the different categories of disease activity, and cutoff points were defined.

Study three: Assessment of the responsiveness of the PUCAI. The index demonstrated excellent responsiveness on 75 children seen twice during the study period (effect size=1.9, standardized response mean=2.2, responsiveness statistics=2.6, correlation with PGA of change=0.84, and area under the ROC curve=0.97 95%CI 0.93-0.99).
Study four was aimed at evaluating the predictive validity of the PUCAI, on a retrospective cohort of 99 children with severe ulcerative colitis admitted for intravenous corticosteroid therapy. The PUCAI, calculated on the third and fifth day of therapy was highly predictive of therapy failure at discharge and one year post discharge (area under the ROC curve 0.84 (95%CI 0.76-0.92).

Study five: a methodological study evaluating the preferred way to determine the minimal clinically important difference (MCID) of health-related outcome measures. This study was conducted using the PUCAI and three other well established instruments. It was concluded that the MCID should be determined primarily by the anchor-based approach using the ROC curve method on the entire cohort, supplemented by calculating the minimal detectable difference beyond statistical error using the standard error of measurement. Small, moderate and large MCID values could be presented based on the degree of expected relevant change.

Together, these studies have contributed to the rigorous development and thorough evaluation of a novel, non-invasive tool for assessing disease activity in pediatric ulcerative colitis clinical studies and practice.
Acknowledgements

I could not have completed this work without the support and assistance of many others. I wish to express my sincere appreciation and gratitude to the following individuals:

To Hillary Steinhart, my thesis supervisor, for providing sage guidance at every step of the process, for providing full support and encouragement to my ideas and thoughts, for shielding me from distractions, for always being available to me and always with a smile and a positive attitude;

To Anne Griffiths, my fellowship supervisor and the one who instigated the idea of this thesis, for being an extraordinary mentor and an incredible guider, for thoughtfully and gradually building me from a young trainee to a researcher capable of experiencing independent studies, for encouraging my independence but not without active guidance, for funding my research initiatives, for being a fierce protector of my best interests, for helping me make the best choices, for putting me forward to the international community, for generating limitless opportunities and fostering my personal and professional growth. For being the one who influenced the most my time spent in Canada, and probably also the years to come;

To Anthony Otley, a member of my thesis committee and a friend, for allowing me to continue his work, for providing me with all the material necessary for succeeding, and for proving important comments along the way;

To Dorcas Beaton, the content expert on my committee, for providing patient and reliable explanations to my many questions and for always keeping the door open to me, for stimulating and challenging my views on measurements and for encouraging me to widen my thoughts beyond the boundaries of the field of Gastroenterology;

To two wise and helpful research coordinators, Krista Uusuoe RN and Robin McLernon, for being so helpful in all aspects of these studies and others, for providing important comments along the way, and for the support and friendship.

To the loves of my life-my amazing wife, Jane, and my three adorable children-Shira, Nadav and Amit, who followed me to Canada for my training years, for being so supportive of my career aspirations, for the willingness of making family sacrifices for it, for their understanding, love and friendship;

To my wise parents who assimilated in me the desire for spiritual and academic achievements, for the encouragements, words of wisdom, and everlasting active support.

To my true friends and colleagues in the fellowship program, Aleixo Muise, Eytan Wine and Thomas Walters, who were always there for me, for reflecting the truth when I needed it, for enriching my academic ideas and providing the best social and professional environment. Nothing would have been as fun without them;

To the Hospital for Sick Children and the doctors, nurses, and secretaries of the Division of Pediatric Gastroenterology, for providing me with an exciting and enjoyable workplace and access to the resources essential for completing this work.

To Gillian Hawker, Ahmed Bayoumi, Amber Gertzbein, Jennifer James, Gladys Honein, and Laurie McQuarrie for guidance with all matters related to the Clinical Epidemiology program;

and, To the Clinical Epidemiology Program and the Department of Health Policy, Management and Evaluation.

iv
# Table of Contents

**TITLE** ........................................................................................................................................................................... i  
**ACKNOWLEDGEMENTS** .................................................................................................................................................... iv  
**TABLE OF CONTENTS** ..................................................................................................................................................... v  
**LIST OF TABLES, FIGURES AND APPENDICES** ............................................................................................................... vii  
**CHAPTER 1: INTRODUCTION** ........................................................................................................................................... 1  
Inflammatory bowel diseases ................................................................................................................................................... 1  
Specificities of health related outcome measures ............................................................................................................. 3  
Review and critique of UC activity indices ......................................................................................................................... 5  
The need for a new measure of UC activity in children ...................................................................................................... 8  
Clinimetric and psychometric measures ............................................................................................................................ 10  
Aims and preliminary work .................................................................................................................................................... 13  
**CHAPTER 2: DEVELOPMENT AND EVALUATION OF THE PUCAI** .................................................................................. 15  
Summary .................................................................................................................................................................................. 16  
Background ............................................................................................................................................................................... 17  
Methods .................................................................................................................................................................................... 18  
Results ...................................................................................................................................................................................... 25  
Discussion ................................................................................................................................................................................ 31  
**CHAPTER 3: PREDICTING OUTCOME OF SEVERE PEDIATRIC UC** ...................................................................................... 36  
Summary .................................................................................................................................................................................. 37  
Background ............................................................................................................................................................................... 38  
Methods .................................................................................................................................................................................... 39  
Results ...................................................................................................................................................................................... 43  
Discussion ................................................................................................................................................................................ 49  
**CHAPTER 4: ROC CURVE AND THE MCID** ........................................................................................................................ 54  
Summary .................................................................................................................................................................................. 55  
Background ............................................................................................................................................................................... 56  
Methods .................................................................................................................................................................................... 58  
Results ...................................................................................................................................................................................... 63
Discussion.............................................................................................................................................67

**CHAPTER 5: DISCUSSION**....................................................................................................................75

Summary of research..................................................................................................................................75

Item generation and consensus methods.................................................................................................76

Item reduction..............................................................................................................................................77

Item weighting...........................................................................................................................................82

Sensibility....................................................................................................................................................84

Reliability..................................................................................................................................................85

Validity.......................................................................................................................................................90

Responsiveness..........................................................................................................................................95

Minimal clinically important difference...................................................................................................101

Limitation of the PUCAI............................................................................................................................114

Future directions.........................................................................................................................................115

**TABLES**...............................................................................................................................................117

**APPENDICES**.......................................................................................................................................131

**REFERENCES**........................................................................................................................................139
List of Tables

Table 1.1: Disease activity indices for ulcerative colitis……………………………………112
Table 1.2: Items and gradations in existing UC activity indices..................................113
Table 1.3: Evaluation phases of a theoretical Disease Activity Index (DAI)..................114
Table 2.1: Items reflecting disease activity in pediatric UC.........................................115
Table 2.2: Patient’s characteristics of the weighting and validation cohorts.................116
Table 2.3: Validation results of the PUCAI....................................................................117
Table 2.4: Responsiveness analysis of the PUCAI.........................................................117
Table 3.1: Characteristics of 99 children admitted for IVCS therapy..........................118
Table 3.2: Distribution of individual items at the 3rd day of IVCS therapy...............119
Table 3.3: Distribution of individual items at the 5th day of IVCS therapy...................120
Table 3.4: Diagnostic utility of indices on days three and five of therapy...............121
Table 4.1: Common methods used for calculating the MID..........................................122
Table 4.2a: Results of the MID for the PUCAI, PCDAI and RQLQ............................123
Table 4.2b: Results of the MID for the CRD domains................................................124
Table 5.1: Studies comparing methods to determine the MID....................................125
List of Figures

Figure 1.1: Basic structure of clinimetric and psychometric instruments.........................11
Figure 1.2: The judgmental process of deriving the initial PUCAI draft.........................14
Figure 2.1: Study design and outline of the PUCAI development.................................19
Figure 2.2a: Cutoff scores of the PUCAI.................................................................28
Figure 2.2b: Cutoff of change scores of the PUCAI..................................................30
Figure 3.1: Outcome of children with severe acute ulcerative colitis................................44
Figure 3.2: ROC curve predicting failure of intravenous corticosteroids.......................46
Figure 3.3: Kaplan Meier curve of time to colectomy in severe acute UC....................47
Figure 3.4: Transverse colon width of pediatric severe UC........................................48
Figure 3.5: Bowel width stratified by short term response to corticosteroids............49
Figure 4.1a-f: MCID of the PUCAI, PCDAI, CRQ, and RQLQ.................................64
Figure 5.1: Percent improvement of the PUCAI by physician global assessment........112
Figure 5.2: ROC curve of percent improvement of the PUCAI.................................112
List of Appendices

Appendix 2.1: Pediatric Ulcerative Colitis Activity Index (PUCAI).........................122
Appendix 2.2: Selected peer-review comments of the PUCAI manuscript.................124
Appendix 3.1: Selected peer-review comments of the severe colitis manuscript........128
Chapter 1: Introduction

INFLAMMATORY BOWEL DISEASE

Definitions and presentation

“Inflammatory bowel disease” (IBD) encompasses two major forms of chronic intestinal inflammation: Crohn's disease (CD) and ulcerative colitis (UC). Their causes have yet to be identified; hence the two conditions are currently defined empirically on the basis of clinical, endoscopic, and histological features. Several genetic polymorphisms confer susceptibility to CD [1], but the pathogenesis of UC is less well understood [2]. Recently, a novel genetic polymorphism in the Protein Tyrosine Phosphatase sigma gene was found to be associated with UC [3]. The inflammation in UC is confined to the colonic mucosa, beginning in the rectum and extending proximally in a continuous fashion. The inflammation in CD is transmural and can involve any part of the gastrointestinal tract, generally in segmental fashion with “skip lesions”. Both UC and CD develop during childhood or adolescence in up to 25% of patients, with UC often having an earlier age of onset than CD [4]. Although adult patients with UC often have disease confined to the distal colon, the majority of patients with pediatric-onset UC have extensive inflammation [4, 5].

The symptoms associated with CD are quite varied, in part because of the different anatomic locations involved. The manifestations of UC, however, are more uniform; typically, recurrent exacerbations of bloody diarrhea, and as the colitis becomes progressively more severe, associated abdominal pain, lethargy, and vomiting. In comparison to pediatric-onset CD, the intestinal inflammation of UC is much less often complicated by linear growth impairment.
Treatment and outcome assessment in IBD

The natural history of IBD is highly variable. The tendency to spontaneous worsening and improvement of intestinal inflammation mandates that the efficacy of therapies be established in the randomized controlled trial (RCT) setting. To date, most RCTs have been conducted among adult patients and efficacy in children has been extrapolated from these data. In UC, several RCTs have affirmed the efficacy of 5-aminosalicylic acid and sulfasalazine in the acute treatment of mild to moderate exacerbations, as well as in the maintenance of clinical remission [6, 7]. There is level I evidence of efficacy of corticosteroids [8, 9] and, recently also anti-tumor necrosis factor alpha drugs in the treatment of moderately severe UC [8, 10, 11]. As new biologic therapies for IBD are rapidly emerging, both adult and pediatric investigators recognize the need to optimize and standardize methodology for assessing disease activity in the clinical trial setting [12].

No single clinical or biochemical parameter consistently reflects activity of intestinal inflammation. Therefore, multi-attribute measures of disease activity have been developed for use in the clinical trial setting. Although other CD activity indices have been used [13], the Crohn's Disease Activity Index (CDAI) is generally accepted as the standard clinical outcome measure in adult CD trials [14]. The Pediatric Crohn's Disease Activity Index (PCDAI) has became the accepted disease activity measure in childhood CD [15]. It includes an item of height velocity, recognizing that linear growth is an important marker of disease activity among children with CD.
In contrast to CD, no single multi-attribute disease activity measure of UC has been consistently employed in either adult or pediatric clinical. Table 1.1 lists the multi-item measures used in UC clinical trials during the past 50 years [12]. These UC activity indices are, in general, not a product of careful item generation, gradation and weighting. Their validation has been an ad-hoc product of ongoing experience within clinical trials, without consistent results. A detailed discussion on each measure follows below.

SPECIFICITIES OF HEALTH-RELATED OUTCOME MEASURES

According to accepted standards of health indices development [16-18], the introduction of a measure for use in clinical research should follow a multi-step process of item generation, item reduction, setting response option (here-forth termed 'grading') weighting (optional) and evaluation. A detailed description of each step is presented in Chapter 5 of this thesis, and briefly mentioned here. The choices of the above evaluation phases are dictated by the conceptual framework, suggested by Kirshner et al. [16], and should be clearly defined prior to the development phase. An index may be developed for discriminative purposes (to distinguish different categories of disease state), evaluative purposes (for measuring longitudinal change, for which responsiveness assessment is crucial) and predictive purposes (predicting concurrent or future criterion). Often, a measure developed for one purpose will subsequently prove to be useful also for other purposes, as will be presented for the PUCAI, in Chapter 3 of this thesis.

After the purpose of the desired instrument has been defined, a pool of all potentially useful items is generated and then reduced to include only the most important ones. Items are graded for scoring and, if required by the conceptual framework,
weighted according to the item’s ability to explain a desired attribute. Once the instrument has been developed, it must be evaluated for sensibility, validity, reliability and if needed, also responsiveness. Sensibility includes an overall impression of the feasibility of the measure and whether it makes general sense (face and content validity). Validity is the degree to which the instrument measures the concept that it purports to measure and includes construct, predictive and criterion validity [19]. Validity is not a property of the instrument but a property of how it is used [17]. Therefore, it is meaningless to say that an instrument is just valid, without specifying the context in which validity was determined (for the PUCAI, specified below). The reliability of an instrument relates to its stability on repeated measures. In other words, it is the amount of inherent error of the measure, both random and systematic [20]. For psychometric multidimensional scales, internal consistency should be assessed, to reflect the inter-item correlation. Responsiveness is the ability of an instrument to accurately detect changes in disease activity over time, when it occurs. The issue is not merely sensitivity to change, but the ability to discriminate between those who improve and those who do not [21]. Highly responsive indices are preferred as they allow clinical trials to be performed with a smaller sample size [21-25]. This evaluative process is illustrated in Table 1.3, using a theoretical instrument aimed at measuring disease activity in pediatric IBD and acting as both discriminative and evaluative tool. The examples provided in the table are a few of many methods available, which are further discussed in Chapter 5.
The last step is to define cutoff scores that correspond to clinically important disease states (e.g. remission, mild, moderate and severe). For an evaluative measure, a definition of “response” (i.e. the minimal important difference), is also required.

REVIEW AND CRITIQUE OF UC ACTIVITY INDICES

The first and most frequently used activity index for UC [26] is the Truelove and Witts' classification that includes only five items (bloody diarrhea, fever, tachycardia, anemia and elevated ESR [8]). The index had two gradations: mild if none of the items were present and severe if all items were present. Secondary to the simple gradation and poor definition of moderate activity, there has been significant ambiguity in defining a longitudinal change in disease activity. Therefore, Truelove’s classification is widely used for identifying patients with severe disease, but not for assessing outcomes. It is considered a conservative classification and many different liberal modifications were introduced over the years. In a systematic review of cohort studies on severe UC, 8 of the 20 studies that used the Truelove and Witts’ classification utilized a liberal modification of the original index [27].

The need for an evaluative index for UC led Lichtiger et al. to develop a new index as part of a clinical trial that demonstrated the efficacy of cyclosporine in steroid-refractory UC (seldom referred to as the "modified Truelove and Witts index") [28]. The scoring system was not rigorously developed and its psychometric properties were never evaluated. Nonetheless, in view of the lack of a better alternative, the non-invasive Lichtiger index is still in common use [12, 27]. In a later trial, an attempt was made to quantify the final score of the Lichtiger index into disease activity categories (remission,
mild, moderate and severe), but these were arbitrarily determined [29]. Moreover, the definitions of the subjective items (i.e. degree of pain and general well-being) were never subjected to reliability assessment. The flaws of Lichtiger’s index apply also to the Clinical Activity Index, developed by Rachmilewitz et al. as part of a study comparing coated mesalazine with sulfasalazine [7]. Although colonoscopy was performed at the same time, it was not compared with the clinical index.

Another attempt to quantify Truelove and Witts’ classification was made by Seo et al. in 1992 [30]. By using multiple stepwise regression analysis, they derived the following equation: Activity Index=60Xblood in stool+13Xbowel movements+0.5XESR-4XHb-15Xalbumin+200. The flaws of this index are many. It was developed on a selected population of patients with severe disease requiring admission, and thus may not perform well in mild to moderate UC. The model was derived from statistical significance screening of 24 items. For Seo’s study, the sample size required for the model, according to the generally accepted guideline of 10 subjects per degree of freedom [31], would have been 250 patients while they included only 85 assessments (including repeated visits). The advantage of performing a univariate step before the multivariable analysis is the simplicity of the procedure and the model. However, items that are not significant in the univariate analysis may prove to be important in the multivariable modeling, and vice versa. Variable screening based on statistical significance involves multiple comparison problems that lead to unreliable models [31-35]. Furthermore, stepwise reduction is a poor method for variable selection as it prioritizes items with larger β estimates. It is, therefore, strongly recommended that the possible predictors be set a priori based on extensive literature review and expert opinion,
or by statistical methods that do not involve the outcome such as principal component analysis or summary indices [31, 34, 36, 37]. Fine tuning of the limited list may be aided by fitting a few models while trying to maximize c-statistics of binary models and $R^2$ of continuous models. The rationale for assessing the predictor variables without knowledge of the patient’s outcome is similar to the rationale for blinding the assessment of the outcome measure [37]. Another limitation is that Seo index was aimed to quantify the Truelove and Witts classification, but the latter is not consistently employed. Predictive validity was subsequently assessed, showing that all patients with a score $<180$ entered remission [38]. Responsiveness was assessed in a trial showing improved scores after infliximab therapy [39]. Despite the more rigorous development of this index in comparison with other UC instruments, it has not gained wide popularity.

Many of the limitations outlined for Seo’s index, also apply for the recently developed Endoscopic-Clinical Correlation Index (ECCI) [40]. Item reduction was based on statistical significance of univariate analysis. However, the criterion used for the item reduction, weighting and also for validation, was colonoscopy performed on 137 adults with UC. This should be regarded as a significant advantage, acknowledging that colonoscopy is the most important variable in assessing disease activity in UC [12]. This index has not been evaluated on a separate cohort, and its reliability, validity and responsiveness are yet to be determined.

Powell-Tuck et al. presented a different index that included symptoms, signs, laboratory data and sigmoidoscopic appearance [10], without rigorous development or psychometric evaluation. Poor correlation was found between this index and endoscopic appearance of the entire colon [41] but it is widely and effectively used in clinical trials,
where it has been responsive to change in clinical status. This index is simple to score and quantitative but is inappropriate for pediatric population since it requires sigmoidoscopy as part of each evaluation. This is also the case for the widely used Mayo-clinic index [6] (and the identical Disease Activity Index [42]) that include sigmoidoscopy and physician global assessment as part of four total items. To avoid inter-observer variation, it would be better to quantify the subjective global assessment of the physician, by means of symptom description. The Mayo score was never formally validated, but based on years of observation and use in clinical trials, the validity of this instrument is self evident [19]. Indeed, this index is currently the most widely used measure to assess outcome in clinical trials of adult UC [12].

Only two measures have been developed exclusively for children. The seldom used Lloyd-Still index [43] does not discriminate between CD and UC despite the major differences in disease manifestations. The index was poorly validated against clinical outcome of a few selected patients from the study group. Moreover, it requires radiological assessment at each evaluation and therefore is not practical. The second measure was introduced by Beattie et al. [44], who developed a Colitis Symptom Score as part of a study to assess mucosal healing following steroid therapy in children. It was not rigorously designed and included four clinical items that did not correlate well with colonoscopic findings. Neither index has been used in any major UC study.

THE NEED FOR A NEW MEASURE OF UC ACTIVITY IN CHILDREN

An instrument designed to measure disease activity should be simple to administer, reproducible (reliable between observers and over time), valid, and responsive (sensitive) to change. Based on the standard requirements of an outcome
measure, no available UC index can be clearly recommended above others [12, 26]. Considerable overlap exists in the items included in the various instruments but they do not share the same grading system (Table 1.2). For example, rectal bleeding is graded variably as “none/occasional/all” or “<50%/>50%/100%” or “little/present” or “streaks/obvious/most”.

In the context of a Pediatric IBD Clinical trials workshop, sponsored by the Crohn's and Colitis Foundation of America (CCFA) in 2004, a group of pediatric IBD experts was assembled to reach consensus concerning outcome assessment in pediatric IBD. Although recommendations for outcome assessment in CD made use of existing measures [15], the panel concluded that for UC, a novel instrument measuring disease activity in pediatric patients should be developed. Clinical trials in adult patients with UC most commonly include post-treatment endoscopic examination as an endpoint, with or without clinical criteria [12], but follow-up colonoscopy is not routinely performed and would not be well accepted at pediatric institutions. The results reported in this thesis concerning the development and evaluation of the Pediatric UC Activity Index (PUCAI) are the product of the research that began at that meeting.

The purpose of the PUCAI is to be used mainly in clinical trials of pediatric UC (hence the requirement of both evaluative and discriminative properties). The PUCAI should be able to detect change also in a short period of a few weeks and should perform well in both in- and outpatient setup including patients with mild to severe disease activity. However, children with fulminant colitis pending surgery are excluded from most trials and thus the PUCAI is not intended to evaluate disease activity in this extreme
side of the spectrum. The PUCAI should perform just as well whether it is used during exacerbation or at new onset of the disease. The PUCAI is not aimed at children with disease confined to the rectum, constituting a small minority of pediatric UC patients. Since the clinical presentation and therapy of these children are often quite different than in more extensive disease, they are typically excluded from most clinical trials.

CLINIMETRIC AND PSYCHOMETRIC MEASURES

Psychometric techniques have been used for many decades to guide the development of psychosocial outcome measures. Since no gold standard exists for measuring psychosocial concepts, such as depression or quality of life, multi-item measures are used. The strong emphasis that the original psychometric techniques placed on statistical methods led Feinstein to introduce clinimetric strategies [45] emphasizing clinical judgment in specifying the importance of clinical data [46].

In classic unidimensional psychometric measures (termed in this thesis “scales” as used by others [17, 47]), items are arranged in domains that, combined, can explain a certain subjective concept in psychosocial science [48]. Within each domain, items are highly correlated with one another and show high level of internal consistency. In contrast, multidimensional clinimetric measures (termed herein “indices” [17, 47]) are constructed so that each item taps a unique aspect of the phenomenon and internal consistency is not important (Figure 1.1). The aim in developing an index is to have only a few items so as to keep the instrument simple for use in clinical practice.
Feinstein argued that what is often missing from psychometric outcome measures is an analytic account of the clinical phenomena that is observed, judged and decided by clinicians or patients themselves [49]. In contrast, Streiner believes that the term “clinimetrics” imposes artificial and unnecessary differentiation from the well established psychometric field [48]. Streiner acknowledges the constructive differences between scales and indices, but he argues that psychometric properties and mathematical modeling could be frequently applied also to clinimetrics. The presentation of clinimetrics as a separate entity that involves only pure clinical input without the aid of statistics is confusing and misleading [50]. On the other hand, items should emerge from strong clinical relevance but may be reduced using judgmental or statistical methods. This is true for both scales and indices. Therefore, in this thesis, there will be no clear
differentiation between 'clinimetrics' and 'psychometrics' while developing and evaluating the PUCAI. Indeed, clinical judgment is increasingly incorporated in the development of psychometric scales [48, 51-53]. Item generation, reduction and grouping in domains of the Quality of Life After Myocardial Infarction Questionnaire (QLMI), a classic unidimensional psychometric scale, was based completely on clinical wisdom and not on factor analysis or other mathematical modeling [54]. Similarly, the Rheumatoid Arthritis Disease Activity Score is a clinical index developed using clear psychometric strategies, including factor analysis [55].

The overlapping strategies may lead to controversy as to how much clinical judgment and statistical consideration should be applied to each part of the measure development. Studies on psychometric scales have compared judgmental to mathematical item reduction [51-53, 56], resulting in different included items. The judgmental approach tends to provide weight scores more similar across items (i.e. regression to the mean). Nonetheless, mathematical modeling is not a sine qua non for accuracy as the psychometric approach was strongly criticized for the lack of face validity and sensibility in developing scales [46]. It seems, therefore, that a thoughtful combination of both mathematical methods with strong clinical consideration should follow the development of both scales and indices. As previously argued by Feinstein and others [46, 51], mathematical methods could be integrated into clinimetrics in order to better quantify clinical judgment, while on the other hand, the strong clinical views of physicians and patients using robust judgmental methods, should aid in the development of scales [48].

Accordingly, the development of the PUCAI was planned using combined
mathematical and judgmental approaches, in an attempt to take the best from each approach.

**AIMS AND PRELIMINARY WORK**

The aim of this thesis was to develop a non-invasive disease activity index, to be used primarily in clinical studies, but also in the description and assessment of patients in clinical practice. Although it was designed *a-priori* to be discriminative and evaluative in purpose, the predictive ability of the PUCAI is also assessed. The work included in this thesis was aimed to provide a ready to use instrument developed through rigorous judgmental and mathematical strategies, with complete psychometric evaluation (i.e. interobserver and test-retest reliability, validity, responsiveness, and cutoff values to categorize disease activity and treatment response). The choice of the change score required to define ‘response’ in clinical trials has a crucial impact on the interpretation of the study results. Therefore, the minimal important difference of the PUCAI was thoroughly evaluated, using different available strategies.

The initial steps of the PUCAI development were carried out before my involvement in the project, by two of my co-supervisors, Drs. Anne M. Griffiths and Anthony R. Otley. Under their direction, a Delphi group was assembled to generate items aided by extensive literature search (with input from Joep de Bruijne, during an undergraduate medical school research elective). Using the same Delphi results, items were reduced by rank order (see details in Chapter 2). Gradations were tabulated from existing measures and sent to the Delphi group for comments. My first involvement in
the PUCAI was in collecting the responses of the Delphi members and finalizing the gradations of the initial draft PUCAI (Figure 1.2).

Figure 1.2: The judgmental process of deriving the initial PUCAI draft
Chapter 2: Development and validation of a Pediatric Ulcerative Colitis Activity Index (PUCAI), and evaluation of reliability and responsiveness: A prospective multicenter study

The purpose of this chapter is to describe:

1) The construction of a Pediatric Ulcerative Colitis Activity Index (PUCAI).
2) The weighting of the draft PUCAI, on a prospective cohort of children with UC.
3) The validation process using a separate prospective cohort of UC children undergoing colonoscopy.
4) The evaluation of interobserver reliability on the weighting and validation cohorts.
5) The evaluation of test-retest reliability on a prospectively enrolled longitudinal cohort.
6) The evaluation of responsiveness of the PUCAI using a prospective longitudinal cohort of UC children.
7) The definition of cutoff points that correspond to remission, mild, moderate and severe disease activity, and the minimal clinically important difference,

Acknowledgement:


Selected comments obtained from the peer-review process (Gastroenterology), together with a point-by-point response, are presented in Appendix 2.2.
SUMMARY

Background: Colonoscopic appearance, the primary measure of disease activity in adult ulcerative colitis, is less acceptable to children.

Aim: To develop and evaluate a non-invasive activity index of pediatric ulcerative colitis.

Methods: Item selection was performed judgmentally using a Delphi group of 36 experts in pediatric IBD. Item weighting was performed by regression modeling using a prospective cohort of 157 pediatric ulcerative colitis patients. Validation was assessed on a separate prospective cohort of 48 children with ulcerative colitis undergoing complete colonoscopy. Responsiveness was evaluated at a follow-up visit of 75 children using effect size statistics, correlational and diagnostic utility approaches.

Results: A list of 41 items was generated and reduced to 11 by rank order. Two physicians completed the pediatric ulcerative colitis activity index (PUCAI) on each of the patients in the weighting cohort. Six clinical items were significant in the regression analysis; the laboratory items and an endoscopic appearance item did not improve the PUCAI performance. In the validation cohort, the PUCAI was highly correlated with physician global assessment (r=0.91, P<0.001), Mayo score (r=0.95, P<0.001) and colonoscopic appearance (r=0.77, P<0.001). Correlations were higher than two non-invasive adult indices calculated concurrently (i.e. the Seo and Lichtiger indices). Inter-observer and test-retest reliability were excellent (ICC=0.95; 95%CI 0.93-0.97 and ICC=0.94; 95%CI 0.87-0.97, respectively). Cutoff points were established using a receiver operator characteristic (ROC) curves on the full cohort. Excellent responsiveness was found at repeated visits (effect size=1.9, standardized response mean=2.2, responsiveness statistics=2.6, correlation with PGA of change=0.84, and area under the ROC curve=0.97 95%CI 0.93-0.99).

Conclusion: The rigorously developed PUCAI is a non-invasive, valid, highly reliable and responsive index with which to assess disease activity in pediatric ulcerative colitis.
BACKGROUND

Investigators recognize the need to optimize and standardize methodology for assessment of disease activity in clinical trials, but no single instrument has been consistently employed in UC [58]. Moreover, with the exception of the Seo index, which has seldom been selected for use in clinical trials, existing multi-item measures in UC have not been rigorously developed and evaluated [6-8, 10, 28, 30, 40, 59].

The most widely used UC activity index in adult clinical trials is the Mayo score that requires sigmoidoscopy at each assessment [12]. Direct examination of the colonic mucosa has become the most important variable for assessing disease activity in adult UC clinical trials [12, 58], but this is less acceptable in pediatrics. Pediatric gastroenterologists participating in the 2004 Crohn’s and Colitis Foundation of America (CCFA) Pediatric IBD clinical trials workshop reviewed the existing measures of UC activity, and concluded that a novel non-invasive instrument for use in pediatric patients should be developed.

According to the accepted standards of health index development [16], the introduction of any new index should follow a multi-step process of item generation, reduction, grading, weighting and then evaluation. An instrument that measures disease activity should be valid for its purpose, reliable and responsive. In this manuscript we describe the development and further assessment of the Pediatric Ulcerative Colitis Activity Index (PUCAI).
METHODS

Study design and overview (figure 2.1)

A Delphi group of pediatric IBD experts was established to aid with item generation, reduction and gradation. A Delphi technique involves an iterative, multi-step process to reach a consensus for a specified question, when the experts are not physically together. The responses of all members of the Delphi group are summarized by an expert panel, processed and resent to the group for further comments until consensus is reached. For the purpose of item weighting and subsequent index evaluation two separate prospective cohorts of UC children were enrolled between November 2005 to November 2006 at five pediatric IBD centers in North America (The Hospital for Sick Children, Toronto; Izaak Walton Killam Hospital, Halifax; Children’s Hospital of Eastern Ontario, Ottawa; Connecticut Children’s Medical Center, Connecticut; Children’s Hospital of Philadelphia, Pennsylvania). UC diagnosis was confirmed using established clinical and radiographic criteria [60]. Since some of the clinical features of UC do not apply to the same extent in ulcerative proctitis, patients with inflammation confined to the rectum were excluded. Left sided and extensive colitis were defined as macroscopic inflammation ending, respectively, distally or proximally to the splenic flexure [61].

Both hospitalized and ambulatory, whether newly or previously diagnosed, UC patients 2-18 years of age were eligible for enrollment. For the validation cohort, a requirement of complete colonoscopy at the assessment time was added to the inclusion criteria. Responsiveness and test-retest reliability were assessed using the subset of patients who were seen twice during the enrollment period. This study was approved by
the institutional review board of each participating center. In addition to informed parental consent, age-appropriate patient consent or assent was also obtained.

**Figure 2.1: Study design and outline of the PUCAI development**

Construct validity is a mini-theory to explain whether a measure acts the way it is expected based on the concept that it represents. Concurrent validity reflects the relationship of the index with another related measures, and convergent validity the relationship with another established index. Extreme group validity reflects whether the measure differentiates extreme sides of the disease spectrum. Inter-observer reliability reflects whether an index is reproducible between different raters at the same time, and test-retest reliability whether a score is reproducible at different times, when the patient remains stable. Responsiveness reflects the ability of an instrument to accurately detect change in disease activity over time, when it occurs.

**Item generation**

Items were first generated through a systematic review of the literature to identify variables used in existing UC indices. The preliminary list of items was distributed to a Delphi group of 48 pediatric gastroenterologists in North America, experienced in IBD, who were asked to comment on the list and identify further indicators.
Item reduction and instrument formatting

The Delphi group members were asked to rank each item on the complete list from 0 (not important) to 3 (extremely important). The results were analyzed using mean ratings and rank order of ratings. The 11 highest ranked variables were sent again to the Delphi group for further ranking, 1 to 11 according to the item's importance in explaining disease activity. This clinimetric strategy has been used previously in the development of numerous scales [51] and found to be robust [62].

In order to minimize inter-observer variability, logical gradations and clear definitions of items were established. Gradations used in other instruments were tabulated and reviewed. Proposed gradation schemes for each item of the PUCAI were distributed to the Delphi group, with the final instrument reflecting consensus opinion.

Item weighting and PUCAI derivation

The aim of this step was to finalize the index by removing unnecessary items, and to weight the gradations using a large prospective cohort. The purpose of a disease activity index is to establish an assessment of overall disease activity that is reproducible between different assessors. Therefore, the primary outcome variable for this step was physician global assessment (PGA), as measured by a visual analogue scale (0-100 mm) with anchors at the extremes (none and fulminant). We used a physician-based rather than a patient-based approach since we wished the outcome to reflect disease activity (in contrast to disease severity), which is best judged by experienced physicians [63].

For each patient, two independent pediatric gastroenterologists completed a PGA of disease activity following complete history taking and physical examination but before
scoring the PUCAI. The mean PGA score served as the dependent variable of the model and the PUCAI items as the explanatory variables. Additional baseline and laboratory data were recorded by a research coordinator. The resulting mathematically weighted PUCAI was sent to the Delphi group and was slightly revised. Inter-observer reliability was assessed by comparing the two independent PUCAI assessments of the two physicians.

**Validation of the derived PUCAI**

In the second cohort, four constructs of disease activity were used in order to achieve a comprehensive evaluation (Figure 2.1): 1) colonoscopic appearance; 2) PGA; 3) a reference invasive index (the most commonly used adult score, the Mayo index [6]); and 4) The Truelove and Witts' classification of disease activity [8]. Activity of macroscopic colitis was scored 0-3 at each of the colon segments (ascending, transverse, descending, sigmoid and rectum) using Beattie’s grading system of quiescent, mild, moderate or severe [44]. A final score (0-15) was calculated by adding all segmental scores [41, 64]. *A priori*, we expected good correlation with macroscopic colonoscopy score, PGA and the invasive Mayo index. Two other non-invasive adult indices (Seo [30] and Lichtiger [28]), were calculated simultaneously for comparison with the PUCAI's performance. The physicians who calculated the PUCAI in the validation cohort followed the same instructions outlined in the PUCAI user guide (appendix 2.1).
**Longitudinal assessment**

Responsiveness and test-retest reliability were established utilizing patients who were seen twice during the study period. At the second visit, in addition to the procedures described for the weighting cohort, the two physicians were also asked to answer the question “how much do you think the disease activity of this patient changed since the last visit?”. The change was graded using a 7-point Likert scale of change (i.e. -3. significantly worse, -2. moderately worse, -1. mildly worse, .0. almost the same, 1. mildly improved, 2. moderately improved and 3. significantly improved [65, 66]). Physicians were encouraged to review the notes from the previous visit, but were blinded to the previous PUCAI score. Our Likert scale of change met the required criteria for use as an external criterion, defined by Guyatt et al [67, 68]. The minimal clinically important difference (MCID) was set as the change score on the PUCAI with the highest combined sensitivity and specificity (obtained with receiver operating characteristic (ROC) curve) to differentiate improved (defined as “at least moderate improvement”, i.e. score of $\geq 2$ on the 7-point Likert scale) versus non-improved patients [66, 69, 70]. Two other change cutoffs, corresponding to physician grading of at least “1. mildly improved” and “3. significantly improved” on the Likert scale of change, were also calculated.

**Statistical analysis**

Data are presented as means ± standard deviation, or medians (interquartile range) and compared using one way analysis of variance (ANOVA) or Kruskal-Wallis on ranks, as appropriate for the distribution of normality. Correlations between individual parameters were sought using Spearman’s or Pearson’s correlations, as appropriate for
the distribution normality. To determine the weighting of each item, a multivariate regression analysis was used with the PUCAI items as the explanatory variable and PGA as the dependant variable. To test the contribution of endoscopic items to the total variance of the PUCAI, a similar model was fitted with a score reflecting an invasive version of the PUCAI as the dependant variable and the clinical items as the explanatory variables. The model $R^2$ was used to determine the added value of the endoscopic score to the total variance. The gradations of “blood in stool” and “stool frequency” followed linear escalation and, thus, were entered as continuous items. The assumptions of the final model were confirmed by plotting the residual distribution.

Serial ROC curves (±95% CI) were used to optimize gradations of individual PUCAI items and to define cutoff scores for categorical disease activity (none, mild, moderate and severe). Construct validation was based on correlation analysis between the index score and the construct. Correlation $r$ of 0-0.25 was considered as lack of correlation, 0.25-0.5-fair, 0.5-0.75 moderate to good and $>0.75$ very good to excellent correlation. Interobserver reliability was assessed on the two PUCAI scores calculated independently by two physicians at the same visit. Test-retest reliability was established by comparing the first and the follow-up PUCAI scores of patients whose disease activity was thought to be unchanged. Intraclass Correlation Coefficient (ICC, using Shrout and Fleiss’s 2,1 two way random ANOVA model ± 95% CI [71]) was used to assess reliability of each item and the total score. ICC $>0.9$ was considered as excellent reliability (and sufficient for individual patient level) and 0.8-0.9 as good reliability.
Changes in scores (labeled as the Greek letter Δ) were determined by subtracting the follow-up score from the initial score. Responsiveness was assessed using three different approaches and six statistical methods [21]:

1. **Effect size statistics**: represent the magnitude of change in the evaluated index [signal (observed change) to noise (some measure of variance) ratio tests]. The following statistical tests were used: i) Paired Student’s t test, comparing the baseline and follow-up PUCAI in the improved group, ii) Standardized response mean (SRM), calculated by dividing the mean ΔPUCAI, by its standard deviation [72], iii) Standardized effect size (SES), calculated by dividing the mean ΔPUCAI score, by the baseline standard deviation [73], and iv) Guyatt’s responsiveness statistic, calculated by dividing the PUCAI’s MCID with the SD of the ΔPUCAI score from the unchanged patients) [74]. A higher effect size statistic indicates a greater change effect and as a general rule a score of >0.8 is considered a large effect, 0.5-0.8 moderate and 0.2-0.5 small [66]. We expected to find a large effect (>0.8) in the group of patients whose disease activity was changed and a small effect (0.2-0.5) in the unchanged group.

2. **Correlational**: correlation with external measure of change (ΔPUCAI with ΔPGA and Likert scale of change).

3. **Diagnostic utility tests**: to differentiate the patients who changed at least moderately from unchanged patients (using ROC curves, sensitivity and specificity [21]). Area under the ROC curve of over 0.7 was considered indicative of ‘fair’ discriminative ability, 0.8 as ‘good’, and over 0.9 as ‘excellent’.

Sample size calculations were based on 10 patients/df [75] for the weighting phase (n=160) and further 47 patients for the validation phase (The correlation between...
PUCAI and the colonoscopy score was expected to be greater than 0.75; a sample size of 47 achieves 80% power to detect a difference of 0.25 between the alternative hypothesis and the null hypothesis correlation of 0.5 using a two-sided hypothesis test with a significance level of 0.05). For the reliability testing, we hypothesized that the ICC will be greater than 0.9, and thus approximately 30 observations are required [76]. Since duplicate assessment was performed on all patients, reliability was assessed on a much larger cohort than required.

Data were entered to a preprogrammed Access database with 3% missing data occurring at random, for which imputation was performed using the hot deck method [77]. No outcome data (i.e. PGA, colonoscopic score, Mayo score) were missing. All comparisons were made using two sided significance levels of P<0.05. Statistical analyses were performed using SAS V9.1 and SPSS V12.0.

RESULTS

Item generation, reduction and grading

Of the 48 pediatric IBD experts contacted, 36 responded and served as the Delphi group. A list of 21 potential items was generated by the expert panel and the Delphi group added another 20. Eleven variables, which were ranked significantly higher than the others, were graded by the Delphi group from 1, important, to 11, least important (Table 1). One item, “rectal urgency”, was removed because of perceived difficulties in applying this variable to young children. C-reactive protein (CRP), although not suggested by the Delphi group, was added to the list due to its importance in the UC literature [44, 78, 79]. The final gradation schemes reflect consensus opinion within the
Delphi group. Based on the above stages, an initial draft PUCAI was used in the weighting and evaluation phase.

**Weighting, reliability and index formatting**

A total of 157 children were recruited for the weighting cohort with a consent rate of 98%. The basic characteristics were typical for a consecutive cohort of pediatric UC (Table 2.2). Following individual item evaluation, three items were removed: “fever” (only two patients scored positive), “abdominal tenderness” (insufficient reliability (ICC=0.74, 95% CI 0.61-0.82) and relatively low univariate correlation with PGA (r=0.65)), and “hematocrit” (difficulties in scoring arose following blood transfusion, and univariate correlation with PGA was low (r=-0.2)). A correlation matrix of the remaining eight items (six clinical and two laboratory) showed no redundancy and good consistency (r>0.4 but less than 0.8 for all correlations).

In the multivariate regression analysis, the eight PUCAI items explained 92% of the PGA variance. The gradations were weighted according to the derived $\beta$ coefficients of the model (Table 2.1), rounded to the nearest 5 (Addendum 2.1: the final PUCAI). The $\beta$ coefficients of CRP and albumin were 0.5 ($p=0.68$) and 1.2 ($p=0.39$) per gradation, respectively, indicating that these items are redundant in the model. These items were, thus, removed from the PUCAI, with a drop of only 0.007 in the $R^2$ and less than 10% change of the other item’s $\beta$ coefficients. Two other versions of the PUCAI were drafted for comparison in the validation and responsiveness phases: PUCAI with the laboratory tests (weights of 5 were arbitrarily assigned to each gradation of the albumin and CRP items), and PUCAI with weights exactly as obtained by the model without rounding.
PUCAI validation

Basic characteristics of the 48 children enrolled for the validation cohort are presented in Table 2.2. Correlations of the constructs with the PUCAI in comparison to the adult contender indices are presented in Table 2.3. The PUCAI showed very good to excellent correlation with colonoscopy score, PGA and the Mayo score. These correlations were higher than those achieved by the Seo and Lichtiger indices with the same constructs. When 11 patients who were commenced on corticosteroids in the previous month were excluded from the analysis, the correlation of the PUCAI with the colonoscopic appearance increased to 0.81 (p<0.001). Neither the addition of CRP and albumin to the index, nor the use of the β coefficients without rounding, improved the PUCAI performance (Table 2.3). To verify that the high correlation achieved with PGA was not biased by the fact that the same physicians scored both the PGA and the PUCAI, we analyzed the correlation of the PGA, scored by one physician, with the PUCAI scored by the other physician, and it was still very high (r=0.89; P<0.001).

There was a significant differentiation of the mean PUCAI scores across the categorical PGA strata of none, mild, moderate and severe (Figure 2.2a; ANOVA, P<0.02 for all post-hoc comparisons). Median PUCAI scores were significantly different between the four groups (0 (0-0) for none, 20 (10-30) for mild, 55 (42.5-60) for moderate and 70 (70-80) for severe; Kruskal Wallis=178, 3 df, P<0.001). Patients were then classified by disease severity according to the Truelove and Witts' index, as modified by Jakobovits and Travis [8, 80]. Heart rate and hemoglobin values were standardized by age and gender [81]. According to this conservative index, eight patients fulfilled the
criteria for severe disease and further eight for mild. The PUCAI, at a cutoff of 40, successfully differentiated 15 of the 16 patients to their appropriate disease activity group (area under the ROC curve 0.98 (0.94-1), P=0.001).

**Figure 2.2a: Cutoff scores of the PUCAI**

<table>
<thead>
<tr>
<th>Definition of disease activity</th>
<th>Area under ROC curve</th>
<th>Sensitivity/Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe: &gt;65</td>
<td>0.91 (0.84-0.98)</td>
<td>79%/85%</td>
</tr>
<tr>
<td>Moderate: 35-64</td>
<td>0.91 (0.84-0.98)</td>
<td>79%/85%</td>
</tr>
<tr>
<td>Mild: 10-34</td>
<td>0.92 (0.87-0.97)</td>
<td>90%/81%</td>
</tr>
<tr>
<td>None: &lt;10</td>
<td>0.99 (0.95-1)</td>
<td>97%/95%</td>
</tr>
</tbody>
</table>

PUCAI score distribution of the combined weighting and validation cohort (n=205), stratified by physician global assessment of disease activity. The median PUCAI scores were significantly different across the four groups (Kruskal Wallis, df=3, p<0.001). ROC, sensitivity and specificity were calculated for ‘none’ vs. ‘mild’, ‘mild’ vs. ‘moderate’, and ‘moderate’ vs. ‘severe’.

**Evaluating the need for sigmoidoscopic assessment**

In order to test the merit of adding sigmoidoscopic appearance to the items of the PUCAI, we fitted a multivariable model with the current six PUCAI items as the explanatory variables and a PUCAI score calculated with an additional sigmoidoscopic item (scored similarly to the Mayo index item) as the dependent variable. The analysis was repeated twice using weights of 5 and then 10 assigned to each gradation of the
sigmoidoscopic severity (summing the total PUCAI score to 100 and 115, respectively). The non-invasive PUCAI explained 99% ($R^2=0.986$) and 96% ($R^2=0.956$) of the variance of the invasive PUCAI using weights of 5 and 10 per gradation, respectively. The correlation between the two versions of seven-item invasive PUCAI with the three constructs were only minimally higher than the six-item non-invasive PUCAI ($r=0.8$ and $0.83$ with colonoscopic score, $r=0.91$ and $0.9$ with PGA, and $r=0.96$ and $r=0.97$ with the Mayo index, respectively for weights of 5 and 10 per gradation in the sigmoidoscopy item). These analyses suggest that the PUCAI reflects disease activity well without the need for sigmoidoscopic appearance item.

**Responsiveness assessment and the MCID**

Seventy five children were evaluated twice during the study period after an interval of 3-8 weeks for outpatients, and as short as one week in hospitalized patients treated with intravenous corticosteroids. Twenty-nine (39%) were unchanged or slightly changed, and 33 (44%) had significant improvement. The rest, (n=13, 17%) showed significant worsening and were not analyzed due to the small sample size. The median change in PUCAI that corresponded to none, mild, moderate and significant improvement were 0 (0-7), 8 (6-19), 28 (19-36) and 43 (37-57) points, respectively (Figure 2b; Kruskal Wallis=42.6, 3 degrees of freedom, P<0.001). The repeated PUCAI score of patients judged to have any improvement was significantly lower than the baseline score ($58\pm21$ vs. $29\pm25$; paired Student’s t test; P<0.001). The PUCAI showed excellent responsiveness as reflected by the several methods used (Table 2.4). The laboratory items did not improve the responsiveness of the PUCAI, which performed
significantly better than the Seo index. It has been suggested that responsiveness should be assessed in a group of patients who are expected to improve following a therapy of proven efficacy, regardless of their actual change [82]. Thus, we repeated the analysis on 46 of the 75 children who were commenced on corticosteroid or high dose 5-ASA at the initial visit, without consideration of the actual change experienced. The PUCAI showed excellent responsiveness also among this subset of patients (Table 2.4). The ΔPUCAI showed excellent ability to discriminate the patients according to the Likert scale of change (Figure 2.2b). Mild improvement in disease activity corresponded to a change in PUCAI score of at least 10 points, moderate improvement to at least a 20 point change and large improvement to at least a 35 point change, as determined by the ROC curve strategy (Figure 2.2b).

Figure 2.2b: Cutoff of change scores of the PUCAI

<table>
<thead>
<tr>
<th>Definition of change</th>
<th>Area under ROC curve</th>
<th>Sensitivity/ Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant improvement ≥35 points</td>
<td>0.92 (0.85-0.99)</td>
<td>76%/ 92%</td>
</tr>
<tr>
<td>Moderate improvement ≥20 points</td>
<td>0.97 (0.93-1)</td>
<td>85%/ 88%</td>
</tr>
<tr>
<td>Small improvement ≥10 points</td>
<td>0.96 (0.91-1)</td>
<td>89%/ 87%</td>
</tr>
<tr>
<td>No significant change: &lt;10 points</td>
<td>0.96 (0.91-1)</td>
<td>89%/ 87%</td>
</tr>
</tbody>
</table>

ΔPUCAI (the difference in the score between baseline and follow-up visits) stratified by physician global assessment of change (on a 7 point Likert scale of change). The median ΔPUCAI was significantly different across the four groups (Kruskal Wallis, df=3, p<0.001).
Reliability

All retained items had excellent inter-observer reliability (ICC>0.87, n=157), as did the total score (ICC= 0.95, 95%CI 0.93-0.97). The 29 patients whose disease activity remained unchanged at a follow-up visit had excellent test-retest reliability (ICC=0.94, 95%CI 0.87-0.97).

Scoring the PUCAI

The best cutoff points to discriminate quiescent and mild, moderate and severe disease activity were derived from the weighting cohort using serial ROC curves and were successfully validated on the validation cohort. Figure 2.2a shows the PUCAI cutoffs among the combined cohort (n=205).

DISCUSSION

We have developed a novel multi-item UC activity index for pediatric patients. The large Delphi group and the multicenter enrollment, serve to enhance its content and face validity. The use of PGA as the outcome measure judged by multiple pediatric IBD experts, provided a good average estimate of disease activity for the weighting [83]. There are multiple precedents for utilizing the PGA in this way. It has been used to weight the items of the widely accepted CDAI [14], and to validate the pediatric Crohn’s disease activity index (PCDAI) [15] as well as other indices [51].

The index proved to be valid and has excellent correlation with the invasive Mayo index, which is the most widely used measure of disease activity in adults [12]. A perfect correlation between colonoscopy and clinical indices cannot be expected since some
Clinical symptoms are not directly related to the mucosal inflammation, macroscopic assessment of the degree of inflammation is subjective, and endoscopic healing tends to lag behind symptom improvement in UC [26, 44]. The increased correlation between the PUCAI and the endoscopic score when excluding the patients who received steroids in the previous month, supports this notion. Nonetheless, colonoscopy is highly important in assessing UC activity [12] and the insufficient correlation of colonoscopic appearance with other non-invasive indices in the past has limited their use in IBD [41, 84-86]. In contrast, the observed correlation of the PUCAI with macroscopic mucosal inflammation is sufficiently strong to allow measurement of disease activity in children without endoscopic assessment.

It is often considered unethical in clinical research settings to perform repeated endoscopy in children, particularly in institutions where general anesthesia is routinely used for such procedures. We have shown that adding an endoscopic item to the PUCAI added very little to the clinical items in terms of validity and explaining the total variance. Identical results were previously published with another invasive UC index, the Powell-Tuck score [87, 88], where the non-invasive items explained 96% of the variability in measuring disease activity. The consistency of these observations strengthens our conclusion that in children, repeated endoscopic assessment should not be an obligatory part of outcomes in clinical trials. This should increase rate of parental and patient consent to participate in pediatric UC clinical trials.

Although the PUCAI was designed a-priori to be non-invasive by virtue of the absence of endoscopy, our mathematical modeling demonstrated that blood tests were redundant, and abdominal examination was also justifiably excluded. The feasibility of
using this instrument for repeated assessments in pediatric clinical trials is, therefore, excellent. Indeed, the PUCAI is easy to score and has clear and accurate gradations and definitions. Inter-observer and test-retest ICC analysis were both over 0.9, implying excellent reliability.

There are two predominant strategies for the development of multi-item health measurement scales: the clinimetric approach, which relies primarily on judgmental input of patients or clinicians, and the psychometric approach, which relies on mathematical modeling [63]. The Seo index and the CDAI were both derived using a purely mathematical (psychometric) approach [30] whereas the PCDAI and most other adult UC indices were derived judgmentally, mostly without rigorous methodology. While these strategies have been traditionally considered mutually exclusive, it has been proposed that they actually may be complementary [51]. Indeed, the psychometric technique’s sole dependency on mathematical modeling has been criticized for its low sensibility and face validity [46], but on the other hand, there is evidence from cognitive psychology that humans are poor at selecting non-redundant discriminating items [89]. We, therefore, followed the recommendation that items should emerge from statistical considerations, but with much “dissected intuition” [48]. We applied a rigorous approach for both techniques using a large expert panel, adequately powered prospective cohorts, and robust statistical and judgmental methods.

Responsiveness denotes not merely sensitivity to change, but the ability to discriminate between those who improve and those who do not [21]. Highly responsive indices allow clinical trials to be performed with a smaller sample size. Many methods have been proposed to assess responsiveness, but none has become standard [22].
Therefore, we used several of the available methods but, similar to previous reports [90, 91], we found the same rank order of responsiveness across all methods used. This strengthens our conclusion that the PUCAI is a very responsive index even during a short term follow up period and, thus, suitable for longitudinal assessments in clinical trials.

A PUCAI cutoff score of <10 differentiated well patients with active versus inactive disease, with an area under the ROC curve of 0.99 (0.98-1), sensitivity 97% and specificity of 95% (Figure 2.2a). This is important for clinical trials since an index that accurately differentiates patients who truly achieved remission from those who did not, increases the power of the statistical comparisons, thus requiring smaller sample size. While defining “remission” based upon the PUCAI was relatively clear, defining “response” required more thought. High placebo response rate was reported in Crohn’s disease clinical trials using a CDAI change score of ≥70 [92] and it was, therefore, suggested that the definition of clinically significant change should be set at ≥100 points [93]. Whereas the minimal important difference has been set in several psychometric scales as the mean change score of the group who reported a small change [70, 94], others have felt that at least a moderate change reflects clinically important response [66, 95]. We, therefore, presented definitions of the PUCAI change that correspond to a small (≥10 points), moderate (≥20 points) and large (≥35 points) response. Learning the lesson from the CDAI, we would recommend using the moderate change (at least 20 points) to define “response” in most circumstances and we, thus, set this cutoff as the MCID of the PUCAI.

In an attempt to avoid selection bias, we enrolled all consecutive children, and thus, the weighting cohort was biased towards a higher proportion of quiescent patients.
The results of the validation cohort, with a more normally distributed disease activity spectrum, suggest that this limitation did not affect the accuracy of the PUCAI. Our sampling, however, included only two patients with fever, an indicator of fulminant disease activity, but rarely found otherwise in pediatric patients with active UC [96]. As constructed, therefore, the PUCAI has a ceiling effect and similar to the Mayo index, will not differentiate well the very severe from the fulminant presentation. In addition, as we excluded proctitis patients, we do not know how well the PUCAI performs in these patients [12].

In summary, the rigorously developed PUCAI can be confidently used to accurately differentiate disease activity states in childhood UC and to assess change over time without the need for colonoscopic assessment. We hypothesize that the PUCAI will also perform very well in adults since none of the retained items in the PUCAI is unique to children and because of the excellent performance of the PUCAI compared with two non-invasive adult indices. Validation in an adult population is underway. The availability of an instrument that could be feasibly employed across all ages would be an asset to clinical trial design.
Chapter 3: Predicting the outcome of severe pediatric ulcerative colitis: what is the optimal timing for second line therapy?

The purpose of this chapter is to:

1) Assess the outcome of severe pediatric ulcerative colitis.

2) Assess the ability of the PUCAI to predict failure of conventional therapy, in comparison to other prediction rules available in adults.

3) To suggest a treatment algorithm based on the prediction of the PUCAI.

Acknowledgement


Selected comments obtained from the peer-review process (GUT), together with a point-by-point response, are presented in Appendix 3.1.
SUMMARY

**Background:** Despite the predominance of extensive disease in children with ulcerative colitis (UC), data concerning severe pediatric UC are sparse. We reviewed rates and outcomes of intravenous-corticosteroid therapy on the largest reported cohort to date, with long-term follow-up. The pediatric UC activity index (PUCAI), and other variables were evaluated for their ability to predict therapy failure.

**Aims:** 1) To evaluate the short (by discharge), medium (1 year) and long term (6 year) outcome of intravenous-corticosteroid therapy in pediatric UC; 2) To identify variables predictive of intravenous-corticosteroid failure; 3) To determine the incidence of admissions of intravenous-corticosteroid therapy and population-based failure rates.

**Methods:** 99 children (49% males; age 2-17 years) were hospitalized for treatment of severe UC (90% extensive; 49% new onset UC) between 1991 and 2000. Clinical, laboratory and radiographic data were reviewed. Predictors of corticosteroid response were analyzed using univariate and multivariate analyses at days 3 and 5 of therapy. Abdominal radiographs were retrieved and reviewed by two independent radiologists. Colectomy rates were calculated using Kaplan-Meier survival analyses.

**Results:** 28% (95% CI 23-34%) of Greater Toronto Area-resident children with UC required admission for intravenous corticosteroid therapy, of whom 53 (53%; 95%CI 44-63%) responded. Several predictors were associated with corticosteroid failure, but in multivariable modeling only C-reactive protein (OR=3.5 (1.4-8.4)) and number of nocturnal stools (OR=3.2 (1.6-6.6)) remained significant at both days 3 and 5. The PUCAI, Travis and Lindgren’s indices strongly predicted non-response. Radiographically, upper range of colonic luminal width was 40mm in children younger than 11 years versus 60mm in older patients. Cumulative colectomy rates at discharge, 1-year and 6 years were 42%, 58% and 61%, respectively.

**Conclusions:** Children with UC commonly experience at least one severe exacerbation. Response to intravenous corticosteroid is poor. The PUCAI, determined at day 3 (>45 points) should be used to screen for patients likely to fail corticosteroids and at day 5 (>70 points) to dictate the introduction of second line therapies.
BACKGROUND

Intravenous corticosteroids have been the mainstay of treatment in severe ulcerative colitis (UC) since the landmark trials of Truelove et al. in 1955 and 1974 [8, 9]. As recently documented in a systematic review of cohort studies, one third of adult patients hospitalized with severe UC fail to respond to such therapy [27]. In comparison to adult-onset UC, pediatric-onset disease is more often extensive [4, 5], suggesting that severe exacerbations might be more common. However, to date only three small retrospective studies (44 patients in total) have reported the short-term corticosteroid response rate of severe colitis occurring in children [4, 57]. Moreover, although several adult investigators have developed indices predictive of steroid failure, including the rules of Travis, Lindgren, Seo and Ho [78, 79, 97, 98], no such analyses have been performed in pediatric populations.

We retrospectively reviewed admissions of children and adolescents with active UC hospitalized for intravenous corticosteroid therapy at The Hospital for Sick Children (SickKids), Toronto, over a 10-year period. Specifically we aimed 1) to evaluate the short-term response rates and longer term outcomes following intravenous corticosteroid therapy; and 2) to evaluate the ability of the PUCAI and other measures to predict treatment failure, thereby guiding timing of second-line therapy in young patients. We hypothesized that the higher prevalence of extensive disease in children compared with adults would result in a high frequency of severe exacerbations necessitating hospitalization, and a high rate of steroid failure.
METHODS

This was a single center, retrospective, longitudinal cohort study approved by the local Institutional Review Board. We searched the hospital electronic database for UC-related admissions during the period 1991-2000, using the International Classification of Disease-ninth revision-Clinical Modification (ICD-9-CM) codes for UC (556.x). The charts of all potential patients were retrieved and reviewed.

UC patients, aged 2-18 years, were included if they were admitted to SickKids for initiation of treatment with intravenous corticosteroids. The diagnosis of UC was confirmed in all patients using established clinical, endoscopic and histological criteria [99, 100]. Patients documented to have an intercurrent enteric infection were excluded. To avoid repeated measures bias, it was decided \textit{a priori} to consider only the first eligible admission for patients with more than one hospitalization. The two major analyses performed here were outcomes with therapy, and identification of predictors of steroid failure.

During the first six years of the study period, no pediatric gastroenterologist in the Greater Toronto Area (GTA) cared for hospitalized IBD patients outside of SickKids, and adult gastroenterologists did not independently manage children with IBD under the age of 15 years [101]. Hence, younger patients with postal codes indicating residence in the GTA, who were treated at SickKids during this time period, approximated a population-based cohort [101]. Patients, 15 years of age and over, and all children with postal codes indicating residence outside of the GTA, may have constituted a tertiary referral cohort, and were excluded from the epidemiologic analysis. To estimate the prevalence of severe exacerbations among children with UC, we expressed the number of hospitalized GTA
children as a percentage of the total number of GTA-resident children with UC followed by the SickKids IBD program during the time period.

Data were extracted by RAM, CMW and CC using explicit standardized forms. All charts were independently reviewed by DT and discrepancies were solved by consensus. Data were entered into a customized database using Access® 2003 Professional (Microsoft Corporation, Redmond, WA). Intravenous corticosteroid therapy at this institution is given either as methylprednisolone 1-1.5 mg/kg/day, usually up to 60 mg daily in two divided doses, or equivalent doses of hydrocortisone (standardized in this report as methylprednisolone-equivalent dosing by a 1:5 conversion ratio). 5-ASA preparations are not prescribed to patients hospitalized for treatment of severe UC and antibiotics are administered to febrile children only. During the study period, available second-line drugs were cyclosporine and tacrolimus as previously described [102].

Clinical and laboratory values were recorded at admission, on the third and fifth day of corticosteroid treatment, and at introduction of second line therapy (if applicable). Outcome data were summarized at short-term (i.e. upon discharge), medium-term (i.e. 1-year follow-up) and long term follow-up (i.e. upon transfer to adult care or most recent follow-up). Disease activity at baseline was determined using the PUCAI. The clinical prediction rules of Travis [79], Lindgren (i.e. the fulminant colitis index) [78] and Seo [97], as well as the PUCAI were calculated on the third and fifth day of corticosteroid therapy. In this pediatric population, it was not possible to calculate Ho’s index [98] as the item of colonic dilatation may be age-dependent and there is no existing nomogram to standardize colonic width according to age. We, therefore, retrieved and analyzed the radiographs of our cohort. Plain abdominal radiographs, performed during the first three
days of therapy, were independently reviewed by two pediatric radiologists blinded to the clinical and outcome data. Each radiologist completed a standardized form including maximal supine transverse colon and small bowel luminal width, presence of free air, air-fluid levels in the upright position, portal venous air, mucosal ulcerations (i.e. thumb-printing) and haustral thickening.

Statistical analyses

The proportion of steroid-refractory children (defined by the requirement of second-line drugs or colectomy during the hospitalization) was calculated with the corresponding 95% confidence interval (CI). The others, considered steroid-responsive, were further classified with respect to steroid dependency (defined by requirement of steroid therapy for more than six months in the subsequent year).

To determine predictors of steroid failure, descriptive univariate analysis was performed on predefined potential predictors at admission, day 3 and day 5 of therapy. Chi square or Fisher’s exact tests were used to compare categorical variables (such as gender and disease extent). Continuous variables (such as steroid dose, and days of bloody diarrhea) are presented as mean ± SD or median (interquartile range (IQR)) and compared using unpaired Student’s t-test or Wilcoxon rank sum test, as appropriate. Unadjusted logistic regression was used to obtain the corresponding odds ratio (OR) for each predictor. Multivariate logistic regression was then modeled to associate predictors with corticosteroid failure. Variable screening for predictive models based on statistical significance leads to unreliable models [34, 35], and it has been strongly recommended that the possible predictors be set a priori based on extensive literature review and expert
opinion [34, 37]. We selected variables that were consistently reproduced in the literature [27], modified for pediatric populations: number of bowel movements, CRP, albumin, weight loss and blood in stool. Heart rate was not selected due to difficulty of the assessment and standardization in young children, and fever was not selected due to rarity in children with UC [57, 96]. Governed by maximizing c-statistics, we tested whether modifications of the selected variables improved the model fit (i.e. counting only stools containing blood, only nocturnal stools, percent weight loss, and change in scores since previous assessment at day 3 and 5). Time to colectomy during the entire follow-up was analyzed by the non-parametric Kaplan Meier survival estimate and strata were compared using the log rank test.

Agreement between the two radiologists was assessed using Intraclass Correlation Coefficient (ICC, using Shrout and Fleiss’s 2, 2 two way random ANOVA model ± 95% CI [71]) for continuous variables (i.e. bowel width and number of air-fluid levels) and kappa statistics for discrete variables.

Sensitivity, specificity, predictive values and likelihood ratios were computed for the previously developed predictive indices (i.e. Travis, Seo, and Lindgren indices) as well as the PUCAI, and compared using a receiver operating curve (ROC). Tests achieving an area under the ROC curve (±95% CI) of over 0.7 were considered fair tests, 0.8 good, and those achieving over 0.9 excellent tests. All comparisons were made using two-sided significance levels of P<0.05. Statistical analyses were performed using SAS V9.1.3 (SAS Institute Inc., Cary, NC) and SPSS V15.0 (SPSS Inc., Chicago, IL).
RESULTS

Patients

A total of 114 children were hospitalized at SickKids for treatment of acute UC during the 10-year study period. Fifteen, in whom an intercurrent enteric pathogen was identified, were excluded. Descriptive statistics of the remaining 99 children are presented in Table 3.1. Mean PUCAI scores at admission were similar in the new onset patients (PUCAI 69±12.6) compared with patients admitted for an exacerbation of the disease (PUCAI 72±11.9). Of the 51 patients admitted with an exacerbation of previously diagnosed UC (median disease duration 14 months (IQR 8-26)), 37 (67%) were treated with oral prednisone prior to admission (median therapy duration 14 days (IQR 8-35)).

Of the included 99 children, 55 were aged less than 15 years at the time of the admission and had postal codes indicating residence in the GTA. During the same time period, 196 children residing in the GTA and under the age of 15 were followed by the IBD program in SickKids. Thus, it can be calculated that 28% (95%CI 23-34%) of GTA young children under follow-up care for UC through the SickKids IBD program required at least one admission for intravenous corticosteroid therapy.

Outcomes with intravenous corticosteroid treatment

Short-term and longer-term outcomes are depicted in Figure 3.1. Fifty three children (53% (95%CI 44-63%) responded to intravenous corticosteroid therapy and were discharged after a median hospital stay of 10 days (IQR 7.5-14). Response rates were not different amongst new-onset UC patients versus those with disease in relapse ($\chi^2$, P=0.4; Table 3.1). The proportion of patients who responded to corticosteroids in a
subgroup analysis of the 55 younger patients who resided within the GTA, was identical to the full cohort (53% (95%CI 41-68%)), suggesting no referral bias. Only one patient (1%) was diagnosed with toxic megacolon using standard criteria [103], and underwent colectomy six days following admission. Forty-six patients (46%) failed corticosteroid therapy and were treated with second-line drugs (one cyclosporine, five tacrolimus) and/or colectomy (n=40; Figure 3.1). Four of the six patients treated with calcineurin inhibitors, responded and were discharged without colectomy.

Figure 3.1: Short, medium and long term outcome of children with ulcerative colitis admitted for intravenous corticosteroid therapy at the Hospital for Sick Children, Toronto, during 1991 to 2000

Twenty-two of the total 57 children, who were discharged without colectomy proved steroid dependent during the subsequent year (22% of the full cohort) (Figure 3.1); azathioprine was initiated in 13. One year after discharge, 16 additional patients
(including 1 tacrolimus short-term responder) required colectomy, bringing the 1-year colectomy rate to 58% (95%CI 49-68%). In the total follow-up period following discharge (6±3.6 years), cumulative colectomy rate was 61% (95%CI 52-71%), with only three additional patients requiring colectomy subsequent to the 1-year mark. Eighty-one children (81%) were followed until transfer to adult care at the age of 18 years.

**Predictors of outcome**

The following variables at both days 3 and 5 of intravenous corticosteroids were significantly associated with short-term steroid failure using univariate analysis (Tables 3.1, 3.2 and 3.3): number of nocturnal stools, number of stools per 24 hours, amount of blood in the stool, albumin, CRP, ESR and weight loss prior to admission. All 14 patients who had >2 nocturnal stools at the third day of steroid therapy (100%) and 12 of 13 patients at the fifth day (92%) required second-line therapy prior to hospital discharge (P<0.001). Similarly, 25 of the 27 patients with no nocturnal stools at day 3 (93%) and 23 of 27 at day 5 (85%) recovered without additional therapy (P<0.001). In a multivariate logistic regression, only the number of nocturnal stools and CRP (mg/dL) remained significant both at day 3 (OR 3.2 (95%CI 1.6-6.6), P=0.002 for nocturnal stools; and OR 2.4 (95%CI 1.01-5.8), P=0.049 for CRP) and at day 5 (OR 2.8 (95%CI 1.4-5.8), P=0.003 for nocturnal stools; and OR 3.5 (95%CI 1.4-8.4), P=0.006, for CRP).

All four evaluated indices (i.e. PUCAI, Lindgren, Seo and Travis) significantly differentiated the responders from the non-responders at both day 3 and 5 of therapy (Tables 3.2 and 3.3). The cross-sectional scores of all indices had a better discriminative performance than their change in score over time (i.e. from baseline to day 3 or 5, and
from day 3 to day 5). ROC curves of the indices were plotted (Figure 3.2) to compare their overall discriminative performance. Since Travis score is a categorical variable it could not be plotted with the others. There were no differences in the accuracy of the prediction rules before and after 1996, the year when the Travis rule was first published, suggesting that the decision to proceed to second line therapy in our pediatric cohort was not confounded by the knowledge of the adult predictors.

The third day of corticosteroid therapy may serve as a screening day to identify non-responders [78, 79]; hence, high sensitivity is desired to prepare selected patients for second line therapies. By the fifth day, second line therapy may be executed and, thus, high specificity is required. Cutoffs were chosen to follow this rationale (Table 3.4),
except for the Travis rule, which is designed as a fixed dichotomous rule at day 3. The prediction of colectomy by the PUCAI was significant not only by discharge, but also on long-term follow up (Figure 3.3).

Abdominal radiographs were performed in 46 children during the first three days of corticosteroid therapy. There were no radiographs with free air, pneumatosis intestinalis or portal venous air and only two with overt mucosal ulcers. There was a good agreement between the two blinded radiologists for transverse colon width (ICC 0.85 (95%CI 0.72-0.92)), small bowel width (ICC 0.76 (0.57-0.87)) and number of air-fluid levels (ICC 0.90 (0.81-0.94)), but only fair agreement for abnormally thickened haustra (kappa 0.42; P=0.004). The mean of the two evaluations was used for further analyses.
The distribution of colon luminal width of children older than 11 years of age was similar to data previously published in adults [104, 105], with width of up to 60 mm (mean 36±19 mm) evidence in the absence of clinical and laboratory criteria of toxic megacolon (Figure 3.4). In contrast, the upper width range in younger children was 40 mm (27±11 mm; P=0.039, Student’s t-test, compared with older children). None of the radiographic variables were associated with response to corticosteroid therapy, including bowel width (Figure 3.5), number of air-fluid levels (P=0.58; Wilcoxon rank sum test), and thickened haustra (P=0.39; X² test).

Figure 3.4: Distribution of transverse colon luminal width of pediatric severe UC, during the first three days of intravenous corticosteroid therapy.
DISCUSSION

We have presented novel data concerning the prevalence of acute severe exacerbations in children with UC, and have examined response and predictors of response to conventional corticosteroid treatment in a sizeable pediatric cohort. Several studies have documented a greater prevalence of extensive colitis in pediatric-onset UC [4, 57] compared to adults, of whom only one-third are reported to have macroscopic disease proximal to the splenic flexure [5]. Exacerbations requiring hospitalization, therefore, might be expected to occur more frequently in children, but this burden of illness among pediatric patients has not been previously assessed. It is generally accepted that the lifetime risk of acute severe UC among adult patients is approximately 15% [106-
In our pediatric population-based sub-cohort, we have documented a greater likelihood of hospitalization for treatment of severe disease (28%).

Equally surprising is the paucity of previously published data concerning corticosteroid response rates among children with acute severe UC. A previous meta-regression reported a weighted short-term colectomy rate of 29% in pooled data concerning 1991 patients hospitalized with acute UC [27]. Only three previous small studies, however, evaluated the outcome of admissions for severe pediatric UC. One (n=11) reported a lower rate of colectomy than in adults (10%); the other two (n=13 and n=20) reported higher rates (45%, 46%) [96, 109, 110]. Our 46% steroid-refractory rate, in a much larger cohort, confirms that response to intravenous steroids may be poorer in children than in adults. Similar to recent outcome studies of pediatric UC treated with corticosteroids (mostly ambulatory), we observed a high proportion of steroid dependency during 1-year follow-up [111, 112]. The poorer response of children compared with adults, may be explained at least in part by the higher proportion of extensive colitis in childhood onset UC. Disease extent has been consistently associated with severe disease course [9, 113-116]. Interestingly, our colectomy rate after one year remained remarkably stable, supporting data from a population-based study that documented most of the UC-related admissions during the first few years of diagnosis [117].

Among this patient cohort, response did not vary with corticosteroid dosage in the range administered. This confirms our heterogeneity-controlled meta-analysis that found virtually no correlation between the steroid dose (at or above equivalent of 60 mg methylprednisolone daily in adults), and the proportion of patients failing therapy [27].
Moreover, clinical trials that assessed continuous versus bolus [118] or pulse dosing [119] of intravenous steroids in UC failed to identify differences in response. Similarly, in ambulatory adult patients with moderate exacerbations of UC, 40 mg of prednisolone was as effective as 60 mg daily with less toxicity [120], and once daily was as effective as multiple daily doses [10]. More research is required to unravel the mechanisms of corticosteroid resistance, as bioavailability of corticosteroid does not seem to play an important role [121].

Untreated severe UC attacks in adults were previously associated with a 24% mortality rate [8, 80], but timely introduction of medical therapy reduced this rate to 1% [27]. Although some case series suggest that prolongation of steroid therapy (i.e. beyond two weeks) is effective in eventually achieving remission, this approach is associated with increased toxicity, discomfort and cost [110, 122]. Adult clinical guidelines, therefore, recommend that second-line therapy be initiated if no response to corticosteroids is noted within several days of initiating intravenous therapy [80, 96, 98, 116, 123, 124].

To give clearer direction concerning implementation of second-line therapy, adult investigators have developed measures predictive of steroid failure [78, 79, 97, 98]. In a retrospective analysis by Travis et al. stool frequency of >8/day or 3-8/day and CRP>45 mg/L on the third day of therapy had a positive predictive value (PPV) of 85% for colectomy [79]. Lindgren et al. [39, 78] developed the fulminant colitis index (stool frequency/day + 0.14XCRPmg/L) with a PPV of ~70% at a cutoff score of >8 at day 3 of therapy. Ho et al. developed an index based on stool frequency, albumin level and colonic dilatation [98]. Ours is the first study to compare these predictive tools head-to-
head. We selected low cutoffs of the prediction rules at day 3 (i.e. high sensitivity), to identify patients to be prepared for second line therapy (for example, for suitability for infliximab therapy, calcineurin inhibitors or surgical consult), and higher cutoffs (i.e. high specificity) on day 5 to guide execution of the planned therapy in the 40-50% of patients in whom the risk for colectomy is >90%. This approach should guide appropriate escalation of therapy earlier in the disease course and shorten the overall hospital stay. The subset of patients not fulfilling these conservative cutoffs may be treated for several more days with continued corticosteroids. We examined the predictive ability of the PUCAI developed originally as an evaluative and discriminative tool [57]. The novel data from the present retrospective analysis suggest that a PUCAI score of >45 on day 3 should dictate planning of second-line therapy and PUCAI >70 on day 5, should prompt the execution of the planned therapy.

This study provides the first descriptive data concerning abdominal X-rays in children with severe UC, but who do not fulfill the clinical criteria for toxic megacolon [103]. The distribution of colonic luminal width in children above 11 years of age follows the published data from adults, in whom width of up to 60 mm may occur without clinical or laboratory criteria for toxic megacolon [104, 105, 125]. In contrast, colon dilatation of more than 35-40 mm was unusual in younger children. Unlike adult studies [98, 113, 114, 126-129], we found no associations between radiographic appearances and outcome. Adolescents (>11 years of age) with a transverse colon luminal width of 40-60 mm were still very likely to respond to therapy.

Childhood-onset extensive UC is an IBD phenotype associated with significant morbidity. This study, although retrospective, has highlighted the short-term response
rates and longer term outcomes in an otherwise understudied group of patients. The population based sub-analyses reduced the risk of referral bias. Documentation of steroid response rates and predictors of response in children are important, so that much needed novel therapies may be tested in appropriate patients. We have generated data concerning the predictive value of the PUCAI and other predictive measures, when employed in children. Cut-off scores suggested by this study will be re-examined in an ongoing prospective multicenter study of severe pediatric UC, wherein additional serologic and genetic factors associated with acuity and response to treatment, will also be assessed.
Chapter 4: Properties of alternative methods to establish the minimal important difference and the minimal detectable change: a lesson from the PUCAI and three other measures

The purpose of this chapter is:

1) To further characterize the minimal important difference of the PUCAI.

2) To identify the preferred method of determining the minimal important difference, using the PUCAI and three other health related outcome measures.

3) To highlight differences in determining the minimal important difference between clinimetric indices and psychometric scales.

Acknowledgement: I wish to acknowledge Gordon Guyatt, Holger Schunemann, Lauren Griffith, Dorcas Beaton, Anne M. Griffths and Jeffery N. Critch who provided me databases of other health-related measures, and assisted to formulate this chapter by providing stimulating questions and comments.
SUMMARY

Background: Studies on health related outcome measures are increasingly using receiver operator characteristic (ROC) curve methods to determine the minimal important difference (MID). However, data to guide the incorporation of the ROC approach are lacking. We aimed to evaluate the ROC approach on four different health related measures.

Methods: Studies of two patient-based quality of life instruments and two physician-based activity indices (the Pediatric Ulcerative Colitis Activity Index (PUCAI), the Pediatric Crohn’s Disease Activity Index (PCDAI), the Chronic Respiratory Questionnaire (CRQ), and the Rhinoconjunctivitis Quality of Life Questionnaire (RQLQ)) provided prospectively collected longitudinal data. For each, the MID was calculated using the ROC methods while dichotomizing the 7 and 15 point global rating of change in multiple ways. The results were compared with the commonly used mean change approach, the distributional approach and the minimal detectable change (MDC), calculated by the various available methods.

Results: The ROC approach yielded MID values that were consistently lower than the mean change approach. Using the entire cohort, rather than just the two groups adjacent to the dichotomization point yielded a more precise and sensible MID estimates. Of methods to calculate the MDC, the 95% limits of agreement method and the reliable change index yielded estimates far larger than the MID. There was no consistency between the MID values obtained by the ROC approach and the distributional-based approach.

Conclusion: When calculating the MID using the ROC method, use of the entire cohort maximizes precision. Of the distribution-based methods of calculating the MDC, the SEM method provides values closest to the MID.
Background

The choice of the change in outcome measure score that defines an important response is likely to impact interpretation of the results in clinical studies. The Minimal Clinically Important Difference (MCID) was defined as the smallest change perceived as beneficial by patients or physicians and that leads to a change in the patient’s management, assuming minimal toxicity and cost [67, 130]. Clearer thinking subsequently led to increased emphasis on the patient’s perspective by revising the MCID to Minimally Important Difference (MID), removing the focus on clinical interpretation of psychometric scales [131]. However, for many therapies toxicity and cost are not minimal, and thus, others left the term MCID, which requires some criterion of clinical significance [25, 70].

Distribution-based and the anchor-based approaches are the two major strategies for determining the MID [132]. The first is an internal property of the measure on a given cohort and result in a series of measures gathered from distributional information (i.e. variance). The latter uses an external criterion (i.e. anchor) to interpret the significance of a particular magnitude of change. The major disadvantage of the distribution based approaches is that they determine a minimum detectable change, but have no way to ascertain whether and by how much this change is important [133]. Only an external criterion can provide a threshold of importance, thus, determining a MCID. Other approaches to determine the MCID include expert’s opinion, data driven prognostic anchors (e.g. change in therapy or medical action) and population based outcome [55, 134].
Crosby et al. has elegantly summarized the statistical methods to establish the distribution and anchor based approaches, with the various available cross sectional and longitudinal anchors [135]. The distribution based approach may utilize the effect size statistics (including Cohen’s effect size (ES) [136], and standardized response mean (SRM) [72], variations of the Standard Error of Measurement (Wyrwich SEM [137], and Jacobson’s Reliable Change Index (RCI) [138]) and the Bland and Altman limits of agreements [139]. As evident in Table 4.1, authors have suggested different values to characterize the MID within the effect size statistics (i.e. 0.2, 0.5 and 0.8) and the SEM methods (i.e. 1, 1.96 and 2.77). The major role of these distributional based strategies is in identifying the minimal detectable change (MDC, also referred to as the smallest detectable difference), which is a boundary of measurement error; it is the smallest change in score that can be detected beyond error of repeated measurement [25, 140, 141]. The methods of Wyrwich and Jacobson (Table 4.1) are based on the same formula, substituting different levels of acceptable type I error. Wyrwich argued that 50% was acceptable at an individual level, leading her work with the 1 SEM. This later was observed to be equivalent to half standard deviation, which we would see as algebraically true if the test-retest reliability were strong (ICC of 0.97). Stratford set 95% (5% error) as acceptable in early work, and later shifted to 90% (10% type I error) for the MDC. The MDC reflects the usual limits for a change score in a stable group. This is often translated into the lower threshold for interpretable change because anything lower than that could be the same that one would observe in stable patients. However, the methods to obtain the MDC are dependent on the number of observers, the scoring methods and the data distribution of the population under study.
For the anchor-based approach, several external criteria have been suggested but the most widely used is a global rating of change (GRC) [142]. The first and most widely used statistical strategy to ascertain the MID from GRC, is to calculate the mean change score of patients who graded themselves as having small change (1-3 points on a 15 point scale (i.e. -7 to 7) or 1 on a 7 point scale (i.e. -3 to 3), where 0 means no change) [130]. Others used the mean change of patients graded as 2-3 [67, 143] or 5 [144, 145] points on a -7 to +7 point scale or used different number of points on the GRC [21, 133, 146, 147]. The vast majority of these studies focused on psychometric scales, and it is not known whether it also applies to clinimetric indices. A concept of ‘quality of life’ may have a different important cutoff on the GRC than a clinimetric concept of ‘disease activity’.

Rather than using the mean change of groups, studies are increasingly using the receiver operator characteristics (ROC) curve method to define the best cutoff change correlating with the MCID [21, 66, 145, 146, 148-152]. Recently, it has been recommended that the ROC strategy should be used to combine the anchor and distribution based approaches in determining the MCID [133, 135]. Since the outcome used for the ROC analysis must be a dichotomous variable, some recorded the criterion for improvement as simply ‘improved’ or not, or whether or not therapy was changed [21, 149, 150]. However, a GRC that includes more categories allows clearer description of the amount of change that occurred [66, 146, 151]. It is not known which cutoff on the GRC should be used for the analysis, and how it compares with the distributional based approach and the MDC.

Here, we evaluate the ROC method to ascertain the MID using two psychometric scales and two clinimetric indices, in order to guide its future use. We compare the results
of the ROC analyses to other commonly used techniques, to benchmark the impact of alternative dichotomization cutoffs.

**METHODS**

**Population**

For this study, we used prospectively collected longitudinal datasets of the following established instruments: the Pediatric Ulcerative Colitis Activity Index (PUCAI) [57], the Pediatric Crohn’s Disease Activity Index (PCDAI) [15, 66], the Chronic Respiratory Questionnaire (CRQ) [67, 153], and the Rhinoconjunctivitis Quality of Life Questionnaire (RQLQ) [154].

*Pediatric UC Activity Index (PUCAI)*

The PUCAI is a 6-item physician-based disease activity index aimed for use in pediatric UC clinical trials (score range 0-85). Development, weighting and validation were performed using combined judgmental and mathematical strategies utilizing a Delphi group and large prospective pediatric cohorts (Chapter 2). A longitudinal cohort of 75 children was used to determine responsiveness, in which two physicians completed a 7-point GRC (i.e. '-3' significantly worse, '-2' moderately worse, '-1' mildly worse, '0' no change, '1' mildly improved, '2' moderately improved and '3' significantly improved). The demographic details of that cohort with the data description are described in Chapter 2.
**Pediatric Crohn’s Disease Activity Index (PCDAI)**

The PCDAI is an 11-item physician-based disease activity index aimed for use in clinical trials of pediatric Crohn’s disease (score range 0-100). Items were selected and weighted judgmentally based on the adult Crohn’s Disease Activity Index [15]. The construct validity and reliability of the index were evaluated on a prospective cohort of children with Crohn’s disease. Responsiveness was assessed on a longitudinal cohort of 25 children using a 7-point global rating scale, as described for the PUCAI [66].

**Chronic Respiratory Questionnaire (CRQ)**

The CRQ is a psychometric scale aimed to evaluate health-related quality of life of patients with chronic respiratory diseases. It includes 20 items across 4 domains: dyspnea (five items), fatigue (four items), emotional functioning (seven items) and mastery (four items), all measured on a 7-point Likert scale (1- worst health to 7-best health). For this study, we used the longitudinal dataset that compared the original and the standardized versions of the CRQ [155]. A total of 281 patients were enrolled, of whom 177 completed the follow-up visit. Change was recorded on a 15 point GRC completed by the patient from -7 (a very great deal worse) through 0 (no change) to +7 (a very great deal better).

**Rhinconjunctivitis Quality of Life Questionnaire (RQLQ)**

The RQLQ was developed in 1989 to assess quality of life of patients with rhinoconjunctival diseases [156]. The instrument has 28 items in seven domains (sleep, non-rhinoconjunctivitis symptoms, practical problems, nasal symptoms, eye symptoms, .
activity limitations, and emotional function). For this study we used the dataset of 61 adults with moderate to severe ragweed pollen-induced rhinoconjunctivitis that participated in a nasal steroid spray clinical trial [154, 157]. At each follow-up visit, patients rated their change in quality of life on a 15-point GRC, as described for the CRQ. We utilized the data obtained from the second evaluation period, as it was the only period that satisfied Guyatt’s criteria for change scales, described below. Sixty eligible patients were included.

**Statistical considerations:**

In order to establish that an external criterion reflects the actual change occurred, Guyatt *et al.* proposed four mandatory criteria to assess the relationship between the GRC and the observed data [67]: the correlation of the GRC with the change of the instrument score should be more than 0.5; there should be a positive correlation between the GRC and the baseline instrument score (i.e. the further the score from the mean, the larger the anticipated change) [145]; there should be a negative correlation between the GRC and the repeated instrument score (i.e. the larger the change, the lower the disease activity following the change) and the correlation of the GRC with the repeated instrument score should be at least 0.2 smaller than with the change of the assessed instrument. The datasets of all measures and domains used here were first assessed according to these four criterions.

On the 15-point GRC (-7 to +7), -1 to 1 was referred to as ‘no change’, 2-3 as ‘small change’. 4-5 as ‘moderate change’ and 6-7 points as ‘large change’ [154]. On the 7-point GRC used in the clinimetric indices (-3 to +3), 0 was referred to as ‘no-change’, 1
as ‘small change’, 2 as ‘moderate change’ and 3 as ‘large change’ [15]. As discussed later in Chapter 5, it seems that MID for improvement may not be similar to the MID of deterioration and thus could not be lumped together in the analysis. Due to the smaller sample size of deteriorating patients in the included studies, this study did not address the MID of worsening scores.

ROC curves were plotted for several cutoffs of the GRC: ‘no change’ versus any improvement, ‘mild improvement’ versus ‘moderate improvement,’ and ‘no change’ and ‘mild improvement’ combined, versus ‘moderate’ and ‘significant improvement’ combined. The number of observations in each category is tabulated in the footnotes of Tables 4.2a and 4.2b. The best cutoff of change, corresponding to the point where the second diagonal crosses the ROC curve (i.e. the point where the shoulder of the curve is closest to the left upper side of the figure) was set as the MCID of the evaluated cutoff. This cutoff provides the best balance between sensitivity and specificity in differentiating subjects that truly changed from those who truly remained unchanged. Although different cutoff points are often selected when using the ROC method in diagnostic tests (i.e. obtaining an unbalanced higher sensitivity or specificity), optimizing sensitivity and specificity is most appropriate for MCID. The MCID values obtained by the different dichotomization strategies were compared with several other anchor and distribution based methods, as tabulated in Table 4.1. Sensitivity, specificity and area under the ROC curve (95%CI) were calculated for the MCID estimate. For the standard error of measurement (SEM) methods, the test-retest Intraclass Correlation Coefficient of the stable patients was used to determine ‘r’ in the equation [23, 158, 159]. This was chosen over the Cronbach's alpha ‘r’ as it is more appropriate when assessing longitudinal
change, and since Cronbach's alpha is not available, nor suitable, in the clinimetric indices. Statistical analyses were performed using SPSS V15.0 (SPSS Inc., Chicago, IL).

RESULTS

All included indices and domains satisfied the criteria for correlations between the GRC and the corresponding domains, thus verifying the validity of the GRC. By the ROC approach, the MID (and thresholds for moderate and large change) values obtained in analyses using all data were, for the most part, similar to the analysis in which only the two adjacent groups to the dichotomization point were included (e.g. only mild versus moderate improvement or none + mild versus moderate + significant improvement; Tables 2a and 2b). This was especially true for the PUCAI and PCDAI. In two of the three cases where the MID values differed (i.e. ‘moderate change’ of the RQLQ, and ‘moderate change’ of the emotional function domain of the CRQ), there was a logical gradient between the small, moderate and large change if the entire cohort, but not when analysis was restricted to the adjacent groups (Tables 2a and 2b). The sensitivity, specificity and area under the ROC curve were higher using the entire data set. Similarly, the precision of the estimates, reflected in narrower confidence intervals, were higher for the entire cohort (Tables 2a and 2b).

The ROC approach yielded MCID values that were nearly always smaller than the mean change strategy within each category of small, moderate and large change (Figure 4.1a-f). There was a significant variability in the MDC between and within the instruments, calculated by the 95% limits of agreement, 1.96 RCI and 1 or 1.96 units of SEM (Figure 4.1a-f).
Small change Moderate change Large change

LOA, limits of agreement; SEM, standard error of measurement; SD, standard deviation; ROC, receiver operator characteristic; RCI, reliable change index Small, moderate and large change scores were determined as in the text and table 4.1.
Figure 4.1c: The minimal important difference (MID) of the dyspnea domain of the Chronic Respiratory Questionnaire, calculated using several distributional and anchor based methods.

LOA, limits of agreement; SEM, standard error of measurement; SD, standard deviation; ROC, receiver operator characteristic; RCI, reliable change index Small, moderate and large change scores were determined as in the text and table 4.1.

Figure 4.1d: The minimal important difference (MID) of the fatigue domain of the Chronic Respiratory Questionnaire (CRQ), calculated using several distributional and anchor based methods.

LOA, limits of agreement; SEM, standard error of measurement; SD, standard deviation; ROC, receiver operator characteristic; RCI, reliable change index Small, moderate and large change scores were determined as in the text and table 4.1.
Figure 4.1e: The minimal important difference (MID) of the emotional-function domain of the Chronic Respiratory Questionnaire (CRQ), calculated using several distributional and anchor based methods.

Figure 4.1f: The minimal important difference (MID) of the Rhinoconjunctivitis Quality of Life Questionnaire (RQLQ, Figure 1f), calculated using several distributional and anchor based methods.

LOA, limits of agreement; SEM, standard error of measurement; SD, standard deviation; ROC, receiver operator characteristic; RCI, reliable change index Small, moderate and large change scores were determined as in the text and table 4.1.
There was also great variability between the distributional-based methods used to calculate the MDC and the MID values obtained by the anchor based methods, although the variability was somewhat less in the PUCAI and PCDAI. The 95% limits of agreement and 1.96 RCI proved the most conservative, very often suggesting larger thresholds even than the MID that corresponded to large change using other methods. The MDC, calculated by 1.96 units of SEM, was larger than the moderate MID in the quality of life instruments, but not in the PUCAI and PCDAI. The value of 0.5 SD sometimes, but not always, approximates the mean change of the small change group.

**DISCUSSION**

There is no consensus regarding the exact definition of the magnitude of change that patients, or clinicians, deem important. Perhaps reflecting this inconsistency, various statistical approaches exist to identify the desired threshold, and these approaches lead to different results. This study highlights the optimal method for the ROC approach, and the limitations of the distributional methods.

**Distributional or anchor-based approach?**

The distribution-based methods are based entirely on statistical criteria and thus depend on the distribution of change score that vary between samples [138]. It is also not established how these cutoffs correspond to changes that are important to patients, or clinicians. The need to establish correspondence with other measures of clear importance to patients and clinicians strongly recommends reliance on external anchors for
establishing the MID. Several clinically relevant anchors have been suggested [21, 150, 160-163]; ideally, the anchor should allow quantification of small, moderate and large change. Despite its limitations [82, 135], the GRC is the most widely used approach, allows ascertainment of MID, moderate and large changes [142], and can facilitate establishing the validity of the scale [145].

Distributional approaches still have a major role in identifying the MDC. We used the major approaches available for determining the MDC, including the distributional (i.e. SEM or RCI) and the limits of agreement methods. Use of distributional approaches is complicated by competing suggestions for the “beyond error” thresholds [70, 149] (Figure 4.1a-f). Childs used 1.96 of SEM to define the MDC, correlating the score to 95% confidence interval (i.e. MDC95) [151]. Stratford used 1.645 SEM (MDC90) to determine the smallest change beyond error of the Roland-Morris Scale [164], and Wyrwich advocated the use of 1 SEM (i.e. MDC67) [165]. Similarly, 1.96 units of the Jacobson’s RCI was also suggested as the MDC95 [138]. Just like the 1.96 RCI and SEM, the 95% limits of agreement is also designed to determine the MDC95 [139]. In this study, it was found to be comparable with the 1.96 RCI and the mathematical equivalent 2.77 SEM. The MDC obtained by these three 95% confidence interval strategies, appear to be higher even than substantial clinical relevant cutoff of moderate or large change. The choice of 1 SEM yields a liberal MDC that is almost always lower than the small clinical change, judged by a GRC. A 1.96 SEM yields a more conservative MDC, higher than the small change but lower than moderate change. Researchers should acknowledge this reflection between the boundaries of error and the MID values, when choosing either value for their studies.
Since the MDC is based on the data distribution, it often varies from study to study [135, 149, 166]. Different MDC values were found using similar methods but on two different populations [141, 149]. Similarly, values differed based on whether the baseline score lied in the extremes or closer to the mean [164]. In our study, 1 SEM always lay between 0.2-0.5 effect size units (both are distributional-based approach), but the relation to the anchor-based methods varied between the different instruments, making generalization impossible.

**Mean score or ROC curve for finding the MCID?**

The first and most widely used external anchor of change utilized the mean change of the group that fulfills the criterion (e.g. mean change of the patients who score small change [130]). However, this technique has several limitations. Mean change is a poor descriptor of non-normally distributed data, which is often the case in change scores, and is susceptible to outlying values. To overcome this limitation it has been suggested that the median rather than the mean should be considered. However neither strategy utilizes all available data and if a small number of values exist within the criterion category, then precision is reduced. Moreover, by using the mean change, individuals who score lower than the mean but higher than the cutoff to the next category on the GRC (i.e. lower than the mean of the small change but still in the category of small change) will be misclassified as not experiencing important change, when they actually have. These limitations are successfully addressed by using the ROC approach to that dichotomizes the entire cohort into two categories that correspond to the boundaries on the GRC. This also describes the magnitude of change by means of sensitivity and
specificity, which is more interpretable to clinicians, and allows the comparison of different approaches using the area under the ROC curve.

**One or several MCID values?**

Like others [145], it was found here that the ROC approach yielded an MCID value that was consistently lower than the mean change approach. It should be acknowledged that changing the latter approach to the ROC method will reduce the MID values, which in turn, may influence clinical trial outcome. The original definition of MCID includes the change that indicates therapy alteration assuming lack of cost and toxicity. The MCID, however, cannot be defined in isolation from the context of its interpretation or use, and in practice, cost and toxicity are crucial components in decision making [167]. This was acknowledged in the revision of the MID definition, including some cost and toxicity in the consideration of the change score [67]. A natural extension of this revision is that several change values could be presented for different levels of costs and toxicities. Moreover, the choice of the MID value may depend on the score distribution at baseline [145]. Larger change scores are expected in severely active patients, compared with identical moderately active patients [73]. This may be explained by the regression to the mean phenomena, by ceiling and floor effects of the instruments, by psychological patterns of response in accordance to baseline severity, and by natural history of the disease under study [69]. If a clinical trial evaluates a very cheap and safe medication (e.g. 5-ASA preparation for UC) even a small change may be considered significant enough to recommend the use of the drug [132], and a small MID should be used, especially if mild to moderately active patients are included. In contrast, a large
change values should define response in a clinical trial of an expensive and toxic medication (e.g. biological therapy) provided to severely active patients. The use of a small MID in inappropriate scenarios for the sake of showing positive results will be counterbalanced by high placebo response rates.

A second limitation of the multiple MID approach is that it may become impossible to compare the results of different studies that use different outcome criteria. To minimize this limitation, clinical studies should report the proportion of patients achieving small, moderate and large response in each group that may correspond to small and definite response. For example, Stratford et al. dichotomized a -7 to +7 GRC using two different cutoffs of +5 and also +3, yielding small and moderate MCID [145]. Salaffi et al. utilized the ROC curve to find the best change score of a VAS of pain on a 5 point GRC scale (-2 to +2) using a cutoff value of both +1 (i.e. slightly better) and +2 (i.e. much better) [168] resulting in different MCID. The same was presented in the development of the PUCAI (Chapter 2).

The different cutoff values used in the literature for the ROC analysis reflect different MID values, as judged by the investigator (i.e. small, moderate and large change scores). A cutoff of 50mm on a 100mm VAS of change (0 = no change 100= very important change), was used to find the MID using the ROC approach [148]. In comparing the MID of four different elbow scoring systems in rheumatoid arthritis, the cutoff for improvement on the ROC curve has been set as at least +1 (i.e. somewhat satisfied with the treatment) on a -2 to +2 GRC scale [146]. Deyo et al. and de Vet et al. dichotomized a 6-point GRC where important improvement was defined as a score of at least +2 on a 6 point scale (where +6 means much worse and +1- cured) [21, 133]. An
effective drug will yield a larger proportion of significant responders, but a less effective cheap and safe drug may still lead to a small clinically significant response.

**Who should be included in the ROC analysis?**

Based on the results found here, we suggest that ROC analysis should always include the entire cohort in the ROC analysis rather than including only the two adjacent groups around the dichotomization point. When the PUCAI manuscript was first submitted for publication, the entire cohort was used for the ROC analyses. Subsequent to a peer-review comment, this was changed to include only the two adjacent groups (addendum 2.2). One other study similarly used adjacent groups in the ROC analysis [146], but most used the entire cohort, without clear guidance in the literature [21, 133, 145, 148, 168]. Including only adjacent groups underestimate the number of false negative and positive cases, as a change score of an individual may not uniformly relate to the appropriate category on the GRC (i.e. extreme categories on the GRC may include individual change values that should lie on the opposite side of the dichotomizing point of the GRC). Moreover, the power of the analysis is maximized by using the entire cohort, which leads, as we found, to more precise estimates with logical gradient across groups. It remains to be determined if a different analysis should be performed for deteriorating and improving patients, reflecting a potentially different change distribution.

**Differences between patient and physician-based instruments**
Most methodological research on the MID was performed on patient-based psychometric scales. Two clinimetric disease activity indices were included here, in which it is the clinician’s perspective that matter in determining the clinically relevant change [51]. The data obtained by the physicians (including the GRC) for the PUCAI and the PCDAI were more homogenous than obtained from patients in the quality of life scales. This was reflected by lower MDC, and convergence of the MDC and MCID values obtained by the various methods for each change category. Although physicians obtain the raw clinical data from the patient, it undergoes subjective processing prior to scoring the instrument and the GRC. The rationale for using a physician’s external measure for disease activity was noted by Jaeschke et al. [130] stating that an experienced clinician would have no difficulty in identifying a change that is clinically important having observed a large number of similar patients. The scoring of the physician is benchmarked by the change observed in many similar patients, whereas patients have no knowledge to reference their disease activity resulting in more heterogeneous responses. It must be emphasized that this advantage holds only for judging disease activity and not disease severity (i.e. quality of life) which should be scored by the patients themselves.

Summary

Using two patient- and two physician-based instruments, we present novel data that should guide future identification of MID values. The use of an anchor that quantifies a gradient of change, such as the 15 or the 7 point GRC, facilitates establishing the validity of the anchor. The calculation of the MDC could also be presented, using the
1 or 1.96 SEM method. In the ROC calculation, use of the entire cohort rather than only the two groups adjacent to the dichotomization increases the precision of MID estimates.
**Chapter 5: Discussion**

**SUMMARY OF RESEARCH**

In a series of studies, I presented the comprehensive development and evaluation of the PUCAI. To date, the PUCAI is the most rigorously developed and evaluated activity index in IBD. In view of the lack of existing similar outcome measures in pediatric populations, the PUCAI was well and rapidly received by the IBD community. During the short time since its introduction, the PUCAI has been incorporated into several clinical studies in pediatric UC. In a late amendment, the PUCAI was incorporated as a secondary outcome measure to an ongoing industry-initiated multi-center clinical trial of infliximab (anti-tumor necrosis factor alpha) in the active treatment and maintenance of remission of pediatric UC. The number of the required endoscopies decreased, as a consequence of this amendment, from three to two per participant. The PUCAI is planned to be the primary outcome in a clinical trial evaluating a once daily preparation of 5-ASA and no endoscopic evaluation is planned. Hence, the introduction of the PUCAI has brought real change to outcome assessment in pediatric UC clinical studies. By eliminating the requirement for invasive tests, ethical concerns are addressed. It is anticipated that enrolment rates will increase and dropout rates decrease as a result of adoption of the PUCAI. In the following pages, I will review the development and evaluation phases of the PUCAI, highlighting challenges, uncertainty and possible solutions, while also reviewing the existing literature.
ITEM GENERATION and CONSENSUS METHODS

The purpose of the item generation phase is to compile a complete set of variables, from which the items for the index will be chosen [169]. It should start with a comprehensive literature search for related indices and consultation with experts. Then, a broader panel of experts should be contacted for additional evaluation.

The three most commonly used consensus methods include the Delphi group, the nominal group, and the focus group techniques [170]. For the PUCAI, the Delphi group technique was used in order to obtain opinions from many pediatric IBD experts in a systematic way. Delphi is an anonymous process in which the participants never meet (or even know of one another) and are asked to comment on the issue in question [170]. It is a widely used technique and has been previously shown to be valid [170-172]. The responses are summarized and tabulated by the primary investigator and the results, reflecting the group's opinion, are sent back to the group for further feedback. These steps are repeated until a convergence of opinions is reached, judged by many available methods (e.g. at least 80% of responders offer no alternative scores, self confidence rating, taking the mean and SD as a reference, considering the highest ranked items, or taking Cronbach's alpha over 0.9 as a reference of response homogeneity [170-172]). Alternatively, the process may be terminated when a point of diminishing response is reached (e.g. new data is obtained from less than 20% of responders). Except for the anonymity, the Delphi resembles the nominal group in structure [171]. The anonymity nature of the process eliminates the potential dominance of influential participants. Moreover, the absence of an obligation to meet greatly improves feasibility and compliance, thus, eliminating constraints on sample size and composition of the group.
The reliability and generalizability of the Delphi group was increased by obtaining a large sample of heterogeneous pediatric gastroenterologists. The Delphi has been criticized for the influence that the primary investigator has over the processing of the response [173]. To limit this bias for the PUCAI, the processing of the responses was carried out by an expert panel rather than by a single investigator.

Patient or physician judgment may be used in generating items, depending on the conceptual framework of the measure. A Delphi group composed of experts in the field was used, because "disease activity" is a concept best judged by experienced physicians who care for IBD patients [130]. After all, it is the physician who decides, on an everyday basis, which therapy is required based on his/her interpretation of the disease activity. Quality of life scales, in contrast, are best judged by the patients themselves, as the individual feelings may not be related to objective assessment of disease activity [63].

ITEM REDUCTION

After the pool of all potentially important items has been generated, it must now be reduced to a feasible list, retaining only the most important variables. The two major approaches for item reduction are judgmental and statistical [174].

Judgmental approach

A variety of judgmental methods have been used for item reduction, ranging from arbitrary decisions of a single investigator to robust group consensus techniques, described above [174]. The purpose of this judgmental phase may be to reduce the variables of the preliminary list to a feasible number for mathematical modeling (to avoid
over-fitting), as was done for the PUCAI. Alternatively, mathematical modeling may not be performed and all retained items would be then forced in the index, as was done with the Pediatric Crohn's Disease Activity Index [15].

In either case, the judgment should be guided by considering both the item importance in explaining the concept of the instrument, and the frequency of the occurrence of the item within the target population. There are two main methods to create this composite is the "frequency-importance" and the "severity-importance" approaches. An example of the latter comes from the judgmental item reduction of the Disabilities of Arm, Shoulder and Hand (DASH) scale [51], and the Asthma Quality of Life Questionnaire (AQLQ) [56]. Items were graded as “extremely important" to "not at all” important, and separately as "extremely severe" to "not at all severe" (on a 5 point Likert scale). The combined score of severity and importance was used to rank order the items. The validity of this combined strategy is questionable and no difference in item reduction was found between the two methods [62]. Therefore, and in order to maintain the simplicity of the Delphi process, only one question was presented to the group: "how important is the item in explaining disease activity?". Since our target population was not patients but experienced physicians we assumed that both frequency and severity will be incorporate in the response. Nonetheless, the weighting cohort was used to format the PUCAI based also on item frequency. Hence fever, that was rarely scored, was subsequently omitted from the PUCAI.

Other judgmental considerations when reducing items should include ease of scoring, reliability, and responsiveness of the item. We omitted the item of hemoglobin
concentration from the PUCAI, as it was difficult to score, and abdominal examination since it was not sufficiently reliable.

**Mathematical approach**

This approach can be utilized in various ways [174]. For clinimetric indices, the most widely used technique is to fit the items in a multivariable regression model, while selecting one important outcome as the dependent variable. In this way, items may be excluded from the model based on non-significant P-value and low β values and/or by improving the overall fit of the model. The main challenge of this approach is selecting the most important outcome measure for the model. Since disease activity is a concept for which a criterion does not exist, the utility of one dependent variable is misleading. As argued in Chapter 2, even colonoscopic assessment, a very important construct, could not be regarded as a gold standard for disease activity in UC. Acknowledging this limitation, physician’s global assessment of disease activity was chosen as the dependent variable in an attempt to quantify as many attributes as possible in the outcome. However, it is likely that using a different outcome (e.g. endoscopy score) may have changed the resulting model.

Item reduction may be performed by including all potential items in a multivariable regression model. However, even in cases where very large databases are available, the validity of this pure mathematical modeling is questionable. As Feinstein argued, these models lack clinical relevance [46] and are exposed to statistical error and selection bias. Regression modeling should be informed by clinical knowledge and not be treated as a "black box" [31, 35].
The simplest way of selecting variables for the model is to perform a univariate analysis of each item with the outcome and then selecting the ones with the lowest P value or the highest correlation coefficient. The authors of the Seo index used Spearman’s rank correlation coefficient to study the relationships between disease activity and 24 clinical variables. Significant correlations (P<0.05) were found in 18 of the 24 initial items, which were then included in a multivariable model. This univariate selection has no implication on the strength of the association (large sample size will also result in a significant P values for a small effect size) and items that are not significant in the univariate analysis may prove important in conjunction with other items. Therefore, despite the simplicity of performing univariate step before the multivariable analysis, it is not recommended for item reduction.

For psychometric scales, in which high inter-item correlation is required, factor analysis may be used to select and arrange related items in appropriate domains. The advantage of using factor analysis is that item selection is not based on the outcome and thus, the problem of multiple statistical comparisons is avoided. Moreover, the sample size required for factor analysis is less than for multivariable regression modeling (approximately 5 subjects per degree of freedom for factor analysis versus 10 subjects per degree of freedom for multivariable analysis [31]).

**Combining the judgmental and mathematical item reduction**

Based on the above discussion and as done for the PUCAI, it is the view of this author that item reduction should be performed in two steps, first judgmental and then mathematical. It is important to employ data reduction technique before doing a
regression analysis with no reference to the response data (i.e. no univariate analysis) [31]. Multivariable models select items based on one outcome, but other considerations may be equally important. The judgmental process should retain all items that fulfill the multiple considerations discussed above (e.g. importance, frequency, responsiveness, feasibility) and then multivariable modeling may be used to identify the most important items in explaining an important outcome.

Clinimetrics and psychometric strategies were previously contrasted for item reduction in psychometric scales only: the Disabilities of Arm, Shoulder and Hand (DASH) scale [51], the QuickDASH [53], the Quality of Life after Myocardial Infarction Questionnaire (QLMI) [52], and the Asthma Quality of Life Questionnaire (AQLQ) [56]. The content of the resulting tools were considerably different in all four studies. Of the four studies, only the QLMI and the QuickDASH compared the performance of the two versions obtained by judgmental and mathematical item reduction [52, 53]. The tools obtained by clinical judgment showed slightly higher concurrent validity and responsiveness, and slightly lower reliability than the mathematical strategies. These studies on psychometric scales may indicate that with a unidimensional instrument the choice of the exact items are of minor significance as long as they all represent the same attribute of the domain. Thus, these results may not be generalized to a clinimetric activity index such as the PUCAI, which is a multidimensional instrument and the identity of the items is more important. Therefore, a special attention should be placed in choosing the best items to a clinimetric outcome measure. The use of robust mathematical modeling, governed by strong judgmental input, may maximize all available research tools to obtain that goal.
ITEM WEIGHTING

Assigning weights of importance to items will not improve performance of psychometric scales since they are composed of many items highly interrelated [51, 175, 176]. In contrast, clinical indices are typically composed of fewer items, each representing a unique aspect of the phenomenon and each with a potentially different importance in explaining it. Most clinical indices, however, assume equal importance of the included items [63]. This saves the resources associated with an item weighting step but actually the assumption of equal importance rarely holds in the clinical arena.

There are two major strategies to determine the relative importance of the included items in indices [51, 175]. The first is to judgmentally assign weights based on theoretical reasoning of clinicians or patients. The Pediatric Crohn’s Disease Activity Index (PCDAI) and others were weighted using this method [15, 175]. Rigorous methodology should be utilized to obtain the clinical knowledge in a systematic way (e.g. nominal or Delphi groups following a thorough literature review). In the second method, multivariable modeling is used to quantify each item according to its importance in explaining the dependent variable. The latter could be patient or physician’s perspective or an important clinical outcome. The Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) and the Crohn’s Disease Activity Index (CDAI) were weighted using this approach, with physician’s assessment as the dependent variable [14, 177]. In order to weight the index mathematically, a large prospective cohort of patients is required to allow for at least 10 patients per degree of freedom in the final model [75]. The judgmental approach is quicker and cheaper even when done in a rigorous way but it is intuitive to assume that a large group of experts would not commonly agree on extreme
weighting of items (e.g. assigning a weight of ‘0’ to the laboratory items of the PUCAI), regardless of the ‘true’ weight. Kirwan et al. in a clinical judgment analysis, asked two rheumatologists to provide a physician’s global assessment of disease activity on paper charts, and then to state how much emphasis they placed on specific items when providing the global assessment [178]. Both physicians placed comparable weighting across five items, but multivariable modeling of the same data showed that, in practice, the decision making relied on only part of the items. Similarly in another study, some of the items thought by physicians to be important in the diagnosis of acute otitis media, received weights of zero in mathematical modeling of the very same physicians [179]. In a different study, models calculated from clinical judgments explained 88% of the variance on a validation set, whereas rheumatologist’s specified judgment policies could explain only 34% [180].

Taken together, these studies support the choice of mathematical weighting of the PUCAI, despite the cost and time involved. Physicians are poor in determining whether items should be excluded when reaching their overall impression by means of explicit judgment policy. We have subsequently showed (data not yet published), that a Delphi group approach to weight the PUCAI retained the laboratory items, thus reaching a very different instrument. Moreover, the prospective weighting of the PUCAI was also used for formatting the index, identifying non-discriminating items and those with low frequency of endorsement (e.g. ‘fever’ which was rarely found, was removed post-hoc from the item list).
SENSIBILITY

Sensibility consists of an aggregate of properties that make up the common sense aspect of an instrument, including *feasibility, face* and *content* validity [181]. Feinstein defined several dimensions of sensibility, each contains several items [182], and herein I will evaluate the sensibility of the PUCAI accordingly.

Content validity is a subjective assessment whether the measure left out any items that most related people would agree that they are important to the measure. It is thus understandable, that with better content validity we can expect also better general validity. The large multicenter consecutive recruitment, and the geographically wide Delphi group ensured generalizability, face and content validity of the PUCAI [78]. Feasibility encompasses both respondent and administrative burden. An instrument is feasible if the participant and researcher report that the instrument is completed within reasonable limits of participant discomfort and both participant and researcher time constraints. The time needed for the completion of the PUCAI is minimal and physical examination, bloodletting or endoscopic examination, are unnecessary. The incorporation of the PUCAI as an outcome measure of recent several large scale clinical trials, reflect the sensibility of the PUCAI. The PUCAI has clear gradations and definitions and together with its user’s guide, is easy to understand.

The purpose of the PUCAI (i.e. concept (disease activity), use (discriminative, predictive and evaluative), and the need for a new measure) is clearly justified in the introduction of this thesis. The population and the setting for using the PUCAI are also presented (i.e. all pediatric UC patients, excluding proctitis and fulminant colitis).
RELIABILITY

Reliability is the fraction of the true variance from the total variance (true variance and the error variance combined), [20]. In other words, reliability reflects the difference in scoring that is related to repeated measures. A score from any instrument is a product of the true value (which can never be determined with certainty), random error (imprecision) and systematic error (bias) [174, 183]. The sources of variability can relate to the patient variability, observer reliability and instrument variability [184]. There is no such thing as 'test reliability'; reliability has meaning only when applied to a specific setting [76]. For the PUCAI we evaluated the inter-observer reliability (i.e. the scoring consistency obtained by two or more raters on the same subject and at the same time) and intra-observer (i.e. the scoring consistency obtained over time on the same individual with stable conditions). Since one of the major aims of the PUCAI is to evaluate response in clinical trials, the demonstration of high inter-observer reliability is of great importance.

Statistical measures to assess reliability

Different authors advocate various statistical methods to assess reliability. In older studies, reliability was often assessed by calculating the proportions of pairs in agreement [30, 185]. This method could overestimate the true reliability since by chance alone some of the ratings will agree. Statistical significance to prove reliability (e.g. one sided or paired Student's t-test) similarly should not be used as it is sample size dependent. Pearson's product moment correlation coefficient is another method often used for expressing reliability, but its use is strongly discouraged [74, 76, 183, 186-188].
Correlation analysis allows incorporation of only two measures, it is dependent on the range of values (i.e. the larger the range, the higher the correlation), and more importantly, even if correlation between two measures is high, reliability may be poor if systematic bias exists (i.e. best fit line is 45° but not on the line of equality). To address this, Bland and Altman presented the 95% limits of agreements (LOA) by calculating the mean difference ± 1.96 SD of the difference [139]. To differentiate fixed from proportional (i.e. change in the magnitude of the difference between the measures across disease severity) biases, the mean score is plotted against the difference in the scores [187]. A linear regression model is fitted to test whether the slope of the mean values is significantly greater than 0 (when change in scores serves as the dependent variable) and then proportional bias is confirmed [189]. If the mean value of the difference is significantly different from zero (tested by one sample Student’s t test) then a fixed bias exists. The LOA was developed to show how much two different method of measurement differ; if the difference is not clinically important, the two methods are said to agree. It is the task of the researcher to judge whether the LOA are narrow enough for the test to be of practical use. Sample size required for extrapolating the LOA results to another population is at least 40 subjects [186]. The LOA is a valid way to assess measurements by two different methods [189]. Some argue that it could be generalized to assess reliability of readings from the same method, albeit with some limitations, [186, 188, 190], while others strongly disagree [76, 191].

A widely used way of assessing reliability is the Intraclass Correlation Coefficient (ICC) which considers the true and the error variance [71]. The ICC is based on repeated measures ANOVA and produces a unitless measure that ranges from 0 to 1 (but it is
possible to receive a negative value). There are numerous versions of the ICC that can yield very different results [183, 188]. In the landmark study by Shrout and Fleiss, six different models of the ICC were presented [71]. Based on \( n \) targets that are rated by \( k \) judges, three different models may be recognized: 1) Each target \( n \) is rated by a different set of \( k \) judges, randomly selected from a population of judges; 2) A random sample of \( k \) judges is selected from a larger population, and each judge rates each target (i.e. each judge rates \( n \) targets); 3) Each target is rated by each of the same \( k \) judges, who are the only judges of interest. Each model requires a different ANOVA model. The first model is analyzed by a one-way random effects ANOVA, the second by two-way random effects ANOVA, and the third by a two-way fixed effects model. An example for the first model, where the effect of the raters is not crossed with the targets, may be classes of students rating the quality of their teachers. Under the second model we wish to generalize to other raters within the sampled population, whereas in the third, we are interested only in the raters under study. The third model does not include variance associated with systematic error, and thus closely approximate the Pearson r, with its limitation as a reliability measure [183]. This discussion should clarify and justify our choice of the second model for evaluating the PUCAI. If the individual ratings are not used, but rather the mean of \( m \) ratings (\( m \) is not required to be equal to \( k \)), then the reliability of the mean rating is of interest. An example to illustrate this may be assessing ratings of a team of physicians. ICC for mean rating is always greater than the individual rating, and typically, mean rating is used when individual rating is unreliable [71]. In chapter 3 the mean rating was used to evaluate reliability between the two radiologists as it was not our aim to assess reliability of the radiology test in general but only to show
that, on average, the radiologists agreed and thus, our data is valid. The unit that defines $m$ should then be determined by a reliability study or on substantive ground (done automatically in SPSS software- reporting "average measures" output). Each of the three models may be used either with individual or mean ratings, summing the possible models to six. McGraw and Wong expanded the Shrout and Fleiss system to two more general forms, incorporated to the SPSS output [183, 192], and non-parametric ICC methods have been proposed [188].

It must be emphasized that the ICC, comparing variability of between to within groups, is highly sensitive to the data distribution [186, 188]. Highly variable data will result in high ICC values even in the context of poor reliability. Therefore, the interpretation of the ICC cannot ignore the distribution of the data under study. It has also been argued that just like the correlation coefficient, the ICC measures association and not agreement [187, 189], while others disagree [76]. In any case, a measure of absolute agreement should be reported alongside the ICC, such as the LOA method [193, 194]. The plot suggested by Bland and Altman is very useful in exploring the data; it provides a graphic view of random, fixed and proportional bias and outlying values are clearly identified [139, 187]. As an alternative, the ICC and a measure of the data distribution could be combined by using the standard error of measurement (SEM) method (i.e. $SD(SQRT(1-ICC))$) [195]. The incorporation of a measure of error with a summary of the data distribution produces more stable results, but it has been shown that it still somewhat depends on the data distribution [186]. The SEM has the same units as the instrument under study, and thus more intuitive. The equation $SEM \times 1.96 \times SQRT (2)$, can be used to determine the difference needed between separate measures on a specific subject for it
to be considered real [183]. The 1.96 is the z-score that corresponds to 95% CI, and the \( \sqrt{2} \), relates to the two repeated measures. In chapter 2 only ICC was presented due to space limitation but I supplement it here with the SEM approach. For the PUCAI, the SEM equals 1.9 points (i.e. 7.7 \( \times \) \( \sqrt{1-0.96} \)), and the difference in score in a specific individual to be considered real is 5.2 points (i.e. 1.9 \( \times \) 1.96 \( \times \) 1.41), further illustrating the high reliability of the PUCAI.

The weighted Kappa statistics on ordinal scale approximates the ICC, and thus similar criticism of the ICC have been also directed at the Kappa [187]. Another drawback of the kappa is that the two raters must always be the same individual for the results to have any meaning [187].

**Strategies to improve reliability**

Reliability is a crucial aspect of an instrument; the higher the reliability the smaller the sample size required to prove a certain effect size in a clinical trial [196], and the higher the maximum possible validity [19]. Several ways were used to enhance reliability of the PUCAI. In addition to the total score, the reliability of individual items was assessed excluding those with low ICC (i.e. abdominal tenderness). Alternatively, one item at a time could have been excluded while looking at the overall ICC changes.

Reliability may be improved significantly by training of the raters and standardizing responses. Individual training of raters is not feasible for the PUCAI as it is intended for global use. However, we provided a user guide (appendix 2.1) to standardize responses, and thus increase consistency of rating.
Increasing the number of items, should also increase reliability [76]. This was not required in our case since the PUCAI was sufficiently reliable as is.

**VALIDITY**

Validity is the concept of assessing whether a measure truly measures what it is meant to measure. An unreliable index is never valid and thus, reliability should be confirmed before assessing validity. Three major types of validity exist: content validity (discussed above under 'sensibility'), criterion validity (including predictive validity) and construct validity (including concurrent, convergent, divergent and known groups/discriminant validity) [19].

The definition of criterion validity is the correlation of a scale with a gold standard of the disorder under study [19]. The gold standard can be concurrent or predictive, depending on the conceptual framework of the scale. Although disease activity is a concept for which concurrent gold standard rarely exists, it has been widely accepted that macroscopic appearance of the inflamed mucosa is the most important measure in UC activity [12]. In fact, the FDA now requires all randomized controlled trials performed in adults with UC to include colonoscopy as the primary endpoint. Nonetheless, due to the reasoning raised in the discussion of chapter 2, we included colonoscopic appearance as a construct and not a criterion.

**Construct validity**

Since disease activity is a concept difficult to explain with a single measure, construct validation was used for the PUCAI. It is a mini-theory to explain the
relationship among various attitudes [197] in a process hypotheses testing: someone who scores X in the index will score Y in a related measure. The hypotheses should be set *a priori* to avoid bias and may be presented as correlation coefficients (rho).

Different types of construct validity exist: concurrent validity (e.g. the PUCAI score was highly correlated with colonoscopic appearance), extreme and known groups (measuring the discriminant ability of a scale in different populations with known disease characteristics; e.g. the PUCAI differentiated well the patients in remission and with severe UC judged by the Truelove and Witts classification), convergent and divergent validity (how our measure is in agreement or disagreement with a different established measure; e.g. the PUCAI was highly correlated with the invasive Mayo score) and the multitrait-multimethod matrix [19, 198, 199]. The latter is a technique looking at convergent and discriminant validity at the same time. It is suitable for a scale with several domains rather than a clinimetric index [200], and thus was not used in the PUCAI evaluation.

**Predictive validity**

This is the ability of the measure to predict an important outcome. Although the ability of an outcome measure to differentiate patients who will deteriorate or not adds to the validity of the measure, the implications of predictive indices are mostly clinical. Calculating the PUCAI during admissions of children with severe UC should aid clinician in the difficult decision making of introducing second line therapy. Indeed, one of the peer-review comments received on the manuscript of chapter 3 was "Most importantly, the paper identifies clinically relevant criteria that can be used in a timely
manner at the bedside to help manage children with severe UC; this is likely to change practice by encouraging early decision making on second line therapy" (addendum 3.1).

Three steps are required before incorporating a clinical prediction rule to clinical practice [37, 201]. The first step includes the derivation of the clinical rule according to the phases outlined above (i.e. item generation, reduction and weighting). Various statistical techniques exist for development of the clinical rule. Harrell et al. compared different strategies for item reduction for clinical prediction rule and concluded that performing regression modeling on clustered summary indices or following incomplete principal component analysis, had the best predictive discrimination [31, 32]. The fundamental flaw of automated regression analysis was discussed in the introduction. The resulting model from stepwise procedure is like the weatherman saying: "...there's a 40% chance of rain, and 10% chance that I know what I am talking about…" [202]. It is also dependent on sample size; the more data we put into the model, the more it will reflect the particular population on which it is based, and the less it will be applicable to new population (which is the meaning of overfitting). Internal validity could be assessed on the same population using bootstrapping or, better, splitting of the cohort into derivation and test cohorts [203].

Prediction rules must show accuracy. A prediction may be inaccurate in two ways; it could be un-calibrated (systematically high or low), and/or it could have an error in discrimination (ranking of individual risk is out of the right order) [202, 204, 205]. While discrimination assesses whether higher estimates of risk are given to individuals who are at higher risk, calibration assesses how close the given estimate is to the true underlying probability of the outcome for a particular individual. For logistic regression,
there are several goodness-of-fit indices that can be used to assess calibration, the most
common is the Hosmer-Lemeshow goodness-of-fit statistic [206]. It measures how far the
estimated probability for a particular case is from the probability obtained from a set of
similar cases. Small values indicate a good calibration while values exceeding 20 indicate
significant lack of calibration (P<.01) [205]. Another way to show calibration of
predictions is to plot observed frequencies against predicted probabilities [207]. These are
best presented at the development phase of the predictive index. The PUCAI was not
developed according to these prediction techniques as it was not aimed initially as a
predictive rule. However, that does not preclude an attempt to assess its prediction utility
on a separate cohort.

Discrimination is more easily calculated. For the PUCAI, it was established by
diagnostic utility tests (i.e. sensitivity, specificity, predictive values, likelihood ratios and
area under the ROC curve). Providing P values and odd ratios as a proof for predictive
validity is wrong and misleading, as only extremely high odd ratios correspond to
reasonable sensitivity and specificity of a diagnostic test [208].

It is important that the examiner who scores the test will be blinded to the
outcome, and that the outcome is determined independently from the predictors [37]. A
major limitation of our retrospective study is that it inevitably violated both of these
requirements. Therefore, a similar study is currently underway using prospective cohort
of children admitted with severe UC. In this study, the predictive ability of the PUCAI
will be validated on different patients, as the accuracy of prediction rules degrades from
the sample it was first developed. Indeed, generalizability must be determined before a
clinical rule is widely accepted (i.e. reproducibility and transportability) [204].
Reproducibility requires the prediction rule to replicate its accuracy in different patients from similar population. Transportability requires the rule to produce accurate predictions in a sample drawn from a different but plausibly related population or in data collected using methods slightly different from the original study. Both requirements will be evaluated in the ongoing prospective study, which is different from the study presented in chapter 2 in the period, geography and method of enrolment.

According to the level of evidence of clinical prediction rules [201], the PUCAI could be currently classified as level IV prediction rule (i.e. index not validated or validated on a retrospective cohort). Following the prospective study, the PUCAI may qualify for level II prediction rule (i.e. validation in one large prospective study), allowing it to be used in various settings with confidence. Level I rules are required to be subjected successfully to impact analysis. This can be done in the RCT setting where the intervention is the different decision making (using cluster or individual randomization), or using historical cohort as a control for the new decision making algorithm. The latter is the simplest, albeit the weakest, in strength of design.

**Validity-final note**

Validity is a matter of degree, not an all-or-none phenomenon, and is an evolving property. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide the use of the test [145]. As Streiner pointed out, the important question of validity is: “do the results of this study allow us to draw the inference that we wish to make?” [19]. I believe that according to the results presented in this thesis, the answer is 'yes' for the PUCAI. Nonetheless, evaluation of any instrument
is a continuing process, in the progress of understanding what the test score means under different scenarios and populations.

**RESPONSIVENESS**

An increasing number of health-related outcome measures are used in longitudinal assessment in clinical trials. For the outcome measure to be used in this way, both longitudinal reliability (test-retest) and responsiveness should be determined. It had been shown that a measure could be valid but not responsive, and vice versa [65]. For instance, the PCDAI has an item of height velocity that is calculated over the preceding 6-12 months [15]. This item may increase the validity of a pediatric activity index, but certainly reduces responsiveness. Despite the fact that responsiveness is considered the most essential property of an evaluative measures [22, 65, 74] the methodology of assessing responsiveness lags behind other psychometric techniques. Twenty five definitions, and 31 different formulas have been found to assess responsiveness [209]. Beaton et al. have made a significant attempt to order the types of possible changes using the “cube classification system of responsiveness” [23, 25, 210].

**Statistical strategies to assess responsiveness**

There are three general approaches to detect responsiveness: effect size statistics that present the magnitude of change in the evaluated index (e.g. tests that calculate the ratio of a signal (observed change) to noise (some measure of variance)), cor relational and diagnostic utility (e.g. ROC curves, sensitivity and specificity) [21]. Each approach emphasizes a slightly different aspect of responsiveness. For the PUCAI, we used all
three approaches, including a variety of summary statistics suggested in the literature for comprehensive evaluation.

*Effect size statistics*

Effect size statistics is also termed 'distributional' and 'internal' responsiveness, highlighting the fact that it measures distributional property of the scores. The most commonly used strategies were utilized here: Cohen's effect size, the standardized response mean and Guyatt's responsiveness statistics [211]. The latter may be superior to the others as it relates the MCID to spurious change of stable patients [74], but the MCID must be available in order to use this index. These values express the magnitude of change in terms of variation. Paired Student’s t test, also in the category of effect size, is sample size dependent and thus not recommended [209, 211, 212]. An exception is when comparing the responsiveness of two measures; then the relative efficacy index, calculated by squaring the ratio of the paired t-tests, may be beneficial [213]. Paired t-test was calculated for the PUCAI for the sake of completion as it is commonly seen in the literature despite its limitations. The main disadvantage of the effect size approach is that it cannot measure responsiveness at an individual patient level [211]. In other words, a score of >0.8 is considered as large effect, but it is not related to the amount of change that actually occurred in the patient, and whether it was significant or not. It should be noted, therefore, that the arbitrary cutoff values of 0.2, 0.5 and 0.8 effect size should be interpreted according to the expected magnitude of change [24, 74]. Therefore, the distributional methods were calculated separately on the improved and the unchanged group, judged by the external anchor. As presented by others, the same order of
responsiveness was found here using either of the effect size statistics; high for the changed group and low for the unchanged group [91].

**Correlational**

Meenan et al. suggested correlating the change score with another established responsive measure [214]. A similar analysis was performed here by correlating the PUCAI change with two external criterions: the 7-point Likert scale of change and the change score in the PGA scored on a 100 mm VAS. The correlational approach preserves the gradient of improvement (i.e. categorization is not necessary) and was found to be consistent with the diagnostic utility approach [21]. It should be acknowledged, however, that correlation analysis measures only linear association (which may not be the case in responsiveness [211]) and is highly sensitive to the range of values and outlying scores. The correlational approach was extended to regression models, but this is not widely used [215].

**Diagnostic utility**

Deyo argued, that responsiveness is not merely sensitivity to change, but the ability of the instrument to differentiate between those who changed and those who did not [21]. The ROC curve is conceptually perfect to assess this rationale. It quantifies, in intuitive values of sensitivity, specificity, and area under the ROC curve, the exact outcome needed from the evaluative property of an index. It is a graph of ‘true positive’ (sensitivity) versus ‘false positive’ (1-specificity) for each of several cut-off points in score change. The area under the curve can be interpreted as the probability of correctly
discriminating between improved and non improved patients. The PUCAI showed high area under the ROC curve, reflecting high responsiveness.

Unlike the internal responsiveness, the ROC approach utilizes an external approach, requiring an external criterion to anchor the calculated change to treatment effect. ROC is increasingly used in longitudinal assessment of health related outcome measures. The likelihood ratio method was elegantly adjusted to estimate the change score of the Roland Morris Back Pain Questionnaire [152]. A comprehensive discussion of the topic was presented in chapter 4 and below under the MCID section.

**Synthesis of the three approaches**

Interestingly, the results of the different statistical formulas and the three different approaches were consistent in any sub-analysis performed. This was also found by others, including the comparison between the ROC approach and the effect size statistics [90, 91]. Since the different approaches represent different conceptual views of responsiveness, it is the view of this author that all three should be presented in responsiveness analysis. No gold standard exists for evaluating responsiveness [22]; but is there a preferred method?

The degree of change in the distributional approach is greatly affected by the type of population under study, the effectiveness of therapy administered, the timing of the repeated evaluation following the therapy (i.e. longer interval allows full effect of the therapy, and regression to the mean phenomena), and which construct of change was utilized [25]. The resulting effect size value reflects not only the generic property of the instrument but also the conditions of the study, and although it has a role in sample size
calculations [74], it does not have a meaning by itself to quantify the degree of responsiveness. Some argue that the difference between change as a whole and change related to clinical improvement is the difference between longitudinal validity and responsiveness [211, 216]. I feel that this differentiation adds very little to the concept and it is rather arbitrary and confusing. Even the internal responsiveness methods require a-priori hypothesis regarding the expected magnitude of change, and thus, the identification of the patients who indeed changed. This complies with the original definition of responsiveness to detect change when it occurs.

Since the correlation coefficient preserves the gradient of the two change scores used, it provides insight into the validity of the degree of change, assuming that the reference change score is sufficiently reliable [217]. However, it does not translate to a meaningful number quantifying responsiveness. In contrast, the ROC approach provides a clear description of responsiveness in measures similar to the ultimate use of the instrument: how likely the measure is to truly differentiate the responders from the non-responders in a given clinical trial. The area under the ROC curve serves as a summary measure of the overall diagnostic accuracy and may be easily used for comparing the responsiveness of different indices. Another advantage of the ROC approach is that it allows the analysis of responsiveness at the individual level, whereas the distributional approach is at the population level. The challenge of interpreting group data at an individual patient data, most often required at clinical trials, has been previously demonstrated [25]. On the other hand, ROC analysis requires that the external outcome criterion is dichotomous rather than preserving the degree of change. To reduce this limitation, it has been suggested in chapter 4 that several MCID will be presented.
reflecting minor, moderate and large response and separate ROC curves should be employed for the different cutoffs and for deterioration and improvement.

Relating the change in score to the observed change may utilize an external criterion of change or a cohort of patients deemed to have changed following therapy of known efficacy [82] [66, 213]. As discussed below the anchor based approach is not without limitations, but using only patients deemed to have changed is also not perfect. The most effective treatment to induce remission in UC is corticosteroids, yet only half will enter remission in 1 month, 30% will improve and 20% will show no response [218]. On the other hand, some patients will improve without any therapy [219]. In addition, the magnitude of responsiveness depends on the efficacy of therapy given (e.g. 5-aminosalicilic acid preparation versus infliximab for UC). We thus assessed the responsiveness of the PUCAI using both approaches (i.e. GRC and patients deemed to have changed), obtaining the same order of magnitude.

All in all, it is the view of this author that the diagnostic utility approach should be considered superior to others approaches. As others found [57, 90, 220], the same degree of responsiveness was found across all methods but this is not universal [221].

**Responsiveness-summary**

It is not surprising that the PUCAI was found to be very responsive. First, the conceptual framework of the PUCAI was evaluative according to Kirshner’s criteria [16], and as such we excluded non-responsive items (e.g. hemoglobin and height were not included). Second, it has been shown that shorter indices are more responsive than longer ones [21]. Third, disease activity in UC may change significantly just in a few days, and
thus the PUCAI measures a responsive concept. The Sickness Impact Profile (SIP) for example, aims at a relatively stable concept by its nature (i.e. functional status) [217] and, thus, responsiveness could be difficult to prove. Fourth, disease specific indices are more responsive than generic ones [221] and finally, clinimetric indices are generally more responsive than psychometric scales [46, 222]. Not only because they are comprised of less items, but also due to the lower inter-item correlation.

It should be acknowledged that in order to generalize the responsiveness of PUCAI, more studies are required on other populations (e.g. children with severe UC). Inherent problems with change score of any instrument relate to 'regression to the mean phenomenon' and to 'floor and ceiling effect' [217]. It is likely that different change scores should be used in accordance with the study design and population under study (e.g. larger change scores for higher baseline scores and vice versa). These limitations and possible solutions are further discussed in the next topic of the minimal important difference in score that should define response.

MINIMAL (CLINICALLY) IMPORTANT DIFFERENCE

The Minimal Clinically Important Difference (MCID) was originally defined as the smallest change in an instrument perceived as beneficial to patients or physicians and that leads to a change in the patient’s management, assuming no toxicity and cost [130]. Later, this definition was revisited to include minimal toxicity and cost in the consideration [67]. The importance of using the MCID is that it is more relevant to present the proportion of patients achieving a particular benefit, rather than reporting the mean difference between two study groups [12, 223]. Of all parts of the PUCAI
development, determining the change score that defines ‘response’ for clinical trials was the most challenging. Therefore, a chapter was devoted to this topic (chapter 4), and will be further discussed.

**Different approaches to define the MCID**

Wells *et al.* found nine different approaches to calculate the MCID [224]. These, may be clustered into anchor-based, distributional, and data-driven [70, 142]. The MCID may also be determined from a judgmental process of experts in the fields reaching consensus, like the Delphi group described above [130, 225]. The following discussion will focus on the first three approaches.

*Anchor based approach*

As discussed in chapter 4, determining the MCID requires clinical relevance, and thus an external anchor is required. Many external anchors have been used, but the most commonly utilized anchor to date is a global rating of change (GRC), scored on a 3-15 point Likert scale [21, 68, 130, 143, 146, 147, 226]. Other anchors include status on an important measure of function [160], the presence of relevant symptoms [161], change of therapy [150], mean score of patients deemed to have changed following therapy of proven efficacy [82], disease severity [162], and outcome criteria such as mortality or resuming full activity [21, 163].

The anchor-based approach requires two steps: one establishing the MCID, and the second examining the proportion of patients who achieved this cutoff [166]. The anchor could be computed by the patient, the physician, the payer or the society,
depending on the concept of the instrument. Since the PUCAI is a measure of disease activity and not quality of life or cost-effectiveness tool, the physician was chosen to determine which cutoff is interpreted as “clinically significant”. This may be criticized by the fact that the physician’s knowledge of the patient’s symptoms comes from the patients and, thus, they should be asked directly [82]. Moreover, physicians often underestimate functional disabilities, pain and discomfort reported by their patients [142]. Indeed, the change should have been reported by the patient if our measure was intended to be patient reported instrument (e.g. health related quality of life). However, the PUCAI is a physician-based clinimetric index, intended for scoring following clinical judgment whether the symptoms are related to disease activity or not. For instance, if a disease is no longer active but the patient still has perianal pain from the previous frequent diarrhea, the patient's score will reflect change in quality of life but not in disease activity. Therefore, the judgment of clinical change should also be judged by clinicians, integrating thorough history taking, physical examination, and laboratory tests.

Having chosen a single anchor, several analytic methods are available to derive the change score that correspond to an important change on the anchor. The first and most widely used technique utilizes the mean change of the group that fulfills the criterion [130]. However, I have shown that this technique has several profound limitations, which are successfully addressed by using the ROC approach to differentiate a clear cutoff within the entire cohort (Chapter 4). This also enables a clear description of the magnitude of the change by means of sensitivity, specificity, and allows the comparison of different approaches using the area under the ROC curve. Indeed, it was
shown in Chapter 4 that the ROC method provides a less variable estimate for the MCID across the three change scores (i.e. small, moderate and large).

The GRC is the easiest and most applicable of all anchors, but it is not without limitations [68, 82, 135, 227]. Recall bias exist in comparing the current state with a previous one. Patients provide retrospective estimates of change that are highly correlated with their present state [82], and are biased by what the interviewer wants to hear. The rater may be influenced by recent good or bad events that falsely alter the judgment of change for the entire period. This limitation may be somewhat less relevant for the PUCAI where physicians rated the GRC. Physicians are more objective in assessing disease activity since they have comparable reference for scoring, having treated many other similar patients. It was shown in Chapter 4 that MCID values scored by physicians (i.e. PUCAI and PCDAI) are less heterogeneous than patient-based rating. Nonetheless, even among physicians there is still large variation between in judging degree of change [167, 228]. Another concern, applicable to all anchor based methods, is how to address MID that falls below the threshold of MDC, as found by us (Figure 4.1) and others [134, 229]. Finally, using different anchors may lead to different MCID values, and there is no consensus which one should be used.

Despite these limitations, the GRC is the most widely used anchor-based technique to ascertain the MID [142]. Several measures were used to minimize bias in using the GRC for the PUCAI. First, two physicians scored the GRC independently and the mean score was considered. Second, to reduce recall bias, we encouraged physicians to review the notes from previous visit, while still blinded to the previous PUCAI score. Finally, the validity of the GRC was established using the four criteria proposed by
Guyatt *et al.* and outlined in chapter 4 [67, 68, 230]. Since our GRC satisfied these criteria it may be concluded that the GRC score in our longitudinal cohort reflected the true change. These criteria can be evaluated only upon the completion of the study and since the correlation of GRC score is often less than 0.5, several anchors should be collected to account for these cases [142, 147].

2. *Distributional based approaches*

The distributional-based approaches relate the change score to some measure of the data variability. External anchors are not required and the calculations are straightforward. Only an external anchor can determine the MCID, but an MID (i.e. avoiding the ‘clinically’ within the definition) may be calculated from the distributional approaches. I argued that the MCID concept should be preferred over the MID, but the concept can be very useful to estimate the MDC (chapter 4). The statistical strategies include the effect size statistics (including Cohen’s effect size (ES) [136], and standardized response mean (SRM) [72], variations of the Standard Error of Measurement (SEM, including Wyrwich SEM [137], and Jacobson’s Reliable Change Index (RCI) [138]) and the Bland and Altman limits of agreements [139]. The SEM is defined as the standard error in an observed score that obscures the true score [142]. As evident in Table 4.1, different values were suggested to characterize the MID within the effect size statistics (i.e. 0.2, 0.5 and 0.8) and the SEM methods (i.e. 1, 1.96 and 2.77).

We compared the various distributional based strategies to determine the MDC within themselves and to the anchor based methods (Figure 4.1). We have shown that no single distributional approach consistently correlated with the anchor based approach,
between the different instruments and across small, moderate and large change score. This is understandable, since the distributional based approaches are greatly influenced by the distribution of data under investigation [136].

Some suggest that the MDC should actually serve as the change score that define response [134]. A strong argument in favor of this approach comes from the original definition of the MCID (i.e. a change that warrants change in therapy assuming minimal costs and toxicity). Any true change (i.e. beyond error), no matter how small, is worth treating with a therapy associated with minimal cost and toxicity. Although the MDC determines the cutoff value in which one can be confident that change really occurred, it does not address whether this change is important. Ironically, however, the MDC is often larger than the MCID [134, 144, 149, 229]. Similar results were presented in this thesis (Figure 4.1) especially for the 95% limits of agreements that was consistently higher than the clinically relevant MID, at times even more than moderate change.

In general, the SEM approach is the preferred method among the distributional-based approaches. By counterbalancing reliability and the inter-observer variability, the SEM is less dependent on the data distribution and produces more stable results than the effect size [135, 137, 142]. Furthermore, since the SEM is sample independent, it should not be dependent on sample size, which is a primary criticism of the other distribution-based methods. The SEM is expressed in the units of the instrument, facilitating its interpretation. Nonetheless, some variation still exists across different populations, and it is generally accepted that although the SEM is the preferred distributional approach strategy [142], the MID will still somewhat differ from study to study [135, 149, 166, 186]. One limitation of the SEM approach is that it assumes constant measurement error.
across the range of scores, but at least in one study, smaller SEM was found at the extremes of scores compared with around the mean [164]. Jacobson’s RCI yields larger values for a given change than Wyrwich SEM approach by a factor of 1.41 (i.e. \(\text{SQRT}(2)\)), as it multiplies the 1-r in the denominator by 2, to account for the test-retest samples [23, 138]. Other modifications of the RCI have been proposed, but are not commonly used [135].

3. Data driven (change observed in a population)

This approach denotes observing a population of interest and correlating the change score with various outcomes. This strategy is referred by Guyatt and colleagues [166] as “population based approach” (unlike the “individual focused approach”). It had been suggested that this is the most accurate (yet most demanding) method to establish the MCID [140]. Nonetheless, it is important to emphasize that, unlike the individual-focused approach, the population based approach does not identify a clear cutoff score that differentiates trivial from important scores, but rather translates observed change score into meaningful outcomes. These data are typically accumulated through extensive use of the instrument in longitudinal clinical studies. Several examples exist of studies that utilized this approach [145, 149].

Comparisons of different MCID methods (Table 5.1)

Only a few papers compared the anchor and distributional-based methods “head to head” on the same data sets, which are tabulated in Table 5.1. Since each study contrasted different methods, it is not possible to generalized a conclusion. One SEM
with a test-retest reliability of 0.75 was found to be comparable with the mean score of the group rated as 'small change' on a GRC scale, and also comparable to an effect size of 0.5 [137, 166]. However, since most outcome measures have higher reliability, this comparison may not be generalized. The association between the mean score of the small change with an effect size of 0.5, was replicated in an unrelated study [132]. Cella et al. found that the anchor-based approach (criterions chosen were symptom-based) and two distributional approaches (one SEM and 0.2-0.3 effect size) yielded comparable MCID values of the Functional Assessment of Cancer Therapy-Lung (FACT-L) on 573 patients with lung cancer [161]. In our study (chapter 4), 1 SEM always lay between 0.2-0.5 effect size units (both are distributional-based approach), but the relation to the anchor-based methods varied.

The ROC method to establish the MCID using rheumatologists’ opinion whether patients had a clinically relevant change or not (2-point scale) yielded larger MCID than the MDC, calculated using the 95% limits of agreements [149]. The best cutoff on an ROC curve to differentiate patients with improvement of at least +3 on a -7 to +7 point GRC scale was comparable with 1.96 SEM of the numerical pain rating scale (NPRS) [151]. These heterogeneous studies support our results that no consistency could be found between the anchor and the distributional-based approaches. In view of the dependence of the distributional approach on the population under study and the different available methods to relate anchors to the observed change, this conclusion should come as a very small surprise.

As a consequence of this lack of correlation it may be reasonable to present the results of both the MDC and the MCID, calculated by the distributional and the anchor-
based approaches, respectively. In this way the performance of the measure across populations and manipulations is well defined [231], while the MDC provides the boundary of error of the MCID. Crosby et al. suggested using the more conservative of the two methods as the MCID [232]. Other methods were also suggested to incorporate both the anchor and distributional approached in one strategy [133, 233].

MID found in one study may not be generalized to another. On the other hand it was argued that the MID of a generic instrument should not vary considerably between populations, as it questions the use of the measure as a generic instrument [234, 235]. Only extensive and ongoing experience can provide fine tuning and recognition of what change of an instrument means in clinical life.

**Utilization of the MCID of the PUCAI**

*Three MCID values*

Like others [70, 167, 168], we argued that different MID values may be presented, corresponding to small, moderate and large change scored. The score corresponding to small change could be determined either by an anchor or by statistical distributional manipulations (reflecting change beyond statistical error), but the other two MCID values *must* be referenced to a clinically significance anchor. Farrar et al. [223] listed seven factors that influence the choice of a change score to define response, most are relative to the population under study (i.e. severity of disease, desired effect, toxicity, measurement units of the outcome, degree of the instrument responsiveness, who is the subject of the instrument, the study question).
MCID dependence on baseline score

It has been shown that the MCID is dependent on the baseline score, partially due to regression to the mean phenomenon, and ceiling/floor effect of the instrument, and partially due to psychological basis where patient’s perception of change depends on how well they previously felt [70, 145, 164]. Klooster et al. confirmed that patient-perceived pain improvement is not uniformly distributed over the range of VAS [236]. It was also shown that the magnitude of a MCID of pain increases as baseline scores increases [237-239].

We could not stratify our PUCAI cohort by the baseline score due to the small sample size of very active patients. The dependence of the PUCAI’s MCID on the baseline values is yet to be determined. Nonetheless, the presentation of small, moderate and large MCID for the PUCAI may be utilized also in this regards. A larger MCID could be used in the case of high baseline values (i.e. including severely active patients). A 10 point difference from 85 to 75, may not be the same between 30 and 20 [69, 133]. It is challenging to determine whether a patient with an initial score of 10 (i.e. mild disease activity) who experienced a reduction in PUCAI to 5 (i.e. remission) should be considered as a treatment success, as the patient did enter remission. On the other hand, this change is smaller even from the MDC.

To overcome these limitations, a percent change from baseline score could be used in assessing response [168]. The percent change of the PUCAI, stratified by the degree of improvement is presented in Figure 5.1. Since ‘no improvement’ clustered with ‘mild improvement’, and ‘moderate’ clustered with ‘large improvement’ a cutoff point of ‘at least moderate' was used to plot an ROC curve of percent change (Figure 5.2). A
change of at least 30% in the PUCAI score reflect moderate MCID, with a sensitivity of
91% specificity of 87% and area under the ROC curve of 0.96 (95% CI 0.91-1).

A recent study found that a 5-10% change of the total instrument score is
comparable with other MID methods (i.e. 0.33-0.5 effect size, 1 SEM, and 0.5 points per
item) [240]. This may correspond to small change only. Klooster et al. used a 7-point
Likert scale of change to identify the MCID of VAS of pain. He found that a change
score of 55% corresponds to the same cutoff that we used (i.e. at least moderately better).
Figure 5.1: Percent improvement of the PUCAI, stratified by physician’s global assessment of change (n=48)

Patients with remission at baseline (PUCAI<10) and those deteriorated were excluded.

Figure 5.2: ROC curve of percent improvement of the PUCAI, to differentiate patients with at least moderate improvement (n=33) from unchanged (n=15)

Area under the curve=0.96 (0.91-1); best cutoff 30% change (sensitivity 91%, specificity 87%).
Patients with remission at baseline (PUCAI<10) and those deteriorated were excluded.
**Improvement versus deterioration**

The MCID of the PUCAI was determined on patients who showed improvement, and this may not be generalized to deterioration. Psychologically, patients admit to deterioration only when a substantial worsening occurs. In contrast, patients often are looking for the smallest difference to reassure themselves of improvement. Asymmetry in MID for improvement and deterioration have been previously reported [241]. A larger cohort of deteriorating patients should be evaluated to determine the PUCAI’s MCID of worsening.

**MCID: Final Notes**

No single method has been shown to ideally define the MCID, and its contrasting conceptual approaches result in extremely variable conclusions. Unfortunately, determining the MCID is also the most important issue raised by investigators designing clinical trials. Therefore, a large section of the thesis was dedicated for determining the MCID of the PUCAI. The different methods available yield a range of values. Which cutoff should be used in which study, or whether percent change score should be also utilized, will hopefully be further clarified through the use of the instrument in research and clinical practice.

In proposing the ‘three MCID’ concept, I recognize the illusion of one generic MCID [69, 70, 167]. Hays wrote [69]: "using the MCID as an absolute threshold (without consideration of cost) is just plain wrong". Following this rationale, the PUCAI’s small, moderate and large MCID may be defined as change of at least 10, 20 and 35 points,
respectively, but a 30% improvement may also define moderate change. Most studies using the PUCAI should find the moderate cutoff most appropriate, but investigators may wish to report the proportion of patients achieving all three change scores (i.e. small, moderate and large).

**LIMITATIONS OF THE PUCAI**

It is important to acknowledge that the PUCAI cannot be confidently used in scenarios outside the framework of this thesis. For example, the PUCAI was not assessed in children with UC confined to the rectum. The PUCAI has a ceiling effect and thus, cannot differentiate between the severe to the fulminant end of the disease activity spectrum. Moreover, the sample size of severely active children in our study was small in the original study. Nonetheless, in a retrospective analysis of children with severe UC, the PUCAI seems to function reasonably well also in this population (Chapter 3). We are in the process of further characterizing the PUCAI in severe UC, using a prospectively enrolled cohort. Lastly, the PUCAI can be confidently used only following the methods utilized here. For example, the PUCAI as it was used in the evaluation was scored by experienced gastroenterologists. It is not known whether the same reliability can be achieved when the scoring is done by research coordinators, nurses, or even by the patient themselves. Similarly, the performance of the PUCAI, presented in Chapter 2, reflects data obtained in a clinic visit and not from charts or over the phone.
FUTURE DIRECTIONS

The evaluation of a health-related outcome measure is an on-going process of learning more about the measure and making new predictions [19]. The PUCAI was developed and evaluated using high standards of methods available, but adjustments may be necessary according to real life experience. I look forward to receiving the data from the ongoing trials that incorporated the PUCAI to their protocol, in order to validate the results of this thesis, on different populations.

None of the items retained in the PUCAI, are specific to children and, thus, we initiated a study to evaluate the PUCAI in adults. It is not reasonable to anticipate that the PUCAI will replace endoscopic evaluation in adults, due to the relative ease of performing sigmoidoscopic evaluation in this population. However, a well validated responsive, non-invasive, disease activity index is invaluable for assessing disease in between sigmoidoscopy visits.

The PUCAI contains only symptom-related items (no physical examination findings and no laboratory values). Therefore, it could potentially be completed by a research coordinator, either in person or via telephone, or even self administered by the patient and caregiver. In clinical trials, it is often necessary to complete an activity index repeatedly; the ability to do so over the telephone would obviate frequent visits, thus improving compliance. Therefore, we plan to assess reliability of the PUCAI when completed by patient/family, by a research coordinator via telephone contact, and by a research coordinator in person.

At SickKids, Toronto, as in many other centers, a clinical nurse is the one who receives first calls from IBD patients, and reviews the active cases with a physician. In
high volume centers, it may be useful to provide the nurse with an assessment tool to aid in the decision making over the phone (i.e. referring to emergency room, scheduling urgent clinic visit, or non-urgent review with the physician). We plan to evaluate whether the PUCAI, scored by an IBD nurse over the phone, could dictate such decision making based on its ability to predict the need for admission, corticosteroid initiation or other change in therapy.

Impact analysis of the PUCAI to guide second line therapy in severe pediatric UC is necessary before implementing the PUCAI as a valid clinical prediction rule in this scenario.

It is mandated that the rapidly emerging novel therapies will be evaluated in the clinical trial setting. Robust outcome measures are of utmost importance in determining the outcome of these trials. We have provided an outcome measure that is suitable for use in clinical trials of pediatric UC, and for determining the timely introduction of second line therapy in severe acute UC. The complete lack of invasive items, the large consensus group involved in the development and the rigorous methods used, place the PUCAI as the leading UC activity index in children. Nonetheless, it has been suggested that the original authors of a measure generally report an optimistic picture of its performance [18], and thus, duplication of the results presented here by independent groups is necessary.
Table 1.1: Disease activity indices for ulcerative colitis

<table>
<thead>
<tr>
<th>Author (Ref)</th>
<th>Year</th>
<th>Name</th>
<th>Pediatric/Adults</th>
<th>Rigorously developed?</th>
<th>Evaluated for psychometric properties?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truelove &amp; Witts [8]</td>
<td>1955</td>
<td>Classification for ulcerative colitis severity</td>
<td>A</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Powell-Tuck [10]</td>
<td>1978</td>
<td>St Mark’s scoring system for UC</td>
<td>A</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Lloyd-Still [43]</td>
<td>1979</td>
<td>Clinical score for chronic IBD in children</td>
<td>P</td>
<td>No</td>
<td>Poorly</td>
</tr>
<tr>
<td>Schroeder and Sutherland [6, 42]</td>
<td>1987</td>
<td>Mayo clinic (Clinical Activity Index)</td>
<td>A</td>
<td>No</td>
<td>Poorly</td>
</tr>
<tr>
<td>Rachmilewitz [7]</td>
<td>1989</td>
<td>Scoring system for UC clinical symptoms</td>
<td>A</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Lichtiger [28]</td>
<td>1990</td>
<td>Symptom score for acute UC</td>
<td>A</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Seo [30, 38]</td>
<td>1992</td>
<td>Complex integrated disease activity index</td>
<td>A</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Beattie [44]</td>
<td>1996</td>
<td>Colitis symptom score</td>
<td>P</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Walmsley [59]</td>
<td>1998</td>
<td>Simple Clinical Colitis Index (for initial clinical use only)</td>
<td>A</td>
<td>Yes</td>
<td>Partially</td>
</tr>
<tr>
<td>Azzolini [40]</td>
<td>2005</td>
<td>Endoscopic-clinical correlation index (ECCI)</td>
<td>A</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Table 1.2: Items and gradations used in existing ulcerative colitis disease activity indices

<table>
<thead>
<tr>
<th>Score range</th>
<th>Truelove and Wits [8]</th>
<th>[10]</th>
<th>[43]</th>
<th>[6]</th>
<th>[7]</th>
<th>[28]</th>
<th>[30]</th>
<th>[44]</th>
<th>[59]</th>
<th>Azzolini [40]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild/severe</td>
<td>0-24</td>
<td>12-100</td>
<td>0-12</td>
<td>0-31</td>
<td>0-21</td>
<td>Vary</td>
<td>0-10</td>
<td>0-20</td>
<td>0-3</td>
<td>Vary</td>
</tr>
<tr>
<td>Stools frequency</td>
<td>&lt;4 &gt;6</td>
<td>0-2</td>
<td>0-3</td>
<td>0-4</td>
<td>0-3</td>
<td>0-4</td>
<td>0-3</td>
<td>0-4</td>
<td>0-3</td>
<td>Vary</td>
</tr>
<tr>
<td>Urgency</td>
<td></td>
<td>0-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stool form</td>
<td></td>
<td>0-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nocturnal diarrhea</td>
<td></td>
<td>0-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood per rectum</td>
<td>Yes/No</td>
<td>0-2</td>
<td>0-3</td>
<td>0-4</td>
<td>0-3</td>
<td>0-4</td>
<td>0-3</td>
<td>1-3</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Abdominal pain</td>
<td></td>
<td>0-2</td>
<td>0-3</td>
<td>0-3</td>
<td>0-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well being</td>
<td></td>
<td>1-10</td>
<td>0-5</td>
<td></td>
<td>0-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anorexia</td>
<td></td>
<td>0-1</td>
<td>2-20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nausea</td>
<td></td>
<td>0-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extra-intestinal</td>
<td></td>
<td>0-2</td>
<td>3-20</td>
<td>0-9</td>
<td></td>
<td></td>
<td></td>
<td>0+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abdominal</td>
<td></td>
<td>0-3</td>
<td>1-10</td>
<td>0-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fever</td>
<td>&gt;37.5</td>
<td>0-2</td>
<td>0-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Tachycardia</td>
<td>&gt;90/min</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>&lt;75%</td>
<td>1-5</td>
<td>0-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>ESR</td>
<td>&gt;30</td>
<td>1-5</td>
<td>0-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Albumin</td>
<td></td>
<td>1-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>WBC</td>
<td></td>
<td>1-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anti-diarrhea</td>
<td></td>
<td>0-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rectal prolapse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0-1</td>
</tr>
<tr>
<td>Endoscopy/ X-ray</td>
<td></td>
<td>0-2</td>
<td>1-15</td>
<td>0-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global assessment</td>
<td></td>
<td>0-3</td>
<td>0-3</td>
<td>0-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ESR, erythrocyte sedimentation rate; WBC, white blood cells
Table 1.3: Evaluation phases of a theoretical Disease Activity Index (DAI) in inflammatory bowel disease (IBD).

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Example from a DAI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VALIDITY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Face and content validity</td>
<td>Most experts in the field will judge the index as sensible and that the included items are the important ones</td>
<td>A group of experts in pediatric IBD approves the completeness and sensibility of the DAI</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>Determines the relationship between the measure and a gold standard</td>
<td>Not applicable to IBD</td>
</tr>
<tr>
<td>Construct validity</td>
<td>A mini-theory to explain whether a measure acts the way it is expected based on the concept it represents</td>
<td></td>
</tr>
<tr>
<td>Concurrent validity</td>
<td>The relationship of the index with other measures that reflect the same attribute</td>
<td>The DAI is highly correlated with endoscopy score, histology score and physician global assessment</td>
</tr>
<tr>
<td>Convergent validity</td>
<td>The relationship between the new index and an established one</td>
<td>The DAI is appropriately correlated with a previously established adult activity index</td>
</tr>
<tr>
<td>Extreme group validity</td>
<td>Whether the measure differentiates extreme sides of the disease spectrum</td>
<td>The DAI differentiated IBD children who were sent home without change in therapy from children admitted to the hospital</td>
</tr>
<tr>
<td>Predictive validity</td>
<td>The ability of a measure to predict important</td>
<td>Not applicable to the DAI since it was not aimed to be a predictive index</td>
</tr>
<tr>
<td><strong>RELIABILITY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-observer reliability</td>
<td>Provides a measure of the index error</td>
<td>Two independent physicians scored similarly the DAI, concurrently on the same individuals</td>
</tr>
<tr>
<td>Test-retest reliability</td>
<td>Whether a score is reproducible at different times, when the patient remains stable</td>
<td>The DAI remained unchanged on repeated visits of patients whose disease activity was judged to be unchanged</td>
</tr>
<tr>
<td><strong>RESPONSIVENESS</strong></td>
<td>The ability of an instrument to accurately detect change in disease activity over time, when it occurs</td>
<td>The DAI dropped significantly over time in a group of IBD children treated with steroids for active disease, and not significantly when no therapy was given.</td>
</tr>
</tbody>
</table>

Methods presented in this table are for illustration only and are several of many (see text for detail).
Table 2.1: Items reflecting disease activity in pediatric UC, ranked highest by the Delphi group and their corresponding β coefficients and P-values from the modeling of the weighting cohort.

<table>
<thead>
<tr>
<th>Item</th>
<th>Delphi’s group rank order¹</th>
<th>β coefficient²</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectal bleeding</td>
<td>2.3</td>
<td>9.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stool frequency</td>
<td>3.1</td>
<td>3.9</td>
<td>0.004</td>
</tr>
<tr>
<td>Nocturnal stools</td>
<td>4.4</td>
<td>7.7</td>
<td>0.002</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>4.5</td>
<td>6.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hemoglobin (excluded a-priori)</td>
<td>4.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Abdominal tenderness (excluded post-hoc)</td>
<td>5.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stool consistency</td>
<td>6.0</td>
<td>3.2</td>
<td>0.05</td>
</tr>
<tr>
<td>Fever (excluded post-hoc)</td>
<td>6.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Albumin</td>
<td>6.6</td>
<td>1.2</td>
<td>0.39</td>
</tr>
<tr>
<td>Rectal urgency (excluded a-priori)</td>
<td>6.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Activity level</td>
<td>7</td>
<td>5.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CRP (forced a-priori into the model)</td>
<td>-</td>
<td>0.5</td>
<td>0.68</td>
</tr>
</tbody>
</table>

¹ experts were asked to rank the items from 1 (most important) to 11 (least important), thus, lower mean scores represent higher ranked items
² for simplicity, the β coefficients represent the score of the item, entered as continuous variable; however, the actual PUCAI weighting was based on explicit dummy variables
Table 2.2: Patient’s characteristics of the weighting and validation cohorts. Medians (interquartile range) or mean (± SD) are presented as appropriate for the data distribution.

<table>
<thead>
<tr>
<th></th>
<th>Weighting cohort (n=157)</th>
<th>Validation cohort (n=48)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>82 (52%)</td>
<td>19 (40%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>12.7 ± 3.8</td>
<td>12.8 ± 3.3</td>
</tr>
<tr>
<td>Range (years)</td>
<td>2-18</td>
<td>3-18</td>
</tr>
<tr>
<td>Weight (Kg)</td>
<td>45.9 ± 17.8</td>
<td>44 ± 16.1</td>
</tr>
<tr>
<td>Z score</td>
<td>-0.02 ± 1.18</td>
<td>-0.30 ± 1.23</td>
</tr>
<tr>
<td>Height (Cm)</td>
<td>151 ± 21</td>
<td>151 ± 16</td>
</tr>
<tr>
<td>Z score</td>
<td>-0.09 ± 1.13</td>
<td>-0.18 ± 1.12</td>
</tr>
<tr>
<td>Disease duration (years)</td>
<td>2.2 (0.7-3.9)</td>
<td>0.2 (0-1.9)</td>
</tr>
<tr>
<td>First attack</td>
<td>32 (20%)</td>
<td>23 (48%)</td>
</tr>
<tr>
<td>Exacerbation</td>
<td>125 (80%)</td>
<td>25 (52%)</td>
</tr>
<tr>
<td>Disease extent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left sided</td>
<td>37 (23%)</td>
<td>13 (27%)</td>
</tr>
<tr>
<td>Extensive</td>
<td>120 (77%)</td>
<td>35 (73%)</td>
</tr>
<tr>
<td>Disease severity (PGA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quiescent</td>
<td>76 (48%)</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>Mild</td>
<td>31 (20%)</td>
<td>11 (23%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>38 (24%)</td>
<td>19 (40%)</td>
</tr>
<tr>
<td>Severe</td>
<td>12 (8%)</td>
<td>13 (27%)</td>
</tr>
<tr>
<td>Concurrent medications&lt;sup&gt;1&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral 5-ASA preparation</td>
<td>86 (55%)</td>
<td>16 (33%)</td>
</tr>
<tr>
<td>Corticosteroids</td>
<td>28 (18%)</td>
<td>11 (23%)</td>
</tr>
</tbody>
</table>

<sup>1</sup>Started prior to visit date
Table 2.3: Validation results of the PUCAI and PUCAI with laboratory tests, compared with Lichtiger and Seo indices. Numbers represent Pearson’s rho correlation coefficient.

<table>
<thead>
<tr>
<th></th>
<th>PGA</th>
<th>Colitis score</th>
<th>Mayo score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUCAI</td>
<td>0.91*</td>
<td>0.76*</td>
<td>0.95*</td>
</tr>
<tr>
<td>PUCAI with labs</td>
<td>0.90*</td>
<td>0.77*</td>
<td>0.92*</td>
</tr>
<tr>
<td>Lichtiger index</td>
<td>0.83*</td>
<td>0.71*</td>
<td>0.89*</td>
</tr>
<tr>
<td>Seo index</td>
<td>0.77*</td>
<td>0.69*</td>
<td>0.84*</td>
</tr>
</tbody>
</table>

*P<0.001; PGA=physician global assessment on a 100 mm visual analogue scale.

Table 2.4: Responsiveness analysis of the PUCAI, PUCAI with laboratory tests and the adult Seo index.

<table>
<thead>
<tr>
<th></th>
<th>PUCAI</th>
<th>PUCAI + labs</th>
<th>Seo index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean change (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved</td>
<td>38.3±17.8</td>
<td>42.2±21</td>
<td>51±40.4</td>
</tr>
<tr>
<td>Stable</td>
<td>6.3±7.7</td>
<td>8.4±6.4</td>
<td>17.7±17.5</td>
</tr>
<tr>
<td>Deemed to have changed</td>
<td>31.6±20.6</td>
<td>35.5±23</td>
<td>44.4±37.6</td>
</tr>
<tr>
<td>SES [change/SD baseline]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved</td>
<td>1.85</td>
<td>1.62</td>
<td>1.26</td>
</tr>
<tr>
<td>Stable</td>
<td>0.14</td>
<td>0.18</td>
<td>0.35</td>
</tr>
<tr>
<td>Deemed to have changed</td>
<td>1.65</td>
<td>1.5</td>
<td>1.24</td>
</tr>
<tr>
<td>SRM [change/SD changed]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved</td>
<td>2.15</td>
<td>2.1</td>
<td>1.3</td>
</tr>
<tr>
<td>Stable</td>
<td>0.45</td>
<td>1.29</td>
<td>1.11</td>
</tr>
<tr>
<td>Deemed to have changed</td>
<td>1.5</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Responsiveness statistics [MCID/SDunchanged]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation with ΔPGA</td>
<td>0.84</td>
<td>0.82</td>
<td>0.55</td>
</tr>
<tr>
<td>AUC of ROC between changed and unchanged</td>
<td>0.97</td>
<td>0.95</td>
<td>0.75</td>
</tr>
</tbody>
</table>

SES, standardized effect size; SD, standard deviation; SRM, standardized response mean; MCID, minimal clinically important difference; PGA, physician global assessment; AUC, area under the curve; ROC, receiver operator characteristics

Effect size statistics (SES, SRM and RS) should be interpreted as follows: 0.2-0.5 small effect, 0.5-0.8 moderate effect, >0.8 large effect.
Table 3.1: Characteristics of 99 children admitted for intravenous corticosteroid (IVCS) therapy during 1991-2000. Count (%), medians (interquartile range) or mean ± SD are presented as appropriate for the data distribution.

<table>
<thead>
<tr>
<th></th>
<th>IVCS Response (n=53)</th>
<th>IVCS failure (n=46)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Males</strong></td>
<td>26 (49%)</td>
<td>21 (46%)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td>11.5 ± 4.1</td>
<td>11.6 ± 4.5</td>
</tr>
<tr>
<td><strong>Range (years)</strong></td>
<td>2-17</td>
<td>2-17</td>
</tr>
<tr>
<td><strong>Disease duration (months)</strong></td>
<td>1.8 (0-13.6)</td>
<td>6.1 (0.2-19)</td>
</tr>
<tr>
<td>First attack</td>
<td>27 (51%)</td>
<td>20 (43%)</td>
</tr>
<tr>
<td>Exacerbation</td>
<td>26 (49%)</td>
<td>26 (57%)</td>
</tr>
<tr>
<td><strong>Disease extent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left sided ¹</td>
<td>6 (11%)</td>
<td>4 (9%)</td>
</tr>
<tr>
<td>Extensive ¹</td>
<td>47 (89%)</td>
<td>42 (91%)</td>
</tr>
<tr>
<td><strong>Steroid type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methylprednisolone</td>
<td>18 (34%)</td>
<td>13 (28%)</td>
</tr>
<tr>
<td>Hydrocortisone</td>
<td>35 (66%)</td>
<td>33 (72%)</td>
</tr>
<tr>
<td><strong>Steroid dose (mg/kg/day)</strong></td>
<td>0.94 (0.8-1.4)</td>
<td>1.05 (0.83-1.5)</td>
</tr>
<tr>
<td><strong>PUCAI at admission</strong></td>
<td>67 ± 13.8</td>
<td>74 ± 9.5</td>
</tr>
<tr>
<td># moderate ³</td>
<td>18 (34%)</td>
<td>7 (15%)</td>
</tr>
<tr>
<td># severe ³</td>
<td>35 (66%)</td>
<td>39 (85%)</td>
</tr>
<tr>
<td><strong>% weight loss</strong> ⁴</td>
<td>4.9 (1.1-7.3)</td>
<td>6.3 (3.2-8.9)</td>
</tr>
<tr>
<td><strong>Prior prednisone therapy</strong></td>
<td>23 (44%)</td>
<td>26 (57%)</td>
</tr>
<tr>
<td><strong>Days of bloody diarrhea</strong></td>
<td>24 (10-60)</td>
<td>25 (14-45)</td>
</tr>
<tr>
<td><strong>Steroid course during the previous year</strong></td>
<td>14 (26%)</td>
<td>8 (17%)</td>
</tr>
</tbody>
</table>

In exploratory univariate analysis, none of the above variables were significantly different between the two groups, except for PUCAI at admission (Student’s t-test; P=0.015) and weight loss (analysis of covariance (ANCOVA) of current weight adjusted for weight one month prior to admission; P=0.023).

¹ According to the Montreal classification [61]
² Standardized as methylprednisolone equivalent
³ As previously defined [57]
⁴ Over one months prior to admission
⁵ At the start of intravenous corticosteroid therapy

PUCAI, Pediatric Ulcerative Colitis Activity Index
Table 3.2: Distribution of individual items at the third day of intravenous corticosteroid (IVCS) therapy. Count (%), medians (interquartile range) or mean ± SD are presented as appropriate for the data distribution.

<table>
<thead>
<tr>
<th>Day 3 variable</th>
<th>IVCS Response (n=53)</th>
<th>IVCS failure (n=46)</th>
<th>Odds ratio (95% CI)¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nocturnal diarrhea (episodes/ per night)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>25 (47%)</td>
<td>2 (4%)</td>
<td></td>
</tr>
<tr>
<td>1-2</td>
<td>28 (53%)</td>
<td>30 (65%)</td>
<td></td>
</tr>
<tr>
<td>&gt;2</td>
<td>0 (0%)</td>
<td>14 (31%)</td>
<td></td>
</tr>
<tr>
<td><strong>Stools per 24 hours</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-2</td>
<td>22 (42%)</td>
<td>4 (9%)</td>
<td></td>
</tr>
<tr>
<td>3-5</td>
<td>25 (47%)</td>
<td>14 (30%)</td>
<td></td>
</tr>
<tr>
<td>6-8</td>
<td>5 (9%)</td>
<td>14 (30%)</td>
<td></td>
</tr>
<tr>
<td>&gt;8</td>
<td>1 (2%)</td>
<td>14 (30%)</td>
<td></td>
</tr>
<tr>
<td><strong>Blood in stool</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None or small amount infrequently</td>
<td>10 (19%)</td>
<td>2 (4%)</td>
<td></td>
</tr>
<tr>
<td>Small amount in majority of stools</td>
<td>24 (45%)</td>
<td>10 (22%)</td>
<td></td>
</tr>
<tr>
<td>Large amount in the majority of stools</td>
<td>19 (36%)</td>
<td>34 (74%)</td>
<td></td>
</tr>
<tr>
<td>**Temperature (&gt;37.8°C)**²</td>
<td>4 (8%)</td>
<td>9 (20%)</td>
<td></td>
</tr>
<tr>
<td><strong>Abdominal tenderness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PUCAI score</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 ± 17</td>
<td>70 ± 14</td>
<td>2.2 (1.5-3.1)²</td>
<td></td>
</tr>
<tr>
<td><strong>Seo score</strong></td>
<td>194 ± 34</td>
<td>226 ± 30</td>
<td>1.4 (1.2-1.6)²</td>
</tr>
<tr>
<td><strong>Lindgren score</strong></td>
<td>4.2 ± 2.3</td>
<td>9.4 ± 4.3</td>
<td>1.6 (1.3-1.9)</td>
</tr>
<tr>
<td><strong>Travis score</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>0 (0%)</td>
<td>17 (38%)</td>
<td>31 (3.9-666)</td>
</tr>
<tr>
<td>Negative</td>
<td>53 (100%)</td>
<td>29 (62%)</td>
<td></td>
</tr>
<tr>
<td><strong>Blood tests</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albumin</td>
<td>33 ± 5.7</td>
<td>30 ± 4.4</td>
<td>0.53 (0.4-0.8)⁵</td>
</tr>
<tr>
<td>CRP (mg/dL)</td>
<td>0.71 ± 0.53</td>
<td>1.87 ± 1.57</td>
<td>6.2 (2.6-14.9)</td>
</tr>
<tr>
<td>ESR</td>
<td>38 ± 22</td>
<td>50 ± 25</td>
<td>1.3 (1.03-1.5)²</td>
</tr>
<tr>
<td>Hemoglobin g/L</td>
<td>101 ± 22</td>
<td>97 ± 17</td>
<td>1 (0.97-1.01)</td>
</tr>
<tr>
<td>Platelets</td>
<td>452 ± 142</td>
<td>496 ± 158</td>
<td></td>
</tr>
</tbody>
</table>

¹ Obtained by a univariable logistic regression, without dummy variables
² Calculated for a 10-unit change
³ Temperature was measured orally
⁴ Judged as at least mild tenderness by the physician
⁵ Calculated for a 5-unit change

PUCAI, Pediatric Ulcerative Colitis Activity Index; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein
Table 3.3: Distribution of individual items on the **fifth** day of intravenous corticosteroid (IVCS) therapy. Seven patients who were discharged between day 3 and 5 were excluded from this analysis. Count (%), medians (interquartile range) or mean ± SD are presented as appropriate for the data distribution.

<table>
<thead>
<tr>
<th>DAY 5 variable</th>
<th>IVCS Response (n=46)</th>
<th>IVCS failure (n=46)</th>
<th>Odds ratio (95% CI)¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nocturnal diarrhea</strong> (episodes/per night)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>23 (50%)</td>
<td>4 (9%)</td>
<td>8.1 (3-22)</td>
</tr>
<tr>
<td>1-2</td>
<td>22 (48%)</td>
<td>30 (65%)</td>
<td></td>
</tr>
<tr>
<td>&gt;2</td>
<td>1 (2%)</td>
<td>12 (26%)</td>
<td></td>
</tr>
<tr>
<td><strong>Stools per 24 hours</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-2</td>
<td>22 (48%)</td>
<td>5 (11%)</td>
<td>4.7 (2.4-9.1)</td>
</tr>
<tr>
<td>3-5</td>
<td>20 (43%)</td>
<td>16 (35%)</td>
<td></td>
</tr>
<tr>
<td>6-8</td>
<td>4 (9%)</td>
<td>15 (33%)</td>
<td></td>
</tr>
<tr>
<td>&gt;8</td>
<td>0 (0%)</td>
<td>10 (22%)</td>
<td></td>
</tr>
<tr>
<td><strong>Blood in stool</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None or small amount infrequently</td>
<td>13 (28%)</td>
<td>1 (2%)</td>
<td>4.2 (2-8.6)</td>
</tr>
<tr>
<td>Small amount in majority of stools</td>
<td>16 (35%)</td>
<td>10 (22%)</td>
<td></td>
</tr>
<tr>
<td>Large amount in the majority of stools</td>
<td>17 (37%)</td>
<td>35 (76%)</td>
<td></td>
</tr>
<tr>
<td><strong>Temperature (&gt;37.8°C)³</strong></td>
<td>3 (7%)</td>
<td>6 (13%)</td>
<td>2.2 (0.5-9.2)</td>
</tr>
<tr>
<td><strong>Abdominal tenderness</strong></td>
<td>4 (9%)</td>
<td>12 (26%)</td>
<td>3.7 (1.1-12.5)</td>
</tr>
<tr>
<td><strong>PUCAI</strong></td>
<td>47 ± 18.8</td>
<td>68 ± 14</td>
<td>2.4 (1.6-3.5)²</td>
</tr>
<tr>
<td><strong>Seo</strong></td>
<td>188 ± 44</td>
<td>227 ± 27</td>
<td>1.4 (1.2-1.6)²</td>
</tr>
<tr>
<td><strong>Lindgren</strong></td>
<td>3.9 ± 2.5</td>
<td>8.7 ± 3.5</td>
<td>1.7 (1.4-2.1)</td>
</tr>
<tr>
<td><strong>Blood tests</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albumin</td>
<td>32 ± 5.5</td>
<td>29 ± 5.3</td>
<td>0.52 (0.36-0.8)⁵</td>
</tr>
<tr>
<td>CRP (mg/dL)</td>
<td>0.67 ± 0.54</td>
<td>1.66 ± 0.99</td>
<td>7.9 (3.3-18.4)</td>
</tr>
<tr>
<td>ESR</td>
<td>36 ± 22</td>
<td>49 ± 26</td>
<td>1.3 (1.04-1.54)²</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>101 ± 22</td>
<td>96 ± 18</td>
<td>1 (0.97-1.01)</td>
</tr>
<tr>
<td>Platelets</td>
<td>424 ± 130</td>
<td>505 ± 153</td>
<td>1 (0.99-1.01)³</td>
</tr>
</tbody>
</table>

¹ Obtained by a univariable logistic regression, without dummy variables
² Calculated for a 10-unit change
³ Temperature was obtained orally
⁴ Judged as at least mild tenderness by the physician
⁵ Calculated for a 5-unit change

PUCAI, Pediatric Ulcerative Colitis Activity Index; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein
### Table 3.4: Diagnostic utility of indices on days three and five of therapy in predicting short-term intravenous corticosteroid failure. Numbers in brackets represent 95% confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>Cutoff</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>+LR</th>
<th>-LR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAY 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUCAI</td>
<td>&gt;45</td>
<td>93 (84-98)</td>
<td>41 (33-46)</td>
<td>58 (52-61)</td>
<td>88 (69-97)</td>
<td>1.6</td>
<td>0.16</td>
</tr>
<tr>
<td>Lindgren</td>
<td>&gt;4</td>
<td>91 (81-97)</td>
<td>57 (48-62)</td>
<td>65 (58-69)</td>
<td>88 (74-96)</td>
<td>2.1</td>
<td>0.16</td>
</tr>
<tr>
<td>Seo</td>
<td>&gt;195</td>
<td>91 (81-97)</td>
<td>43 (34-48)</td>
<td>59 (52-62)</td>
<td>85 (67-95)</td>
<td>1.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Lindgren</td>
<td>&gt;8</td>
<td>64 (54-70)</td>
<td>92 (83-97)</td>
<td>88 (74-96)</td>
<td>75 (67-79)</td>
<td>8.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Travis</td>
<td>-</td>
<td>38 (30-40)</td>
<td>100 (93-100)</td>
<td>88 (74-96)</td>
<td>75 (67-79)</td>
<td>8.2</td>
<td>0.4</td>
</tr>
<tr>
<td><strong>DAY 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUCAI</td>
<td>&gt;70</td>
<td>44 (35-49)</td>
<td>93 (84-98)</td>
<td>87 (68-97)</td>
<td>63 (56-66)</td>
<td>5.1</td>
<td>0.6</td>
</tr>
<tr>
<td>Lindgren</td>
<td>&gt;9</td>
<td>36 (27-38)</td>
<td>98 (89-100)</td>
<td>94 (72-100)</td>
<td>60 (55-62)</td>
<td>16</td>
<td>0.7</td>
</tr>
<tr>
<td>Seo</td>
<td>&gt;240</td>
<td>27 (18-32)</td>
<td>93 (85-98)</td>
<td>80 (54-95)</td>
<td>56 (51-59)</td>
<td>4</td>
<td>0.8</td>
</tr>
<tr>
<td>Travis</td>
<td>-</td>
<td>22 (14-24)</td>
<td>100 (91-100)</td>
<td>99 (67-100)</td>
<td>56 (52-56)</td>
<td>10.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

PUCAI, Pediatric Ulcerative Colitis Activity Index; PPV, positive predictive value; NPV, negative predictive value; LR, likelihood ratio
Table 4.1: Common methods used for calculating the minimal important difference (MID)

<table>
<thead>
<tr>
<th>Method (ref)</th>
<th>Description</th>
<th>Cutoffs used</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anchor based approaches</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC curve approach [21]</td>
<td>The MID is the Δscore that corresponds to the point where the second diagonal crosses the ROC curve (i.e. at the most left-top corner)</td>
<td>Different cutoffs may be used on the anchor for dichotomization of the cohort to patients who improved and those who did not</td>
</tr>
<tr>
<td>Mean change [130]</td>
<td>The MID is the mean Δscore of patients who improved</td>
<td>Mean Δscore of patients who were judged to have small, moderate or large change</td>
</tr>
<tr>
<td><strong>Distribution based approaches</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect size [136, 242]</td>
<td>MID = \frac{Δ\text{score}}{\text{SD}_{\text{baseline}}}</td>
<td>MID = Δ\text{score} while setting the equation at 0.2 (small effect [147, 231, 243]), 0.5 (moderate [132, 244]) or 0.8 (large effect)</td>
</tr>
<tr>
<td>Wyrwich SEM [137]</td>
<td>MID = \frac{Δ\text{score}}{\text{SD}_{\text{baseline}} \sqrt{1-r}}</td>
<td>MID = Δ\text{score} while setting the equation at 1 (small effect [165, 245, 246]), 1.96 (moderate [159]) or 2.77 (large effect [159, 247])</td>
</tr>
<tr>
<td>Jacobson’s Reliable Change Index [138]</td>
<td>MID = \frac{Δ\text{score}}{\text{SD}_{\text{baseline}} \sqrt{2\times(1-r)}}</td>
<td>MID = Δ\text{score} while setting the equation at 1.96 (moderate [138])</td>
</tr>
<tr>
<td>Bland and Altman 95% limits of agreements [139]</td>
<td>MID is the mean Δ\text{score}±1.96 SD_{\text{Δ\text{score}}}</td>
<td>-</td>
</tr>
<tr>
<td>0.5 SD approach [94]</td>
<td>The MID is 0.5 SD of the Δ\text{score}</td>
<td>-</td>
</tr>
</tbody>
</table>

SD, standard deviation; SEM, standard error of measurement; MID, minimal important difference; ROC, receiver operating characteristic
Table 4.2a: Results of the minimal important difference (MID) for the PUCAI, PCDAI and RQLQ, obtained by ROC analysis of different cutoffs of the global rating of change.

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>PUCAI&lt;sup&gt;1&lt;/sup&gt;</th>
<th>PCDAI&lt;sup&gt;1&lt;/sup&gt;</th>
<th>RQLQ summary score&lt;sup&gt;4&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MID Sens Spec AUC (95%CI)</td>
<td>MID Sens Spec AUC (95%CI)</td>
<td>MID Sens Spec AUC (95%CI)</td>
</tr>
<tr>
<td>Small change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 &lt; -1+2+3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 &lt; -1</td>
<td>≥10 89/87 0.96 (0.91-1)</td>
<td>≥9 75/67 0.83 (0.64-1)</td>
<td>≥0.53 75/76 0.81 (0.69-0.93)</td>
</tr>
<tr>
<td>0 &lt; -1</td>
<td>≥10 71/64 0.82 (0.65-0.98)</td>
<td>≥9 17/67 0.44 (0.0-0.9)</td>
<td>≥0.62 88/78 0.84 (0.68-0.99)</td>
</tr>
<tr>
<td>Moderate change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0+1 &lt; -2+3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 &lt; -2</td>
<td>≥20 85/88 0.97 (0.93-1)</td>
<td>≥12 93/89 0.97 (0.91-1)</td>
<td>≥0.68 76/69 0.74 (0.59-0.90)</td>
</tr>
<tr>
<td>1 &lt; -2</td>
<td>≥20 82/71 0.88 (0.71-1)</td>
<td>≥12 86/89 0.94 (0.83-1)</td>
<td>≥1.09 78/50 0.50 (0.16-0.82)</td>
</tr>
<tr>
<td>Large change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0+1+2 &lt; -3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 &lt; -3</td>
<td>≥35 76/92 0.92 (0.85-0.99)</td>
<td>≥20 99/94 0.99 (0.96-1)</td>
<td>≥0.77 77/70 0.76 (0.58-0.94)</td>
</tr>
<tr>
<td>2 &lt; -3</td>
<td>≥35 76/75 0.77 (0.63-0.94)</td>
<td>≥20 99/86 0.98 (0.92-1)</td>
<td>≥0.75 50/70 0.64 (0.37-0.92)</td>
</tr>
</tbody>
</table>

<sup>1</sup>0, no change; 1, small change; 2, moderate change; and 3, large change.

<sup>2</sup>59 patients were stable or improved: 19 were stable, 7 mildly improved, 12 moderately improved and 21 significantly improved; 16 who deteriorated were excluded.

<sup>3</sup>23 patients were stable or improved: 3 were stable, 6 mildly improved, 7 moderately improved and 7 significantly improved; 2 who deteriorated were excluded.

<sup>4</sup>57 patients were stable or improved: 32 patients remained unchanged, 9 with mild improvement, 6 with moderate improvement, and 10 with significant improvement; 3 patients who deteriorated were excluded; 3 who deteriorated were excluded.

ROC, Receiver Operator Characteristic; Sens, sensitivity; Spec, Specificity; AUC, area under the receiver characteristic curve; CI, confidence interval
Table 4.2b: Results of the different minimal important difference (MID) for the CRQ.

<table>
<thead>
<tr>
<th>Cutoffs</th>
<th>Dyspnea</th>
<th></th>
<th></th>
<th>Fatigue</th>
<th></th>
<th></th>
<th>Emotional function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MID Sens AUC/Spec (95%CI)</td>
<td></td>
<td></td>
<td>MID Sens/Spec AUC (95%CI)</td>
<td></td>
<td></td>
<td>MID Sens/Spec AUC (95%CI)</td>
</tr>
<tr>
<td>Small change</td>
<td>0 &lt; -&gt; 1+2+3</td>
<td>≥0.37 78/65 0.78 (0.71-0.86)</td>
<td></td>
<td>≥0.38 70/62 0.71 (0.63-0.79)</td>
<td></td>
<td>≥0.18 63/52 0.64 (0.56-0.72)</td>
<td></td>
</tr>
<tr>
<td>0 &lt; -&gt; 1</td>
<td>≥0.37 77/65 0.67 (0.54-0.81)</td>
<td></td>
<td>≥0.38 71/62 0.70 (0.57-0.83)</td>
<td></td>
<td>≥0.21 35/52 0.50 (0.34-0.65)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate change</td>
<td>0+1 &lt; -&gt; 2+3</td>
<td>≥0.63 70/74 0.79 (0.72-0.86)</td>
<td></td>
<td>≥0.63 66/65 0.68 (0.59-0.76)</td>
<td></td>
<td>≥0.36 61/60 0.68 (0.59-0.76)</td>
<td></td>
</tr>
<tr>
<td>1 &lt; -&gt; 2</td>
<td>≥0.73 67/71 0.68 (0.55-0.81)</td>
<td></td>
<td>≥0.63 63/52 0.51 (0.35-0.65)</td>
<td></td>
<td>≥0.21 68/61 0.67 (0.51-0.82)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large change</td>
<td>0+1+2 &lt; -&gt; 3</td>
<td>≥1.10 71/72 0.76 (0.65-0.87)</td>
<td></td>
<td>≥1.0 68/66 0.70 (0.57-0.82)</td>
<td></td>
<td>≥0.50 54/64 0.65 (0.53-0.78)</td>
<td></td>
</tr>
<tr>
<td>2 &lt; -&gt; 3</td>
<td>≥1.40 64/52 0.65 (0.51-0.79)</td>
<td></td>
<td>≥1.1 58/60 0.60 (0.45-0.76)</td>
<td></td>
<td>≥0.50 54/51 0.54 (0.40-0.69)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 65 patients remained unchanged, 18 with mild improvement, 52 with moderate improvement, and 28 with significant improvement; 20 patients who deteriorated were excluded.

2 74 patients remained unchanged, 21 with mild improvement, 43 with moderate improvement, and 20 with significant improvement; 25 patients who deteriorated were excluded.

3 85 patients remained unchanged, 17 with mild improvement, 41 with moderate improvement, and 26 with significant improvement; 14 patients who deteriorated were excluded.

ROC, Receiver Operator Characteristic; Sens, sensitivity; Spec, Specificity; AUC, area under the receiver characteristic curve; CI, confidence interval
Table 5.1: Studies comparing methods to determine the minimal important difference (MID).

<table>
<thead>
<tr>
<th>Study (ref)</th>
<th>Instrument</th>
<th>Method compared</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruynesteyn 2001 [149]</td>
<td>Radiological progression of joint damage</td>
<td>ROC vs. 95% limits of agreements</td>
<td>ROC method using dichotomized outcome of clinically relevant change or not (scored by rheumatologists) &gt; 95% limits of agreements</td>
</tr>
<tr>
<td>Childs 2005 [151]</td>
<td>numerical pain rating scale</td>
<td>ROC vs. SEM</td>
<td>ROC method of at least 3 on a -7 to +7 point GR scale of change= 1.96 SEM</td>
</tr>
<tr>
<td>Norman 2003 [94]</td>
<td>Several existing scales</td>
<td>0.5 standard deviation vs. mean change</td>
<td>0.5 standard deviation= mean change of the group with small change (rated by patients on GR scale)</td>
</tr>
<tr>
<td>Norman 2001 [132]</td>
<td>Simulation studies</td>
<td>Points on the response scale vs. effect size</td>
<td>0.5 on a 7-point scale=effect size of 0.5</td>
</tr>
<tr>
<td>Guyatt 2002; Wyrwich 1999 [137, 166].</td>
<td>Chronic Respiratory Questionnaire (CRQ)</td>
<td>SEM vs. mean change</td>
<td>1 SEM (with r=0.75)= mean change of the group with small change (rated by patients on GR scale)= effect size of 0.5</td>
</tr>
<tr>
<td>Walters 2003 [248]</td>
<td>SF-6D</td>
<td>0.5 standard deviation vs. standardized response mean</td>
<td>0.5 standard deviation=standardized response mean of 0.3</td>
</tr>
<tr>
<td>Jaeschke 1989 [130]; Juniper 1994 [143]</td>
<td>Chronic Respiratory Questionnaire (CRQ), the Chronic Heart failure Questionnaire (CHQ); The Asthma Quality of Life Questionnaire</td>
<td>Points per items vs. mean change group</td>
<td>0.5 per item in each domain= mean change of the group with small change (rated by patients on GR scale)</td>
</tr>
<tr>
<td>Cella 2002 [161]</td>
<td>Functional Assessment of Cancer Therapy-Lung</td>
<td>Effect size vs. SEM</td>
<td>Effect size of 0.3-0.5= 1 SEM= Symptom based clinical change</td>
</tr>
</tbody>
</table>

ROC, Receiver Operator Characteristic; GR, global rating; SEM, standard error of measurement
## Appendices

### Appendix 2.1: Pediatric Ulcerative Colitis Activity Index (PUCAI)

<table>
<thead>
<tr>
<th>ITEM</th>
<th>POINTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Abdominal pain:</strong></td>
<td></td>
</tr>
<tr>
<td>No pain</td>
<td>0</td>
</tr>
<tr>
<td>Pain can be ignored</td>
<td>5</td>
</tr>
<tr>
<td>Pain cannot be ignored</td>
<td>10</td>
</tr>
</tbody>
</table>

| **2. Rectal bleeding**                    |        |
| None                                      | 0      |
| Small amount only, in less than 50% of stools | 10    |
| Small amount with most stools             | 20     |
| Large amount (>50% of the stool content)  | 30     |

| **3. Stool consistency of most stools**   |        |
| Formed                                    | 0      |
| Partially formed                          | 5      |
| Completely unformed                       | 10     |

| **4. Number of stools per 24 hours**     |        |
| 0-2                                       | 0      |
| 3-5                                       | 5      |
| 6-8                                       | 10     |
| >8                                        | 15     |

<table>
<thead>
<tr>
<th><strong>5. Nocturnal stools (any episode causing wakening)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>10</td>
</tr>
</tbody>
</table>

| **6. Activity level**                              |        |
| No limitation of activity                          | 0      |
| Occasional limitation of activity                  | 5      |
| Severe restricted activity                         | 10     |

| **SUM OF PUCAI (0-85)**                            |        |
PUCAI user guide

Most items contained in the PUCAI can be scored using the instructions provided within the instrument. The following issues require additional clarification:

Time period for evaluation

- Answers should reflect a daily average of the last two days.
- However, if clinical conditions are changing rapidly (e.g. during intense intravenous therapy), the most recent 24 hours should be considered.
- For patients undergoing colonoscopy, answers should reflect the two days before bowel cleanout was started.

Rectal bleeding

- “Large amount” should be selected if large amount of blood is present in most stools.

Number of stools per 24 hours

- Clustered several small stools over a very short period of time that could be related to tenesmus or incomplete evacuation, should be considered as ONE stool.

Activity level

- Occasional limitation of activity= could attend school or equivalent, but reduced activity (e.g. attends school but does not play at breaks).
- Severe restricted activity=could not attend school or equivalent activity.
Appendix 2.2: Selected comments obtained from the peer review process of submitting the PUCAI manuscript to Gastroenterology, with a point by point response

Article Review
3/20/07
Primary Author: Turner
Title: Development, validation, and evaluation of a PUCAI

Significance:

This is an important, methodologically strong study that addresses an important gap in clinical trials in pediatric UC. There are no validated noninvasive indices for children with UC. Current standards require regular invasive measurement with lower endoscopy, and IRBs and potential subjects are very reluctant to approve or participate in trials that require repeated lower endoscopy. While this is a strong study, a critical audience will be the FDA, which can effectively kill this instrument by refusing to base approvals on trials that rely on it as a primary endpoint. Therefore, convincing the FDA that this is the best possible study is important for its long-term impact. Some improvements in the methodology can be made.

1. The authors assume that the instrument should be physician-centered, rather than patient-centered. This is appropriate for very young children, but may be less so for adolescents. It is not demonstrated that physician assessment correlates with patient assessment. It is also unclear whether the symptoms that are important to physicians are the ones that are important to patients. There are many diseases in which there are significant differences between patient and physician assessment. An abstract on adults with UC accepted for DDW2007 suggests that this is the case in adult UC. This instrument may not measure some symptoms that are quite important to patients. This choice of a physician-centered rather than patient-centered approach (not an unreasonable choice for pediatrics) should be addressed and justified, either in the methods or the discussion.

Reply: When our group developed the disease-specific HRQOL instrument (the IMPACT questionnaire) we used a patient-oriented approach, since the concept of QOL is subjective and best judged by the patient. It is our opinion, however, that the treating pediatric gastroenterologist is better able to objectively assess the degree of “disease activity” (e.g. in comparison to other patients). As suggested by the reviewer, in the revised manuscript we have further justified the use of a physician-oriented rather than a patient-oriented approach (page 6, first paragraph and in the Discussion, first paragraph).

2. There are many, many strengths here - recursive item generation, actual instructions for use of the instrument to improve reliability, committing to the PGA before scoring the PUCAI,
blinding to the previous PUCAI score, defining a robust MCID, etc.

**Reply:** We thank the reviewer for these very favorable comments concerning the methodology we employed in the instrument development.

3. **Can you provide a better justification for throwing out urgency?** This is a very important symptom to patients, and it is difficult to justify throwing it out without *testing* it.

**Reply:** The development of the PUCAI was designed to combine both clinimetric and psychometric techniques. As such, item generation and reduction were performed *judgmentally* in an iterative, multi-step process involving many pediatric IBD experts. Although ‘urgency’ was scored within the first 10 items, it was low on the list, and the expert panel decided to exclude it based on their experience with children. Specifically, the PUCAI was developed for a wide range of pediatric age groups and urgency cannot be accurately assessed by young children or their parents. It can be seen that all retained items of the PUCAI and their gradations are easily scored by a parent of a four year old, based on observation only. In accordance to our protocol, we did not collect data on excluded items and, thus, no statistical analysis is provided for ‘urgency’.

4. **Can you provide a better justification for throwing out hemoglobin/hematocrit?** It was significant in the Seo index derivation. It is difficult to justify throwing it out without *testing* it. You have the data (hemoglobin for the Seo index) - why not test it? Or was the goal to minimize use of expensive laboratory tests that require a needlestick? If that is the case, say so.

**Reply:** In contrast to ‘urgency’ we indeed recorded hemoglobin data on all children. Hemoglobin showed low univariate correlation with PGA (r=-0.18, P=0.18). We believe that these results further justify the exclusion of hemoglobin from the PUCAI. This analysis, which was not included in the article for space limitation, is now added as suggested by the reviewer (Results page 11, last paragraph). In the validation phase, hemoglobin also showed low correlations with colonscopic score (r=-0.23, P=0.11) and with the Mayo score (r=-0.21, P=0.15).

5. **ROC curves and differentiation of levels of disease - it seems like you are presenting only the best data (0.98!) - almost any instrument should be able to differentiate mild from severe with a good ROC (even the original T&W can do that), but what about the ROC curves to differentiate mild from moderate? Moderate from severe? If you are going to present ROC values, present all of them. Give us an idea of how well the instrument does with more difficult gradations, even if it is somewhat bad news (these are very difficult gradations to separate).
Reply: The reviewer rightly questions the exact comparisons used to calculate the ROC. We can reassure the reviewer, that we did not show only the best data (i.e. comparing only the mild to the severe) but we rather categorized ALL groups into two for the analysis (i.e. none vs mild+moderate+severe, none+mild versus mod+severe, severe vs none+mod+severe). We used this approach, believing that if the question is, for example, how does the PUCAI differentiate active from inactive patients (as commonly used in clinical trials) then all active patients (just like in clinical trials) compose one group, and inactive patients the second group. The same applies also for the other comparisons (severe disease from the rest, etc). However, following the reviewer’s suggestion, we have changed the analysis to compare only the adjacent categories, and we have added a clear description of what was done to the legend of figure 2. The AUC of the ROC and the corresponding sensitivity and specificity have been changed accordingly. Despite this conservative analysis, the AUC of the ROC are still very high, as expected from eyeballing the box plot (Figure 2a): the PUCAI differentiates very well between the individual categories, especially between remission and mild disease. We feel that this strengthens our results since remission is an important outcome in clinical trials and the clear differentiation can increase the power of the study. That has also been added to the manuscript in the Discussion (page 17, last paragraph).

6. Consider the possibility (largely unavoidable) that assessing the PUCAI immediately after the PGA likely improved the correlation of the items with the PGA. Please discuss this as a limitation. When were the Lichtiger and Seo indices done in relation to these? These may help as comparators.

Reply: In response to the reviewer’s concern that the PGA and the PUCAI were highly correlated since they were scored by the same physician, we performed a second analysis, comparing the PGA scored by one of the physicians with the PUCAI scored by the other physician. The order of the physician was randomly determined (using SPSS software). In this analysis, the PUCAI remained very highly correlated with the PGA of the other physician (r=0.89, P<0.001). This analysis strengthens the validity of our results and was added to the ‘Results’ section (page 12, last paragraph). The Lichtiger and Seo indices were not scored as a block, but the questions were intermixed in order to avoid the identification the indices and, thus, to minimize bias.

7. The endoscopy issue - this will be a sticking point for the FDA. You need to strengthen this argument. You should strongly consider testing whether adding the colonoscopy score to your model makes any significant improvement in the model. If it does not, that is a stronger argument for excluding endoscopy. You may also want to cite Am J Gastroenterol. 2005 Feb;100(2):355-61, which also found endoscopy added little to the standard indices in UC.
Reply: We thank the reviewer for the comments on the importance of colonoscopy to the FDA decision. We have added the suggested analysis of the PUCAI (‘Methods’ in the statistical analysis first paragraph; ‘Results’ page 13, last paragraph continuing to page 14). We have cited the suggested article and have expanded the discussion, justifying further the exclusion of endoscopic item (Discussion, page 16 last paragraph continuing to page 17). We trust that, as a result of these additions, our conclusion that endoscopy is not needed in pediatric clinical trials is now more convincing.
Appendix 3.1: Selected comments obtained from the peer review process of submitting the severe colitis manuscript to GUT, with a point by point response

REVIEWER 1

General:

This is an informative evaluation of acute severe colitis in children, an area which in geographical terms is a virtual blank on the map. It contains useful epidemiologic data, new data that defines the radiographic limits of colonic diameter above and below the age of 11 years, the first comparative analysis of predictive indices that will be relevant to adult gastroenterologists as well as pediatric gastroenterologists, and describes the outcome of treatment. The attention to detail is impressive, including the parameters for epidemiologic analysis, interobserver analysis for radiologic features, ROC curves for the predictive indices, and assessment of whether publication of previous indices might have introduced inadvertent bias into the timing of colectomy. For the reasons described by the authors about the extent and presentation of ulcerative colitis in children, it comes as little surprise that the outcome of acute severe colitis is somewhat worse than in adults. Most importantly, the paper identifies clinically relevant criteria that can be used in a timely manner at the bedside to help manage such children: this is likely to change practice by encouraging early decision making on second line therapy. The need for this is manifest, since only 6/99 received calcineurin inhibitors.

Specific comments are few:

1. Nocturnal stool frequency: how was 'nocturnal' defined?

   Reply: For this retrospective review we were obliged to define “nocturnal” as stools occurring during the sleep hours. The individualization of ‘sleep hours’ was possible since in our institution the sleep/awake status is hourly recorded on the patient’s chart, by the bedside nurse.

2. PUCAI: this is prospectively validated and substantive, but will be new to most readers. A table of the criteria and calculation would be helpful, since the Gastroenterology paper is only just available.

   Reply: We agree that inclusion of the PUCAI would be helpful. We have provided it as an addendum.

REVIEWER 2

General:

Very little is known about the natural history of childhood-onset ulcerative colitis (UC). To date only three small retrospective studies (44 patients in total) have reported the short-term
corticosteroid response rate of severe colitis occurring in children. The authors have attempted to study the short term and long term responses to corticosteroid in 99 children over a 10 year period. This is a retrospective analysis, and the data collection and analysis is subjected to the inherent problems associated with such data collection (missing data, poor documentation etc).

Minor points:

1. How was the disease extent determined? Did all of them have complete colonoscopies?

   Reply: We applied the Montreal classification of UC extent. Disease was considered extensive either if colonoscopy revealed macroscopic disease extending proximal to the splenic flexure OR if colectomy specimen showed gross disease in that distribution. Colonoscopy was not complete in all patients; the Montreal classification does not require a distinction between pancolitis and extensive colitis. All patients classified as “left-sided colitis” were colonoscoped until normal mucosa was reached distal to the splenic flexure.

2. It is well known that C. difficile infection (it appears these were excluded assuming intercurrent infections) can be seen in many ulcerative colitis patients, and the mere presence of C- difficile may or may not be significant to the acuteness of the disease process. How many were excluded because of C. difficile infection?

   Reply: We agree with the reviewer’s comment regarding C. difficile co-infection. We excluded infected patients, including C. difficile, because it might influence disease response, and in clinical practice we expect physicians to employ appropriate antibiotics in addition to standard treatment of a UC exacerbation. Moreover, the knowledge that the patient is infected with C. difficile may influence the physician’s decision on timing and type of second line therapy, justified or not. Five patients were excluded because of identification of C. difficile.
REFERENCES


[134] van der Heijde D, Lassere M, Edmonds J, Kirwan J, Strand V, Boers M. Minimal clinically important difference in plain films in RA: group discussions, conclusions,


Brozek JL, Guyatt GH, Schunemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes. 2006;4:69.

Schunemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005 Apr;40(2):593-7.


Jensen CB, Grossman SJ, Jollow DJ. Improved method for determination of cellular thiols, disulfides and protein mixed disulfides using HPLC with


