Validating a Global Measure of Severity in Children with Chronic Conditions

by

Braden K. Tompke

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Science

in

Public Health and Health Systems

Waterloo, Ontario, Canada, 2019

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is the work of Braden Tompke with the collaboration of his supervisor, Dr. Mark Ferro. Appropriate actions were taken to ensure confidentiality of the participants in the REACH Study dataset.

Abstract

Background: Approximately 20% of children live with a chronic physical condition, such as asthma, epilepsy, or diabetes. These conditions place considerable burden on children, their families, clinicians, and the health system. However, these burdens are reduced when conditions are effectively managed, typically accomplished by appropriately monitoring the severity and progression of the condition. Several condition-specific scales exist for measuring severity in children but are limited in their clinical utility for general practitioners or pediatricians who care for children with different conditions.

Objectives: This study aimed to validate the Global Assessment of Severity of Illness (GASI)—a single-item scale that can be used to measure severity in children with different chronic physical conditions. Study objectives were to examine the construct validity, test-retest reliability, responsiveness, and sensitivity/specificity of the GASI.

Methods: Clinicians assessed severity of asthma, food allergy, epilepsy, diabetes, and juvenile arthritis in 56 children using the GASI, Duke Severity of Illness Scale (DUSOI; the external clinical anchor), and a general visual analogue scale (VAS). Parents reported on child health-related quality of life using the KIDSCREEN-27. Kendall's *Tau-c* and area under the receiver operating characteristic curve (AUC) determined the strength of association between measures. Fisher's Exact test indicated whether the GASI could discriminate between children with and without multimorbidity. McNemar's test, the Kappa coefficient, and weighted Kappa assessed stability in GASI ratings over time. The standardized response mean and Guyatt's responsiveness index examined internal and external responsiveness, respectively. AUC

determined sensitivity/specificity. Clinician characteristics, as potential confounders, were investigated within AUC regression models.

Results: The GASI demonstrated strong correlations with the DUSOI composite score (τ_c= 0.57-0.63; AUC= 0.83-0.96) and VAS (τ_c= 0.78) and weak correlation with health-related quality of life (τ_c < 0.1). Lack of discrimination between children with and without multimorbidity was indicated by Fisher's Exact test (p-value > 0.05). Moderate to substantial test-retest reliability was supported by McNemar's test (p-value > 0.05), Kappa (κ= 0.79; CI= 0.51-1.00), and weighted Kappa (κ_w= 0.57; CI= 0.36-0.78). The GASI was largely responsive (Cohen's d= 0.84; CI= 0.68-1.11) and the magnitude of sensitivity/specificity was low to moderate (AUC= 0.62-0.81). Construct validity was excellent regardless of whether regression modeling accounted for type of diagnosis, clinician, or child age.

Conclusion: Initial evidence supports validity of the GASI to make meaningful comparisons of severity between different chronic conditions in children. Future research using larger samples should aim to replicate these findings and test inter-rater reliability between different health professionals. Such work is needed to fill knowledge gaps in comparative pediatric research and, potentially, simplify clinical practice.

Acknowledgments

I would first like to thank my supervisor, Dr. Mark Ferro, for his tremendous generosity and encouragement throughout my graduate studies. I am especially thankful for his efforts to challenge my abilities and channel my passion to improve health care. I would also like to thank Dr. Ashok Chaurasia and Dr. Chris Perlman for lending their expertise on my thesis committee. Their contributions were essential to my growth as a researcher.

I would like to thank the children, families, and clinicians who contributed to the REACH Study, and also Jessica Zelman who coordinated the study. Without them this thesis would not have been possible.

I am grateful for my colleagues in the School of Public Health and Health Systems, especially my fellow ARCH Lab members who supported me through this challenging program. Their friendship has brought me much joy in this stage of my education.

I owe many thanks to my family for their continual prayers and support throughout this time. I also would like to thank my friend, Ben Ellis, whose music gave me peace and focus while completing this thesis. I would like to thank my wife, Shayna, for her many sacrifices to support my graduate studies. Her love, encouragement, and joy were crucial to my completion of this program. Above all, I am grateful for Christ who has been with me in every step of my education.

Table of Contents

List of Figures

List of Tables

Introduction and Overview

Approximately 20% of children live with a chronic physical condition, such as asthma, epilepsy, or diabetes.[1] These conditions place considerable burden on children, their families, and clinicians. At a systems level, chronic conditions account for 42% of health care costs among children.[2] Moreover, children with chronic conditions are at increased risk for mental disorder and their families experience more stress and financial hardship than those of their healthy peers.[3] However, these burdens are reduced when conditions are effectively managed, typically accomplished by appropriately monitoring the severity and progression of the condition.[4–8]

Monitoring severity often involves assessment using a health measurement scale, a number of which have been developed that accurately and reliably measure condition severity. These scales are developed in various forms, with some using multiple items to represent the latent construct of severity, and others using only a single item. The latter method is based primarily upon global judgment of the rater, and can improve upon limitations of the former.

Several condition-specific severity scales for children exist but are limited in their clinical utility for general practitioners or pediatricians who care for children with different conditions. Clinical utility can be improved with a quick and easy to use scale that can be used across conditions. Reasons why clinicians forego routine measurement include lack time and lack of scale versatility.[9] From a research standpoint, a scale that could be used across conditions would also be useful for making group-based comparisons.

Adapted from the Global Assessment of Severity of Epilepsy (GASE),[10,11] the Global Assessment of Severity of Illness (GASI) is a single-item scale that can be used to measure severity in children with different chronic physical conditions. This study investigates the validity, reliability, responsiveness and sensitivity/specificity of the GASI. Because the GASI

requires little time of clinicians and addresses various complications in measuring severity of different conditions, it has potential to improve measurement, and possibly, management of chronic conditions in children.

This thesis begins with a review of research and health care concerning children with chronic conditions, including the management of these conditions with the help of severity rating scales. I then explain the rationale behind development and validation of the GASI, and how it improves upon the limitations of current severity scales. Afterward, I describe how clinical data was collected using the GASI and the statistical methods used to assess its psychometric properties. I present the results from psychometric testing and discuss the implications of these findings for research and clinical practice. Finally, I end by reviewing the strengths and weaknesses of this study in addition to proposing important directions for future research.

Literature Review

Children with Chronic Physical Conditions

In children, chronic physical conditions are prevalent, burdensome, and difficult to manage. A chronic physical condition (CPC) is a disorder that has a biologic basis, will last at least one year, and produces at least one of the following sequelae: (a) limitation of physical function in comparison with healthy peers; (b) dependency on medications, special diet, medical technology, assistive devices, or personal assistance; or (c) need for ongoing medical care/accommodation. This modified definition[12] excludes mental disorders[1,13] to clarify the unique factors associated with CPCs in children.

The most common CPCs in children are asthma, food allergy, epilepsy, diabetes and hypertension.[14] The burdens associated with these conditions include those mentioned in the above definition and also include comorbidity. In a large Canadian population-based cohort, Ferro et al.[3] found that children with CPCs are at risk for increased symptoms of anxiety and depression, which corroborates evidence from a large longitudinal study in British children.[15] Pinquart et al.[16] conducted a meta-analysis and found that children with CPCs had elevated levels of anxiety and depression compared to those without CPCs. In addition to outcomes of psychopathology, Varni et al.[17] investigated child and parent reports and found that children with CPCs experienced worse health-related quality of life (HRQL) than healthy children. Burdens of childhood CPCs extend to parents and siblings and include distress surrounding the child's health and safety,[18–20] anxiety related to caregiving responsibilities[21,22] (e.g., building relationships with clinicians),[23] increased levels of interpersonal stress,[24] and financial burden[25] compared to families of children without chronic conditions. The burden of mortality is great, with CPCs being among the top five leading causes of childhood deaths.[26] Many of the burdens

experienced by children with CPCs are moderated by condition specific factors, including severity,[16,17,27,28] making severity measurement an important component of managing CPCs.

Prevention and Management of Chronic Conditions

Measurement of patient outcomes is essential to the work of many health professionals and researchers. In the clinical setting, children with CPCs are typically diagnosed and treated on the basis of outcomes including biologic markers, symptoms experienced, and responses to intervention.[29] Often with the use of scales, this information is obtained by clinicians,[30–32] caregivers[7,8] and patient proxies,[33–35] and patients themselves.[36] Routinely monitoring these outcomes over time is essential for tailoring individual stepped care[29] for people with chronic conditions and is crucial to the success of the Chronic Care Model in children[37] and adults.[38]

In the past few decades, the call for outcome measurement in children[8,9,37,39] and adults[40,41] with chronic conditions has been primarily limited to physiologic monitoring of select conditions[42] such as diabetes[43] and cancer.[30] Although physiologic measurement is often necessary with these conditions, it may not be sufficient for tracking overall patient progress. For example, an individual with cancer may provide a blood sample indicating a reduction in cancer cells, but such a measure will not examine whether functional capacity of the individual has improved. Dimensions of health such as functional capacity, pain, or condition severity must be subjectively measured and such measurements are crucial to monitoring chronic conditions. Kelley et al.[39] refers to measurement of multiple health dimensions as multidimensional measurement or monitoring, a practice which is increasingly emphasized in general medical practice,[44–46] including care for children with CPCs.[47] Bickman et al.[48] also found that many clinicians value receiving regular multidimensional reports on the progress of their patients. There is evidence that routine outcome monitoring using multidimensional measurement can

predict deterioration in the health of patients[9] and improve treatment outcomes[41,49] when implemented appropriately.[50]

In clinical research, the measurement of patient outcomes has also been useful for improving understanding of chronic conditions in children. Cross-sectional investigation, for instance, has been important for studies examining HRQL outcomes in children with CPCs.[28] Likewise, longitudinal research has been used to investigate questions such as whether depressive symptoms in parent-proxies affect their reports on health outcomes in their child with epilepsy.[51] Outcome measurement also plays an important role in public health research. Disease surveillance[52,53] uses routine outcome monitoring data to enhance program planning, accountability, and disbursement of funding for the prevention of chronic conditions.[54]

Health Measurement Scales

The type of outcome that can be obtained from a health measurement scale is determined by how the scale was developed. Initial development of a health measurement scale involves quantifying estimates of healthiness by assigning numerical scores or ordinal categories to subjective clinical judgments.[55] When objective measurement of a health outcome (e.g., severity) is not possible, a subjective process is required. Subjective assessment of health by use of a scale has shown to be valid and reliable.[55] To assess a health construct that is non-observable (i.e., a latent construct), many scales use multiple items or questions to measure observable variables that are related to the latent health construct. Ideally, combining measurements of these variables will provide a more accurate assessment of the latent construct versus assessing the latent construct directly.[56] Using a multi-item scale also provides information about how different items specifically contribute to the measurement of a latent construct. However, single-item scales have been considered to be better measures of latent constructs in a number of situations.[55,57]

Completion of a single-item scale relies on the informant's expertise of the measured construct because additional items are not provided to guide judgment. When multi-item scales measure complex constructs, there is the possibility that relevant elements of the construct are neglected. With a single-item scale, the informant is not limited by specific items but is free to consider all elements relevant to the latent construct. There is evidence that clinicians can adequately estimate latent constructs in this way.[58] Centrally reliant on clinician judgment, single-item scales often perform similarly to multi-item scales,[55] and can outperform multi-item scales when measuring certain constructs.[57] Most importantly, single-item scales have been shown to demonstrate many forms of validity and reliability.[55]

The most common forms of single-item scales include the visual analogue scale (VAS), Likert, and numeric rating scale. The rating format of the Likert and numeric rating scale are considered ordinal, as opposed to continuous, because the response options are separated into distinct ordinal categories.[59,60] The VAS is a continuous line, often 100 mm in length, anchored by descriptions indicating minimum and maximum endpoints of the scale between which the rater places a mark. Some argue that the VAS is not truly continuous because raters still place their mark as if the scale were composed of different categories,[61] while other studies report that this only occurs in VASs with intermittent numbers or symbols.[62] With no categories to guide comparison of different ratings on the VAS, interpretation of meaningful change risks bias.[63] Traditionally, each response option on a Likert scale is anchored by adjectives or descriptive phrases, while numeric rating scale categories are labeled by numbers and anchored with descriptions at the minimum and maximum ends of the scale.[61] Some single-item scales are graphical, such as The Faces Scale,[55] and represent a construct such as mood or pain along a continuum of different facial expressions. The graphical scale is often useful with young children and in cases where language barrier prevents patients from being able to read scale

descriptions.[64] Each single-item scale format has advantages and disadvantages, but all have been found useful in clinical practice and research.[59]

The Feasibility and Utility of Scales for Measuring Patient Outcomes

Implementation of routine multidimensional measurement in clinical practice has proven difficult.[9,31,50,65–70] Because scales contribute to this measurement, identifying issues of scale feasibility and utility may help improve scale development and the success of measurement implementation. The following three issues are commonly found among health measurement scales: 1) The time required for completion is often not practical for busy clinics;[9] 2) Scales have not been validated for measurement across different conditions;[71,72] and 3) Information obtained from scales are typically useful for the clinician, and rarely to additional stakeholders, such as administrators.[9]

Issue 1: The time required to complete a scale primarily depends on the length and complexity of the scale. Clinicians are rightly concerned about giving up current rhythms of practice to adopt those which will accost time from their schedule. Already clinicians do not have enough time to meet practice guidelines for chronic care.[73] In the U.S., some insurance reimbursement policies have led clinicians to limit visits to ten minutes.[74] In consideration of time constraints, scales that require less time from clinicians are better for maintaining desired workflow. Ideal incentivization of multidimensional monitoring will rely on factors other than money. In the U.S., most clinicians are not financially reimbursed for multidimensional monitoring, while they are often reimbursed for running physiological tests.[9] Although Canada has implemented financial incentives for physicians requiring additional time for clinical assessments, such incentives may not be sustainable.[75] Alternative incentives include quick and simple initiatives that improve patient outcomes.

Evidence suggests that implementation of clinical activities, such as outcome documentation,[76,77] is more successful when less time is required from individuals involved.[68,78] Moreover, clinical initiatives are more likely to have long-term success when they are easy to understand.[79] While many scales are being developed with fewer items and shorter length,[55,80] it is important to remember how these characteristics affect the utility and psychometric integrity of the scale. If clinicians are to consider the utility of scales that are short and simple, scale developers must ensure that such scales retain the validity and reliability integral to their use.

Issue 2: Development of scales that can be used for multiple conditions should be considered for the following reasons: 1) Determining eligibility of a scale for multiple patients with different conditions can be an overwhelming process,[67] and so the availability of condition-generic scales can reduce the number of scales that clinicians need to review; 2) Multimorbidity assessment is simplified by using one scale across conditions;[81] and 3) Standardized measurement is needed for valid cross-condition comparison.[82]

Issue 3: Implementation of an activity will typically be more successful when it benefits multiple stakeholders in a system.[78] Scale developers should be cognizant of this principle when designing scales to be implemented for routine measurement. For example, in addition to clinicians, administrators and payers are also stakeholders in routine measurement because they need access to actionable information[9] to predict healthcare utilization and assess overall quality of care.[70,83] Because severity scales provide useful information for these objectives,[84–86] efforts should be made to improve the quality and utility of such scales.

Measuring Severity

Severity scales have been developed for different purposes. When the purpose is not explicit, a scale might be used to measure constructs other than which it was intended. In the words of Ruth Stein, "an appropriate method [of measurement] cannot be selected without

knowledge of purpose."[87] For example, it is unclear whether many asthma scales are assessing asthma severity or asthma control, an important distinction within this condition, especially among children.[88] Evidently, the definitions of severity are as diverse as the purposes for measuring severity. A popular understanding of the different types of severity considers three categories: "physiological or morphological severity; functional severity; and burden of illness."[87] When a scale is designed to predict organ failure[89] or mortality,[90] it is essentially designed to measure and predict "physiological/morphological" severity. When the goal is to measure global severity, the scale should measure all three categories in a weighted or unweighted manner.

Severity scales exist in multi- and single-item form and most often use patient-based (i.e., specific to the patient) rather than condition-based metrics (i.e., specific to the condition).[91] The majority of patient-based scales have been developed to measure severity in a specific condition[10,92–97] or a specific subset of conditions.[76,98,99] Scales limited to the measurement of specific conditions can be referred to as categorical scales, while non-categorical scales allow the measurement of virtually any condition.[12]

While non-categorical multi-item scales are useful for prompting the consideration of various aspects of condition severity, important aspects are often neglected.[57,100] This problem is sometimes ameliorated by limiting the scale to a single item.[56,101] Unfortunately, literature surrounding single-item scales is sparse, and include a number of reports where clinicians used single-item severity scales that were not validated.[93,102] In three different studies on arthritis, including juvenile arthritis, clinicians used the same single-item scale that had no evidence for validity.[103–105] Such scales can potentially misinform research because there is no evidence that they are measuring what they purport to measure, the severity of arthritis. These cases speak to

the need for further development and validation of single-item scales that measure condition severity.

Severity Scales for Chronic Physical Conditions in Children

Despite the rapidly growing evidence-base for the validity and reliability of brief scales,[80,106] only two scales are relevant for clinician global assessment of severity in children with different CPCs: *The Severity of Illness Index* and *The Duke Severity of Illness Scale*.

The Severity of Illness Index (SII): The SII was initially developed as a non-categorical generic severity scale for hospital inpatients.[107] While the SII has undergone a series of alterations to improve its measurement specificity and validity in children, the original and alternative versions continue to be commonly used. The original SII is a seven-item scale with items measuring stage of principal diagnosis, complications, interactions, dependency, procedures, response to therapy, and remission of symptoms, and requires 2 to 15 minutes to complete.[107] Response options range from 1 to 4 with each option labeled by severity criteria specific to the item.[107] Most SII validation studies have not specified age, but at least one is known to include children.[108] The SII has demonstrated excellent interrater agreement (90.8%-97.7%), good face validity as agreed between clinicians, and predictive validity with regard to resource use.[109] Interrater reliability of the SII varies with the type of health professional rating the condition (weighted Kappa= 0.69-0.79).[107] The SII is less reliable in individuals with moderate condition severity versus extreme severity.[108]

An updated, seven-item version of the SII is known as the Comprehensive/Computerized Severity Index (CSI).[110] Though still a non-categorical scale, condition-specific descriptions are provided when a diagnosis is specified. Condition descriptions are enabled by computer algorithms built in the CSI. Because of its success in adults,[110–112] a version of the CSI was developed specifically for use in children. The Pediatric CSI[85] has been shown to predict and

discriminate mortality (Hosmer-Lemeshow tests: p-value= 0.41-0.98; AUC= 0.89-0.99, p-value < 0.001) and explain variation in length of stay and cost of services (R^2 = 0.13-0.67; R^2 = 0.08-0.73, p-value < 0.005).[85]

The Duke Severity of Illness Scale (DUSOI): The DUSOI measures severity of various conditions, using four items that assess symptoms, complications, prognosis, and treatability.[113] Each item is a five-point numeric rating scale. While a composite score of these four items provides a global assessment of condition severity, the DUSOI also includes a single-item global assessment in the form of a horizontal VAS.[113] Using the DUSOI composite score, five studies provided evidence for interrater reliability (ICC= 0.45-0.79)[86,113–116] and two demonstrated intra-rater reliability for individuals with CPCs (ICC= 0.67-0.89).[114,116] There is also evidence for agreement between the DUSOI composite score and the VAS (ICC= 0.61, p-value < 0.001).[113] Although little effort has been made to assess concurrent validity of the DUSOI, the scale shows good clinical face validity[116] and has demonstrated predictive validity in its ability to predict future health service charges (R^2= 0.05).[86] Only a subset of DUSOI validation studies included samples with infants, children, or adolescsents.[115,116] Although the DUSOI takes only one to two minutes to complete,[115,116] complexity of administering the DUSOI makes it less feasible than alternative severity scales.[81] For example, in a study where thirty clinicians used the DUSOI to assess severity, nearly 30% of clinicians reported having difficulty using the scale.[116]

Although these scales present a number of benefits to measuring severity of CPCs, the needs of many clinicians and researchers remain unmet. In sum, the SII is currently limited by its response time (up to 15 minutes), scant evidence for valid use in children, and inability to measure outpatients and individuals with moderate condition severity. However, revisions of the SII resulted in some improvements: The computerized version has been shown to take only 2 minutes to complete, and the Pediatric CSI has demonstrated valid use among a large sample of

hospitalized children. Development of an outpatient version of the Pediatric CSI has been reported,[85] but has not been validated.

Limitations of the DUSOI include complexity of use, average response times ranging over one minute, and clinicians reporting its lack of utility for child health examination.[115] Moreover, clinicians participating in the current study have reported that items on the DUSOI complicated assessment of severity. The VAS that was specifically validated for use with the DUSOI may ameliorate such time and complexity issues. However, the DUSOI VAS has not been validated in a child sample and VASs have, at times, been considered impractical for the clinical setting because they lack categories to assist quick interpretation of the differences between ratings.[59]

Rationale

The Gap in Child Severity Measurement

In response to the limitations of current scales, I performed initial validation of the Global Assessment of Severity of Illness (GASI), a scale specifically developed for the needs of researchers and clinicians caring for children with various CPCs. The GASI is a single-item scale that allows for quick and simple global assessment of severity. As a single-item Likert-type scale, the GASI is expected to improve upon the limitations of similar severity scales (e.g., SII/Pediatric CSI, DUSOI), and provide a step toward standardized measurement of severity across children with different chronic conditions.

Study Objectives

Validation of the GASI included the following tests: 1) Construct validity: whether the scale measures what it purports to measure—the overall severity of a condition; 2) Test-retest reliability: whether the scale returns similar severity measurements at different points in time in the subgroup of children whose condition did not change according to an established clinical measure; 3) Responsiveness: whether the scale is able to detect clinically important changes in the severity; and 4) Sensitivity/Specificity: the probability that the GASI will correctly measure change and no change on an external criterion.[117]

Hypotheses

To achieve these objectives the following hypotheses were tested sequentially according to convention in scale validation research:[55]

1. Construct validity:
    a. *Concurrent validity*: The GASI will have at least moderate correlation ($\tau_c \geq 0.3$),[118] and demonstrate strong relationships (AUC ≥ 0.7),[119] with two established measures of global severity, the VAS and the DUSOI.

b. *Discriminant validity*: The GASI will be moderately correlated ($\tau_c > 0.30$) with HRQL domains on the KIDSCREEN-27 that represent the severity construct, and correlate weakly ($\tau_c = 0.10\text{-}0.30$) with HRQL domains that do not represent condition severity. Additionally, because the GASI is a global assessment, its relationships with the VAS and the DUSOI composite score will be stronger than with individual items on the DUSOI.[120]

c. *Discriminative validity*: GASI ratings will be higher for children with multimorbidity (comorbid mental disorder) versus children without multimorbidity (p-value < 0.05).

2. Test-retest reliability: GASI ratings will not change from baseline to six months in stable subgroups (p-value > 0.05) and will demonstrate adequate test-retest reliability ($\kappa \geq 0.7$).
3. Responsiveness: The GASI will demonstrate a moderate to large magnitude of responsiveness (Cohen's $d > 0.5$) in both a distribution-based assessment (no clinical anchor) and an anchor-based assessment.
4. Sensitivity/Specificity: The GASI will demonstrate at least moderate sensitivity/specificity ($AUC \geq 0.7$)[119] with the DUSOI composite score as an external clinical anchor.

Methods

Study Sample

Data come from the Researching Adolescent and Child Health study (REACH), a six-month prospective pilot study that aimed to assess mental disorders in children newly diagnosed with a CPC.[121] In addition to appraising the feasibility of a larger follow-up study, goals of the REACH pilot study were to assess the prevalence of child multimorbidity, identify factors correlated with multimorbidity in children and parents, and assess the effects of multimorbidity on changes in child quality of life and parental psychosocial outcomes over six months. Health professionals recruited families from two pediatric academic hospitals in Ontario, Canada with the aim of recruiting 60 children and families over 12 months. Recruitment targeted families at four outpatient clinics where a child had been recently diagnosed with a CPC. Participating clinicians were the first to have contact with eligible families, providing them with study details for participation.

Study inclusion criteria required that the child was aged 6 to 16 years, was diagnosed with asthma, diabetes, epilepsy, food allergy or juvenile idiopathic arthritis no more than 6 months before recruitment, and that at least one parent could read English. The study aimed to recruit 12 children per chronic condition, for a total of 60 children with their families. Minimum age criterion was specified based on the minimum age that was valid for the study measures, and maximum age criterion was specified to ensure children did not transition into adult care during the study. Inclusion criterion for diagnoses of child chronic conditions was specified to represent the most common CPCs in children.[14] English skills in parents were required because not all study measures have been validated in other languages. Children diagnosed with a degenerative neurological disorder were excluded.

Data Collection

Study investigators followed up with eligible consenting families to schedule a convenient time for a telephone interview to assess child mental disorder, and surveys were mailed to parents at baseline and six months to measure psychosocial outcomes and demographic characteristics. Child mental disorder was assessed again at six months by telephone interview. All parents provided proxy reports for their children, and children who were at least 11 years of age self-reported for mail surveys and telephone interviews. Specific details about how study measures were used are explained below, and more details about the REACH study are documented elsewhere.[121]

Analyses in the current study were conducted on data from parent and clinician reports.[121] Though 62 families were contacted to participate in the REACH study, 50 participated and 44 were retained. Parent reports provided data for 50 children at baseline and 44 at the six-month follow-up. Participation and retention were better among clinicians; their reports provided data for 55 children at baseline and 51 at follow-up. While an appropriate sample size for validating single-item scales has not been substantiated, generally increasing the number of items in a scale also increases the sample size required for robust validation testing.[122] Moreover, pilot studies, such as the REACH study, involving initial scale development and validation do not require the same level of power as a comprehensive scale analysis.[123,124] Guyatt et al.[125] suggested that a sample size of n = 34 (paired observations) is sufficient to establish responsiveness if the scale is predicted to be moderately responsive ($d \geq 0.5$).[126] Terwee et al.[119] consider a sample size of at least n= 50 adequate for most validation tests. For validation in pilot studies, Johanson et al.[123] recommend having samples at least n=24, and they support the recommendation of Hertzog et al.[127] n= 30 to n= 40 if study objectives primarily involve group comparisons, rather than intervention. Validation of single-item scales has involved samples as small as n = 9,[128] n = 35,[129] n= 40,[124] n = 71[130] and n = 75.[131] While scientific consensus on computing *a priori* sample

size is lacking,[119,132] especially for single-item scales, these findings provide a helpful context for appraising practical significance of sample sizes for single-item validation studies. Notably, the sample here provided by the REACH study meets nearly all these sample size recommendations for initial scale validation. Further comments on statistical power of the tests in this study are provided in the analysis and discussion sections of this thesis.

Measures

Clinician Report: Immunologists completed severity assessments for children with asthma and food allergies, endocrinologists for children with diabetes, neurologists for children with epilepsy, and rheumatologists for children with juvenile idiopathic arthritis. These clinicians were asked to carefully read instructions before using the study severity measures, and otherwise, no further measurement training was provided. The clinician most familiar with the condition of the child completed the severity reports. Severity was measured using the GASI, the DUSOI and the VAS. The GASI is a single-item 7-point Likert scale that asks clinicians to rate the severity of a condition given a range of response options from 'Not at all severe' to 'Extremely severe' (see Figure 1). The GASI was adapted from the single-item Global Assessment of Severity of Epilepsy (GASE) scale, which is valid and reliable[10,11] and was specifically designed to improve upon existing multi-item scales by measuring all aspects of epilepsy severity.[10] Unlike the GASE, the GASI uses the term "disease" rather than "epilepsy" when asking, "Taking into account all aspects of this patient's [disease], how would you rate its severity as his/her last visit?"

Assessment of Illness Severity:

Taking into account all aspects of this patient's disease, how would you rate its severity at his/her last visit? (check only one)

☐	☐	☐	☐	☐	☐	☐
Extremely severe	Very severe	Quite severe	Moderately severe	Somewhat severe	A little severe	Not at all severe

Figure 1. The Global Assessment of Severity of Illness

The DUSOI includes four individual components of severity (symptoms, complications, prognosis, and treatability) and a composite score of the four severity components (see Figure 2). The four components of severity are each assessed using five-point scales, and the composite score is calculated as the summed four ratings divided by the total summed score possible. The DUSOI has been used in patients aged 4 months to 89 years,[115] has demonstrated high reliability and clinical face validity,[113,114] and demonstrated greatest clinical value for individuals with CPCs versus other conditions.[115,116] Because extensive research surrounds validation of the DUSOI composite score, this score was used as the clinical anchor for condition severity in all final models and tests in this study.

The VAS is a 50 mm horizontal line where the distance measured from the leftmost part of the line to the rater's mark is converted into a score out of 100 (see Figure 3). The 0 mm endpoint is anchored by the phrase "lowest severity" and 50 mm by "highest severity".[113] The VAS measures multidimensional constructs[133] and characteristics on a continuum[134] while demonstrating consistent precision over time.[135] Although commonly a self-report tool, there is evidence for valid use of the VAS by external raters. In one study, investigators used the VAS and a simple descriptive scale to measure functional capacity of patients with rheumatoid arthritis completing two different tasks, and there were significant correlations between the two scales for assessment of both tasks ($\rho = 0.42$; $\rho = 0.54$).[136] When used by clinicians, the VAS demonstrated agreement with the DUSOI composite score (ICC= 0.61, p-value < 0.001).[113]

Assessment of Illness Severity:

	None	Questionable	Mild	Moderate	Severe	
1. Symptoms (since last visit)	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	
2. Complications (since last visit)	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	
		Disability				
3. Prognosis	None	Mild	Moderate	Major	Threat to Life	
	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	
	Need for Treatment					
	No	Questionable	**If Yes →**	Good	Questionable	Poor
4. Treatability	☐ 0	☐ 1		☐ 2	☐ 3	☐ 4

Figure 2. The Duke Severity of Illness scale

Please mark with an X the appropriate place along the line below to indicate how you would rate this patient's overall severity of illness since their last visit.

LOWEST SEVERITY _____ **HIGHEST SEVERITY**

(Please mark one X along the line)

Figure 3. The visual analogue scale

Parent Report: Parents used the KIDSCREEN-27 to report on child HRQL and study investigators administered the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID) with parents to screen for child mental disorder. The KIDSCREEN-27 is a multidimensional HRQL instrument with 27 items measuring five domains: physical well-being, psychological well-being, parent relations and autonomy, social support and peers, and school environment.[137] Domain scores are generated from five-point Likert scale items and are converted to T-scores with a mean of 50 and a standard deviation of 10. Higher T-scores indicate better HRQL. In a large international sample, the KIDSCREEN-27 displayed excellent internal reliability ($\alpha > 0.78$) and item discriminant validity (IDV > 80%), and reasonable structural validity (RMSEA= 0.069).[137] Further testing in samples of children with and without CPCs confirmed acceptable test-retest reliability (ICC= 0.61-0.74), satisfactory criterion validity (r= 0.71-0.96), and acceptable convergent and discriminant validity with other HRQL instruments.[138] The KIDSCREEN-27 has demonstrated low to moderate informant agreement between children with CPCs and their parents, with agreement improving over time.[139] The KIDSCREEN-27 has also been found to demonstrate partial measurement invariance in a clinical sample of children with mental disorder and their parents.[140]

The MINI-KID is a structured diagnostic interview that uses screening questions and skip patterns to screen for 24 child and adolescent mental disorders contained in the DSM-IV and ICD-10.[141] The interview is conducted with children aged 6 to 17 years, their parents, or both, and takes approximately 30 minutes to administer.[141] Not all MINI-KID modules were used in the REACH study. Rather, modules were used that screen for the most common mental disorders in children. The MINI-KID has been validated using the Schedule for Affective Disorders and Schizophrenia for School Aged Children-Present and Lifetime Version and demonstrated very good interrater reliability (AUC ≥ 0.89), acceptable to excellent test-retest reliability (AUC ≥

0.75), acceptable to high sensitivity (0.43-1.00), and substantial to high specificity (0.73-1.00).[142] Confirmatory factor analysis provides evidence for convergent and discriminant validity of the MINI-KID using latent factors from the Brief Child and Family Phone Interview (BCFPI),[142] which is a validated measure of child mental disorder.[143] Diagnostic agreement between the standard version of the MINI-KID and the parent-proxy version (MINI-KID-P) was higher in a sample containing primarily outpatients (κ= 0.46-0.94)[141] versus a primarily population-based sample (κ= 0.05-0.33).[142] Recent work comparing the validity of various diagnostic tools, including self-completed problem checklists like the KIDSCREEN-27, discusses how structured interviews are a suitable means for classifying child mental disorder in clinical research.[144]

Analysis

The following analyses were performed in SAS Studio 9.0.4. As would be expected in an outpatient sample, initial exploration of the data revealed that clinicians did not use the full range of ratings on the GASI scale (see Appendix B, Figures B1-B2). Clinician ratings of severity clustered primarily among the lowest 5 ordinal outcomes and were characterized by a positively skewed distribution. Therefore, final tests and models were conducted using analytic methods best suited for ordinal categorical outcomes and nonparametric distributions. Granted most studies involving 7-point Likert scales use tests assuming continuous outcomes, Appendix C contains results of the following analyses using tests that assume a continuous outcome. Including results of the continuous outcome analyses extends the metric for comparison with other studies and further substantiates the findings of final models and tests contained in the main body of this thesis.[56]

Construct Validity: Concurrent validity of the GASI was assessed by measuring the correlation of GASI ratings with scores on the DUSOI and VAS using Kendall's *Tau-c* (τ_c)[145–148] correlation coefficient. The *Tau-c* is ideal for large frequency tables and is recommended for

tables that are not square. Correlations were calculated using scores at baseline and six months. Cohen's conventions were used to interpret τ_c= 0.10-0.30 as weak correlation, τ_c= 0.30-0.50 as moderate, and $\tau_c > 0.50$ as strong,[118] while also keeping in mind that *Tau-c* has been considered by some to be an overly conservative estimate of correlation.[149]

Furthermore, strength of association with the VAS and DUSOI were measured using generalized linear modeling and logistic regression. In these models, area under the receiver operating characteristic curve (AUC) demonstrated the overall strength of association and regression coefficients evaluated the contribution of individual terms in the model. AUCs less than 0.5 were interpreted as the association being no better than chance, 0.5 to 0.7 as a weak association, 0.7 to 0.9 as moderate, and over 0.9 as strong.[56] Final models accounted for (a) random variance introduced by time (baseline to six months), (b) random variance correlated among like diagnoses, which simultaneously accounted for the type of treating clinician, and (c) the potential confounding effects of child age. Age of a child can potentially affect how a clinician approaches the severity assessment and possibly influence complexity of the assessment. For example, among the chronic conditions included in this sample, age has shown to be associated with the type of symptoms experienced by children.[150,151] Furthermore, some symptoms that emerge at these ages are very difficult to discern.[152] With the other model adjustment, cluster sizes in the random effects statement were unbalanced after nesting children within their diagnoses. Potential bias from unbalanced clusters was accounted for by both the internal SAS syntax of PROC GLIMMIX[153–157] and by the Kenward-Roger correction for denominator degrees of freedom.[153]

Discriminant validity of the GASI was assessed using the KIDSCREEN-27 and the DUSOI. The KIDSCREEN-27 is suitable for discriminant validation because some domains on this instrument are intrinsically related to condition severity and others less so. Kendall's *Tau-c*

was used to measure correlation of GASI ratings with scores on different domains of the KIDSCREEN-27, testing functional bounds of the GASI.[158] It was hypothesized that the GASI would correlate weakly (τ_c= 0.10-0.30) with the KIDSCREEN-27 domains representing unrelated constructs (Parent Relations and Autonomy, Social Support and Peers, School Environment). Weak correlation with these domains would support divergent validity. Evidence for convergent validity required at least moderate correlation ($\tau_c > 0.30$) with the KIDSCREEN-27 domains more representative of condition severity (Psychological Well-being, Physical Well-being). Additionally, because the GASI is a global assessment, correlation with the VAS and DUSOI composite score was hypothesized to be stronger than correlation with individual items on the DUSOI.[120]

"Discriminative" validation tested whether the scale was able to discriminate between children with multimorbidity versus children without multimorbidity. Multimorbidity was coded as screening positive for mental disorder using the MINI-KID. Fisher's Exact test was performed on GASI ratings from the group of children with multimorbidity and from the group with no multimorbidity. Tests were performed using baseline and six-month ratings. A p-value less than 0.05 was required to demonstrate ability of the GASI to discriminate between patients with and without multimorbidity.

Test-retest Reliability: Support of test-retest reliability required that GASI ratings did not change from baseline to six months in the subgroup of patients whose conditions remained stable throughout the study according to the DUSOI and VAS, the established clinical measures.[11] Establishing reliability provides evidence for minimal measurement error when using the GASI, and is a necessary prerequisite for assessing the responsiveness of a scale.[56] In order to understand which dimensions of severity the GASI measures reliably over time, multiple tests were conducted using different items from the DUSOI. Patient subgroups were classified as

"stable" according to the clinical anchor used in each test-retest analysis. Table 1 reports the "stable" criteria that were applied to the data in order to conduct test-retest analyses using the Kappa coefficient, weighted Kappa, and the intraclass correlation coefficient (ICC).

Table 1 – Criteria for Stable Severity as Defined by Study Measures

	Criteria when computing Kappa	Criteria when computing Weighted Kappa and ICC
"Stable" defined by the DUSOI	When using DUSOI items *(symptoms, complications, prognosis, treatability)*: ○ Item rating must not change from low (<2) to high (≥2), or vice versa When using DUSOI composite score: ○ Item rating must not change from low (<40%) to high (≥40%), or vice versa	When using DUSOI items *(symptoms, complications, prognosis, treatability)*: ○ Must have no change in item rating When using DUSOI composite score: ○ Must have <8.3% change in score
"Stable" defined by the VAS	Rating must not change from low (<40%) to high (≥40%), or vice versa	Must have <10.9% change in the rating

ICC= intraclass correlation coefficient

As dichotomization is necessary for computing the Kappa coefficient, the criterion used to classify stable patients using the DUSOI items was primarily based on the descriptions of high and low severity that were evident in the scale options. Because severity of the study sample clustered at the lower end of all the severity scales, cut-points for stability in the DUSOI composite score and VAS were placed toward the lower end of the scale to ensure reasonable sizes of dichotomized groups. Group sizes were afterward verified by assessing the distributions of severity ratings. For GASI ratings, it seemed appropriate to dichotomize by aggregating

ratings from “Not at all severe” to “A little severe” (Low severity) and “Somewhat severe” to “Extremely severe” (High severity), in consideration of the right-skewed distribution of severity in the sample.

For analysis with the weighted Kappa and ICC, the rational for criteria used to classify stable patients using individual DUSOI items was based on evidence that clinically important change can be represented by approximately half a point change in a 7-point Likert scale.[159] Therefore, any change observed in the DUSOI item would be clinically meaningful. Additionally, response options for these items possess clinical descriptions that intrinsically demonstrate clinical importance with a mere one-point change. Clinical stability is not intuitively observed in the DUSOI composite and VAS ratings and was estimated using other methods. Common cut-point estimates for clinical stability include half a standard deviation in scale scores[160] or 0.5 change in a 7-point Likert scale.[159] The latter cut-point estimate was redefined as an 8.3% change because the DUSOI composite and VAS are not 7-point Likert scales, and was performed using an equidistant transformation of the cut-point

$$\left\{m = categories;\ \left(\left(\frac{x+1}{m-1} - \frac{x}{m-1}\right) \div 2\right) = \left(\left(\frac{x+1}{6} - \frac{x}{6}\right) \div 2\right) * 100\% \approx 8.3\%\right\},$$

as suggested by Svensson.[63] The primary indicator of a valid cut-point is that it yields rates similar to the referenced norm.[56,57] Presented below is the rational for using different cut-points when assigning clinically stable subgroups using the DUSOI composite score and the VAS.

The referenced norm cut-point for identifying clinical stability was “no change in individual DUSOI items”. Because the DUSOI is an established clinical measure its use as the referenced norm improved the quality of reliability testing[119] and potentially prevented false-negative test-retest results.[161] The four individual DUSOI items yielded clinically stable subgroups with a mean of n= 32.5 and a median of n= 32, roughly representing the norm size for a stable subgroup in this sample. In comparison, the half a standard deviation cut-point estimate

yielded n=8 for the DUSOI composite and n=28 for the VAS, while the 0.5 change on 7-point scale (~8.3% change) estimate yielded n=29 for the DUSOI composite and n=24 for the VAS. In sum, the latter estimate better resembles the norm rates when using the DUSOI composite, and the former estimate better resembles the norm rates when using the VAS. The half a standard deviation estimate is sometimes too stringent when most individuals in the sample have low to moderate levels of impairment, in this case severity, and are less likely to make large improvements.[160] This effect may be stronger in composite scores compared to single ratings like the VAS, and may explain why half a standard deviation was not a suitable cut-point for the DUSOI composite score in this sample.

Using the clinically stable subgroups, test-retest reliability was first assessed using the Generalized McNemar test.[148,162] Generalized McNemar statistics with a p-value greater than 0.05 provided initial evidence that GASI ratings in the stable subgroup did not change from baseline to six months. Next, the Kappa and weighted Kappa coefficients with 95% confidence intervals were used to assess reliability of measurements over time.[56,126] A Kappa coefficient less than or equal to 0 is typically interpreted as poor agreement, .01 to .20 as slight, .21 to .40 as fair, .41 to .60 as moderate, .61 to .80 as substantial, and .81 to 1 as almost perfect.[163] A Kappa coefficient of at least 0.70 was necessary to establish reliability of the GASI.[126,164] Weighted Kappa was computed using quadratic (Fleiss-Cohen) weights[165] as they are appropriate for ordinal outcomes with potentially large tables[166] and allow for meaningful comparison with the ICC.[56,167] Indeed, when the sample is large enough ($n \geq 40$)[165] the ICC and quadratic weighted Kappa are identical.[56] Because meaningful interpretation of weighted Kappa is often lost by differential weighting schemes,[163] this study advocates the use of Kappa quadratic weights for future reliability testing of the GASI when the ICC cannot be used. The test-retest ICC is not appropriate for analyzing ordinal categorical outcomes when the sample is not large.[163,165,168]

Fortunately, both Kappa and weighted Kappa are measures of absolute agreement[167,169] and therefore meet criteria for robust test-retest designs.[170,171]

Responsiveness: Internal responsiveness was first assessed using a distribution-based approach where the standardized response mean[117] (SRM) was computed. Distribution-based assessment of responsiveness was required[172] because there is no *ideal* gold-standard severity scale to function as an anchor for assessing responsiveness. The SRM was calculated by dividing mean change in GASI ratings from 0 to 6 months by standard deviation of change ratings during that time. Using the SRM, as opposed to *t*-tests, removes reliance on sample size[117] and provides a within-person assessment of change.[172] In the absence of a normal distribution, a bootstrap procedure was implemented to obtain an approximate distribution from which to compute the SRM point estimate and 95% confidence intervals with interpretation using Cohen's conventions (d= 0.2; 0.5; and 0.8) for small, medium and large magnitude of responsiveness.[173]

Next, an anchor-based approach was used to measure responsiveness. Using a clinically anchored subset of "changed" scores (DUSOI scores that did not qualify as "stable"),[11] Guyatt's responsiveness index[125] (GRI) was calculated to assess ability of the GASI to detect clinically important changes in condition severity.[117] The GRI is the ratio of average change scores of changed patients divided by the standard deviation of the change scores in stable patients, and was interpreted using Cohen's effect size conventions as previously mentioned.

Sensitivity/Specificity: The AUC, calculated by generalized linear and logistic regression models, indicated ability of GASI to discriminate between "stable and "changed" patients using change criteria defined by the DUSOI.[11,117] An AUC less than 0.5 was interpreted as sensitivity/specificity being no better than chance, 0.5 to 0.7 as low, 0.7 to 0.9 as moderate, and over 0.9 as high.[56,174] Model adjustments for child characteristics were identical to those applied in the models assessing concurrent validity.

Results

Sample Characteristics

The mean age of children in the study sample was 11.3 years (SD= 3.3) and, overall, both sexes were almost equally represented (47.3% males). At baseline (n= 55), the majority of children were diagnosed with asthma (n= 16), and then followed by juvenile arthritis (n= 12), diabetes (n= 11), epilepsy (n= 8), and food allergy (n= 8). Among these children, 58.2% screened positive for mental disorder. At the six-month follow-up, minimal attrition was observed (n= 4), and while the proportions of CPCs represented in the sample was nearly unchanged, there was a lower prevalence of multimorbidity (42.9%). Median severity ratings for food allergy were higher than or equal to all other diagnoses for all severity scales at baseline and six months, while median severity ratings for diabetes were always the lowest. Table 2 contains additional details of the study sample.

Construct Validity

Concurrent Validity: Table 3 contains results of Kendall's *Tau-c* correlation between the GASI and established severity measures. All correlations with the GASI were significant ($p \leq .001$) at baseline (τ_{c1}) and six months (τ_{c2}), with the exception of the DUSOI items *Symptoms* and *Complications* which were not statistically significant. As hypothesized, correlations were strongest with the DUSOI composite score (τ_{c1}= 0.63, CI= 0.51-0.76; τ_{c2}= 0.57, CI= 0.46-0.69) and the VAS (τ_{c1}= 0.78, CI= 0.67-0.88; τ_{c2}= 0.78, CI= 0.65-0.91), providing evidence for convergent validity. The correlation with *Prognosis* was also strong (τ_{c1}= 0.64, CI= 0.50-0.77; τ_{c2}= 0.68, CI= 0.56-0.80). Nearly identical patterns of correlation magnitude and statistical significance were found when this analysis was performed using Spearman's rank correlation (see Appendix C, Table C1).

Table 2 – Sample Characteristics and Severity Scoring at Baseline and 6 Months

	Full Sample	Food Allergy	Asthma	Diabetes	Epilepsy	Juvenile Arthritis
Baseline						
n	55	8	16	11	8	12
Age, mean (SD)	11.3 (3.3)	10.2 (3.4)	10.3 (3.3)	13.7 (1.7)	12.8 (2.1)	11.5 (2.8)
Male, n (%)	26 (47.3)	7 (87.5)	5 (31.25)	3 (27.3)	7 (87.5)	4 (33.3)
Multimorbid, n (%)	32 (58.2)	5 (62.5)	10 (62.5)	6 (54.6)	4 (50.0)	7 (58.3)
Condition Severity						
GASI, median (IQR)	16.7 (33.3)	41.7 (25.0)	16.7 (25.0)	0.0 (0.0)	33.3 (33.3)	16.7 (41.7)
DUSOI, median (IQR)	37.5 (12.5)	37.5 (0.0)	37.5 (9.4)	18.8 (18.8)	31.3 (21.9)	37.5 (18.8)
VAS, median (IQR)	16.3 (41.0)	45.3 (30.7)	16.7 (52.6)	0.0 (4.4)	33.8 (38.7)	18.2 (42.3)
Six Months						
n	51	8	15	10	7	11
Multimorbid, n (%)	21 (42.9)	2 (25.0)	9 (60.0)	1 (11.1)	4 (50.0)	5 (55.6)
Condition Severity						
GASI, median (IQR)	16.7 (33.3)	83.3 (0.0)	16.7 (16.7)	0.0 (0.0)	16.7 (33.3)	16.7 (50.0)
DUSOI, median (IQR)	31.3 (18.8)	43.8 (0.0)	37.5 (12.5)	12.5 (12.5)	25.0 (6.3)	31.3 (12.5)
VAS, median (IQR)	19.8 (47.0)	85.9 (5.3)	15.6 (26.0)	7.6 (17.4)	17.8 (17.6)	18.2 (36.8)

For comparison, all severity scores have been standardized to a scale of 0 (lowest severity) to 100 (highest severity).

IQR= interquartile range

Table 3 – Concurrent Validity Assessed by Kendall's *Tau-c* Correlation

	Baseline			6 Months		
Severity Measure	**τ_c (95% CI)**	**p-Value**	**n**	**τ_c (95% CI)**	**p-Value**	**n**
DUSOI: Symptoms	0.24 (-0.00, 0.47)	0.051	51	-0.04 (-0.28, 0.21)	0.753	51
DUSOI: Complications	0.05 (-0.04, 0.14)	0.267	50	0.06 (-0.07, 0.19)	0.353	51
DUSOI: Prognosis	0.64 (0.50, 0.77)	<.001	53	0.68 (0.56, 0.80)	<.001	51
DUSOI: Treatability	0.26 (0.11, 0.41)	0.001	52	0.38 (0.20, 0.56)	<.001	51
DUSOI: Composite	0.63 (0.51, 0.76)	<.001	49	0.57 (0.46, 0.69)	<.001	51
VAS	0.78 (0.67, 0.88)	<.001	52	0.78 (0.65, 0.91)	<.001	51

Kendall's *Tau-c* correlation coefficients (τ_c) are reported with 95% confidence intervals at baseline and 6 months for correlation of the GASI with individual DUSOI items, the DUSOI composite score, and the VAS.

Table 4 contains the results of unadjusted and adjusted regression models assessing the strength of association between the GASI (dependent variable) and the DUSOI composite score (independent variable) at baseline (AUC_1) and six months (AUC_2). The unadjusted models demonstrated moderate to strong associations between the GASI and DUSOI composite score (AUC_1= 0.88, CI= 0.79-0.98; AUC_2= 0.83, CI= 0.70-0.95). After adjusting for correlated random effects within diagnosis groups (nested effects), the baseline AUC increased (AUC_1= 0.94, CI= 0.85-1.00), but the AUC in the six-month model decreased (AUC_2= 78, CI= 0.64-0.92). Further adjusting the model for child age resulted in improvement in the baseline model (AUC_1= 0.96, CI= 0.92-1.00) and in the six-month model (AUC_2= 0.86, CI= 0.75-0.97).

In the longitudinal models (see Table 5), the unadjusted AUC again demonstrated a moderate to strong association between the GASI and the DUSOI composite score (AUC= 0.85, CI= 0.77-0.93). Magnitude of the AUC was further increased after adjusting for nested random effects (AUC= 0.94, CI= 0.89-0.99), but the AUC was unaffected by the addition of the covariate for child age. Because the DUSOI composite score is the most validated and provides the most parsimonious model, its results are shown here and provide the main support for concurrent validity of the GASI. Results for the joint model with all DUSOI items and for the model with independent associations with severity scales also demonstrated strong associations overall (see Appendix A, Tables A1-A3). This reflects what is also found for associations measured using multiple linear regression (see Appendix C, Tables C2-C4) and further supports concurrent validity of the GASI.

Table 4 – Concurrent Validity Assessed by AUC in Unadjusted and Adjusted Regression Models

		Baseline			6 Months		
Model	**Independent Variables**	**Est. (SE)**	**n**	**AUC (95% CI)**	**Est. (SE)**	**n**	**AUC (95% CI)**
1	DUSOI	23.07 (2.88)†	49	0.88 (0.79, 0.98)	11.20 (3.68)†	51	0.83 (0.70, 0.95)
2 (nested)	DUSOI	23.05 (7.65)†	49	0.94 (0.85, 1.00)	11.21 (3.71)†	51	0.78 (0.64, 0.92)
3 (nested)	DUSOI	31.39 (11.79)*	44	0.96 (0.92, 1.00)	11.27 (3.83)†	49	0.86 (0.75, 0.97)
	Child Age	0.49 (0.22)*	--	--	0.07 (0.12)	--	--

Area under the receiver operating characteristic curve (AUC) is reported for each step of constructing the fully adjusted model. AUCs demonstrate strength of association between the GASI (dependent variable; dichotomized as Low/High severity) and the DUSOI composite score. The GASI was dichotomized by aggregating ratings from “Not at all severe” to “A little severe” (Low severity) and “Somewhat severe” to “Extremely severe” (High severity). Models 2 and 3 include a nested random effects statement identifying children within their treating clinician. Parameter estimates are only significant where noted. *p<.05, †p<.01

Est.= Estimate

SE= Standard error

Table 5 – Concurrent Validity Assessed by AUC in Longitudinal Unadjusted and Adjusted Regression Models

Model	Independent Variables	Estimate	SE	p-Value	n	AUC (95% CI)
1	DUSOI	11.15	2.64	<.001	100	0.85 (0.77, 0.93)
2 (nested)	DUSOI	13.80	3.48	<.001	100	0.94 (0.89, 0.99)
3 (nested)	DUSOI	14.96	3.74	<.001	93	0.94 (0.89, 0.99)
	Child Age	0.20	0.11	0.069	--	--

Area under the receiver operating characteristic curve (AUC) is reported for each step of constructing the fully adjusted model. AUCs demonstrate strength of association between the GASI (dependent variable; dichotomized as Low/High severity) and the DUSOI composite score by assessing both GASI and DUSOI measurements performed at baseline and 6 months while controlling for time (baseline to six months). GASI dichotomization and nested random effects are equivalent to Table 4.

SE= Standard error

Discriminant Validity: Table 6 displays the results of discriminant validity testing for correlation (τ_c) of the GASI with individual domains on the KIDSCREEN-27. All correlations were nonsignificant at baseline and six months (p-value > 0.05). Divergence was observed among baseline tests, with weak correlations ranging from τ_c= -0.004 for *Parents and Autonomy* to τ_c= -0.10 for *Social Support and Peers.* Negative correlations were observed for all baseline measures, indicating that as condition severity increased, HRQL decreased. At six months, correlations with all KIDSCREEN-27 domains were weak (τ_c= -0.12-0.08) and the direction of relationships differed across KIDSCREEN-27 domains. These same patterns of correlation strength and direction of relationship were found when analysis was performed using Spearman's rank correlation (see Appendix C, Table C5).

Discriminative Validity: The results of Fisher's Exact tests for discriminative validity of the GASI are reported in Table 7. Because the initial frequency table had low cell counts, a second table was created with aggregated GASI ratings to ensure associations between the GASI and presence of multimorbidity were adequately tested. GASI ratings did not discriminate between children with and without multimorbidity (p-value > 0.05). This finding is also supported by results from the Mann-Whitney U Test (see Appendix C, Table C6).

Table 6 – Discriminant Validity Assessed by Kendall's *Tau-c* Correlation

KIDSCREEN-27 Domain	Baseline			6 Months		
	τ_c (95% CI)	p-Value	n	τ_c (95% CI)	p-Value	n
Physical Well-being	-0.02 (-0.25, 0.20)	0.851	48	0.08 (-0.14, 0.30)	0.483	44
Psychological Well-being	-0.07 (-0.27, 0.13)	0.485	49	-0.08 (-0.36, 0.19)	0.556	43
Parents and Autonomy	-0.004 (-0.22, 0.21)	0.968	48	0.05 (-0.18, 0.29)	0.665	44
Social Support and Peers	-0.10 (-0.30, 0.11)	0.345	49	-0.08 (-0.33, 0.18)	0.546	44
School Environment	-0.08 (-0.29, 0.14)	0.490	46	-0.12 (-0.38, 0.15)	0.381	43

Kendall's *Tau-c* correlation coefficients (τ_c) are reported with 95% confidence intervals at baseline and 6 months for correlation of the GASI with individual KIDSCREEN-27 domains.

Table 7 – Discriminative Validity Assessed by Fisher's Exact Test of Independence

	Baseline			6 Months		
Frequency Table	**Table Probability (P)**	**p-Value**	**n**	**Table Probability (P)**	**p-Value**	**n**
(a) Multimorbidity (yes/no) x GASI (0 – 4)	<.001	0.096	55	0.003	0.527	51
(b) Multimorbidity (yes/no) x GASI (0 – 2)	0.028	0.476	55	0.022	0.343	51

The results of Fisher's Exact test for two frequency tables are reported. Frequency tables were analyzed at baseline and six months using data in (a) raw and (b) aggregated format. Frequency table details: (a) Two columns pertain to presence of multimorbidity (No Multimorbid, Yes Multimorbid) and five rows pertain to GASI ratings (0=Not at all severe; 1=A little severe; 2=Somewhat severe; 3=Moderately severe; 4=Quite severe/Very severe); (b) Two columns pertain to presence of multimorbidity (No Multimorbid, Yes Multimorbid) and three rows pertain to GASI ratings (0=Not at all severe; 1=A little severe, Somewhat severe; 2=Moderately severe, Quite severe/Very severe).

Test-retest Reliability

Tables 8-10 contain results for test-retest reliability. Generalized McNemar's test demonstrated no change in GASI ratings from baseline to six months for all stable subgroups defined by individual DUSOI items, the DUSOI composite score and the VAS (p-value > 0.05). These findings are equivalent to results from the analysis using Wilcoxon signed rank test (see Appendix C, Table C7).

According to the Kappa coefficient, which examined the reliability of change in the GASI for ratings that change from low to high, or high to low, strength of agreement in GASI ratings ranged from moderate (κ= 0.57, CI= 0.32-0.82) for stable subgroups defined by *Treatability* to almost perfect (κ= 0.87, CI= 0.69-1.00) for stable subgroups defined by the VAS. Similar to the VAS subgroup, substantial agreement was demonstrated using GASI ratings from the stable subgroup defined by the DUSOI composite score (κ= 0.79, CI= 0.51-1.00). Similar to the tests for construct validity, the magnitude of Kappa coefficients were highest when incorporating severity scales more related to global severity.

Weighted Kappa tests, which examined the reliability of smallest changes possible in GASI ratings, generated smaller reliability coefficients than Kappa tests. Agreement ranged from moderate (κ_w= 0.45; CI= 0.15-0.75) in the *Treatability* subgroup to almost perfect (κ_w= 0.81, CI= 0.64-0.99) in the VAS subgroup. Agreement between GASI ratings at baseline and six months was moderate for the subgroup defined as stable according to the DUSOI composite score (κ_w= 0.57, CI= 0.36-0.78). These weighted Kappa results closely reflect findings from the test-retest analysis using bootstrapped ICCs (see Appendix C, Table C8).

Table 8 – Test-retest Reliability Assessed by Generalized McNemar's Test of Homogenous Distributions

Measure Defining Stable Subgroup	Test Statistic (GMN)	DF	p-Value	n
DUSOI: Symptoms	4.32	3	0.229	27
DUSOI: Complications	8.88	4	0.064	38
DUSOI: Prognosis	1.14	3	0.767	30
DUSOI: Treatability	2.37	4	0.667	34
DUSOI: Composite	9.17	4	0.057	29
VAS	2.67	3	0.446	28

The results of Generalized McNemar's Test for six frequency tables are reported. Frequency tables were constructed using GASI ratings from children in the Stable subgroup. Stable is defined by individual DUSOI items, the DUSOI composite score, and the VAS (see Table 1).

DF= Degrees of freedom

Table 9 – Test-retest Reliability Assessed by Kappa Coefficient

Measure Defining Stable Subgroup	κ (95% CI)	n
DUSOI: Symptoms	0.63 (0.37, 0.90)	37
DUSOI: Complications	0.64 (0.39, 0.88)	43
DUSOI: Prognosis	0.64 (0.39, 0.90)	35
DUSOI: Treatability	0.57 (0.32, 0.82)	45
DUSOI: Composite	0.79 (0.51, 1.00)	29
VAS	0.87 (0.69, 1.00)	36

Kappa coefficients (κ) with 95% confidence intervals (CI) are reported. Coefficients were computed using GASI ratings from children in the Stable subgroup. Stable is defined by individual DUSOI items, the DUSOI composite score, and the VAS (see Table 1). GASI dichotomization is equivalent to the description in Table 4.

Table 10 – Test-retest Reliability Assessed by Weighted Kappa Coefficient

Measure Defining Stable Subgroup	κ_w (95% CI)	n
DUSOI: Symptoms	0.62 (0.43, 0.81)	27
DUSOI: Complications	0.53 (0.31, 0.75)	38
DUSOI: Prognosis	0.63 (0.46, 0.80)	30
DUSOI: Treatability	0.45 (0.15, 0.75)	34
DUSOI: Composite	0.57 (0.36, 0.78)	29
VAS	0.81 (0.64, 0.99)	28

Kappa coefficients with quadratic weights (κ_w) and 95% confidence intervals (CI) are reported. Coefficients were computed using GASI ratings from children in the Stable subgroup. Stable is defined by individual DUSOI items, the DUSOI composite score, and the VAS (see Table 1).

Responsiveness

Results for responsiveness of the GASI are found in Table 11. Distribution-based responsiveness, as measured by the SRM, demonstrated ability of GASI ratings to respond to meaningful changes in condition severity where meaningful change was defined by the distribution of change magnitude observed in the sample. Therefore, based on a bootstrapped distribution of GASI ratings at baseline and six months, the GASI demonstrated a large magnitude of responsiveness (SRM= 0.84, CI= 0.68-1.11). Anchor-based responsiveness, as measured by the GRI, demonstrated ability of GASI ratings to respond to change in severity where change was defined by external anchors. As explained in the methods section, the external anchors that provided change definitions were the individual DUSOI items, the DUSOI composite score, and the VAS. According to the change that occurred in these anchors, the GASI demonstrated a medium (GRI= 0.77, CI= 0.24-1.47) to large (GRI= 3.83, CI= 2.65-6.27) ability to detect change in condition severity. Importantly, the magnitude of responsiveness in the GASI was large when change was defined by the clinical anchor, the DUSOI composite score (GRI= 0.83, CI= 0.29-1.70).

Sensitivity/Specificity

Table 12 describes the models and results for regression analysis of sensitivity/specificity of the GASI. In the unadjusted model, the AUC confidence interval dipped just below the null value (AUC= 0.5), suggesting that the ability of the GASI to discriminate between change and no change in condition severity is no better than chance (AUC= 0.62, CI= 0.46-0.77). Adjusting for diagnosis (nesting child in clinician) improved magnitude of this estimate (AUC= 0.78, CI= 0.63-0.92), and resulted in a model that satisfied the *a priori* requirement for adequate sensitivity/specificity. Further adjusting the model for child age increased the AUC again (AUC= 0.81, CI= 0.68-0.94).

Table 11 – Responsiveness Assessed by Standardized Response Mean and Guyatt's Responsiveness Index

Responsiveness Test	Test Statistic (95% CI)	n
Standardized Response Mean	0.84 (0.68, 1.11)	51
GRI when change defined by:		
DUSOI: Symptoms	1.09 (0.51, 1.96)	47
DUSOI: Complications	0.77 (0.24, 1.47)	46
DUSOI: Prognosis	1.01 (0.56, 1.64)	49
DUSOI: Treatability	1.74 (0.83, 2.94)	48
DUSOI: Composite	0.83 (0.29, 1.70)	45
VAS	3.83 (2.65, 6.27)	48

The results of distribution- and anchor-based responsiveness statistics with bootstrapped 95% confidence intervals (CI) are reported as the standardized response mean and Guyatt's Responsiveness Index (GRI), respectively.

Table 12 – Sensitivity/Specificity Assessed by AUC in Unadjusted and Adjusted Regression Models

Model	Independent Variables	Estimate	SE	p-Value	n	AUC (95% CI)
1	∆GASI	0.29	0.21	0.179	45	0.62 (0.46, 0.77)
2 (nested)	∆GASI	0.29	0.22	0.196	45	0.78 (0.63, 0.92)
3 (nested)	∆GASI	0.28	0.23	0.223	43	0.81 (0.68, 0.94)
	Child Age	-0.04	0.11	0.747	--	--

Area under the receiver operating characteristic curve (AUC) is reported for each step of constructing the fully adjusted model. AUCs demonstrate the magnitude of sensitivity/specificity of the GASI when identifying “change” or “no change” in condition severity. Condition severity was considered “changed” if it did not meet criteria for stability as defined by the DUSOI composite score (see Table 1; “Computing Weighted Kappa and ICC”). Models 2 and 3 include a nested random effects statement identifying children within their treating clinician.

∆GASI= (GASI ratings at six months) – (GASI ratings at baseline)

SE= Standard error

Discussion

In response to the limitations of current severity assessments, this study provided evidence for validity of the GASI—a scale developed to improve brief global assessment of severity in children with CPCs. In brief, the GASI was found to be valid, reliable, and responsive. The scale demonstrated appropriate associations with established severity measures and was able to detect changes in condition severity, doing so with minimal measurement error. As intended in its development, the GASI had a precise scope of measurement. Its ratings reflected condition severity rather than other potentially related constructs such as HRQL or presence of mental disorder.

Construct Validity: The GASI demonstrated excellent construct validity, including concurrent and discriminant features, establishing that it measures what it purports to measure—the global severity of CPCs. Concurrent validity was robust regardless of model adjustments for diagnosis (treating clinician) and child age, providing evidence that using the GASI across children aged 6-17 years with different CPCs is valid.

The GASI did not have a perfect association with the DUSOI composite score, which suggests the two scales measure severity somewhat differently. This could potentially be explained by the fact that the GASI was designed to measure global severity where the weighting of individual aspects of severity is inherently performed by the clinician. In contrast, the DUSOI composite score measures severity constructs that are both specific (e.g., complications) and global (e.g., prognosis), and does so in an unweighted manner. Indeed, the GASI converged with *Prognosis* and diverged with the DUSOI's potentially less "global" items, *Symptoms* and *Complications*. In the current sample, the DUSOI item *Symptoms* may also be less relevant because (a) the study clinicians found its meaning ambiguous for the conditions being assessed,

and (b) clinicians have previously ranked it as less helpful for assessing chronic conditions compared to acute conditions.[116]

The strongest relationships among scales were observed between the GASI and the VAS. This could potentially be explained by them both being single-item scales, neither being distorted by items or indicators less related to global severity. The small degree of weakness observed in the relationship between the GASI and the VAS could be explained by a number of factors. For example, the area to place ratings for the GASI is about twice as large as that for the VAS, which may have caused clinicians perceived the magnitude of severity to be greater at the endpoints of the GASI compared to the endpoints of the VAS. Research also shows that VASs are prone to "end-of-scale" effects, where ratings trend toward ends of the scale, and that Likert scales are prone to "middle-of-scale" effects, where ratings trend toward middle of the scale,[175] which would temper associations between these scales. In addition, some discrepancies between the GASI and VAS could be attributed to random error.[56]

Divergent validity of the GASI is demonstrated by the absence of strong correlations with any domain of child HRQL. The overall lack of correlation observed corroborates findings from the recent pan-Canadian study on pediatric epilepsy where family factors were more relevant than severity when modeling HRQL.[176,177] Although severity of a CPC may be a poor indicator of HRQL overall, observing no correlation with the HRQL domain relating to physical well-being is somewhat surprising because there is a theoretical relationship between these constructs. In the current study, factors that may contribute to lack of correlation with the physical well-being domain include (a) the potential masking/confounding effects of multimorbidity,[178] for which investigation by regression was beyond the scope of this study, and (b) discrepancies known to exist between parent and clinician perspectives on child HRQL,[179] which may also be relevant to physical and psychological domains.

Fisher's Exact test of independence demonstrated inability of the GASI to discriminate between children classified as multimorbid and not multimorbid. This is not surprising considering clinicians were directed to rate severity of children's physical, not mental, conditions. Severity was not associated with multimorbidity in this sample, which contrasts with a recent report that the presence of mental disorder in youth was associated with higher levels of disability compared to youth with only physical conditions.[180]

Test-retest Reliability: This study provided evidence to support test-retest reliability of the GASI. First, reliability was supported by the generalized McNemar's test, which provided reason to proceed with more rigorous reliability testing. Next, depending on which clinical cut-point was used, the Kappa coefficient demonstrated a substantial to almost perfect magnitude of agreement. This means that in an outpatient population of children with CPCs, change observed from "low severity" to "high severity", or vice versa, is meaningful and should not be attributed to measurement error.[56] Furthermore, test-retest analyses using the weighted Kappa and the stability definition derived from the VAS demonstrated that *any* change in the GASI is meaningful. However, the same cannot be said when stability is defined by the DUSOI composite score, the study's main clinical anchor. Moreover, because of small diagnosis subgroups in the weighted Kappa analyses, generalizability of these findings to various CPCs is limited. Therefore, current interpretation of the GASI should rely on low versus high severity ratings (as described above) rather than on minimum changes in the scale until additional research in this area is conducted.

Responsiveness: The GASI is highly capable of detecting meaningful change in condition severity. Some global severity scales have been known to demonstrate more responsiveness than measures tapping individual domains,[57,120] which may explain the notably high responsiveness observed in the GASI. With test-retest reliability having established that changes observed in the

GASI are meaningful, responsiveness findings establish that the GASI can detect such changes in the severity of a CPC when they occur.[56]

Sensitivity/Specificity: The findings suggest that when the MCID is applied across different diagnoses, the GASI is unable to discriminate between change that is or is not clinically important. However, because the sensitivity/specificity regression models were adequately powered for making comparisons ($n \geq 44$),[56] comparison among these models indicates that this discriminative ability would be restored if factors related to diagnosis were controlled for. It is possible that the definition of MCID was not appropriate for every type of CPC in the sample. Criteria for MCID in a scale are often informed by characteristics of the diagnosed condition.[181] If future studies are unable to establish a MCID that is generalizable across multiple CPCs, MCID cut-points for the GASI will need to be established for individual diagnoses. However, the confidence interval computed in this study is wide and includes appropriate magnitudes of the AUC for sensitivity/specificity. This suggests that the current finding should not be considered conclusive, and that using the current MCID definition in a larger sample may demonstrate adequate sensitivity/specificity in the GASI.

Implications

The findings from this study have important implications for research and clinical practice. The GASI is able to compare severity across a number of different childhood CPCs and can therefore help fill gaps in comparative pediatric research. For example, in this study the GASI provided evidence that, on average, mental disorder had no effect on severity of CPCs in this sample. Follow-up research using the GASI could aim to replicate these findings and test whether the effect varies across different CPCs. When choosing a tool to assess severity, the GASI may better reflect global properties of severity than the DUSOI[10] and may provide easier interpretation of severity than the VAS which has no categories to explain the different meanings

ratings.[59] As a result, the GASI may provide a more suitable overall assessment of severity to, for instance, help clinicians understand the progress of chronic conditions,[9] or help patients understand the effects of self-management.[182,183] The GASI can also be used by researchers and a variety of healthcare stakeholders.

With regards to its use in the clinical setting, the GASI is a simple and ultra-brief severity assessment that is ideal for busy practices. As such, the GASI could contribute to a solution surrounding the systems-level issue of deficient routine outcome monitoring. Routine outcome monitoring is advocated in the pediatric Chronic Care Model[37] and has been shown to improve patient outcomes and reduce burdens on clinicians and healthcare systems.[9] However, this activity is often hindered because clinical information systems essential to the Chronic Care Model are either missing, not used, or misused.[184] In a study of 108 care teams that manage CPCs across the United States, the majority of care teams did not have a condition registry with which to track the progress of patients toward clinical goals.[42] This problem also limits public health practice that uses data on patient outcomes over time.[52,53] At a more nuanced level, even when clinical information systems are utilized, problems still exist at the point of clinician documenting.[77,185–187] In a retrospective cohort study of 2,109 Canadian patients hospitalized for myocardial infarction, an investigation of clinician documentation quality found that only 58% of patient charts contained information on whether the patient had a previous history of heart failure, which should always be included in charts for these patients.[185] The GASI could equip clinical information systems with a versatile scale allowing quick and simple outcome measurement and incentivize proper documenting of outcomes by clinicians.

Study Strengths and Limitations

Strengths of this study include a wide range of analyses useful for comparison with future investigations of the GASI. Evidence for the current findings is strengthened by convergence of

a priori hypotheses for the main analyses.[56] The same can be said for the alternative analyses that were performed assuming continuous outcomes (see Appendix C). Tests were conducted on data from real clinical settings, compared to the common alternative of case vignettes or patient charts, and therefore supports use of the GASI in outpatient practices. The regression adjustments for child and clinician characteristics assisted interpretation of results and informed future directions for validating the GASI. This use of regression modeling is a demonstration of the validity generalization methodology developed by Hunter et al.[188] Because this methodology is not widely used in scale validation, the present study is an important example of advanced applications in validation science, especially for studies with small samples.

The following limitations may be found in this study. The form that was completed by clinicians contained all three severity measures, allowing for potential priming effects between scales. However, the effect of priming is likely to be minimal based on the different scale formats, the different question prompts across the scales, and the differential functioning observed between the scales (i.e., discriminant validity). Because there is no 'gold standard' severity scale for this population, the DUSOI is not a perfect clinical anchor. Ideal gold standards are rare[172] and appropriate gold standards are difficult to find for most validation studies in the healthcare field.[67] This means that construct validation requires additional forms of testing, as performed in this study using discriminant and discriminative validation[56] with multiple severity scales[87] and *a priori* hypotheses.[55]

Generalizability of the findings may be limited by the relatively small sample. However, the sample size was typical for initial validation of single item scales and statistical power was adequate for all final models and tests, with the exception of the weighted Kappa analyses. These analyses, at times, had incomplete representation of diagnoses and sample sizes were just below the threshold (n= 30)[170] recommended for interpreting conventional reliability indices. In

addition, missing data may bias the results of statistical analyses, especially because missing data bias has a stronger effect in small samples. Rates of missingness, and potential missing data biases, were greatest for discriminant analyses with the KIDSCREEN-27 (11%-16%) and for the fully adjusted AUC models (15%-20%) testing concurrent validity and sensitivity/specificity. However, missing data did not exceed 5% in the longitudinal concurrent validity AUC model or in any test that used only the six-month data. Multiple imputation was not used because little is known about its validity when missingness is less than 20% in small samples.[189] Based on the pattern of missingness and on comments made by study clinicians, missing data is best explained by clinicians preferring not to use certain items on the DUSOI.

External validity of the diagnosis subgroups is also limited. Although this study included the most common childhood CPCs,[122] which is ideal for validation of a generic severity scale, only one female was present in each of the epilepsy and food allergy subgroups. However, it is not apparent that equal sex representation would yield findings different from the current study. For children with epilepsy, type of seizure is the only variable consistently related to sex,[190,191] and research in childhood food allergy has found that sex is not significantly related to prevalence or severity.[192] With regards to age, both the epilepsy and diabetes subgroups included children who were ten years of age and older, and these children were more likely to be experiencing puberty.[193] While puberty has negligible effect on epilepsy,[194] it may increase diabetes related complications.[150] However, there was no bias toward increased *Complications* or *Symptoms* (measured by the DUSOI) in the study subgroup with diabetes. Rather, like the other subgroups the diabetes subgroup had very low ratings for *Complications* and *Symptoms*.

Additionally, limitations exist in the assessment of test-retest reliability. Retest after six months is reasonable for children with epilepsy whose conditions are expected to remain stable during this time.[11] However, this may not be applicable for every CPC in the sample. I

ameliorated this issue by defining stability as having equivalent severity scores at baseline and six months on a highly reliable clinical anchor. This definition of stability meant that the *overall* contribution of severity determinants was equivalent at baseline and six months, but it did not guarantee that the *proportional* contribution of these determinants was equivalent at both times, which is typically assumed with shorter test-retest periods. Though this is a limitation of the current test-retest analysis, it is not likely to have a large affect findings surrounding the reliability of global ratings, such as in the GASI, in the same way it would affect ratings of individual aspects of severity. Finally, a long delay before retest has the benefit of preventing recall bias from invalidating test-retest results, especially for short scales[56] like the GASI.

Future Considerations

The priority for future work is to further evaluate reliability of the GASI. Currently, the reliability findings only support use the GASI where ratings are interpreted in a binary sense (low versus high severity), which limits its utility. Future testing of inter-rater reliability may support interpretation of the full range of GASI ratings[56] and also provide evidence for whether other clinicians, such as nurse practitioners, can reliably use the scale. Knowledgeable informants of condition severity also include children with CPCs and their parents, and future work should investigate whether they can reliably complete the GASI and whether those assessments are useful for research and routine clinical practice. Such analyses would ideally incorporate reliability generalization methodology to identify variables that affect the magnitude of reliability.[56]

Additionally, further testing should examine whether current definitions for MCID can be used to establish sensitivity/specificity of the GASI in a larger sample, or whether diagnosis-specific definitions for MCID are necessary. Overall, future validation studies will benefit from using larger samples where there is better representation of the age and sex within each diagnosis

subgroup. Understanding external validity of the GASI will also be improved by including only the GASI on study forms completed by clinicians (i.e., avoid priming effects) and through the collection of more diagnostic details surrounding diabetes (e.g., Type 1, Type 2) and epilepsy (e.g., temporal, complex partial, generalized). Future studies should assess missing data and acknowledge these diagnostic details when doing so.

Effort should also be devoted to examining whether the GASI is a valid scale for other physical conditions in children. Furthermore, future work should also consider whether the GASI improves upon the limitations of current severity assessments in children with diagnosed mental disorders, with undiagnosed conditions, or with acute conditions.

Finally, future research should gather evidence on feasibility and examine whether implementation of the GASI helps achieve patient-centered and systems-level goals in clinical practice. Initial evidence of feasibility is found in the current study as the GASI had fewer missing data than all other study measures. Future feasibility testing may include asking clinicians how long it took them to complete the GASI. Implementation research with the GASI would be suitable in a chronic care model.[38] For example, because of the increasing use of patient satisfaction as a healthcare performance measure,[195] pediatric outpatient practices using the chronic care model could investigate whether including the GASI within the clinical information system[182] could support patient-provider discussion such that improvements are seen in child and parent (a) understanding of progress of the condition, (b) understanding of the care plan, and (c) satisfaction with the care plan.[183] Such research would make a valuable contribution to the evidence for chronic care models, as robust evaluations are lacking.[196]

With regards to implementation in clinical practice, discussions should begin surrounding the risks of using the GASI so that consequential validity[56] can be established. Consequential validity is especially relevant in cases where a scale influences clinical decisions such as whether

or not a child meets program eligibility or should receive a certain medical intervention.[12] Physicians' consensus at the Third Conference on Advances in Health Status Assessment was that measurement best informs practice when generic scales are used first and afterward supplemented by diagnosis- or problem-specific scales.[197] Following such guidelines may reduce the risk of negative consequences of scale use. Additional considerations on how to interpret and respond to severity data have been previously published[55] and also provided by generic scales such as the Global Assessment Scale (GAS)[76] and Severity of Illness Score (SIS).[110] Though some have argued that the consequences of using a subjective scale should be primarily attributed to the clinician,[198,199] knowledgeable informants such as children and parents are also responsible for the outcomes of care in some ways.[183]

Consequential validation should also consider how non-systematic use of a single-item global scale can result in variability of construct measurement, especially when the construct has previously been defined in different ways. Such measurement issues can be largely circumvented by paying careful attention to the question prompt of a scale. For example, numerous scales have been developed to measure both condition 'control' and condition 'severity.' With asthma, for instance, former definitions of severity provided by U.S. national asthma guidelines were narrow, defining severity as "the level of control in the unmedicated state."[88] This definition was irrelevant for the majority of asthma patients, only useful for initial consultations where asthma had not yet been treated. Addendums to this definition have since attempted to improve the usefulness of severity assessment,[200] but the definition is still minimally relevant for the majority of outpatient visits.

Fortunately, a definition of global severity is evident within the question prompt of the GASI and is relevant for routine clinical assessment of conditions like asthma. The question prompt of the GASI (see Figure 1) indicates that the severity rating should encompass all aspects

of a patient's condition, not just a few, and not aspects less relevant to severity. Evidence that this occurs was demonstrated by discriminant validation in the current study. In fact, our findings showed that the GASI was significantly correlated with *Treatability*, the DUSOI item that is equivalent to condition control. Hence, in contrast to typically less relevant severity assessments of conditions like asthma, the GASI offers a useful severity assessment for the preponderance of clinical visits, including visits where condition control is an important component of the assessment. This study provides evidence that the severity definition of the GASI achieves what is recommended by Stein et al.,[87] that each severity scale should be clear about its goal in measuring the severity construct. That said, careful attention to the question prompt will preserve utility of the GASI and minimize deviation in the construct that is measured.

With these recommendations in mind, the GASI should be considered for a variety of activities in the clinical setting. For example, a clinician could use the GASI to track conditions over time by rating severity of a CPC at each visit and occasionally reviewing the trend of severity ratings. An increase in severity ratings could notify the clinician of the child's deterioration and highlight the need to administer a more problem-specific assessment or reevaluate the child's care plan. Similar action could also be taken if opportunities for improvement were available, but severity ratings remained unchanged over time. In contrast, if when reviewing GASI ratings the clinician notices a trend of decreasing severity or severity remaining stable at the level of "not at all severe," they should consider whether additional changes in the child's care plan (e.g., reducing pharmacotherapy) would improve the child's quality of life without compromising stability of the CPC.

Similar guidelines for scales have been followed in routine care for mental disorders[9] and have been highly effective for complex systems-level healthcare coordination. For instance, according to a donor support coordinator at the Donor Network of Arizona (Tompke AA 2018,

email communication, 26th Nov), Arizona's successful coordination of organ donation has relied heavily on systematic use of the Glasgow Coma Scale[201] (GCS) for nearly 30 years. The GCS ranges from 3 (severe coma) to 15 (no impairment). The Donor Network of Arizona uses the GCS almost daily to (a) track ventilated patients with impaired consciousness and (b) facilitate communication between donor support coordinators, nurse practitioners, physicians, and organ procurement teams. If coordinators are notified of ventilated patients with an initial GCS score ≤ 5, they immediately refer the patient for assessment with the organ recovery coordinator. However, if the GCS is above 5 they assign another team to track worsening or improvement of the condition over five days. Coordinators then make a follow-up phone call to the patient's nurse and repeat this protocol.

These GCS guidelines have demonstrated long-term success in supporting the work of numerous professionals coordinating health care across different settings. Elements of these guidelines could be translated for outpatient application of the GASI to improve systems-level activities, such as patient referrals. Resources for development and evaluation of clinical guidelines[78,202] will make a valuable contribution to future explorations of the utility of the GASI.

Conclusion

In conclusion, researchers and clinicians should be confident that rating severity using the GASI is valid and reliable when interpretation is limited to a dichotomous outcome of low or high severity. Initial evidence supports this approach when using the GASI among children with select chronic conditions in the outpatient setting. Additional research, ideally in a larger clinical sample, will be required to support interpretation of the full range of GASI ratings. The GASI can be used for monitoring severity over time and for making valid comparisons of severity between children with asthma, food allergy, diabetes, epilepsy, and juvenile arthritis. The GASI

presents numerous advantages over current scales that assess severity in children with chronic conditions and demonstrates potential to help reduce burdens on the healthcare system and improve the health of children.

References

1. van der Lee JH, Mokkink LB, Grootenhuis MA, Heymans HS, Offringa M. Definitions and measurement of chronic health conditions in childhood: a systematic review. JAMA. 2007;297(24):2741–51.
2. Newacheck PW, Kim SE. A national profile of health care utilization and expenditures for children with special health care needs. Arch Pediatr Adolesc Med. 2005;159(1):10.
3. Ferro MA, Boyle MH. The impact of chronic physical illness, maternal depressive symptoms, family functioning, and self-esteem on symptoms of anxiety and depression in children. J Abnorm Child Psychol. 2015;43(1):177–87.
4. Katon W, Von Korff M, Lin E, Simon G. Rethinking practioner roles in chronic illness: the specialist, primary care physician, and the practice nurse. Gen Hosp Psychiatry. 2001;23:138–44.
5. Von Korff M, Glasgow RE, Sharpe M. ABC of psychological medicine: organising care for chronic illness. BMJ. 2002;325(7355):94.
6. Von Korff M, Gruman J, Schaefer J, Curry SJ, Wagner EH. Collaborative management of chronic illness. Ann Intern Med. 1997;127:1097–102.
7. Knafl K, Breitmayer B, Gallo A, Zoeller L. Family response to childhood chronic illness: description of management styles. J Pediatr Nurs. 1996;11(5):315–26.
8. Hafetz J, Miller VA. Child and parent perceptions of monitoring in chronic illness management: a qualitative study. Child Care Heal Dev. 2010;36(5):655–62.
9. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychother Res. 2015;25(1):6–19.
10. Speechley KN, Sang X, Levin S, Zou GY, Eliasziw M, Smith M Lou, et al. Assessing

severity of epilepsy in children: preliminary evidence of validity and reliability of a single-item scale. Epilepsy Behav. 2008;13(2):337–42.

11. Chan CJ, Zou G, Wiebe S, Speechley KN. Global assessment of the severity of epilepsy (GASE) scale in children: validity, reliability, responsiveness. Epilepsia. 2015;56(12):1950–6.

12. Stein REK, Bauman LJ, Westbrook LE, Coupey SM, Ireys HT. Framework for identifying children who have chronic conditions: the case for a new defintion. J Pediatr. 1993;122(3):342–7.

13. Pless IB, Douglas JWB. Chronic illness in childhood: part I. epidemiological and clinical characteristics. Pediatrics. 1971;47(2):405–14.

14. Miller GF, Coffield E, Leroy Z, Wallin R. Prevalence and costs of five chronic conditions in children. J Sch Nurs. 2016;32(5):357–64.

15. Pless IB, Power CC, Peckham CS. Long-term psychosocial sequelae of chronic physical disorders in childhood. Pediatrics. 1993;91(6):1131–6.

16. Pinquart M, Shen Y. Anxiety in children and adolescents with chronic physical illnesses: a meta-analysis. Acta Paediatr. 2011;100(8):1069–76.

17. Varni JW, Limbers CA, Burwinkle TM. Impaired health-related quality of life in children and adolescents with chronic conditions: a comparative analysis of 10 disease clusters and 33 disease categories/severities utilizing the PedsQL^{TM} 4.0 Generic Core Scales. Health Qual Life Outcomes. 2007;5.

18. Uhl T, Fisher K, Docherty SL, Brandon DH. Insights into patient and family-centered care through the hospital experiences of parents. J Obstet Gynecol Neonatal Nurs. 2013;42(1):121–31.

19. Leventhal-Belfer L, Bakker AM, Russo CL. Parents of childhood cancer survivors: a

descriptive look at their concerns and needs. J Psychosoc Oncol. 1993;11(2):19–41.

20. Pinquart M. Parenting stress in caregivers of children with chronic physical condition—a meta-analysis. Stress Heal. 2018;34(2):197–207.

21. Sharpe D, Rossiter L. Siblings of children with a chronic illness: a meta-analysis. J Pediatr Psychol. 2002;27(8):699–710.

22. Murphy NA, Christian B, Caplin DA, Young PC. The health of caregivers for children with disabilities: caregiver perspectives. Child Care Health Dev. 2006;33(2):180–7.

23. Smith J, Cheater F, Bekker H. Parents' experiences of living with a child with a long-term condition: a rapid structured review of the literature. Heal Expect. 2013;18:452–74.

24. Reichman NE, Corman H, Noonan K. Effects of child health on parents' relationship status. Demography. 2004;41(3):569–84.

25. Zan H, Scharff RL. The heterogeneity in financial and time burden of caregiving to children with chronic conditions. Matern Child Health J. 2015;19:615–25.

26. Hamilton BE, Hoyert DL, Martin JA, Strobino DM, Guyer B. Annual summary of vital statistics: 2010-2011. Pediatrics. 2013;131(3):548–58.

27. Bompori E, Niakas D, Nakou I, Siamopoulou-Mavridou A, Tzoufi MS. Comparative study of the health-related quality of life of children with epilepsy and their parents. Epilepsy Behav. 2014;41:11–7.

28. Zashikhina A, Hagglof B. Health-related quality of life in adolescents with chronic physical illness in northern Russia: a cross-sectional study. Health Qual Life Outcomes. 2014;12(1):12.

29. Von Korff M, Tiemens B. Individualized stepped care of chronic illness. West J Med. 2000;172:133–7.

30. Hannan C, Lambert MJ, Harmon C, Nielsen SL, Smart DW, Shimokawa K, et al. A lab

test and algorithms for identifying clients at risk for treatment failure. J Clin Psychol. 2005;61(2):155–63.

31. Jette DU, Halbert J, Iverson C, Miceli E, Shah P. Use of standardized outcome measures in physical therapist practice: perceptions and applications. Phys Ther. 2009;89(2):125–35.

32. Dutton M. Chapter 1: who are physical therapists, and what do they do? In: Kearns B, editor. McGraw-Hill's NPTE (National Physical Therapy Exam). 2nd ed. 2012.

33. Upton P, Lawford J, Eiser C. Parent-child agreement across child health-related quality of life instruments: a review of the literature. Qual Life Res. 2008;17(6):895–913.

34. Ravens-Sieberer U, Herdman M, Devine J, Otto C, Bullinger M, Rose M, et al. The European KIDSCREEN approach to measure quality of life and well-being in children: development, current application, and future advances. Qual Life Res. 2014;23:791–803.

35. Varni JW, Limbers CA, Burwinkle TM. How young can children reliably and validly self-report their health-related quality of life?: an analysis of 8,591 children across age subgroups with the PedsQL™ 4.0 Generic Core Scales. Health Qual Life Outcomes. 2007;5:1–13.

36. Wagner EH, Austin BT, Korff M Von, Wagner EH, Austin BT. Organizing care for patients with chronic illness. Milbank Q. 1996;74(4):511–44.

37. Lail J, Schoettker PJ, White DL, Mehta B, Kotagal UR. Applying the chronic care model to improve care and outcomes at a pediatric medical center. Jt Comm J Qual Patient Saf. 2017;43(3):101–12.

38. Coleman K, Austin BT, Brach C, Wagner EH. Evidence on the chronic care model in the new millennium. Health Aff. 2009;28(1):75–85.

39. Kelley SD, Ph D, Bickman L. Beyond outcomes monitoring: Measurement Feedback

Systems (MFS) in child and adolescent clinical practice. Curr Opin Psychiatry. 2010;22(4):363–8.

40. Basch E, Deal AM, Kris MG, Scher HI, Hudis CA, Sabbatini P, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol. 2016;34(6):557–65.

41. Carlier IVE, Meuldijk D, Van Vliet IM, Van Fenema E, Van Der Wee NJA, Zitman FG. Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. J Eval Clin Pract. 2012;18(1):104–10.

42. Bonomi AE, Wagner EH, Glasgow RE, Vonkorff M. Assessment of Chronic Illness Care (ACIC): a practical tool to measure quality improvement. Health Serv Res. 2002;37(3):791–820.

43. Shah BR, Hux JE, Laupacis A, Zinman B, Van Walraven C. Clinical inertia in response to inadequate glycemic control: do specialists differ from primary care physicians? Diabetes Care. 2005;28(3):600–6.

44. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. Routine use of patient reported outcome measures in healthcare settings. BMJ. 2010;340(7744):464–7.

45. Miller SD, Hubble MA, Chow D, Seidel J. Beyond measures and monitoring: realizing the potential of feedback-informed treatment. Psychotherapy. 2015;52(4):449–57.

46. Gilbody SM, House AO, Sheldon TA. Outcomes research in mental health. Br J Psychiatry. 2002;181:8–16.

47. Michaud PA, Suris JC, Viner R. The adolescent with a chronic condition. Part II: healthcare provision. Arch Dis Child. 2004;89(10):943–9.

48. Bickman L, Douglas SR, De Andrade ARV, Tomlinson M, Gleacher A, Olin S, et al. Implementing a measurement feedback system: a tale of two sites. Adm Policy Ment Heal

Ment Heal Serv Res. 2016;43(3):410–25.

49. Shimokawa K, Lambert MJ, Smart DW. Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. J Consult Clin Psychol. 2010;78(3):298–311.

50. Wolpert M, Deighton J, De Francesco D, Martin P, Fonagy P, Ford T. From 'reckless' to 'mindful' in the use of outcome data to inform service-level performance management: perspectives from child mental health. BMJ Qual Saf. 2014;23(4):272–6.

51. Ferro MA, Avison WR, Campbell MK, Speechley KN. Do depressive symptoms affect mothers' reports of child outcomes in children with new-onset epilepsy? Qual Life Res. 2010;19(7):955–64.

52. Perlman SE, McVeigh KH, Thorpe LE, Jacobson L, Greene CM, Gwynn RC. Innovations in population health surveillance: using electronic health records for chronic disease surveillance. Am J Public Health. 2017;107(6):853–7.

53. Frieden TR. Asleep at the switch: local public health and chronic disease. Am J Public Health. 2004;94(12):2059–61.

54. Porterfield DS, Rogers T, Glasgow LM, Beitsch LM. Measuring public health practice and outcomes in chronic disease: a call for coordination. Am J Public Health. 2015;105:S180–8.

55. Mcdowell I. Measuring health: a guide to rating scales and questionnaires. Third Edit. New York: Oxford University Press; 2006.

56. Streiner, David L., Norman, Geoffrey R., Cairney J. Health measurement scales: a practical guide to their development and use. 5th Editio. Oxford: Oxford University Press; 2015. 399 p.

57. Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C. Assessing the clinical

significance of single items relative to summated scores. Mayo Clin Proc. 2002;77(5):479–87.

58. Charlson ME, Sax FL, MacKenzie CR, Fields SD, Braham RL, Douglas RGJ. Assessing illness severity: does clinical judgment work? J Chronic Dis. 1986;39(6):439–52.

59. Hasson D, Arnetz BB. Validation and findings comparing VAS vs. Likert scales for psychosocial measurements. Int Electron J Health Educ. 2005;8:178–92.

60. Price DD, Bush FM, Long S, Harkins SW. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. Pain. 1994;56(2):217–26.

61. van Tubergen A, Debats I, Ryser L, Londoño J, Burgos-Vargas R, Cardiel MH, et al. Use of a numerical rating scale as an answer modality in ankylosing spondylitis-specific questionnaires. Arthritis Care Res (Hoboken). 2002;47(3):242–8.

62. Pincus T, Bergman M, Sokka T, Roth J, Swearingen C, Yazici Y. Visual analog scales in formats other than a 10 centimeter horizontal line to assess pain and other clinical data. J Rheumatol. 2008;35(8):1550–8.

63. Svensson E. Concordance between ratings using different scales for the same variable. Stat Med. 2000;19:3483–96.

64. Tomlinson D, von Baeyer CL, Stinson JN, Sung L. A systematic review of faces scales for the self-report of pain intensity in children. Pediatrics. 2010;126(5):e1168–98.

65. Kieling C, Baker-Henningham H, Belfer M, Conti G, Ertem I, Omigbodun O, et al. Child and adolescent mental health worldwide: evidence for action. Lancet. 2011;378(9801):1515–25.

66. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013;346:f167.

67. Greenhalgh J, Long AF, Brettle AJ, Grant MJ. Reviewing and selecting outcome measures for use in routine practice. J Eval Clin Pr. 1998;4(4):339–50.

68. Snyder CF, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Qual Life Res. 2012;21(8):1305–14.

69. Mellor-Clark J, Cross S, Macdonald J, Skjulsvik T. Leading horses to water: lessons from a decade of helping psychological therapy services use routine outcome measurement to improve practice. Adm Policy Ment Heal Ment Heal Serv Res. 2016;43(3):279–85.

70. Georgiou A, Pearson M. Measuring outcomes with tools of proven feasibility and utility: the example of a patient-focused asthma measure. J Eval Clin Pract. 2002;8(2):199–204.

71. Kazdin AE. Evidence-based assessment for children and adolescents: issues in measurement development and clinical application. J Clin Child Adolesc Psychol. 2005;34(3):559–68.

72. Achenbach TM. Advancing assessment of children and adolescents: commentary on evidence-based assessment of child and adolescent disorders. J Clin Child Adolesc Psychol. 2005;34(3):541–7.

73. Østbye T, Yarnall KSH, Krause KM, Pollak KI, Gradison M, Michener JL. Is there time for management of patients with chronic diseases in primary care? 2005;3(3):209–14.

74. Busner J, Targum SD. The clinical global impressions scale: applying a research tool in clinical practice. Psychiatry (Edgmont). 2007;4(7):28–37.

75. Lavergne RM, Law MR, Peterson S, Garrison S, Hurley J, Cheng L, et al. A population-based analysis of incentive payments to primary care physicians for the care of patients with complex disease. CMAJ. 2016;188(15):E375–83.

76. Endicott J, Spitzer RL, Fleiss JL, Cohen J. The Global Assessment Scale: a procedure for

measuring overall severity of psychiatric disturbance. Arch Gen Psychiatry. 1976;33:766–71.

77. Spurgeon A, Hiser B, Hafley C, Litofsky NS. Does improving medical record documentation better reflect severity of illness in neurosurgical patients? Clin Neurosurg. 2011;58:155–63.

78. Fixsen DL, Naoom SF, Blase K a, Friedman RM, Wallace F. Implementation research: a synthesis of the literature. Tampa: National Implementation Research Network; 2005. 1-119 p.

79. Francke AL, Smit MC, De Veer AJE, Mistiaen P. Factors influencing the implementation of clinical guidelines for health care professionals: a systematic meta-review. BMC Med Inform Decis Mak. 2008;8:1–11.

80. Zimmerman M, Ruggero CJ, Chelminski I, Young D, Posternak MA, Friedman M, et al. Developing brief scales for use in clinical practice: the reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. J Clin Psychiatry. 2006;67(10):1536–41.

81. Huntley AL, Johnson R, Purdy S, Valderas JM, Salisbury C. Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide. Ann Fam Med. 2012;10(2):134–41.

82. Lezzoni LI, Ash AS, Coffman GA, Moskowitz MA. Predicting in-hospital mortality: a comparison of severity measurement approaches. Med Care. 1992;30(4):347–59.

83. Bhadoria P, Bhagwat AG. Severity scoring systems in paediatric intensive care units. Indian J Anaesth. 2008;52(5):663–75.

84. Lundh A, Forsman M, Serlachius E, Långström N, Lichtenstein P, Landén M. Psychosocial functioning in adolescent patients assessed with Children's Global

Assessment Scale (CGAS) predicts negative outcomes from age 18: a cohort study. Psychiatry Res. 2016;242:295–301.

85. Horn SD, Torres A, Willson D, Dean JM, Gassaway J, Smout R. Development of a pediatric age- and disease-specific severity measure. J Pediatr. 2002;141(4):496–503.

86. Parkerson GR, Harrell FE, Hammond WE. Characteristics of adult primary care patients as predictors of future health services charges. Med Care. 2001;39(11):1170–81.

87. Stein REK, Gortmaker SL, Perrin EC, Perrin JM, Pless IB, Walker DK, et al. Severity of illness: concepts and measurements. Lancet. 1987;1506–9.

88. Vollmer WM. Assessment of asthma control and severity. Ann Allergy, Asthma Immunol. 2004;93(5):409–14.

89. Lezzoni LI, Moskowitz MA. A clinical assessment of MedisGroups. JAMA. 1988;260:3159–63.

90. Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease: a case-mix measurement. JAMA J Am Med Assoc. 1984;251(5):637–44.

91. Nissim N, Boland MR, Tatonetti NP, Elovici Y, Hripcsak G, Shahar Y, et al. Improving condition severity classification with an efficient active learning based framework. J Biomed Inform. 2016;61:44–54.

92. Rich P, Scher RK. Nail Psoriasis Severity Index: a useful tool for evaluation of nail psoriasis. J Am Acad Dermatol. 2003;49(2):206–12.

93. Otley A, Loonen H, Parekh N, Corey M, Sherman PM, Griffiths AM. Assessing activity of pediatric Crohn's disease: which index to use? Gastroenterology. 1999;116(3):527–31.

94. McLellan AT, Kushner H, Metzger D, Peters R, Smith I, Grissom G, et al. The fifth edition of the Addiction Severity Index. J Subst Abuse Treat. 1992;9(3):199–213.

95. Kroenke K, Spitzer RL. The PHQ-9: validity of a brief depression severity measure. J Gen

Intern Med. 2001;16:606–13.

96. Endicott J, Spitzer RL. A diagnostic interview: the schedule for affective disorders and schizophrenia. Achives Gen Psychiatry. 1978;35(7):837–44.

97. Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med. 2001;2(4):297–307.

98. Derogatis LR, Unger R. Symptom checklist-90-revised. Corsini Encycl Psychol. 2010;1–2.

99. Guy W. ECDEU assessment manual for psychopharmacology. Kensington: Department of Health, Education, and Welfare; 1976.

100. Cramer JA. Assessing the severity of seizures and epilepsy: which scales are valid? Curr Opin Neurol. 2001;14(2):225–9.

101. Gardner DG, Cummings LL, Dunham RB, Pierce JL. Single-item versus multiple-item measurement scales: an empirical comparison. Educ Psychol Meas. 1998;58(6):898–915.

102. Eder L, Thavaneswaran A, Chandran V, Cook R, Gladman DD. Factors explaining the discrepancy between physician and patient global assessment of joint and skin disease activity in psoriatic arthritis patients. Arthritis Care Res. 2015;67(2):264–72.

103. Ravelli A, Viola S, Ruperto N, Corsi B, Ballardini G, Martini A. Correlation between conventional disease activity measures in juvenile chronic arthritis. Ann Rheum Dis. 1997;56(3):197–200.

104. Ravelli A, Viola S, Migliavacca D, Pistorio A, Ruperto N, Martini A. Discordance between proxy-reported and observed assessment of functional ability of children with juvenile idiopathic arthritis. Rheumatology. 2001;40(8):914–9.

105. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis.

Am Coll Rheumatol. 1995;38(6):727–35.

106. Rider LG, Werth VP, Huber AM, Alexanderson H, Rao AP, Ruperto N, et al. Measures of adult and juvenile dermatomyositis, polymyositis, and inclusion body myositis. Arthritis Care Res (Hoboken). 2011;63(S11):S118–57.

107. Horn SD, Chachich B, Clopton C. Measuring severity of illness: a reliability study. Med Care. 1983;21(7):705–14.

108. Schumacher DN, Parker B, Kofie V, Munns JM. Severity of Illness Index and the Adverse Patient Occurrence Index: a reliability study and policy implications. Med Care. 1987;25(8):695–704.

109. Horn SD, Horn RA. Reliability and validity of the Severity of Illness Index. Med Care. 1986;24(2):159–78.

110. Horn SD, Horn RA. The Computerized Severity Index: a new tool for case-mix management. J Med Syst. 1986;10(1):73–8.

111. Ryser DK, Egger MJ, Horn SD, Handrahan D, Gandhi P, Bigler ED. Measuring medical complexity during inpatient rehabilitation after traumatic brain injury. Arch Phys Med Rehabil. 2005;86(6):1108–17.

112. Ryser DK, Horn SD, Russo AA, Madrid CA, Barker LH. Rating the severity of acute injury and illness in traumatic brain injury patients using the Computerized Severity Index. Arch Phys Med Rehabil. 1995;76:1042.

113. Parkerson GR, Broadhead WE, Tse C-KJ. The Duke Severity of Illness Checklist (DUSOI) for measurement of severity and comorbidity. J Clin Epidemiol. 1993;46(4):379–93.

114. Parkerson GRJ, Michener JL, Wu LR, Finch JN, Muhlbaier LH, Magruder-Habib K, et al. Associations among family support, family stress, and personal functional health status. J

Clin Epidemiol. 1989;42(3):217–29.

115. Parkerson GR, Hammond WE, Yarnall KSH. Feasibility and potential clinical usefulness of a computerized severity of illness measure. Arch Fam Med. 1994;3:968–73.

116. Parkerson GR. J, Bridges-Webb C, Gervas J, Hofmans-Okkes I, Lamberts H, Froom J, et al. Classification of severity of health problems in family/general practice: an international field trial. Fam Pract. 1996;13(3):303–9.

117. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53:459–68.

118. Cohen J. Statistical power analysis for the behavioral sciences. Second Edi. New York: Lawrence Erlbaum Associates; 1988.

119. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.

120. Shaffer D, Gould MS, Brasic J, Ambrosini P, Fisher P, Bird H, et al. A Children's Global Assessment Scale (CGAS). Arch Gen Psychiatry. 1983;40:1228–31.

121. Butler A, Van Lieshout RJ, Lipman EL, Macmillan HL, Gonzalez A, Gorter JW, et al. Mental disorder in children with physical conditions: a pilot study. BMJ Open. 2018;8(1):9–11.

122. Worthington RL, Whittaker TA. Scale development research: a content analysis and recommendations for best practices. Couns Psychol. 2006;34(6):806–38.

123. Johanson GA, Brooks GP. Initial scale development: sample size for pilot studies. Educ Psychol Meas. 2010;70(3):394–400.

124. McKinley S, Coote K, Stein-Parbury J. Development and testing of a faces scale for the assessment of anxiety in critically ill patients. J Adv Nurs. 2003;41(1):73–9.

125. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171–8.

126. Guyatt G, Rennie D, Meade MO, Cook DJ. Users' guides to the medical literature: a manual for evidence-based clinical practice. Second Edi. New York: McGraw-Hill; 2008.

127. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health. 2008;31:180–91.

128. Russell JA, Weiss A, Mendelsohn GA. Affect Grid: a single-item scale of pleasure and arousal. J Pers Soc Psychol. 1989;57(3):493–502.

129. Cunny KA, Perri MI. Single-item vs multiple-item measures of health-related quality of life. Psychol Rep. 1991;69:127–30.

130. Abdel-Khalek AM. Measuring happiness with a single-item scale. Soc Behav Personal an Int J. 2006;34(2):139–50.

131. DeSalvo KB, Fisher WP, Tran K, Bloser N, Merrill W, Peabody J. Assessing measurement properties of two single-item general health measures. Qual Life Res. 2006;15(2):191–201.

132. Anthoine E, Moret L, Regnault A, Sbille V, Hardouin J-B. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcome measures. Health Qual Life Outcomes. 2014;12:176.

133. Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. Res Nurs Health. 1990;13(4):227–36.

134. Gould D, Kelly D, Goldstone L, Gammon J. Examining the validity of pressure ulcer risk assessment scales: developing and using illustrated patient simulations to collect the data. J Clin Nurs. 2001;10(5):697–706.

135. Paul-Dauphin A, Guillemin F, Virion JM, Briancon S. Bias and precision in visual

analogue scales: a randomized controlled trial. Am J Epidemiol. 1999;150(10):1117–27.

136. Huskisson EC, Jones J, Scott PJ. Application of visual-analogue scales to the measurement of functional capacity. Rheumatol Rehabil. 1976;15(3):185–7.

137. Robitail S, Ravens-Sieberer U, Simeoni MC, Rajmil L, Bruil J, Power M, et al. Testing the structural and cross-cultural validity of the KIDSCREEN-27 quality of life questionnaire. Qual Life Res. 2007;16:1335–45.

138. Ravens-Sieberer U, Auquier P, Erhart M, Gosch A, Rajmil L, Bruil J, et al. The KIDSCREEN-27 quality of life measures for children and adolescents: psychometric results from a cross-cultural survey in 13 European countries. Qual Life Res. 2007;16:1347–56.

139. Qadeer RA, Ferro MA. Child–parent agreement on health-related quality of life in children with newly diagnosed chronic health conditions: a longitudinal study. Int J Adolesc Youth. 2017;23(1):99–108.

140. Tompke BK, Ferro MA. Measurement invariance and informant discrepancies of the KIDSCREEN-27 in children with mental disorder. (Under Review). Appl Res Qual Life. 2019;

141. Sheehan D V, Sheehan KH, Shytle RD, Janavs J, Bannon Y, Rogers JE, et al. Reliability and validity of the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID). J Clin Psychiatry. 2010;71(3):313–26.

142. Duncan L, Georgiades K, Wang L, Van Lieshout RJ, Macmillan HL, Ferro MA, et al. Psychometric evaluation of the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID). Psychol Assess. 2017;1–13.

143. Cook S, Leschied AW, St Pierre J, Stewart SL, den Dunnen W, Johnson AM. BCFPI validation for a high-risk high-needs sample of children and youth admitted to tertiary

care. J Can Acad Child Adolesc Psychiatry. 2013;22:147–52.

144. Boyle MH, Duncan L, Georgiades K, Bennett K, Gonzalez A, Van Lieshout RJ, et al. Classifying child and adolescent psychiatric disorder by problem checklists and standardized interviews. Int J Methods Psychiatr Res. 2017;26:e1544.

145. Puka L. Kendall's Tau. Int Encycl Stat Sci. 2011;713–5.

146. Croux C, Dehon C. Influence functions of the Spearman and Kendall correlation measures. Stat Methods Appl. 2010;19(4):497–515.

147. Atila G, İşçi Ö. A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metod Zv. 2011;8(1):17–37.

148. Agresti A. Analysis of ordinal categorical data. 2nd ed. Hoboken: Wiley; 2010.

149. Ghent AW. Examination of five tau variants suited to ordered contingency tables, from the viewpoint of biological research. Am Midl Nat. 1984;112(2):332.

150. Silverstein J, Klingensmith G, Copeland K, Plotnick L, Kaufman F, Laffel L, et al. Care of children and adolescents with type 1 diabetes: a statement of the American Diabetes Association. Diabetes Care. 2005;28(1):186–212.

151. Hauser WA. Seizure disorders: the changes with age. Epilepsia. 1992;33:6–14.

152. Watanabe K. Recent advances and some problems in the delineation of epileptic syndromes in children. Brain Dev. 1996;18(6):423–37.

153. McNeish D, Stapleton LM. Modeling clustered data with very few clusters. Multivariate Behav Res. 2016;51(4):495–518.

154. Milliken GA, D E J. Analysis of messy data, volume I: designed experiments. 2nd ed. New York: Chapman and Hall/CRC; 2009.

155. Searle SR. Linear models for unbalanced data. New York: John Wiley and Sons Inc.; 1987.

156. Searle SR, Casella G, McCulloch CE. Variance components. New York: John Wiley and Sons Inc.; 1992.

157. Stroup WW. Generalized linear mixed models: modern concepts, methods and applications. New York: Chapman & Hall/CRC; 2012.

158. Clark LA, Watson D. Constructing validity: basic issues in scale development. Psychol Assess. 1995;7(3):309–19.

159. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Control Clin Trials. 1989;10:407–15.

160. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41(5):582–92.

161. Paiva CE, Barroso EM, Carneseca EC, de Pádua Souza C, Dos Santos FT, Mendoza López R V, et al. A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. BMC Med Res Methodol. 2014;14:8.

162. Sun X, Yang Z. Generalized McNemar's test for homogeneity of marginal distributions. Stat Data Anal. 2008;382:1–10.

163. Sim J, Wright CC. The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2017;85(3):257–68.

164. Streiner DL. Research methods in psychiatry: a checklist for evaluating the usefulness of rating scales. Can J Psychiatry. 1993;38(March):140–8.

165. Fleiss J, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33:613–9.

166. Brenner H, Kliebsch U. Dependence of weighted Kappa coefficients on the number of

categories. Epidemiology. 1996;7:199–202.

167. Schuster C. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. Educ Psychol Meas. 2004;64(2):243–53.

168. Vanacore A, Pellegrino MS. RRep: a composite index to assess and test rater precision. Qual Reliab Eng Int. 2018;34(7):1352–62.

169. Hartmann DP. Considerations in the choice of interobserver reliability estimates. J Appl Behav Anal. 2006;10(1):103–16.

170. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.

171. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1(1):30–46.

172. Zou GY. Quantifying responsiveness of quality of life measures without an external criterion. Qual Life Res. 2005;14(6):1545–52.

173. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care. 1990;28(7):632–42.

174. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med. 2003;29(7):1043–51.

175. Lantz B. Equidistance of Likert-type scales and validation of inferential methods using experiments and simulations. Electron J Bus Res Methods. 2013;11(1):16–28.

176. Ferro MA, Avison WR, Karen Campbell M, Speechley KN. The impact of maternal depressive symptoms on health-related quality of life in children with epilepsy: a prospective study of family environment as mediators and moderators. Epilepsia. 2011;52(2):316–25.

177. Ferro MA, Huang W, Smith ML, Speechley KN, Levin SD, Camfield CS, et al. Quality of life in children with new-onset epilepsy: a 2-year prospective cohort study. Neurology. 2012;79(15):1548–55.

178. Sawyer MG, Whaites L, Rey JM, Hazell PL, Graetz BW, Baghurst P. Health-related quality of life of children and adolescents with mental disorders. J Am Acad Child Adolesc Psychiatry. 2002;41(5):530–7.

179. Coghill D, Danckaerts M, Sonuga-Barke E, Sergeant J, Group AEG. Practitioner review: quality of life in child mental health - conceptual challenges and practical choices. J Child Psychol Psychiatry. 2009;50(5):544–61.

180. Tompke BK, Tang J, Oltean II, Buchan MC, Reaume S V., Ferro MA. Measurement invariance of the WHODAS 2.0 across youth with and without physical or mental conditions. Assessment. 2018;(December).

181. Hilliard ME, Lwrence JM, Modi AC, Anderson A, Crume T, Dolan LM, et al. Identification of minimal clinically important difference scores of the PedsQL in children, adolescents, and young adults with type 1 and type 2 diabetes. Diabetes Care. 2013;36(7):1891–7.

182. Bodenheimer T, Wagner EH, Grumbach K. Part 1: improving primary care for patients with chronic illness. JAMA Intern Med. 2002;288(14):1775–9.

183. van Dulmen AM. Children's contributions to pediatric outpatient encounters. Pediatrics. 2004;102(3):563–8.

184. Smith SA, Shah ND, Bryant SC, Christianson TJH, Bjornsen SS, Giesler PD, et al. Chronic care model and shared care in diabetes: randomized trial of an electronic decision support system. Mayo Clin Proc. 2008;83(7):747–57.

185. Cox JL, Zitner D, Courtney KD, MacDonald DL, Paterson G, Cochrane B, et al.

Undocumented patient information: an impediment to quality of care. Am J Med. 2003;114(3):211–6.

186. Gidwani R, Nguyen C, Kofoed A, Carragee C, Rydel T, Nelligan I, et al. Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial. Ann Fam Med. 2017;15(5):427–33.

187. Stetson PD, Morrison FP, Bakken S, Johnson SB. Preliminary development of the physician documentation quality instrument. J Am Med Informatics Assoc. 2008;15(4):534–41.

188. Hunter JE, Schmidt FL. Methods of meta-analysis: correcting error and bias in research findings. Sage; 2004.

189. Barnes SA, Lindborg SR, Seaman JW. Multiple imputation techniques in small sample clinical trials. Stat Med. 2006;25(2):233–45.

190. Scharfman HE, MacLusky NJ. Sex differences in the neurobiology of epilepsy: a preclinical perspective. Vol. 72, Neurobiology of Disease. 2014. p. 180–92.

191. Christensen J, Kjeldsen MJ, Andersen H, Friis ML, Sidenius P. Gender differences in epilepsy. Epilepsia. 2005;46(6):956–60.

192. Branum AM, Simon AE, Lukacs SL. Among children with food allergy, do sociodemographic factors and healthcare use differ by severity? Matern Child Health J. 2012;16(1):S44–50.

193. Parent A-S, Teilmann G, Juul A, Skakkebaek NE, Toppari J, Bourguignon J-P. The timing of normal puberty and the age limits of sexual precocity: variations around the world, secular trends, and changes after migration. Endocr Rev. 2003;24(5):668–93.

194. Wheless JW, Kim HL. Adolescent seizures and epilepsy syndromes. Epilepsia. 2002;43(SUPPL. 3):33–52.

195. Koné Péfoyo AJ, Wodchis WP. Organizational performance impacting patient satisfaction in Ontario hospitals: a multilevel analysis. BMC Res Notes. 2013;6(1).

196. Bodenheimer T, Wagner EH, Grumbach K. Part 2: improving primary care for patients with chronic illness. Jama. 2002;288(15):1909–14.

197. Lohr KN. Applications of health status assessment measures in clinical practice: overview of the Third Conference on Advances in Health Status Assessment. Med Care. 1992;30(5).

198. Popham WJ. Consequential validity: right concern-wrong concept. Educ Meas Issues Pr. 1997;16(2):9–13.

199. Reckase MD. Consequential validity from the test developer's perspective. Educ Meas Issues Pract. 1998;17(2):13–6.

200. Expert Panel Report 3 (EPR-3): guidelines for the diagnosis and management of asthma–summary report 2007. J Allergy Clin Immunol. 2007;120(5):S94–138.

201. Teasdale G, Maas A, Lecky F, Manley G, Stocchetti N, Murray G. The Glasgow Coma Scale at 40 years: standing the test of time. Lancet Neurol. 2014;13(8):844–54.

202. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. Can Med Assoc J. 2010;182(18):E839–42.

203. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8.

Appendix A

Supplementary Psychometric Analysis

(Assuming Ordinal/Nominal Outcomes)

Table A1 – Concurrent Validity Assessed by AUC: Independent Associations

	Baseline					6 Months				
Model Predictor	**Estimate**	**SE**	**p-Value**	**n**	**AUC**	**Estimate**	**SE**	**p-Value**	**n**	**AUC**
D. Symptoms	0.19	0.25	0.456	51	0.59	-0.27	0.26	0.304	51	0.58
D. Complications	0.86	1.00	0.390	50	0.52	0.32	0.36	0.387	51	0.54
D. Prognosis	1.53	0.44	<.001	53	0.90	1.37	0.39	<.001	51	0.88
D. Treatability	2.47	1.12	0.028	52	0.66	**13.95**	**191.40**	**0.942**	**51**	**0.77**
D. Composite	23.07	7.56	0.002	49	0.88	11.20	3.68	0.002	51	0.83
VAS	**41.21**	**97.22**	**0.672**	**52**	**0.99**	2.53	0.79	0.001	51	0.88

Area under the receiver operating characteristic curve (AUC) is reported. AUCs demonstrate strength of association between the GASI (dichotomized as Low/High severity) and individual DUSOI items, the DUSOI composite score, and the VAS. The GASI was dichotomized by aggregating ratings from “Not at all severe” to “A little severe” (Low severity) and “Somewhat severe” to “Extremely severe” (High severity). Bolded statistics come from models that did not converge because of quasi-complete separation of data and should be interpreted with caution.

D.= DUSOI

Table A2 – Concurrent Validity Assessed by AUC: Joint Model (GASI Outcome= Low/High Severity)

	Baseline			6 Months		
Parameter	**Estimate**	**SE**	**p-Value**	**Estimate**	**SE**	**p-Value**
Intercept	-14.64	5.75	0.011	**-28.59**	**384.6**	**0.006**
DUSOI Symptoms	1.15	0.59	0.049	**-0.06**	**0.45**	**0.896**
DUSOI Complications	1.42	1.88	0.452	**0.49**	**0.47**	**0.302**
DUSOI Prognosis	2.40	0.75	0.001	**1.20**	**0.60**	**0.047**
DUSOI Treatability	3.20	1.98	0.106	**12.81**	**192.3**	**0.004**
	Model: n= 49; AUC= 0.95			**Model: n= 51; AUC= 0.93**		

Area under the receiver operating characteristic curve (AUC) is reported. AUCs demonstrate strength of association between the GASI (dichotomized as Low/High severity) and the DUSOI, controlling for effects of each DUSOI item. GASI dichotomization is equivalent to Table 2. Bolded statistics come from models that did not converge and should be interpreted with caution.

Table A3 – Concurrent Validity Assessed by AUC: Joint Model (GASI Outcome= No/Some Severity)

	Baseline			6 Months		
Parameter	**Estimate**	**SE**	**p-Value**	**Estimate**	**SE**	**p-Value**
Intercept	**-29.42**	**362.20**	**0.935**	-9.26	6.52	0.155
DUSOI Symptoms	**1.49**	**0.68**	**0.027**	0.09	0.56	0.878
DUSOI Complications	**11.92**	**343.3**	**0.972**	1.00	1.00	0.315
DUSOI Prognosis	**1.81**	**0.58**	**0.002**	2.94	1.07	0.006
DUSOI Treatability	**12.19**	**181.10**	**0.946**	2.90	3.06	0.343
	Model: n= 49; AUC= 0.93			Model: n= 51; AUC= 0.93		

Area under the receiver operating characteristic curve (AUC) is reported. AUCs demonstrate strength of association between the GASI (dichotomized as No/Some severity) and the DUSOI, controlling for effects of each DUSOI item. The GASI was dichotomized by distinguishing the rating “Not at all severe” (No severity) from ratings ranging from “A little severe” to “Extremely Severe” (Some severity). Bolded statistics come from models that did not converge and should be interpreted with caution.

Table A4 – Sensitivity/Specificity Assessed by AUC: Single Predictor Per Model

Measure Defining Stable vs. Changed Subgroup	GASI Sensitivity/Specificity (Maximum Likelihood Estimates)				
	Estimate	SE	p-Value	n	AUC
DUSOI: Symptoms	0.12	0.20	0.540	47	0.56
DUSOI: Complications	0.04	0.26	0.878	46	0.55
DUSOI: Prognosis	0.27	0.21	0.205	49	0.67
DUSOI: Treatability	-0.71	0.27	0.008	48	0.72
DUSOI: Composite	-0.29	0.21	0.179	45	0.62
VAS	-0.47	0.22	0.031	48	0.61

Area under the receiver operating characteristic curve (AUC) is reported. AUCs demonstrate ability of the GASI to discriminate between Stable and Changed subgroups. Stable/Changed is defined by individual DUSOI items, the DUSOI composite score, and the VAS.

Appendix B

Exploratory Data Analysis

(Assuming Ordinal/Nominal Outcomes)

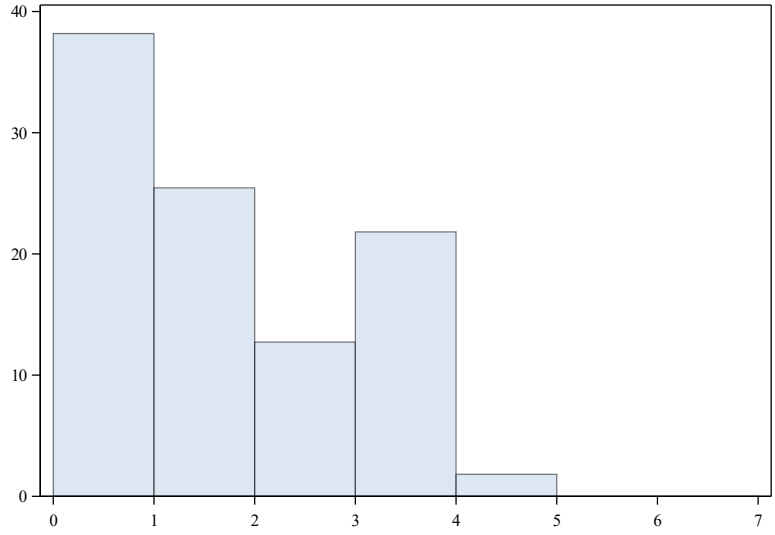

Figure B1. Distribution of GASI ratings at baseline where x-axis represents all possible scale ratings and y-axis represents percent of children in the sample.

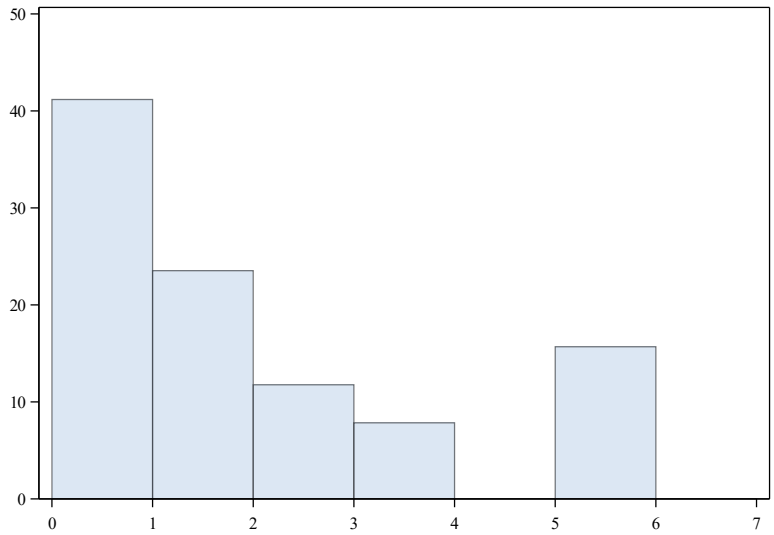

Figure B2. Distribution of GASI ratings at six months where x-axis represents all possible scale ratings and y-axis represents percent of children in the sample.

Table B1 – Exploring Clinician Characteristics as Potential Confounders Using Fisher's Exact Test

	Baseline			6 Months		
Frequency Table	**Table Probability (P)**	**p-Value**	**n**	**Table Probability (P)**	**p-Value**	**n**
Rater Confidence (Low/High) x ClinType (Imm, End, Neur, Rheu)	0.003	0.095	53	<.001	0.004	50

The results of Fisher's Exact test are reported for the frequency table with the row/column variables rater confidence (dichotomized as Low/High) and type of clinical specialist measuring severity (four types).

Rater confidence= the confidence that clinicians reported having in their severity assessment upon completing the clinician form (i.e., one confidence rating pertains to all three severity scales)

ClinType= type of clinical specialist; Imm= immunologist; End= endocrinologist; Neur= neurologist; Rheu= rheumatologist

Table B2 – Exploring Validity with Clinician Characteristics as Covariate Using the AUC

	Baseline			6 Months		
Parameter	**Estimate**	**SE**	**p-Value**	**Estimate**	**SE**	**p-Value**
Intercept	-9.40	3.09	0.002	-5.55	1.80	0.002
DUSOI Composite	25.40	8.62	0.003	12.19	4.03	0.003
ClinType	-0.32	0.36	0.371	0.31	0.28	0.278
	Model: n= 48; AUC= 0.90			Model: n= 50; AUC= 0.83		

Area under the receiver operating characteristic curve (AUC) is reported. AUCs demonstrate strength of association between the GASI (dependent variable; dichotomized as Low/High severity) and the DUSOI Composite score (independent variable) where the clinician characteristic "ClinType" is a covariate. GASI dichotomization is equivalent to Table 2.

ClinType= Type of clinical specialist (Immunologist, Endocrinologist, Neurologist, or Rheumatologist)

Table B3 – Exploring Validity with Clinician Characteristics as Covariate Using the AUC

	Baseline			6 Months		
Parameter	**Estimate**	**SE**	**p-Value**	**Estimate**	**SE**	**p-Value**
Intercept	-8.69	3.14	0.006	-2.94	2.28	0.197
DUSOI Composite	23.29	7.70	0.003	9.72	3.75	0.010
Rater Confidence	-0.29	0.94	0.755	-0.64	0.80	0.426
	Model: n= 48; AUC= 0.89			Model: n= 50; AUC= 0.82		

Area under the receiver operating characteristic curve (AUC) is reported. AUCs demonstrate strength of association between the GASI (dependent variable; dichotomized as Low/High severity) and the DUSOI Composite score (independent variable) where the clinician characteristic "Rater Confidence" is a covariate. GASI dichotomization is equivalent to Table 2.

Table B4 – Exploring Relationships Between Sample Characteristics and GASI Ratings

Test	**Baseline**		**6 Months**	
Fisher's Exact Test (GASI x Diagnosis)	Table Probability: <.001	p-Value: 0.005	Table Probability: <.001	p-Value: <.001
Fisher's Exact Test (GASI x Sex)	Table Probability: 0.004	p-Value: 0.495	Table Probability: <.001	p-Value: 0.202
Kendall's τ_c Correlation (GASI x Age)	Coefficient (τ_c): 0.09	p-Value: 0.445	Coefficient (τ_c): -0.08	p-Value: 0.535

Results of Fisher's Exact test and Kendall's *Tau-c* correlation are reported.

x= "relationship with"

Appendix C

Alternative Psychometric Analysis

(Assuming Continuous Outcomes)

Table C1 – Concurrent Validity Assessed by Spearman-Rank Correlation

	Baseline			6 Months		
Severity Measure	**ρ (95% CI)**	**p-Value**	**n**	**ρ (95% CI)**	**p-Value**	**n**
DUSOI: Symptoms	0.28 (0.00, 0.51)	0.047	51	-0.09 (-0.35, 0.20)	0.550	51
DUSOI: Complications	0.15 (-0.13, 0.41)	0.295	50	0.11 (-0.17, 0.37)	0.446	51
DUSOI: Prognosis	0.74 (0.58, 0.84)	<.001	53	0.82 (0.69, 0.89)	<.001	51
DUSOI: Treatability	0.41 (0.15, 0.61)	0.002	52	0.63 (0.42, 0.77)	<.001	51
DUSOI: Composite	0.76 (0.61, 0.86)	<.001	49	0.72 (0.55, 0.83)	<.001	51
VAS	0.87 (0.78, 0.92)	<.001	52	0.86 (0.76, 0.91)	<.001	51

Spearman-Rank correlation coefficients (rho) are reported with 95% confidence intervals at baseline and 6 months for correlation of the GASI with individual DUSOI items, the DUSOI composite score, and the VAS.

Table C2 – Concurrent Validity Assessed by Multiple Linear Regression: Joint Model

	Baseline				6 Months			
Parameter	**Estimate**	**SE**	**p-Value**	**95% CI**	**Estimate**	**SE**	**p-Value**	**95% CI**
Intercept	-0.67	0.45	0.14	(-1.58, 0.24)	-1.37	0.56	0.02	(-2.49, -0.24)
D. Symptoms	0.24	0.11	0.03	(0.02, 0.46)	-0.06	0.11	0.57	(-0.28, 0.16)
D. Complications	0.53	0.40	0.19	(-.28, 1.33)	0.48	0.15	<.01	(0.18, 0.78)
D. Prognosis	0.64	0.10	<.01	(0.44, 0.84)	0.67	0.10	<.01	(0.47, 0.86)
D. Treatability	0.20	0.23	0.38	(-.26, 0.67)	0.86	0.26	<.01	(0.33, 1.39)
	Model: n= 49; Adjusted R^2= 0.53				Model: n= 51; Adjusted R^2= 0.78			

Multiple linear regression coefficients are reported where all DUSOI items are included in the model. R^2 indicates amount of variation in the GASI that can be explained by variation in the DUSOI.

D.= DUSOI

SE= Standard error

Table C3 – Concurrent Validity Assessed by Multiple Linear Regression: Model Includes DUSOI Composite Score

	Baseline				6 Months			
Parameter	**Estimate**	**SE**	**p-Value**	**95% CI**	**Estimate**	**SE**	**p-Value**	**95% CI**
Intercept	-1.02	0.37	<.01	(-1.76, -0.29)	-1.15	0.56	0.04	(-2.27, -0.03)
DUSOI Composite	6.79	1.06	<.01	(4.67, 8.91)	7.93	1.56	<.01	(4.80, 11.06)
	Model: n= 49; R^2= 0.47				Model: n= 51; R^2= 0.35			

Multiple linear regression coefficients are reported. R^2 indicates amount of variation in the GASI that can be explained by variation in the DUSOI Composite score.

SE= Standard error

Table C4 – Concurrent Validity Assessed by Multiple Linear Regression: Model Includes VAS

	Baseline				6 Months			
Parameter	**Estimate**	**SE**	**p-Value**	**95% CI**	**Estimate**	**SE**	**p-Value**	**95% CI**
Intercept	0.04	0.10	0.72	(-.17, 0.24)	-0.30	0.13	0.03	(-0.57, -0.03)
VAS	0.95	0.06	<.01	(0.84, 1.07)	1.14	0.06	<.01	(1.01, 1.26)
	Model: n= 52; R^2= 0.84				Model: n= 51; R^2= 0.87			

Multiple linear regression coefficients are reported. R^2 indicates amount of variation in the GASI that can be explained by variation in the VAS.

SE= Standard error

Table C5 – Discriminant Validity Assessed by Spearman-Rank Correlation

	Baseline			6 Months		
KIDSCREEN-27 Domain	**ρ (95% CI)**	**p-Value**	**n**	**ρ (95% CI)**	**p-Value**	**n**
Physical Well-being	-0.02 (-0.30, 0.26)	0.878	48	0.11 (-0.20, 0.39)	0.486	44
Psychological Well-being	-0.09 (-0.36, 0.20)	0.543	49	-0.07 (-0.36, 0.23)	0.649	43
Parents and Autonomy	-0.003 (-0.29, 0.28)	0.983	48	0.07 (-0.24, 0.36)	0.667	44
Social Support and Peers	-0.12 (-0.39, 0.16)	0.396	49	-0.10 (-0.39, 0.20)	0.507	44
School Environment	-0.09 (-0.37, 0.20)	0.538	46	-0.13 (-0.42, 0.17)	0.392	43

Spearman-Rank correlation coefficients (rho) are reported with 95% confidence intervals at baseline and 6 months for correlation of the GASI with individual KIDSCREEN-27 domains.

Table C6 – Discriminative Validity Assessed by Mann-Whitney U Test (Wilcoxon Rank Sum Test)

	Baseline			6 Months		
Severity Subgroups	**Test Statistic**	**p-Value**	**n**	**Test Statistic**	**p-Value**	**n**
Multimorbidity vs. No Multimorbidity	563	0.151	55	557	0.833	51

Test statistics for the Mann-Whitney U Test are reported.

Table C7 – Test-retest Reliability Assessed by Wilcoxon Signed-Rank Test

Measure Defining Stable Subgroup	**Test Statistic (S)**	**p-Value**	**n**
DUSOI: Symptoms	-28	0.046	27
DUSOI: Complications	-29.5	0.204	38
DUSOI: Prognosis	-25.5	0.052	30
DUSOI: Treatability	16.5	0.400	34
DUSOI: Composite	-2.5	0.905	29
VAS	10	0.234	28

Test statistics for the Wilcoxon Signed-Rank Test are reported.

Table C8 – Test-retest Reliability Assessed by Intraclass Correlation Coefficient

Measure Defining Stable Subgroup	ICC (95% CI)	n
DUSOI: Symptoms	0.60 (0.52, 0.68)	27
DUSOI: Complications	0.53 (0.45, 0.69)	38
DUSOI: Prognosis	0.61 (0.50, 0.74)	30
DUSOI: Treatability	0.44 (0.21, 0.51)	34
DUSOI: Composite	0.52 (0.34, 0.76)	29
VAS	0.87 (0.65, 0.95)	28

Bootstrapped intraclass correlation coefficients [ICC (A, 1)] with 95% confidence intervals (CI) are reported. Coefficients were computed using GASI ratings from children in the Stable subgroup. Stable is defined by individual DUSOI items, the DUSOI composite score, and the VAS. Bootstrapped ICCs were calculated using a 2-way mixed effects model requiring absolute agreement because clinicians rated condition severity in the same patients[170,203] at baseline and at 6 months. It should be noted that this ICC is computationally equivalent to the random effects ICC [ICC (2,1)],[170] but is not termed “random” in this context because inter-rater reliability is not being formally tested (i.e., children were not randomized to different clinicians).