Neuropsychiatric Disease and Treatment

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

The ESSENCE-Questionnaire in Medical Records Screening for Neurodevelopmental Symptoms/Problems: Utility and Clinical Validity

Valdemar Landgren[1,2], Zohar Raanan Soltis[1], Emma Svensson[3], Michail Theodosiou[2,4], Magnus Landgren[2], Rajna Knez[2,5]

[1]Department of Psychiatry, Skaraborg Hospital, Skövde, Sweden; [2]Gillberg Neuropsychiatry Centre, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [3]Department of Child Psychiatry, Skaraborg Hospital, Skövde, Sweden; [4]School Health Services, City of Gothenburg, Gothenburg, Sweden; [5]Department of Pediatrics, Skaraborg Hospital, Skövde, Sweden

Correspondence: Valdemar Landgren, Department of Psychiatry, Skaraborg Hospital Skövde, Lövängsvägen, Skövde, 541 45, Sweden, Tel +46702960450, Email valdemar.landgren@gu.se

Purpose: Determine the prevalence of symptoms of neurodevelopmental problems (NDPs) with a semi-structured review of fourth grade students' medical records, its interrater agreement and validity as compared with clinical assessment.

Methods: A school-based sample of 11-year-old children provided child health care (CHC) records and school health care (SHC) records. A pediatric neurologist, child psychiatrist and an adult psychiatrist scored the records, with the "Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations-Questionnaire" (ESSENCE-Q, 12 items scored 0–2, summary score range 0–24). Agreement was measured with model-based kappa and intraclass correlation coefficient (ICC). Ratings were validated against a multidisciplinary assessment involving a physician, psychologist, teacher- and parental behavioral rating scales rendering a clinical global impression severity rating (CGI-S, range 1–7) of NDPs.

Results: Out of 223 participants, medical charts were available from 201, of whom 169 were rated by all three raters. Kappa agreement was moderate/strong (~0.8) for 7 of the 12 questionnaire items. Measured with the ICC, concordance in the summary score was good for agreement (~0.8) and excellent (~0.9) for consistency. Test–retest reliability was excellent (ICC = ~0.9). Area under the curve for the ESSENCE-Q in predicting clinical-level problems (CGI ≥4) was ~80% for all three raters, albeit with differing optimal cutoffs.

Conclusion: Using the ESSENCE-Q as a template, NDPs appear to be common in medical records, are identified reliably, and predict clinical-level concern. Medical records screening may facilitate a structured review of medical records in work-ups or be applied in conjunction with other screening measures for neurodevelopmental disorders. However, differences in calibration currently preclude defining a universal cutoff for using the ESSENCE-Q for medical records screening.

Keywords: neurodevelopmental disorders, medical records, abnormal development, child health, ESSENCE

Plain Language Summary

- Although medical records are often reviewed for early symptoms of neurodevelopmental disorders, the quality of information and usefulness of records review for this purpose has not been studied.
- When physicians practicing in three different specialties (child neurology, child psychiatry, adult psychiatry) screened children's medical records with the semi-structured ESSENCE-Q, signs of possible neurodevelopmental problems (NDPs) were common.
- Agreement in the ESSENCE-Q between three raters varied somewhat between different developmental areas but was good to excellent in the total score of the questionnaire.
- Ratings in the ESSENCE-Q could detect NDPs warranting clinical attention and were correlated with the severity of the NDPs. This suggests that rating medical records with the ESSENCE-Q may serve as a screening method and provide reliable supportive findings in the clinical work-up for neurodevelopmental disorders.

Received: 21 March 2022
Accepted: 20 September 2022
Published: 3 November 2022

Introduction

Spitzer proposed the gold standard psychiatric diagnostic process through the acronym LEAD; emphasizing the importance of a longitudinal perspective, expert knowledge, and incorporation of all available data.[1] Although longitudinal assessment of all data in the work-up for neurodevelopmental disorders (NDDs) includes medical records, the prevalence of symptoms of NDDs in medical records has not been reported in the literature. Additionally, the impact of specialty training on the clinician reviewing the records in the context of NDDs is largely unknown.

ESSENCE, the acronym for Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations, is a concept highlighting that children with NDDs generally present to various health care professionals at an early age with nonspecific symptoms.[2] It further emphasizes that different neurodevelopmental problems (NDPs) co-occur, and the importance of considering severity as well as the total amount of functional deficits. The co-aggregation of symptoms of NDDs has been corroborated in a host of family-, twin and genetic studies.[2–5] The ESSENCE-concept emphasizes the importance of a careful review of perinatal and developmental history, hereditary factors, behavioral symptoms and co-occurring chronic medical conditions such as diabetes, obesity and epilepsy. Consequently, developmental deviations or prodromal signs, albeit unspecific, could be expected in medical records of children with NDDs.

Although child health care screens children for developmental delay and deviations, only a minority of children diagnosed with NDDs are detected before school-entry. For example, a register study found that the mean age of childhood attention-deficit/hyperactivity disorder (ADHD) diagnosis in Sweden was 12 years.[6] A Danish study found that the incidence of ADHD and Asperger syndrome peaked at age 17 and 16 respectively for girls.[7] Thus, some patients present for a clinical work-up of NDDs as adolescents or even adults, past the time period during which symptoms are postulated to have first emerged, or when no reliable collateral information can be obtained.[8–11] In the pediatric setting, medical records could therefore provide a second wave of screening, and in the adult setting it could prove an informational source unaffected by recall bias lending independent support for early onset of symptoms.

Medical records review in clinical research is used to ascertain exposures (eg, diagnosis or symptoms) or outcomes (eg, death, hospitalization and complications). The method is constrained by two major limitations. Firstly, medical records may be incomplete and of low quality. Time constraints and work overload are frequent situations faced by health staff performing tasks involving data management, that may affect data quality and influence the diagnostic process and treatment decision-making.[12] Lack of information may result in misclassification and potential bias, and is a cause of disagreement between research and clinical diagnoses (ie, false-negative cases lower agreement between research diagnoses and clinical diagnoses).[13–15] Secondly, inter-rater variability among physician reviewers may vary widely in the command of medical records, due to differences in competence or lack of clear definitions of exposures. Although a host of screening tools for detection of NDDs exists, there is a lack of tools developed for the assessment of medical records.[16,17]

Aim

Our aim was to 1) describe the prevalence of ESSENCE and/or ESSENCE-related problems in medical records, 2) determine the agreement between raters from different medical specialties, 3) explore areas of disagreement and 4) ascertain the predictive validity for impairing NDD symptoms and optimal cutoff level of the instrument in a non-clinical sample of 11-year-old children.

Methods

Setting and Participants

The study was performed in regular (public) schools in western Sweden between 2018 and 2020 to examine the association of pupil's health with school performance. The validity of parent-rated ESSENCE-Q in this cohort has been reported previously, where a more detailed methodological description of the clinical assessments can be found.[18] Here, we report on a sub-study examining the usefulness of using the ESSENCE-Q as a template for experts/clinicians in screening medical records for symptoms of NDDs. A flow diagram of participants and measures are provided in Figure 1.

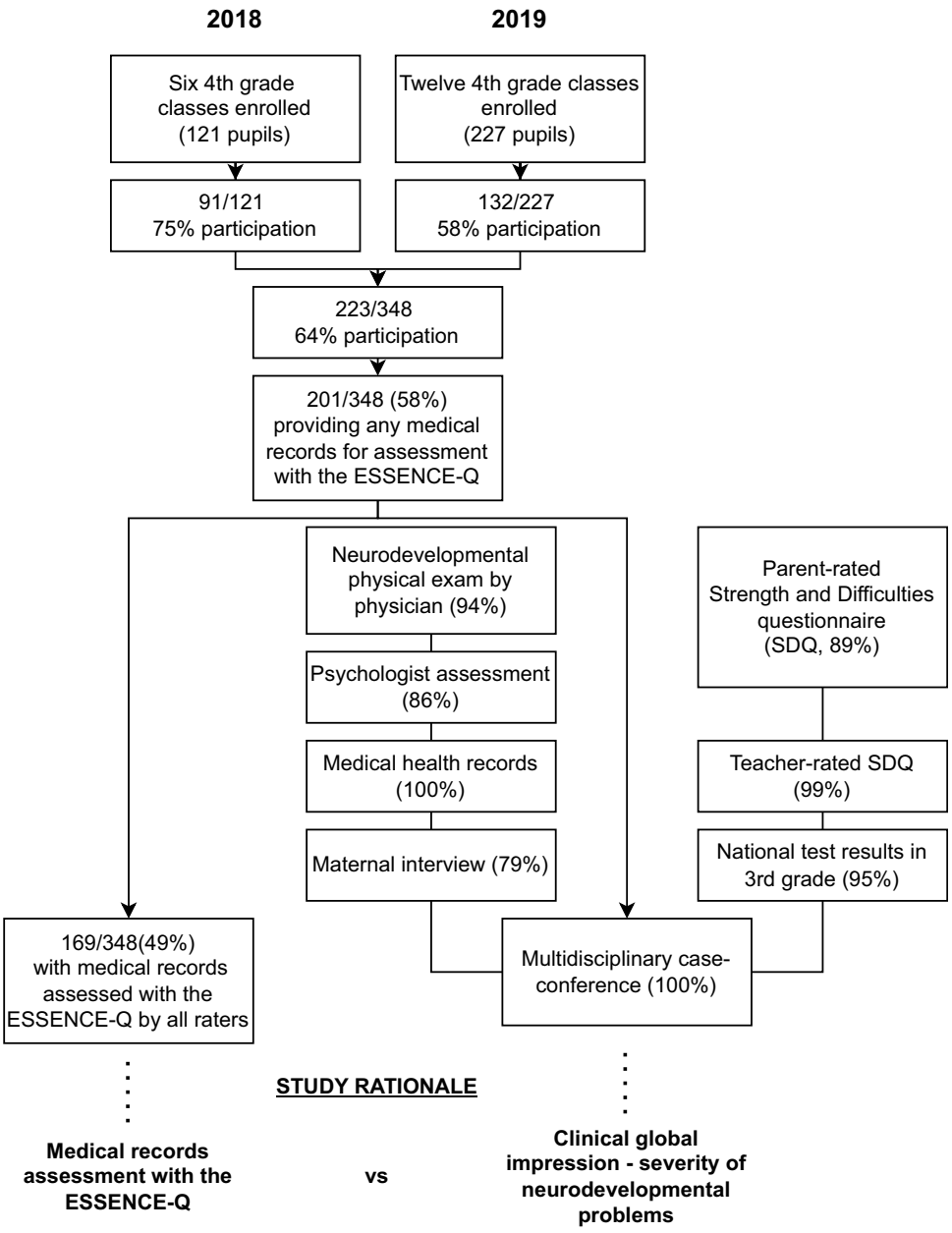

Figure 1 Overview of participants and measures used in the study.

In short, a convenience sample of six schools based on the investigators ability to engage school principals was selected, and parents of all pupils attending fourth grade at the school were invited to participate. On each site, one or two academic years of pupils were recruited. None of the schools had pupils attending special education. Participation constituted an add-on to the existing routine health check-up in fourth grade. The add-on consisted of clinical assessments by a physician (standardized procedure assessing motor performance, interaction, dysmorphology and brief history-taking for 20 minutes), neuropsychologist (Leiter-III non-verbal IQ test, 60 minutes), behavioral ratings from caregivers and teachers (Strength and Difficulties Questionnaire, SDQ), maternal interview, review of academic achievements and medical records.[19,20] A clinical synopsis of all available information was performed in a case-conference attended by a pediatric neurologist, psychologist and a psychiatrist. Based on the synopsis, the severity of NDPs for each participant was rated with the clinical global impression – severity instrument (CGI-S, range 1–7).[21] All case-conferences and ESSENCE-Q ratings were made independently, meaning that no raters were present in the conferences, and no

ESSENCE-Q ratings were available at the conference. All caregivers and children signed informed consent. The study was approved by the ethical review board at Gothenburg university (No. 852-17) and conducted in accordance with the Declaration of Helsinki.

Data Sources

In Sweden, the Child Health Care (CHC) and School Health Care (SHC) have the responsibility to monitor children's health and development regularly. The medical records are therefore a tool to facilitate the documentation of gathered information, for decision-making, support and for follow-up. The national guidelines for child health surveillance in Sweden are followed closely by regional CHCs.[22,23] The program includes supporting parents, providing information and education concerning childcare, health promotion, health check-ups with developmental screenings, anthropometric measurements, control of sight and hearing, as well as a program for immunization.[24] CHCs are either integrated into primary care centers employing public health nurses and general practitioners (75%) or into special units with paediatric nurses and a greater involvement by paediatricians (25%).[23] Although different professionals such as nurses, physicians and psychologists work in the CHCs, the key person is the child health nurse. He/she is a registered nurse with at least 1 year of special training in paediatric development and healthcare, or in public healthcare.[25] The child health program encourages frequent contact with the nurse during the infant's first months, and this contact includes scheduled and unscheduled visits to the clinic, domiciliary visits, telephone consultations, and parental education classes. In addition, the physician sees all children for health check-ups and developmental screenings at certain "key ages".[24] Developmental surveillance is supported at every health supervision visit, as is the administration of standardized screening tests at the 9-, 18-, and 30-month visits. Developmental concerns elicited on surveillance at any visit are to be followed by standardized developmental screening testing or direct referral to intervention and specialty medical care. Special attention to surveillance is recommended at the 5- to 6-year well-child visit, prior to entry into elementary education, with screening completed if there are any concerns.[22]

Measures and Procedure

Exposures

The ESSENCE-questionnaire (ESSENCE-Q) is a tool based on the ESSENCE concept. It is a screening questionnaire developed to speed up the detection of children in need of work-up for an NDD of any sort.[26,27] It has been applied in two ways: to be self-administered by parents and/or caregivers, or expert-administered as a brief interview by a clinician/trained nurse. It is intended for use in both clinical practice and population research. It covers 12 developmental areas (1. General development, 2. Motor development, 3. Sensory reactions, 4. Communication, 5. Activity, 6. Attention, 7. Social interaction, 8. Behaviour, 9. Mood, 10. Sleep, 11. Feeding, 12. Funny spells) scored as No problem=0, Maybe/a little=1, Yes=2. It has been validated as a parental-questionnaire for preschool children undergoing public health check-ups, and children referred for neurodevelopmental assessment.[28,29] The instrument requires no specific training and is intended for screening purpose. It does not render a diagnosis, but may highlight the need for further investigations of functional areas negatively affected in NDDs.

In this study, we examined the utility of a novel method to administer the ESSENCE-Q. We used it as a template for reviewing medical records for possible ESSENCE problems indicative of NDDs. It was completed independently by three raters representing different specialties: Pediatric Neurology, Child Psychiatry and Adult Psychiatry. None of the raters had previous contact with, or knowledge about diagnostic status of the participants. For the purpose of comparing a possible impact of specialty training received by the raters and not the influence from familiarity with NDDs and the ESSENCE concept, they did not receive any specific training in the concept of ESSENCE or completing the ESSENCE-Q in the frame of the study.

Outcomes

Because the study aimed to evaluate a screening measure, we reasoned that the highest validity (given reality of resources at hand) would only be attained if the whole group was assessed (symptomatic and non-symptomatic children alike). For this reason, making full diagnostic assessments of each participant was unfeasible, as it is a very resource-intensive effort in a research study, and therefore done very rarely. This choice came at the cost of diagnostic precision, as no interview

formally assessing diagnostic criteria could be performed. At the same time, we argued that the wide range of informants and sources collected would allow for a credible estimation of the clinical gestalt, with regard to broad symptom areas and degree of impairment. As described previously, we therefore rated each participant with the clinical global impression – severity instrument (CGI-S).[18,21] The CGI-S reflects the clinician's impression of degree of impairment with reference to the distribution of the specific condition under study, based on all available information about the participant. In a clinical setting based on our experience, participants with CGI 1–3 would not be referred for work-up in regular health care (child psychiatry), whereas participants with CGI 4–7 would likely have symptoms/impairments warranting clinical attention and likely diagnosis (For a detailed description and examples of the CGI-S rating, see Landgren et al).

Statistics

Interrater Agreement

Since the kappa statistic was introduced by Cohen, multiple measures of agreement have been developed including Fleiss' kappa, intraclass correlation coefficient (ICC), and model-based kappa.[30] We chose a model-based approach developed by Nelson et al,[31,32] because when applicable it was the most flexible and methodologically robust approach. Whereas the original kappa statistic handles binary data, the model-based kappa accommodates ordinal data, allows for multiple raters, incomplete cases, as well as unbalanced data, and is unaffected by underlying prevalence rates of the condition at study (Cohen's kappa is influenced by the prevalence rate of the classifications).[32,33]

The specific measures used were model-based kappa (range 0–1) for binary classifications, and model-based association (range 0–1), which is the equivalent of kappa for ordinal (>2 levels) classifications. By taking the distance between discordant ratings into account, the association measure "credits" similarity across ratings (A discordant rating of 4 and 5 gets more "credit" than a rating of 3 and 5), typically making the association larger than the absolute agreement, which only "credits" perfect concordance.[30]

We tested the model-based kappa association for the 12 items (range 0–2). We also made three dichotomized analyses with the model-based kappa for binary classifications of ESSENCE-Q comprising none vs ≥ one "maybe", none vs ≥ one rating "Yes", and none vs ≥ one either "yes" or "maybe".

Agreement between all raters in the summary score of the ESSENCE-Q (range 0–24) was measured with the ICC. The ICC measures the strength of inter-rater agreement with continuous or ordinal rating scales, assuming values ranging from 0 to 1, with 1 indicating perfect correlation and 0 indicating no correlation. We analyzed a two-way random-effects model as described by Shrout and Fleiss[34] where both subjects and raters are handled as random effects.[35] We did analyses tailored both for agreement (ICC2,k) and consistency (ICC3,k), separately. The ICC for absolute agreement captures the degree to which raters assign the same score, whereas ICC for consistency captures whether a systematic error may account for low agreement, and if so renders the ICC for consistency higher than that of agreement. Lastly, we tested the test–retest reliability of the summary score rated 1 year apart for one rater with a two-way mixed effects model of agreement (ICC 3,1).[35] The Kappa coefficient was interpreted according to McHugh[36] and ICC according to Koo and Li (<0.5 = poor, 0.5–0.75 = moderate, 0.75–0.9 = good, >0.90 = excellent).[35] All interrater analyses were performed in R version 3.6.3,[37] using the psych[38] and modelkappa[39] package.

Outcome

The predictive validity of the ESSENCE-Q was tested against a binary definition where a CGI-S of 1–3 was negative, and CGI-S 4–7 was positive for clinically relevant NDPs. The AUC of the ESSENCE-Q as rated by each rater was reported, along with receiver operating characteristic (ROC) curves. We reported classification tables with each diagnostic cutoff, including sensitivity, specificity, accuracy (proportion accurately classified, [true positive + true negative/all positive and negative]) negative predictive value (NPV=true negative/all negative) positive predictive value (PPV=true positive/all positive), Youden index (sensitivity + specificity −1), positive (LR+) and negative likelihood ratio (LR-) (LR+=true positive/false positive, LR-=true negative/false negative). Analyses were performed with the pROC package.[40]

Results

Out of 348 eligible children, 223 (63%) participated in the study, and the school nurses provided medical records for 201 participants (Figure 1). There were no data available for dropout analysis. However, because non-participation varied considerably across academic years and classes, it was considered reflective of the school principals' degree of engagement with the project, rather than specific parental or child characteristics. Out of the 201 participants providing medical records, 198 provided CHC medical records, 198 provided SHC medical records and 196 provided both. Because raters received the medical records sequentially during the study period, some cases were not available to all raters. All three raters reviewed 173 CHC medical records, 192 SHC medical records and of those 169 provided medical records from both CHC and SHC.

ESSENCE-Q Rated Records

Distributions of ESSENCE-Q summary scores by rater are described in Figure 2 and Table 1. Distributions were skewed, with the majority of participants assessed with no or very little concern in CHC medical records by all raters, but for a substantial proportion (22–42%) "some" concern was noted (Figure 2, Table 1). The proportion of participants of whom any concern was noted increased greatly in ratings of the SHC medical records (44–93%) and were the largest when reviewing both medical records sources combined (46–95%).

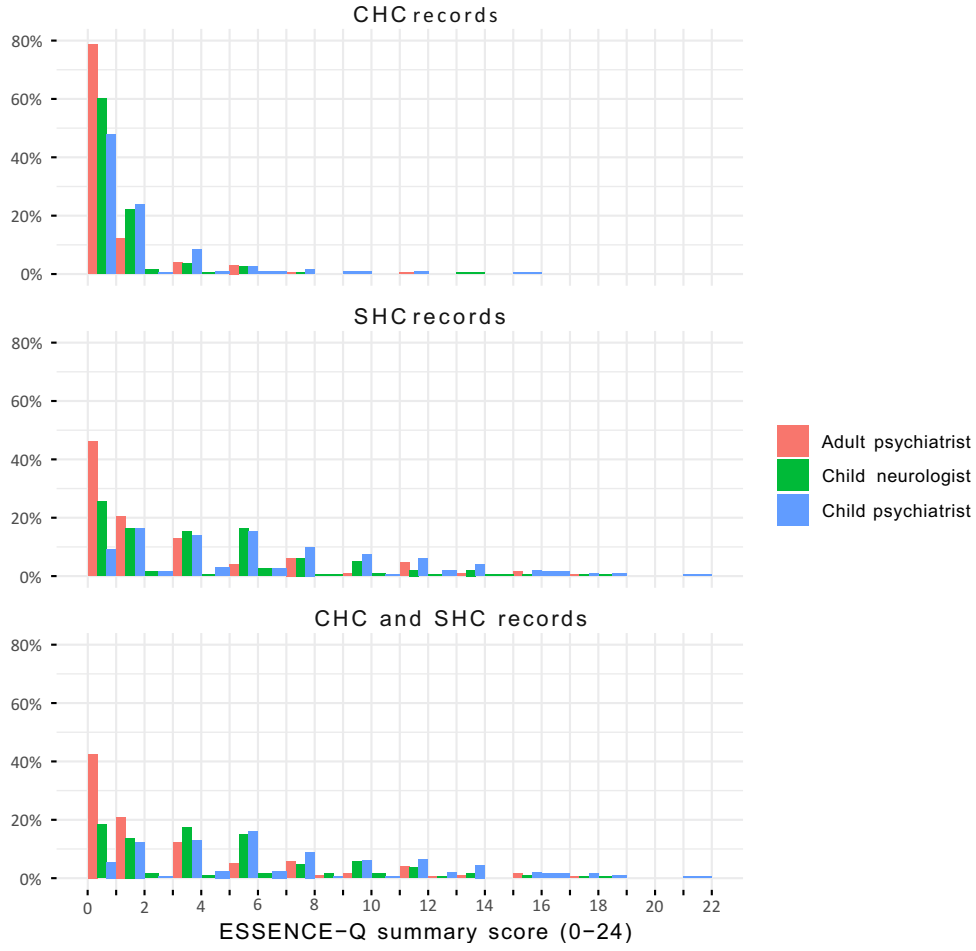

Figure 2 Distributions of ESSENCE-Q summary scores by rater and record source.
Notes: Percentage of ESSENCE-Q summary scores (range 0–24) by rater and record source.
Abbreviations: ESSENCE-Q, Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations-Questionnaire; CHC, child health care; SHC, school health care.

Table 1 Distributions of ESSENCE-Q Summary Scores by Rater and Record Source

ESSENCE-Q Summary Score (Range 0–24)	Child Psychiatrist	Child Neurologist	Adult Psychiatrist
Child health care records			
Mean (SD)	1.7 (2.6)	1.0 (1.8)	0.7 (1.7)
Distribution, n (%)	n=179	n=178	n=193
0	96 (54)	117 (66)	153 (79)
1–2	48 (27)	43 (24)	24 (12)
3–6	25 (14)	16 (9)	14 (7)
>6	10 (6)	2 (1)	2 (1)
School health care records			
Mean (SD)	6.5 (4.8)	4.4 (4.1)	2.8 (4.0)
Distribution, n (%)	n=197	n=192	n=193
0	18 (9)	50 (26)	90 (47)
1–2	34 (17)	33 (17)	41 (21)
3–6	68 (35)	66 (34)	33 (17)
>6	77 (39)	43 (22)	29 (15)
Child health and school service records			
Mean (SD)	7.1 (4.8)	4.9 (4.2)	3.0 (4.1)
Distribution, n (%)	n=176	n=175	n=190
0	11 (6)	36 (21)	83 (44)
1–2	25 (14)	27 (15)	42 (22)
3–6	64 (36)	68 (39)	34 (18)
>6	76 (43)	44 (25)	31 (16)

Notes: Distributions of ESSENCE-Q summary scores (range 0–24) by rater and record source. Three raters, child psychiatrist, child neurologist and adult psychiatrist, scored medical records for signs of possible neurodevelopmental symptoms.

Abbreviation: ESSENCE-Q, Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations-Questionnaire.

As displayed in Figure 3A–C, proportions scored as abnormal varied considerably between items and raters. In CHC records, communication (item 4) was the item most frequently eliciting an abnormal response. In review of both SHC and CHC + SHC medical records combined, there was a large increase in abnormal responses in general, with feeding (item 11), sleeping (item 10), mood (item 9), attention (item 6) and sensory reactions (item 3) most commonly scored as abnormal. The adult psychiatrist had the least proportion of ratings scored as abnormal, whereas the child psychiatrist had the highest. The child neurologist had the greatest tendency to rate items as "maybe", whereas the adult psychiatrist tended to rate items as either "No" or "Yes".

Item-Specific Agreement

Figure 4A–C displays model-based kappa coefficients with 95% confidence interval (CI) of agreement between all three raters. For the CHC medical records, agreement was moderate/strong (~0.8) for most items, except for sensory reactions, for which there was no agreement (~0.1). For the SHC medical records and CHC+SHC medical records combined agreement remained moderate to strong for most items, but decreased to minimal/weak (~0.3) for the general development, mood, sleep and feeding items and sensory reactions showed no agreement. There was no agreement for at least one rating of "maybe" (~0.1), whereas at least one rating "yes" and one rating either "maybe" or "yes" showed weak agreement (~0.5) for CHC records, and minimal agreement (~0.3) for SHC records and CHC+SHC records combined.

ESSENCE-Q Summary Score Interrater Agreement

Measured with the ICC, concordance in the summary score was good for agreement (~0.8) and excellent (~0.9) for consistency (Table 2), a difference indicating minor systematic measurement error between raters (Figure S1A–C).

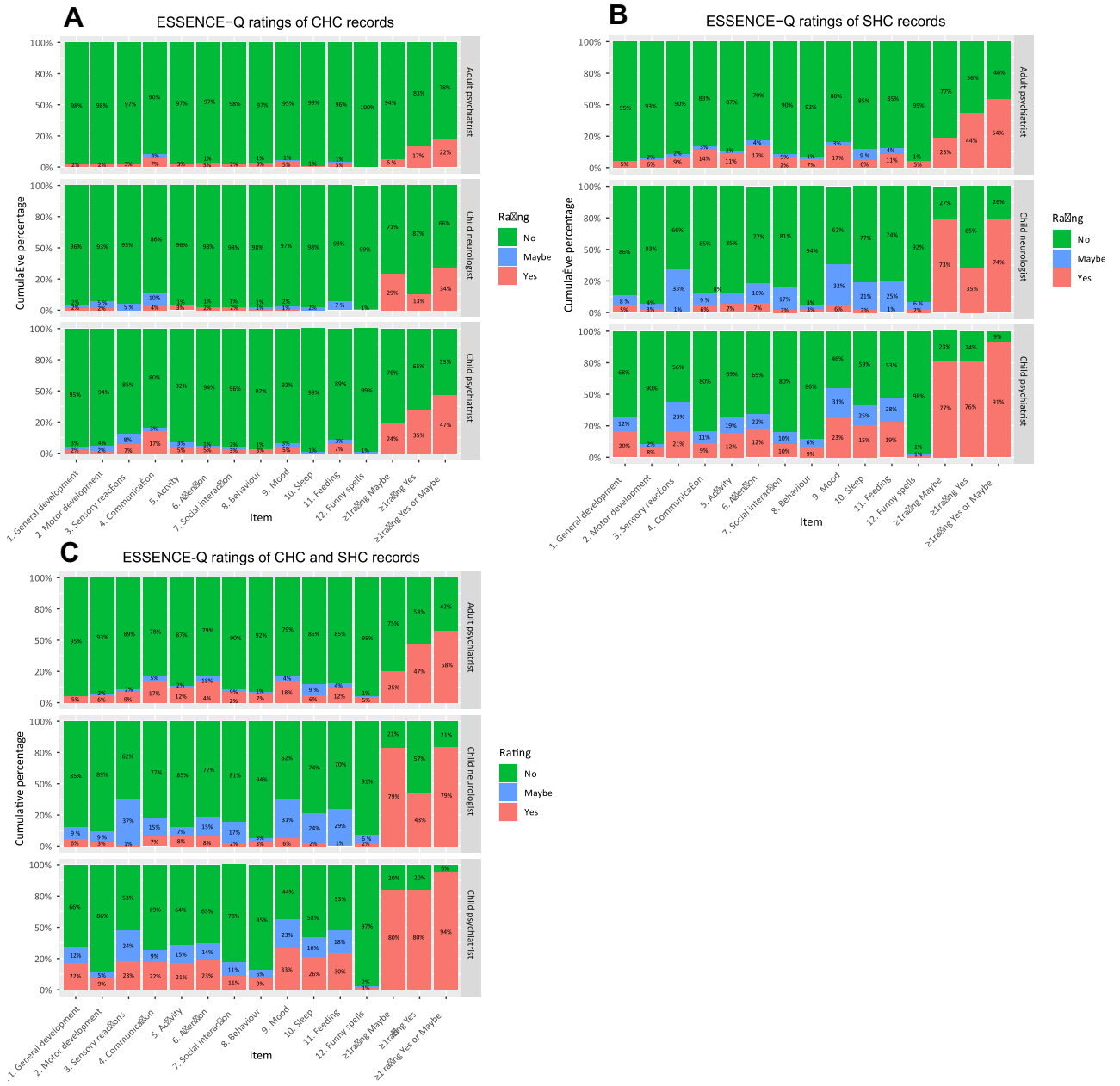

Figure 3 (**A**–**C**) Proportions of responses by rater and ESSENCE-Q item in CHC records (**A**), SHC records (**B**) and CHC+SHC records (**C**).
Notes: Proportion of responses "No", "Maybe" and "Yes" for each item (1–12) in the ESSENCE-Q for possible ESSENCE-related problems, followed by any occurrence of ≥1 rating "Maybe", ≥1 "Yes" or ≥1 either "Maybe" or "Yes" in any of the 12 items.
Abbreviation: ESSENCE-Q, Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations-Questionnaire.

ESSENCE-Q Summary Score Intra-Rater Reliability

The intra-rater test–retest reliability of the ESSENCE-Q rating for the child psychiatrist (n=20) was excellent for CHC medical records (ICC1 = 0.97, 95% CI 0.94–0.99), SHC medical records (ICC1= 0.91, 95% CI 0.83–0.96) and both combined (ICC1 = 0.91 95% CI 0.82–0.96) and can be inspected in Figure S2.

Outcomes

Out of the 201 participants rated by any rater, 75 (37%) were judged to have clinically relevant NDPs (CGI-S 4–7), and 126 (63%) as having scant or negligible NDPs. ESSENCE-Q-ratings of CHC + SHC records were positively correlated

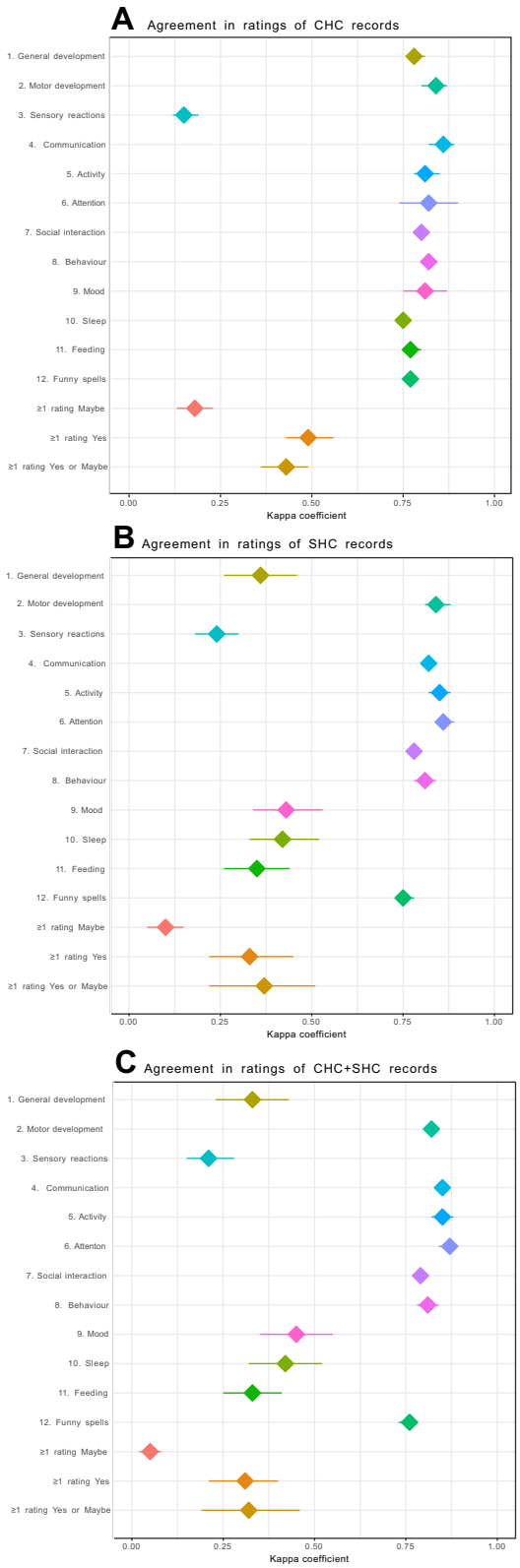

Figure 4 Interrater agreement in scoring the ESSENCE-Q from CHC records (**A**), SHC records (**B**) and CHC+SHC records (**C**).
Notes: Rhombus depicts point estimate and lines 95% confidence interval of model-based kappa association for items 1–12, and model-based kappa coefficient for binary ratings (≥1 rating "maybe", ≥1 rating "Yes", ≥1 rating "Yes" or "maybe").
Abbreviation: ESSENCE-Q, Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations-Questionnaire.

Table 2 Intraclass Correlation Coefficients for the ESSENCE-Q Summary Score

Source	ICC Formula	ICC (95% Confidence Interval)	Observations
CHC records	ICC2,k	0.88 (0.84–0.91)	173
CHC records	ICC3,k	0.89 (0.86–0.91)	173
SHC records	ICC2,k	0.86 (0.67–0.92)	192
SHC records	ICC3,k	0.92 (0.91–0.94)	192
CHC+SHC records	ICC2,k	0.85 (0.64–0.91)	169
CHC+SHC records	ICC3,k	0.92 (0.90–0.93)	169

Notes: ICC for the ESSENCE-Q summary scores of the three raters (adult psychiatrist, pediatric neurologist, child psychiatrist) for agreement (ICC3,k) and consistency (ICC2,k).
Abbreviations: ICC, Intraclass correlation coefficient; CHC, child health care; SHC, school health care; ESSENCE-Q, Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations-Questionnaire.

with the clinical impairment as measured with the CGI-S (CGI-S vs Child psychiatrist Spearman $r = 0.65$ [n=175], CGI-S vs Adult psychiatrist $r = 0.60$ [n=190], CGI-S vs Child neurologist $r = 0.59$, [n=176], all $p<0.001$, Figure 5A–C).

ROC-curves of the capacity of medical records-rated ESSENCE-Q in predicting clinically relevant NDPs (CGI-S of 4–7) are displayed in Figure 6A–C (Corresponding classification tables provided in Table S1A–C). The AUC of the raters was similar (child psychiatrist 0.84 95% confidence interval [CI] 0.78–0.91 [n=176], child neurologist 0.81 95% CI 0.74–0.88 [n=175], adult psychiatrist 0.84 95% CI 0.78–0.90 [n=190]). Cutoffs with the highest accuracy (proportion accurately classified, [true positive + true negative/all positive and negative]) differed between the raters (child psychiatrist ≥8, 81% accuracy, child neurologist ≥3, 79% accuracy, adult psychiatrist ≥2, 79% accuracy). The predictive value for NDPs from ESSENCE-Q rated SHC-records only was low, with an AUC ~ 0.6 for all three raters.

Discussion

Key Results

We found that signs of NDPs are readily identifiable in CHC medical records and increasingly so in SHC records when rated with the ESSENCE-Q by a physician. Interrater agreement in the ESSENCE-Q summary score of medical record ratings measured with the ICC was good to excellent. Further, the ESSENCE-Q ratings based on the medical records were moderately correlated with clinical impairment measured with the CGI-S, and the AUC of the ratings in predicting clinical-level impairment was not negligible.

Interpretation

Prevalence

Even by the most restrictive rater (adult psychiatrist), some concern was noted in the CHC records for almost one-fifth of participants, increasing to half of the study population when including SHC records. We stress that concerns elicited are not specific for NDDs. For example, feeding or mood problems may be due to medical conditions or social circumstances. Nevertheless, signs of possible neurodevelopmental problems are identifiable by scrutinizing medical records and are common. This implies that there is potentially useful information pertaining to NDDs to be found upon scrutiny of records. It has been described that clinically significant ESSENCE concerns present before 3–5 years of age, with a duration of at least 3 months and is evident in 10% of the general population.[2] The estimates herein were higher and may be inflated. On the contrary, multiple studies with inclusive definitions of neurodevelopmental and/or general mental health problems in children find them to be common if assessed longitudinally, and that lifetime prevalence increases as more developmental ages are taken into account, approaching numbers in the present study.[41–43]

Item-Specific (Dis)agreement

Several items in the questionnaire (motor function, communication, attention, social interaction, behavior) showed strong agreement (~0.8), whereas some (general development, mood, sleep, feeding and sensory reactions) consistently showed poor (~0.3) or almost no agreement (~0.1). One reason may be that the former items are hallmark symptoms of NDDs having been

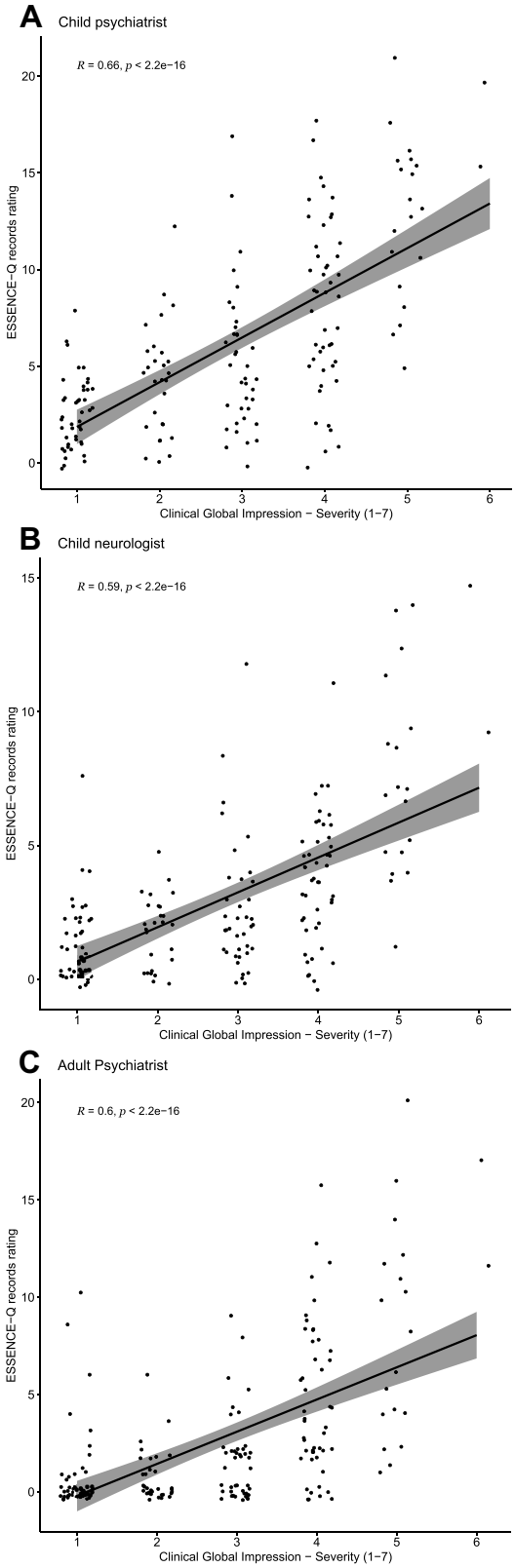

Figure 5 Correlation plot of ESSENCE-Q (range 0–24) rating and Clinical Global Impression – Severity (range 1–7) for the three raters. (**A**) Child psychiatrist (n=175), (**B**) child neurologist (n=176), (**C**) adult psychiatrist (n=190).

Note: *R* denotes Spearman's rank correlation between the two variables.

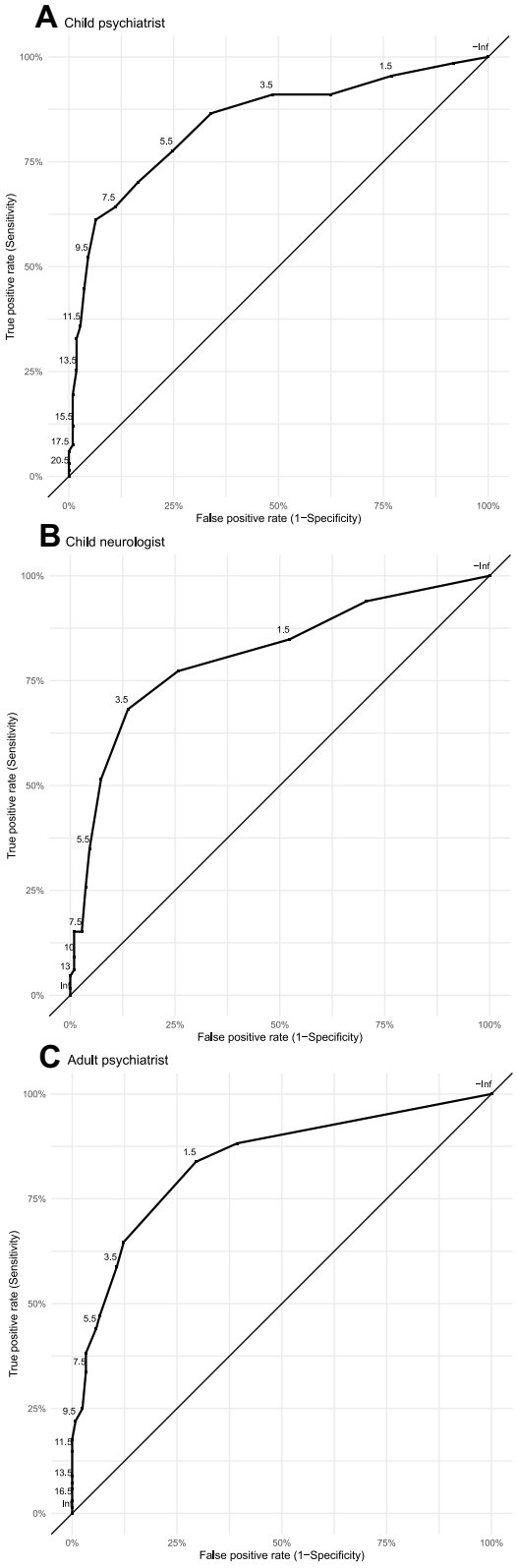

Figure 6 Receiver operating characteristic curves of the classification accuracy of the ESSENCE-Q in detecting clinically impairing neurodevelopmental problems (a Clinical Global Impression-Severity rating of 4–7) for the three raters. (**A**) Child psychiatrist (n=176), (**B**) child neurologist (n=176), (**C**) adult psychiatrist (n=190).
Notes: Points and numbers denote selected cutoffs in the ESSENCE-Q. Clinically relevant neurodevelopmental problems were found in 75 of 201 participants with records rated by any rater.

in clinical use for many years, whereby clinicians are familiar with assessing them.[44] Problems with attention are the defining characteristic of attention-deficit/hyperactivity disorder, and social interaction problems are the core symptom of autism spectrum disorder (ASD).[45] In contrast, some of the latter items (mood, sleep and feeding) are less specific for NDDs. While problems with mood and sleep often precede the typical manifestations of NDDs and frequently co-occur,[2,46] they are not part of diagnostic criteria for NDDs and they may sometimes be better explained by psychosocial factors (eg, problems with family or peers causing insomnia and anxiety) or medical problems (eg, allergy or gastrointestinal conditions causing problems with feeding). Atypical sensory processing has in recent years been increasingly recognized as part of the NDD-spectrum and it is a diagnostic criterion for ASD as well as avoidant/restrictive food intake disorder in the diagnostic and statistical manual of mental disorders – fifth version (DSM-5).[45,47–49] Due to its fairly recent inclusion as a diagnostic criterion and the fact that clinicians in Sweden still in parallel with DSM-5 rely on the International Classification of Diseases – version 10 for assigning diagnoses, the familiarity with inferring indirect signs of atypical sensory processing may vary between clinicians more than that of, eg, social interaction or attention problems.

Additionally, some observations may be consistent with several items in the questionnaire and decrease their agreement. Eating problems may be interpreted as problems with feeding, or as problems with sensory reactions, whereby the same observation is rated under different items. If this was the case, this could be reflected in higher agreement for the occurrence per se of positive ratings than in the specific items of interest. Looking at the data, there was virtually no agreement in occurrence per se of "maybe/a little" (1). However, the occurrence of "Yes" (2), as well as either "Maybe/a little" or "Yes" (1 or 2) had weak agreement. This is consistent with at least some item substitution in the classifications (the same observation coded by raters under different items). As we argue below, absence of agreement in ratings of "Maybe/a little" (1) may in view of the high agreement in the summary score imply a calibration effect (ie, raters differ in their vigilance for clinically significant concern).

Summary Score

Interrater agreement in the summary score vas good to excellent, and indicated minor systematic measurement error. The ICC 2,k formula penalizes the agreement for systematic measurement error (raters scoring higher or lower than other raters, but in a consistent manner), whereas the ICC 3,k formula for consistency describes the degree to which ratings are correlated "in an additive manner", ignoring systematic measurement error and therefore provides the most favorable measure.[35] Substantial difference between these two types of ICC identifies systematic bias as an important contributor to low agreement.[50] As reflected in the difference in mean scores between raters (Table 1), the systematic bias is likely an effect from raters being calibrated differently in sensitivity for signs of ESSENCE-related problems. Although largely agreeing, they differ in how many points are assigned to the concern noted.

Outcomes

As discussed in Landgren et al, the rate of clinically significant problems in this cohort (CGI $\geq$ 4 in 37%) was higher than expected and may be due to several factors.[18] Selection effects (families experiencing problems may be more prone to participate and those without problems more prone to decline), a more demanding school environment that make "borderline cases" with NDPs clinically impaired, or that studies relying on statistically motivated cutoffs may underestimate the prevalence of NDPs in this age group may be reasons. The AUC of the ESSENCE-Q rated medical records were in the range of ~0.80, which is similar to that of other screening measures for NDDs in this age-group, such as the SDQ.[17]

Limitations and Generalizability

The study has some limitations. Interrater agreement was tested for only three raters and to what extent their personality impacted their rating, and whether their competence is generalizable to physicians of their specialties is unknown.

The amount of information in the records reflect habits of Swedish health services and may not be generalizable to other countries. For example, the amount of information may yield less in a similar screening in a setting using non-computerized records.

Although the wide inclusion criterion was the strength of the study, the high attrition rate (51%) is considerable, and findings warrant corroboration, ideally in another cohort. There were a wide range of sources available for the CGI-S assessments, but formal diagnostic interviews for specific NDDs could not be performed.

As mentioned above, the differences in calibration between raters preclude the selection of one optimal cutoff for the instrument, in spite of acceptable AUCs from each individual rater's prediction. That the statistically optimal cutoff for each rater differed (2, 3, and 8) underlines the possibility that clinicians may be differentially calibrated in their vigilance for NDPs. Because the diagnosis of a NDD is based on symptoms and clinical judgement, a single instrument and cutoff with perfect discrimination is unlikely to emerge. Rather, it is the judicial use of multiple sources of information (self-report, collateral information, clinical examination, medical records) that form the basis for well-grounded diagnosis of NDDs. Within such a context, use of the ESSENCE-Q as a template in medical records screening may provide a scaffold for incorporating the findings into clinical decision-making.

Future Research

Generalizability of the method would improve by having a larger pool of raters and professions (eg, nurse, psychologist) and would ideally be tested in another cohort. Whether reliability could be improved further by recording key observations of special significance (eg, early referrals, anthropometric measurements) or applying a scoring guide should be investigated. This must be weighed against the simplicity of using the instrument as it is, which requires no specific training. Given proper clinical validation with regard to NDPs and clinical diagnoses, algorithms derived from machine learning could potentially assist in the screening process.

Conclusion

The use of ESSENCE-Q for medical records screening was proven feasible with regard to identification of signs and interrater agreement, and was associated withthe degree of NDPs. It may provide heuristic for screening and an additional tool work-up of NDDs in a consistent manner and warrants future validation studies.

Acknowledgments

We thank Jari Martikainen from the Bioinformatics Core Facility at the Sahlgrenska Academy, University of Gothenburg, for assistance in statistical analyses and code auditing.

Author Contributions

Rajna Knez and Valdemar Landgren conceptualized and designed the study. Magnus Landgren and Valdemar Landgren acquired the data and Emma Svensson, Michail Theodosiou and Zohar Raanan Soltis performed the ratings. Valdemar Landgren performed the statistical analyses. Zohar Raanan Soltis and Valdemar Landgren drafted the initial manuscript. All authors interpreted the results and took part in revising and critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

Funding was provided by grants from Research and Development of Region Västra Götaland (VGSKAS-931754) and the research fund at Skaraborg Hospital (VGSKAS-967692, VGSKAS-939541).

Disclosure

Dr Magnus Landgren reports grants from Research and Development of Region Västra Götaland, grants from Fund at Skaraborg Hospital, during the conduct of the study. The authors report no conflicts of interest.

References

1. Spitzer RL. Psychiatric diagnosis: are clinicians still necessary? *Compr Psychiatry*. 1983;24(5):399–411. doi:10.1016/0010-440X(83)90032-9
2. Gillberg C. The ESSENCE in child psychiatry: early symptomatic syndromes eliciting neurodevelopmental clinical examinations. *Res Dev Disabil*. 2010;31(6):1543–1551. doi:10.1016/j.ridd.2010.06.002

3. Mulligan A, Anney RJL, O'Regan M, et al. Autism symptoms in attention-deficit/hyperactivity disorder: a familial trait which correlates with conduct, oppositional defiant, language and motor disorders. *J Autism Dev Disord*. 2009;39(2):210–211. doi:10.1007/s10803-008-0640-0
4. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet*. 2013;45(9):984–994. doi:10.1038/ng.2711
5. Pettersson E, Anckarsäter H, Gillberg C, Lichtenstein P. Different neurodevelopmental symptoms have a common genetic etiology. *J Child Psychol Psychiatry*. 2013;54(12):1356–1365. doi:10.1111/jcpp.12113
6. Bahmanyar S, Sundström A, Kaijser M, von Knorring AL, Kieler H. Pharmacological treatment and demographic characteristics of pediatric patients with attention deficit hyperactivity disorder, Sweden. *Eur Neuropsychopharmacol*. 2013;23(12):1732–1738. doi:10.1016/j.euroneuro.2013.07.009
7. Dalsgaard S, Thorsteinsson E, Trabjerg BB, et al. Incidence rates and cumulative incidences of the full spectrum of diagnosed mental disorders in childhood and adolescence. *JAMA Psychiatry*. 2020;77(2):155–164. doi:10.1001/jamapsychiatry.2019.3523
8. Caye A, Rocha TB-M, Anselmi L. Attention-deficit/hyperactivity disorder trajectories from childhood to young adulthood: evidence from a birth cohort supporting a late-onset syndrome. *JAMA Psychiatry*. 2016;73:705. doi:10.1001/jamapsychiatry.2016.0383
9. Cooper M, Hammerton G, Collishaw S, et al. Investigating late-onset ADHD: a population cohort investigation. *J Child Psychol Psychiatry*. 2018;59(10):1105–1113. doi:10.1111/jcpp.12911
10. Taylor MJ, Larsson H, Gillberg C, Lichtenstein P, Lundström S. Investigating the childhood symptom profile of community-based individuals diagnosed with attention-deficit/hyperactivity disorder as adults. *J Child Psychol Psychiatry*. 2019;60(3):259–266. doi:10.1111/jcpp.12988
11. Riglin L, Wootton RE, Thapar AK, et al. Variable emergence of autism spectrum disorder symptoms from childhood to early adulthood. *Am J Psychiatry*. 2021;178(8):752–760. doi:10.1176/appi.ajp.2020.20071119
12. Zegers M, de Bruijne MC, Spreeuwenberg P, Wagner C, Groenewegen PP, van der Wal G. Quality of patient record keeping: an indicator of the quality of care? *BMJ Qual Saf*. 2011;20(4):314–318. doi:10.1136/bmjqs.2009.038976
13. Worster A, Haines T. Advanced statistics: understanding medical record review (MRR) studies. *Acad Emerg Med*. 2004;11(2):187–192. doi:10.1111/j.1553-2712.2004.tb01433.x
14. Jensen-Doss A, Youngstrom EA, Youngstrom JK, Feeny NC, Findling RL. Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *J Consult Clin Psychol*. 2014;82(6):1151–1162. doi:10.1037/a0036657
15. Ludvigsson JF, Andersson E, Ekbom A, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011;11:450. doi:10.1186/1471-2458-11-450
16. Marlow M, Servili C, Tomlinson M. A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: recommendations for use in low- and middle-income countries. *Autism Res*. 2019;12(2):176–199. doi:10.1002/aur.2033
17. Mulraney M, Arrondo G, Musullulu H, et al. Systematic review and meta-analysis: screening tools for attention-deficit/hyperactivity disorder in children and adolescents. *J Am Acad Child Adolesc Psychiatry*. 2021. doi:10.1016/j.jaac.2021.11.031
18. Landgren V, Svensson L, Knez R, et al. The ESSENCE-questionnaire for neurodevelopmental problems – a Swedish school-based validation study in 11-year-old children. *Neuropsychiatr Dis Treat*. 2022;Volume 18:2055–2067. doi:10.2147/NDT.S374930
19. Roid GH, Pomplun M, Martin JJ. *Leiter International Performance Scale − 3rd Edition*. 3rd ed. Stoelting; 2013. Available from https://hogrefe.se/klinisk-psykologi/leiter-3/.
20. Malmberg M, Rydell Margret A, Smedje H. Validity of the Swedish version of the Strengths and Difficulties Questionnaire (SDQ-Swe). *Nord J Psychiatry*. 2003;57(5):357–363. doi:10.1080/08039480310002697
21. Busner J, Targum SD. The clinical global impressions scale. *Psychiatry Edgmont*. 2007;4(7):28–37.
22. Örtenstrand A, Waldenström U. Mothers' experiences of child health clinic services in Sweden. *Acta Paediatr*. 2005;94(9):1285–1294. doi:10.1111/j.1651-2227.2005.tb02090.x
23. Kornfält R. Survey of the pre-school child health surveillance programme in Sweden. *Acta Paediatr Oslo Nor*. 2000;89(434):2–7. doi:10.1111/j.1651-2227.2000.tb03088.x
24. Swedish National Board of Health and Welfare. Health care for mothers and children within the primary health care system. Allmänna råd från Socialstyrelsen 1981. 1981:4.
25. Swedish National Board of Health and Welfare. Kvalitetssäkring av Barnhälsovården. Att skydda skyddsnätet [Quality assurance of child health care. Protecting the safety net]. SoS-rapport 1994. 1994:19.
26. University of Gothenburg. ESSENCE-Q screening questionnaire. Available from: https://www.gu.se/en/gnc/gncs-resources/screening-questionnaires/essence-q-screening-questionnaire. Accessed November 4, 2021.
27. Hatakenaka Y, Ninomiya H, Billstedt E, Fernell E, Gillberg C. ESSENCE-Q - used as a screening tool for neurodevelopmental problems in public health checkups for young children in south Japan. *Neuropsychiatr Dis Treat*. 2017;13:1271–1280. doi:10.2147/NDT.S132546
28. Stevanovic D, Knez R, Zorcec T, et al. ESSENCE-Q: Slavic language versions for developmental screening in young children. *Neuropsychiatr Dis Treat*. 2018;14:2141–2148. doi:10.2147/NDT.S171359
29. Hatakenaka Y, Maeda M, Ninomiya H, Hachiya K, Fernell E, Gillberg C. ESSENCE-Q obtained in routine Japanese public child health check-ups may be a valuable tool in neurodevelopmental screening. *Acta Paediatr*. 2020;109(4):764–773. doi:10.1111/apa.15029
30. Mitani AA, Freer PE, Nelson KP. Summary measures of agreement and association between many raters' ordinal classifications. *Ann Epidemiol*. 2017;27(10):677–685.e4. doi:10.1016/j.annepidem.2017.09.001
31. Nelson KP, Edwards D. Measures of agreement between many raters for ordinal classifications. *Stat Med*. 2015;34(23):3116–3132. doi:10.1002/sim.6546
32. Mitani A, Nelson K. Modeling agreement between binary classifications of multiple raters in R and SAS. *J Mod Appl Stat Methods*. 2017;16(2):277–309. doi:10.22237/jmasm/1509495300
33. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46. doi:10.1177/001316446002000104
34. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428. doi:10.1037/0033-2909.86.2.420
35. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012
36. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22(3):276–282. doi:10.11613/BM.2012.031
37. RStudio Team. RStudio: integrated development environment for R; 2020. Available from: http://www.rstudio.com/. Accessed October 10, 2022.

38. Revelle W. psych: procedures for psychological, psychometric, and personality research; 2020. Available from: https://CRAN.R-project.org/package=psych. Accessed October 10, 2022.
39. Mitani A. *AyaMitani / modelkappa: v1*. Zenodo https://github.com/AyaMitani/modelkappa/ Accessed14 October 2022; 2019. doi:10.5281/zenodo.3546381
40. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform*. 2011;12(1):77. doi:10.1186/1471-2105-12-77
41. Caspi A, Houts RM, Ambler A, et al. Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin birth cohort study. *JAMA Netw Open*. 2020;3(4):e203221. doi:10.1001/jamanetworkopen.2020.3221
42. Eyre O, Hughes RA, Thapar AK, et al. Childhood neurodevelopmental difficulties and risk of adolescent depression: the role of irritability. *J Child Psychol Psychiatry*. 2019;60(8):866–874. doi:10.1111/jcpp.13053
43. Carballal Mariño M, Gago Ageitos A, Ares Alvarez J, et al. Prevalence of neurodevelopmental, behavioural and learning disorders in paediatric primary care. *An Pediatría Engl*. 2018;89(3):153–161. doi:10.1016/j.anpede.2017.10.005
44. American Psychiatric Association (APA). *Diagnostic and Statistical Manual of Mental Disorders*. Third ed. American Psychiatric Association (APA); 1980.
45. American Psychiatric Association (APA). *Diagnostic and Statistical Manual of Mental Disorders*. Fifth ed. American Psychiatric Association (APA); 2013.
46. Shelton AR, Malow B. Neurodevelopmental disorders commonly presenting with sleep disturbances. *Neurotherapeutics*. 2021;18(1):156–169. doi:10.1007/s13311-020-00982-8
47. Brimo K, Dinkler L, Gillberg C, Lichtenstein P, Lundström S, Johnels JÅ. The co-occurrence of neurodevelopmental problems in dyslexia. *Dyslexia*. 2021;27:277–293. doi:10.1002/dys.1681
48. Dinkler L, Bryant-Waugh R. Assessment of avoidant restrictive food intake disorder, pica and rumination disorder: interview and questionnaire measures. *Curr Opin Psychiatry*. 2021;34(6):532–542. doi:10.1097/YCO.0000000000000736
49. Smith B, Rogers SL, Blissett J, Ludlow AK. The relationship between sensory sensitivity, food fussiness and food preferences in children with neurodevelopmental disorders. *Appetite*. 2020;150:104643. doi:10.1016/j.appet.2020.104643
50. Liljequist D, Elfving B, Roaldsen KS. Intraclass correlation – a discussion and demonstration of basic features. *PLoS One*. 2019;14(7):e0219854. doi:10.1371/journal.pone.0219854