Assessing procedural pain in infants: a feasibility study evaluating a point-of-care mobile solution based on automated facial analysis

Kreshnik Hoti, Paola Teresa Chivers, Jeffery David Hughes

Summary

Background The management of procedural pain in infants is suboptimal, in part, compounded by the scarcity of a simple, accurate, and reliable method of assessing such pain. In this study, we aimed to evaluate the psychometric properties of the PainChek Infant, a point-of-care mobile application that uses automated facial evaluation and analysis in the assessment of procedural pain in infants.

Methods Video recordings of 40 infants were randomly chosen from a purposely assembled digital library of 410 children undergoing immunisation as part of their standard care in Prishtina, Kosovo, between April 4, 2017, and July 11, 2018. For each infant recording, four 10 s video segments were extracted, corresponding to baseline, vaccine preparation, during vaccination, and recovery. Four trained assessors did pain assessments on the video segments of 30 infants, using PainChek Infant standard, PainChek Infant adaptive, the Neonatal Facial Coding System-Revised (NFCS-R) single, the NFCS-R multiple, and the Observer administered Visual Analogue Scale (ObsVAS), on two separate occasions. PainChek Infant's performance was compared to NFCS-R and ObsVAS using correlation in changes in pain scores, intra-rater and inter-rater reliability, and internal consistency.

Findings 4303 pain assessments were completed in two separate testing sessions, on Aug 31, and Oct 19, 2020. The study involved videos of 40 infants aged 2·2–6·9 months (median age 3·4 months [IQR 2·3–4·5]). All pain assessment tools showed significant changes in the recorded pain scores across the four video segments ($p \leq 0·0006$). All tools were found to be responsive to procedure-induced pain, with the degree of change in pain scores not influenced by pre-vaccination pain levels. PainChek Infant pain scores showed good correlation with NFCS-R and ObsVAS scores ($r=0·82–0·88$; $p<0·0001$). PainChek Infant also showed good to excellent inter-rater reliability (ICC=0·81–0·97, $p<0·001$) and high levels of internal consistency ($\alpha=0·82–0·97$).

Interpretation PainChek Infant's use of automated facial expression analysis could offer a valid and reliable means of assessing and monitoring procedural pain in infants. Its clinical utility in clinical practice requires further research.

Funding PainChek.

Copyright © 2021 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Lancet Digit Health 2021; 3: e623–34

Published Online September 1, 2021 https://doi.org/10.1016/S2589-7500(21)00129-1

See Comment page e613

Faculty of Medicine, University of Prishtina, Pristina, Kosovo (K Hoti); **Institute for Health Research, The University of Notre Dame Australia, Fremantle, WA, Australia** (P T Chivers PhD); **School of Medical and Health Sciences, Edith Cowan University, Joondalup, WA, Australia** (P T Chivers); **Curtin Medical School, Faculty of Health Sciences, Curtin University, Bentley, WA, Australia** (J D Hughes PhD)

Correspondence to: Prof Jeffery David Hughes, Curtin Medical School, Faculty of Health Sciences, Curtin University, GPO Box U1987, Perth, WA, 6845, Australia **j.d.hughes@curtin.edu.au**

Introduction

Procedural pain is acute pain associated with investigations, treatments, or procedures done in the course of delivering health care.[1,2] As such, pain can arise from any procedure causing actual or potential tissue damage.[1] Such interventions include simple procedures such as intravenous cannulation, venepuncture, finger and heel pricks, immunisations, and dressing changes, to more invasive procedures such as lumbar punctures or bone marrow biopsies.[1,2] Procedures occur in a variety of settings, from hospitals or day surgery centres to ambulatory care clinics, general practice, dental clinics, and the home care environment.[1]

Although procedural pain might be associated with an isolated event, it is not uncommon for children to have multiple painful procedures daily when being cared for in hospital or ambulatory settings.[3] Within emergency departments, procedures represent one of the most common sources of acute painful stimulus in children, with studies showing up to 80% of children undergoing painful diagnostic procedures.[4] Unfortunately, there are numerous studies reporting that pain, especially procedural pain, in children including infants, is often poorly managed.[3,5] These findings are despite the optimisation of the management of paediatric pain being a key health-care priority of WHO and leading paediatric and pain societies.[3]

Poorly managed procedural pain can have short-term and long-term consequences.[1,5] Repeated procedures warrant careful pain management as insufficient pain relief could lead to anxiety and distress during subsequent episodes.[3,5,6] In infants, a major challenge to managing pain is the ability to assess it accurately and reliably. The generally accepted standard for pain assessment is self-report; however, in preverbal children who cannot communicate their pain, age-appropriate behavioural or observational pain assessment tools are recommended.[6]

Research in context

Evidence before this study

In this study, we aimed to validate a novel pain assessment tool for infants, PainChek Infant. Initially, we reviewed currently available evidence on pain assessment tools in infants. In doing so, we searched the following databases: MEDLINE, Embase, PsycInfo, and Joanna Briggs Institute Database from inception through to Oct 31, 2020, using keywords agreed between three reviewers (["infant", "pain measurement", "pain assessment", "pain scale", or "rating pain"], with ["review", "systematic review", or "meta-analysis"]). We also individually searched leading paediatric journals. Our key criteria for literature selection related to identifying and reviewing currently available systematic reviews that reported pain assessment in infants, and were available in English. PRISMA protocol was consulted in reviewing the relevant literature. Following the literature search, we selected the records which fulfilled our criteria and conducted a qualitative analysis of the full papers. We concluded that there are over 40 pain assessment tools that have been developed thus far that are based around observer identification and evaluation of specific biomarkers indicative of pain. However, to our knowledge there is no automated pain assessment tool used in infants. All currently available tools rely on users making their decisions based on subjective observations of the child. Furthermore, our review of the literature showed that the presence of so many pain assessment tools for infants is marred by an absence of universal standardisation, which has resulted in a variety of criteria being used by clinicians in order to assess pain in infants. In practice, this could result in variation in interpretations of the presence and intensity of pain, which could go on to have an effect on how pain is treated.

Added value of this study

This study describes the evaluation of the psychometric properties of a technology-based solution for the assessment of procedural pain in infants. PainChek Infant uses automated facial evaluation and analysis to detect six facial action units (AUs) indicative of the presence of pain. In this study, we showed that, when compared to manual assessment using either the Neonatal Facial Coding System-Revised (a validated facial action recognition tool based observational pain assessment) or an observer administered Visual Analogue Scale, PainChek Infant showed moderate to excellent validity, inter-rater and intra-rater reliability, and internal consistency. Further, in accordance with previous research, the AUs detected represent pain intensity and pain-related distress. Using automated facial evaluation and analysis PainChek Infant provides a rapid (assessments take 3 s), valid, and reliable means of assessing procedural pain.

Implications of all the available evidence

Although there are currently many available tools to assess procedural pain in infants, they are often underused and, as a result, infants continue to undergo painful procedures without adequate pain management, resulting in short-term and long-term negative outcomes. Artificial intelligence in this field can be used to support clinician decision making, allow curiosity-driven care, remove the need to complete mundane tasks, improve communication, and facilitate collaboration. Evaluation of the face is central to all observational pain assessment tools, as the face is highly accessible and facial expressions are considered the most encodable feature of pain. However, decoding of the face by clinicians and parents alike is difficult when trying to discern pain from other causes of distress. It is therefore important that new tools developed to assess pain in this vulnerable population take advantage of technology to objectify and simplify the process, thus allowing users to better identify and quantify pain. This study adds to the body of evidence to support the use of artificial intelligence in the assessment of pain in the infant population, and its advantages over humans in decoding pain behaviours.

Common to the multitude of existing observational pain assessment tools is the evaluation of the child's facial expression.[7] However, as a result of practical difficulties with manual decoding of facial expressions, automated facial expression analysis is a topic of broad research.[8] In children, there have been a number of studies exploring automated pain facial expression recognition.[8,9] However, we are not aware of any automated pain assessment tools for infants that have been tested against existing paper-based validated tools and that are available for use at point-of-care in a mobile application (app).

The aim of this study was to evaluate the psychometric properties of the PainChek Infant, a point of care app that utilises automated facial expression analysis to detect pain in infants aged 1–12 months, against two scales: the Neonatal Facial Coding System (NFCS),[10–14] which has been designed for procedural pain assessment, and the Observer administered Visual Analogue Scale (ObsVAS) pain scale, which is used to quantify procedural pain intensity. In doing so, we question whether it is feasible to use the automated facial expression analysis of PainChek Infant to detect and quantify pain.

Methods

Study design and population

In this feasibility study, we evaluated the PainChek Infant against the Neonatal Facial Coding System-Revised (NFCS-R) and the ObsVAS to investigate whether PainChek's automated facial expression analysis on a series of videos of infants undergoing immunisation can accurately detect and quantify pain. The infant videos used in this study came from a purposely assembled digital library of children undergoing routine immunisation at an immunisation clinic in Prishtina, Kosovo. The parents or guardians of the children provided informed written consent for their child's immunisation

to be video recorded and used for research and development purposes. The digital library consisted of 410 children aged 0–12 years, of whom 329 were infants. After reviewing these infant videos for suitability, on the basis of the infant's face not being obstructed from view, we randomly chose 40 of them, using an electronic randomiser. In cases when there was a substantial obstruction, the next video in line was chosen. Each infant was recorded for approximately 60 s before and 90 s after the vaccination. This study was approved by human research ethics committees of the Faculty of Medicine, University of Prishtina, Kosovo (approval number: 3812/17) and Curtin University, Perth, WA, Australia (approval number: HRE2020-0315).

Procedures

PainChek Infant uses artificial intelligence (AI) for the automated recognition and analysis of an infant's face, allowing detection of six facial action units (AUs) indicative of the presence of pain: AU4 (brow lowering), AU9 (wrinkling of nose), AU15 (lip corner depressor), AU20 (horizontal mouth stretch), AU25 (parting lips), and AU43 (eye closure). These facial actions represent specific muscle movements (contractions or relaxations) as classified by the Baby Facial Action Coding System.[13] Each of the six AUs is scored using a binary scale (0=absent, 1=present), yielding a total potential score of 6. The tool has been specifically designed to assess pain in infants (aged 1–12 months), taking into account the facial actions commonly associated with pain in this population. The algorithms created during the PainChek Infant development were trained on images corresponding to this age group. The algorithms for the app were developed by data scientists from the Universida de Castilla-La Mancha, Spain, using a non-proprietary database of videos of infants undergoing routine immunisations. A five-fold cross validation methodology was used, using independent training and validation datasets. The automated facial analysis could be completed either using a fixed video duration (standard 3 s video mode) or using a fixed number of minimum valid images (video adaptive mode). The video adaptive mode has been developed to address the potential for increased head movements that often accompany pain or distress in infants.

The NFCS-R, one of the two PainChek Infant comparators in this study, utilises the same construct as PainChek Infant in that it uses facial actions as indicators of pain, but is assessor-rated rather than automated. The original NFCS contained ten facial actions; however, a 1996 study showed that reducing the detection of facial actions to five improved the specificity for pain assessment, while maintaining sensitivity and validity.[15] The NFCS-R contains the following five facial actions:[14] brow bulge, eye squeeze, nasolabial furrow, horizontal mouth stretch, and taut tongue. The scoring system in the NFCS-R is also binary, with final pain score between 0 and 5 depending on the presence of each of the above facial actions. The NFCS has been used and evaluated in procedural pain[16] and postoperative pain[15] in infants and it can evaluate the effect of treatment, as well as discriminate between tissue insult and non-tissue insult procedures.[13,15,16] Overall, the NFCS has been shown to have good inter-rater reliability,[14,17–19] construct validity,[13,15,16] and convergent validity.[18,19] More details on the characteristics of the NFCS-R have been published elsewhere.[14]

For more on the **electronic randomiser** see www.randomizer.org

The ObsVAS, the second PainChek Infant comparator, is a commonly used tool that measures and quantifies pain and distress.[20] The scale consists of a 100 mm line on which 0 mm represents no pain or distress and 100 mm represents the worst possible pain or distress. The level of pain or distress is determined by the distance from the 0 mm point. The ObsVAS scale was included to gain an estimate of the level of pain or distress that assessors perceived the infant to be experiencing during the phases of the procedure. A 2009 study reported ObsVAS to have good-to-excellent intra-rater reliability (intraclass correlation coefficient [ICC] 0·69 to 0·91) and inter-rater reliability (ICC 0·55 to 0·97), and strong criterion validity compared with the Modified Behavioural Pain Scale (Pearson's rho 0·81–0·94) in infants undergoing vaccination.[20]

The video recordings for each of the 40 infants were divided into segments to show different phases of the procedure: baseline (before any attempt to prepare the infant for the procedure was made [ie, while still in their parent's arms]), preparation (while the infant's arm was prepared and swabbed), during vaccination (the painful part of the procedure [ie, the 10 s after needle insertion]), recovery (after the painful procedure [ie, between 10 s and 40 s after the needle insertion). It was presumed that any behavioural change suggestive of distress in the infant in baseline and preparation segments was most likely non-pain related, as these infants were yet to have their injection. 160 video segments were prepared for review. Each video segment was 10 s in length (the first 10 s segment without substantial obstruction was chosen) as required for NFCS-R video analysis.[12]

Four assessors used the three pain assessment instruments to assess the pain or distress experienced by the infant. Assessors were blinded to each other's results and did their assessments and data entry remotely via a purposefully designed electronic data management system (EDMS, version 1). The assessors accessed the EDMS with their unique study identification number to ensure that they only assessed the infants allocated to them. Every assessor was assigned 120 video segments on which to do the pain assessments, from one of two testing session sets. Each video testing set included 30 infants, chosen from the pool of 40 infants, ten of whom were unique to that testing session set. This assignment method ensured independent, paired pain assessments were completed by at least two assessors on each video segment. Each testing session dataset was assigned to

one clinically experienced assessor (ie, paediatric nurses) and one clinically naive assessor (ie, nursing students who have not yet started their paediatric placements). To minimise recall bias, each assessor completed two separate testing sessions 4 weeks apart. PainChek Infant facial assessments were completed using version 1.3 (V17) installed onto an iPad Mini-4 (Cupertino, CA, USA, IOS version 13.6.1). The NFCS-R was incorporated into the EDMS so that facial actions observed could be entered whilst the assessors watched the video segments, and a slider bar was incorporated into the EDMS so that the assessor could directly enter their ObsVAS scores. Results obtained from the three pain assessment scales were entered into the EDMS. In case of PainChek Infant, after recording the data, the assessment was cancelled in the app so that the results were not retained on the device.

See Online for appendix

During the first testing session, the video segments, without audio, were loaded into the EDMS for the assessor to view and record their results. Assessors were required to complete their pain assessments using the tools in this order: NFCS-R single video viewing (NFCS-R single), NFCS-R multiple video viewing (NFCS-R multiple), ObsVAS, PainChek Infant video adaptive mode (PainChek Infant adaptive), and PainChek Infant standard 3 s video mode (PainChek Infant standard). This order was chosen to minimise bias that could occur by using automatic PainChek Infant before other tools that require assessor rating. However, the order in which the 30 subjects appeared in both testing session sets was randomly allocated by the research team. All four video segments of each of the 30 infants was assessed using the same tool, and once the assessment of all four video segments was completed for an individual infant and submitted into the EDMS, they could no longer be viewed by the assessor. This approach was adopted to reduce recall bias. After completing assessments of all the videos in their testing session set using one tool, the assessors repeated the assessments using the next allocated tool. To further avoid recall bias, the order in which video segments were presented for assessment was automatically and randomly assigned by the EDMS for each pain assessment method. In the second testing session, 4 weeks later, each assessor completed assessments for the same video segments as they had done in the first testing session, without access to their previous results. In the second testing session, assessors did not repeat the PainChek Infant assessments, rather these assessments were completed independently through a PainChek Infant simulator.

To ensure competency in the use of the pain assessment tools, assessors were required to complete training in the use of PainChek Infant, the NFCS-R,[12] and ObsVAS before data collection began. The assessors were also familiarised with the use of the EDMS.

Statistical analysis

The sample size was based on a minimum of five patients needed to evaluate every item included in the tool.[21] As the PainChek Infant facial domain has six items, a minimum sample size of 30 was necessary. For practical reasons, a sample of 40 was chosen to allow instances in which matched pain assessments could not be completed (eg, if the automated facial analysis failed after two attempts).

IBM SPSS Statistics version 26 was used for the data analysis, unless otherwise stated, with statistical significance set at $p<0·05$, two-tailed. PainChek Infant results were described using frequency and percentage. A binomial logit link generalised linear mixed model (GLMM) assessed each facial action (dependent outcome) with video segments (baseline, preparation, during, recovery), assessors (fixed factor), and infants treated as a random factor (see appendix p 1 for model specifications).

Pain and distress scores were not normally distributed, as assessed by a Kolmogorov-Smirnov normality test. Pain and distress scores were described using mean, SD, median, and IQR for all assessment methods and video segments. A gamma distribution with log link generalised estimating equation (GEE) model was used to examine any differences in pain scores across video repeats, controlling for repeated infant, video segment, assessor, and repeat occasion (see appendix p 1). For all GEE models, test of model effects (Wald [χ^2]), parameter estimates, and Bonferroni corrected pairwise comparisons were reported.

A change score was calculated to assess the responsiveness of the scales to change. The painful procedure change score was calculated for each method by computing the pain score during vaccination minus the baseline score. The non-painful procedure change score was calculated for each method by computing the pain score at preparation minus the baseline score. Painful and non-painful change scores were described using mean, SD, 95% bias corrected and accelerated (BCa) confidence intervals (CI). BCa CI were computed using 1000 samples with sampling stratified on assessors. We considered responsiveness shown if the change in scores was more than 2 points for PainChek Infant methods, or more than 33% for alternative methods, and the rationale for these boundaries is given in the appendix (p 1). A change in scores should not be seen for non-painful procedures. For each method, a linear mixed model (LMM) was used to examine the relationship of change in pain scores for painful and non-painful procedures. The LMMs examined pain score as a continuous outcome, with procedure (painful/non-painful) and assessors treated as fixed factors, and infants treated as a random factor (appendix p 1).

Responsiveness of the scale for pain was also evaluated by grouping infants by score at baseline or preparation phase using the PainChek Infant method into those with low pain scores (<3) and those with high pain scores (≥3). The grouping of infants into low and high pain scores using the ObsVAS and percentage NFCS-R and NFCS (mean score [ie, sum of items observed divided by sum of items visible] multiplied by 100) used a cut-point score of

50% or higher for the high pain score category and scores below 50% for the low pain score category. A LMM with pooled session data examined responsiveness using pain score as a continuous outcome, with fixed effects pain (high or low pain during baseline or preparation), procedure (painful or non-painful), session (first or second), assessment method, assessors, and infants treated as a random factor.

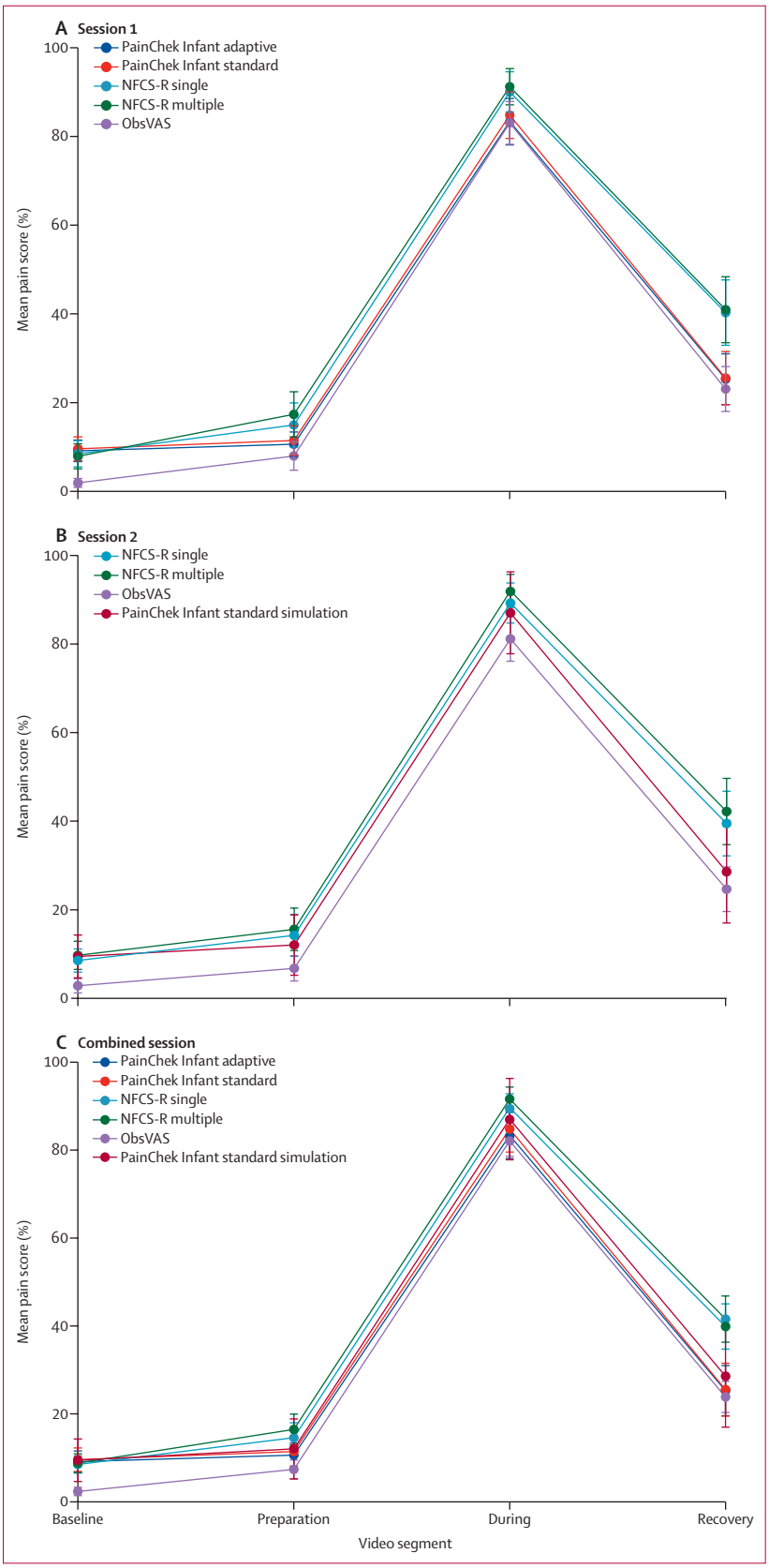
A repeated measures correlation was done to examine whether the scales were measuring the same construct, using the rmcorr function in RStudio version 1.3.1093 with 1000 repetition bootstrap 95% CIs reported. We considered scales with $r>0\cdot75$ as measuring the same construct. Repeated measures Bland-Altman plots[22] were used to assess the level of agreement between methods, using MedCalc software, version 19.5.3. This procedure is based on the calculations described by Zou.[23] We calculated Z scores from the first testing session, and compared Z scores of ObsVAS, NCFS-R multiple, and NCFS-R single viewing with PainChek Infant adaptive and PainChek Infant standard.

Pain score inter-rater differences were examined with ICC with 95% CI separately for testing session, method, and video segment. ICC investigations were also done separately for assessor type. ICC values were interpreted as poor (<0·50), moderate (≥0·50–0·75), good (>0·75–0·90), and excellent (>0·90).[24] Cronbach alpha was used to assess the internal consistency between the four assessors for each method with α greater than 0·70 considered acceptable. Pain scores (ObsVAS, NCFS-R multiple, and NCFS-R single) for the first and second testing sessions were combined (matching infant video, segment, method, and assessor) to examine intra-rater differences. Mean pain scores for PainChek Infant video adaptive mode, PainChek Infant standard 3 s video mode, and PainChek Infant standard 3 s video mode simulated

	First testing session				Second testing session			
	Baseline	Preparation	During	Recovery	Baseline	Preparation	During	Recovery
ObsVAS								
n	118	118	118	118	120	119	120	120
Mean	1·9%	8·0%	83·0%	23·1%	2·9%	6·8%	81·2%	24·6%
SD	5·5%	17·7%	26·4%	27·8%	8·8%	15·7%	27·9%	27·8%
Median	0%	0%	95%	14·5%	0%	0%	95%	15%
IQR	0–0%	0–10%	80–100%	0–40%	0–0%	0–10%	70–100%	0–40%
NFCS-R multiple*								
n	119	119	119	119	120	120	120	119
Mean	7·9%	17·4%	91·2%	40·9%	9·7%	15·6%	92·0%	42·2%
SD	15·5%	28·0%	22·5%	41·0%	17·6%	26·5%	20·5%	41·3%
Median	0%	0%	100%	25%	0%	0%	100%	25%
IQR	0–0%	0–25%	100–100%	0–80%	0–20%	0–20%	100–100%	0–80%
NFCS-R single*								
n	120	120	120	120	120	120	120	119
Mean	8·5%	15·0%	90·0%	40·3%	8·5%	14·2%	89·3%	39·5%
SD	16·4%	27·4%	25·1%	40·8%	14·4%	26·2%	25·0%	40·2%
Median	0%	0%	100%	25%	0%	0%	100%	20%
IQR	0–20%	0–20%	100–100%	0–80%	0–20%	0–20%	100–100%	0–80%
PainChek Infant adaptive†								
n	115	116	114	114	NA	NA	NA	NA
Mean	10·0%	10·0%	83·3%	25·0%	NA	NA	NA	NA
SD	13·3%	15·0%	28·3%	30·0%	NA	NA	NA	NA
Median	0%	0%	100%	17%	NA	NA	NA	NA
IQR	0–1%	0–1%	4–6%	0–2%	NA	NA	NA	NA
PainChek Infant standard†								
n	113	115	115	113	37	36	36	35%
Mean	10·0%	11·7%	85·0%	25·0%	10·0%	11·7%	86·7%	28·3%
SD	15·0%	18·3%	28·3%	31·7%	15·0%	20·0%	26·7%	33·3%
Median	0%	0%	100%	17%	0%	0%	100%	17%
IQR	0–1%	0–1%	5–6%	0–2%	0–1%	0–1%	5–6%	0–3%

n=number of assessments. *Scores ranged from 0 to 5 and are presented here as percentage of maximum score. Scores were calculated using only valid datapoints (ie, missing datapoints were ignored); however, if all facial features were not visible then the pain score was considered missing. †Scores ranged from 0 to 6 and are presented here as percentage of maximum score, and were simulated during the second testing session.

Table 1: **Description of pain and distress scores for each video segment, by assessment method, for the first and second testing sessions**



were combined (matching infant video, segment, and method) to examine intra-rater differences. All models used two-way mixed effects with absolute agreement.

Role of the funding source

The funder of the study is a previous or current employer of authors KH, JDH, and PTC who have had a role in the study design, data collection, data analysis, data interpretation, and writing of the report.

Results

Using 40 individual videos of White infants undergoing immunisation, 4303 pain assessments were completed in two separate sessions, starting on Aug 31, and Oct 19, 2020. Infants were aged 2·2 to 6·9 months, with a mean age of 3·6 months (SD 1·3) and a median age of 3·4 months (IQR 2·3–4·5). 24 (60%) of 40 infants were female. 2384 assessments were conducted using the five pain assessment methods during the first testing session, including 119 using NFCS-R multiple, 120 using NFCS-R single, 120 using PainChek Infant adaptive, 119 using PainChek Infant Standard, and 118 using ObsVAS. All assessors completed 30 assessments per video segment using each assessment method, apart from one assessor who only completed 29 assessments using NFCS-R multiple, 29 assessments using PainChek Infant Standard, and 28 assessments using ObsVAS Pain scores because this assessor failed to save the completed assessments in the EDMS before moving to the next set of video segments. During the second testing session, 1439 assessments were done using three methods (table 1), and an additional 160 assessments were done by simulation (ie, no assessor) independently from the main study using the PainChek Infant standard simulator.

The presence of facial action peaked during vaccination, and were slightly higher at recovery compared with baseline levels, according to PainChek Infant adaptive and PainChek Infant standard (first testing session), and PainChek Infant standard simulated (second testing session; figure 1; appendix p 2). The GLMM indicated that there was a significant difference between video segments for the presence of each of the facial actions ($p<0·001$; appendix p 2). In the first testing session, across all assessments using PainChek Infant adaptive and PainChek Infant standard, there were 41 (4%) failures out of a total of 956 recorded assessments (including 21 [4%] failures of 480 assessments using PainChek Infant adaptive and 20 [4%] failures of 476 assessments

Figure 1: **Pain score changes for each video segment by assessment method, during the first testing session (A), second testing session (B), and combined testing session scores (C)**
PainChek Infant adaptive and PainChek Infant standard (scored on a scale of 0–6) and NFCS-R single and NFCS-R multiple (scored on a scale of scale 0–5) are presented here as percentages. The error bars represent 95% CIs.
NFCS-R=Neonatal Facial Coding System-Revised. ObsVAS=Observer administered Visual Analogue Scale.

using PainChek Infant standard). Reported failures commonly related to poor video quality, posture of the child, or obstruction of the child by the nurse or parent. Excessive head movement contributed to failure in only five occasions. PainChek Infant standard simulated recorded 17 (11%) failures out of 160 segment assessments attempted. These failures were due to the automatic start of the assessment by the system, which meant that the assessor was not able to effectively judge the most suitable time to start the assessment. For failures, only the single segment datapoint was treated as missing, with remaining segment datapoints retained in the analysis. The GEE analysis reported a significant difference between video segments for the presence of facial action frequencies for NFCS-R single and NFCS-R multiple (appendix p 3).

The pain scores for each assessment method at each video segment are described in table 1 for both testing session one and session two. GEE models controlling for assessor and video segment did not detect a significant difference in pain scores across the two testing sessions, using either NFCS-R single (Wald χ^2=0·1; p=0·75), NFCS-R multiple (Wald χ^2=0·1; p=0·78), or ObsVAS (Wald χ^2=0·8; p=0·39). When the PainChek Infant standard simulated data pain scores were compared with the mean (across assessors) PainChek adaptive and PainChek standard scores controlling for video segment, no significant difference was detected between methods (Wald χ^2=1·3 p=0·51).

We examined the responsiveness of the scales using painful and non-painful change scores with LMM results described in table 2. All three PainChek Infant methods reported a clinically significant change (a change in score of more than 2 points) for the painful procedures, with the LMM confirming a significant effect for the responsiveness to a painful procedure of 4·4 (95% BCa CI 4·0–4·7) using the PainChek Infant adaptive and the PainChek Infant standard methods, and 4·6 (95% BCa CI 3·7–5·6) using the PainChek Infant standard simulated method. Similarly, the NFCS-R and ObsVAS methods also reported clinically significant (>33% change) and statistically significant responsiveness (table 2; figure 1). Responsiveness proportions for each method at baseline or preparation indicated a low frequency of infants with high pain or distress levels: PainChek Infant adaptive recognised nine (4%) of 231 infants having high baseline pain or distress levels, PainChek Infant standard recognised 13 (6%) of 238, NFCS-R single recognised 18 (8%) of 240 in the first testing session and 14 (6%) of 240 in the second, NFCS-R multiple recognised 18 (8%) of 238 in the first testing session and 18 (8%) of 240 in the second, ObsVAS recognised five (2%)of 236 in the first testing session and five (2%) of 239 in the second, and PainChek Infant standard simulation recognised six (8%) of 80. LMM did not detect a significant difference in the change in pain scores between those infants reporting low pre-vaccination pain or distress compared with those reporting high pre-vaccination pain or distress (F=1·6; p=0·211).

Repeated measures correlation reported significant correlations between PainChek Infant adaptive and the following methods: ObsVAS (r=0·88, 95% CI 0·85–0·90; p<0·0001), NFCS-R single (r=0·83, 95% CI 0·79–0·86; p<0·0001), and NFCS-R multiple (r=0·82, 95% CI 0·78–0·85; p<0·0001); and between PainChek Infant standard and these methods: ObsVAS (r=0·88, 95% CI 0·86–0·90; p<0·0001), NFCS-R single (r=0·82, 95% CI 0·79–0·85; p<0·0001), and NFCS-R multiple (r=0·83, 95% CI 0·79–0·85; p<0·0001; appendix p 4). Repeated

	Painful event					Non-painful event					Linear mixed model*		
	Mean	SD	95% BCa CI	Median	IQR	Mean	SD	95% BCa CI	Median	IQR	β estimate†	95% BCa CI	p value
First testing session													
ObsVAS	80·3	27·8	74·4 to 85·1	93·5	70 to 100	6·0	18·7	3·1 to 9·5	0·0	0 to 10	74·3	69·2 to 79·4	0·0010
NFCS-R multiple‡	83·0%	25·5%	78·1 to 87·4%	100·0%	75 to 100%	9·7%	31·0%	4·7 to 15·7%	0·0%	0 to 20%	73·4%	66·7 to 80·9%	0·0010
NFCS-R single‡	81·4%	28·4%	75·9 to 86·0%	100·0%	75 to 100%	6·4%	29·9%	1·5 to 11·1%	0·0%	0 to 5%	75·0%	68·8 to 82·0%	0·0010
PainChek Infant adaptive§	4·6	1·8	4·2 to 4·9	5·0	4 to 6	0·2	1·1	0·0 to 0·4	0·0	0 to 1	4·4	4·0 to 4·7	0·0010
PainChek Infant standard§	4·6	1·8	4·2 to 4·9	5·0	4 to 6	0·2	1·1	0·0 to 0·3	0·0	0 to 0	4·4	4·0 to 4·7	0·0010
Second testing session													
ObsVAS	78·3	29·1	73·2 to 83·3	94·0	60 to 100	3·9	17·4	1·0 to 6·9	0·0	0 to 5	74·4	68·7 to 79·8	<0·0001
NFCS-R multiple‡	82·3%	25·9%	77·9 to 86·6%	100·0%	60 to 100%	5·9%	29·1%	0·9 to 10·5%	0·0%	0 to 0%	76·4%	70·0 to 83·3%	<0·0001
NFCS-R single‡	80·7%	28·5%	75·4 to 85·3%	100·0%	60 to 100%	5·7%	28·0%	1·4 to 10·8%	0·0%	0 to 0%	75·0%	68·3 to 82·2%	<0·0001
PainChek Infant standard simulated§	4·8	1·7	4·2 to 5·5	6·0	5 to 6	0·2	1·4	–0·2 to 0·8	0·0	0 to 1	4·6	3·7 to 5·6	0·0020

BCa CI=bias corrected and accelerated confidence intervals. IQR=interquartile range. NFCS-R=Neonatal Facial Coding System-Revised. ObsVAS=Observer administered Visual Analogue Scale. *Linear mixed model includes random effect for infant and fixed effect for assessor with bootstrap estimates. †Compared with non-painful procedure in which parameter is set to 0. ‡Scores ranged from 0 to 5 and change in scores are presented here as percentages. §Scores range from 0 to 6. Scores were calculated using only valid datapoints (ie, missing datapoints were ignored), however; if all facial features were not visible then the pain score was considered missing.

Table 2: **Change in pain scores between painful and non-painful events for each assessment method**

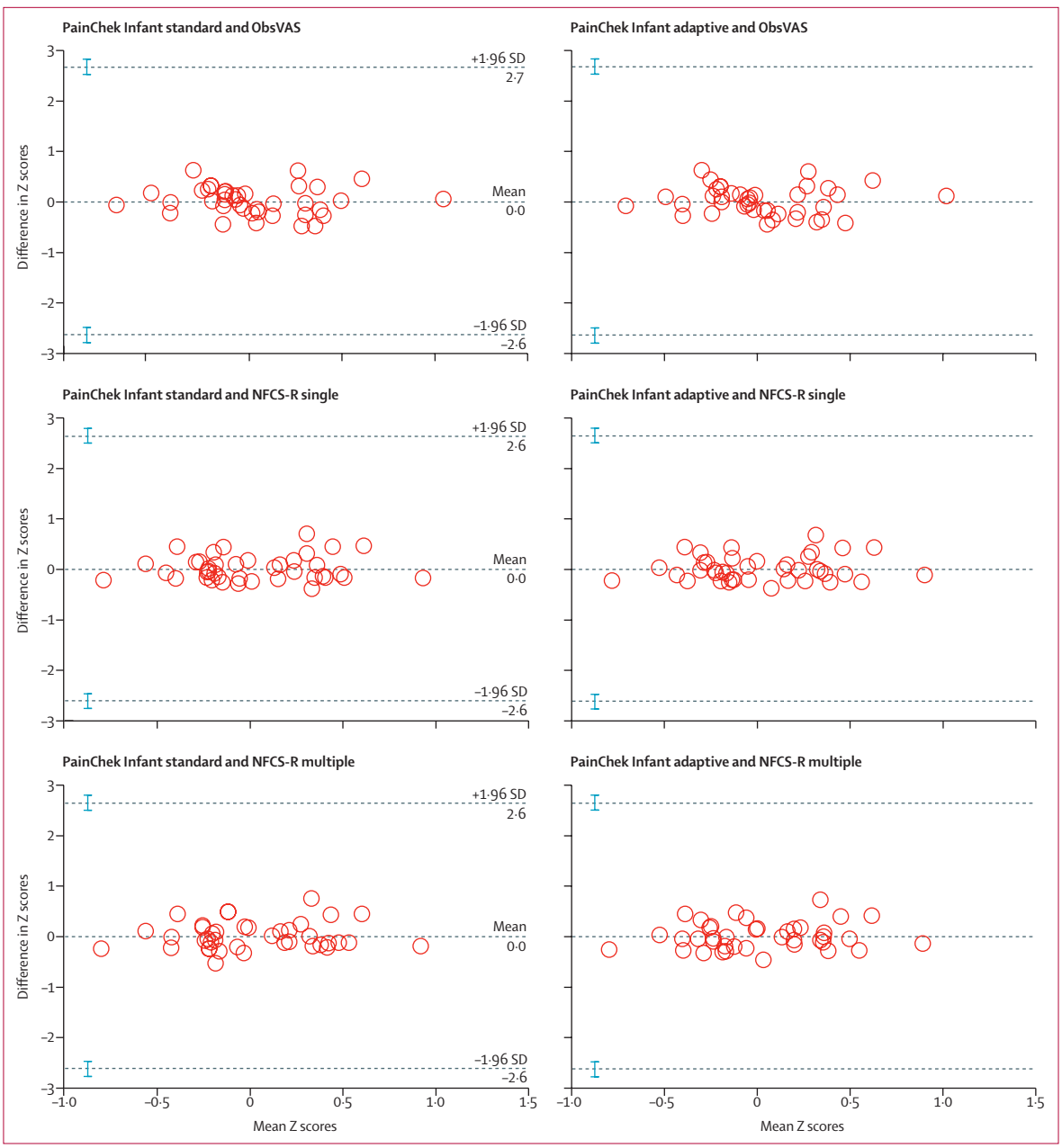


Figure 2: **Repeated measures Bland-Altman plots for both testing sessions**
The dashed lines represent the upper and lower limits of agreement, and the error bars the 95% CIs around the limits of agreement.

measures Bland-Altman plots for both testing sessions found no systematic differences between the compared measures (figure 2). For each comparison a mean difference of 0·02 was reported (except the comparison between NFCS-R multiple and PainChek Infant adaptive, which had a mean difference 0·01), all within the ±1·96 limits of agreement indicating agreement between methods (appendix p 5).

ICC results generally indicated moderate to excellent agreement across all assessors for each method and segment (table 3). Inter-rater reliability was typically better during the vaccination and recovery segments, ranging from moderate (NFCS-R multiple at second testing had an ICC 95% CI lower bound of 0·51), to excellent (ObsVAS and PainChek Infant adaptive at first testing had an ICC 95% CI upper bound of 0·99). Inter-rater reliability in assessing baseline and preparation segments ranged from poor (NFCS-R single at second use had an ICC 95% CI lower bound of 0·08) to excellent (ObsVAS at first use had an ICC 95% CI upper bound of 0·96). PainChek Infant adaptive and PainChek Infant standard consistently reported moderate to excellent (≥0·81) inter-rater

	First testing session				Second testing session			
	Baseline	Preparation	During	Recovery	Baseline	Preparation	During	Recovery
ObsVAS								
ICC	0·91	0·64	0·97	0·97	0·73	0·62	0·94	0·96
ICC 95%CI	0·81–0·96	0·30–0·84	0·94–0·99	0·93–0·99	0·48–0·88	0·29–0·83	0·88–0·97	0·93–0·99
p value	<0·0001	0·0001	<0·0001	<0·0001	<0·0001	0·0008	<0·0001	<0·0001
α	0·91	0·72	0·98	0·97	0·74	0·66	0·94	0·97
NFCS-R multiple								
ICC	0·80	0·83	0·91	0·96	0·58	0·76	0·75	0·94
ICC 95% CI	0·59–0·91	0·62–0·93	0·82–0·96	0·92–0·98	0·22–0·81	0·51–0·90	0·51–0·89	0·85–0·98
p value	<0·0001	<0·0001	<0·0001	<0·0001	0·0004	<0·0001	<0·0001	<0·0001
α	0·82	0·87	0·91	0·97	0·69	0·82	0·78	0·96
NFCS-R single								
ICC	0·67	0·58	0·93	0·93	0·48	0·66	0·89	0·93
ICC 95% CI	0·38–0·85	0·21–0·81	0·85–0·97	0·85–0·97	0·08–0·75	0·35–0·85	0·79–0·95	0·85–0·97
p-value	0·0001	0·0019	<0·0001	<0·0001	0·0057	0·0003	<0·0001	<0·0001
α	0·73	0·63	0·93	0·95	0·58	0·69	0·90	0·95
PainChek Infant adaptive								
ICC	0·81	0·83	0·97	0·92	NA	NA	NA	NA
ICC 95% CI	0·61–0·92	0·65–0·93	0·93–0·99	0·83–0·97	NA	NA	NA	NA
p value	<0·0001	<0·0001	<0·0001	<0·0001	NA	NA	NA	NA
α	0·82	0·85	0·97	0·92	NA	NA	NA	NA
PainChek Infant standard								
ICC	0·87	0·83	0·97	0·94	NA	NA	NA	NA
ICC 95% CI	0·72–0·95	0·66–0·93	0·93–0·99	0·88–0·98	NA	NA	NA	NA
p value	<0·0001	<0·0001	<0·0001	<0·0001	NA	NA	NA	NA
α	0·90	0·84	0·97	0·95	NA	NA	NA	NA

α=Cronbach alpha. ICC=intraclass correlation coefficient. NFCS-R=Neonatal Facial Coding System-Revised. ObsVAS=Observer administered Visual Analogue Scale.

Table 3: **Inter-rater reliability and internal consistency outcomes for each pain and distress method of assessment at each segment for both testing sessions**

	ObsVAS				NFCS-R multiple				NFCS-R single				PainChek Infant			
	ICC	ICC 95% CI	p value	α	ICC	ICC 95% CI	p value	α	ICC	ICC 95% CI	p value	α	ICC	ICC 95% CI	p value	α
Overall	0·98	0·98–0·99	<0·0001	0·98	0·97	0·96–0·97	<0·0001	0·97	0·96	0·96–0·97	<0·0001	0·96	0·99	0·98–0·99	<0·0001	0·99
Baseline	0·85	0·79–0·90	<0·0001	0·86	0·81	0·73–0·87	<0·0001	0·82	0·81	0·73–0·87	<0·0001	0·81	0·93	0·88–0·96	<0·0001	0·93
Preparation	0·95	0·93–0·97	<0·0001	0·95	0·90	0·85–0·93	<0·0001	0·90	0·86	0·80–0·90	<0·0001	0·86	0·84	0·72–0·91	<0·0001	0·84
During	0·94	0·91–0·96	<0·0001	0·94	0·87	0·81–0·91	<0·0001	0·87	0·94	0·92–0·96	<0·0001	0·94	0·98	0·97–0·99	<0·0001	0·98
Recovery	0·95	0·93–0·97	<0·0001	0·95	0·95	0·93–0·97	<0·0001	0·95	0·94	0·91–0·96	<0·0001	0·94	0·97	0·94–0·98	<0·0001	0·97

Intra-rater reliability was based on repeat measures from the first and second testing session for ObsVAS and NFCS-R, whereas intra-rater reliability for PainChek Infant was based on repeat measures from Painchek Infant adaptive (first testing session), Painchek Infant standard (first testing session), and Painchek Infant standard simulated (second testing session). α=Cronbach alpha. ICC=intraclass correlation coefficient. NFCS-R=Neonatal Facial Coding System-Revised. ObsVAS=Observer administered Visual Analogue Scale.

Table 4: **Intra-rater reliability and internal consistency outcomes for each pain and distress method of assessment at each segment**

agreement across all segments for the first session. Internal consistency between assessors (table 3) was acceptable for all methods and segments except for first testing when assessing the preparation segment for NFCS-R single (α=0·63); second testing when assessing baseline segment NFCS-R single (α=0·58) and NFCS-R multiple (α=0·69), and when assessing preparation segment for ObsVAS (α=0·66). PainChek Infant adaptive and PainChek Infant standard methods reported high values of internal consistency ranging from α=0·82 when assessing the baseline segment with PainChek Infant adaptive, to α=0·97 when assessing the during vaccination segment with PainChek Infant standard.

Intra-rater reliability between the first and second testing sessions overall and separated for each video segment was excellent for ObsVAS, NFCS-R multiple, and NFCS-R single (table 4). Comparison of the three PainChek Infant methods reported good-to-excellent intra-rater reliability both overall and separated for each video segment (table 4).

Discussion

This study evaluated the psychometric properties of PainChek Infant compared with the NFCS-R and ObsVAS, to determine its suitability of use in the assessment of procedural pain in infants. Both comparator tools rely on the user to observe and use their judgement related to the presence of pain or the intensity of the pain or distress being displayed, which could have an effect on how pain is treated. These user-reliance properties are common in other existing observational pain assessment tools. In comparison, PainChek Infant, resulting from advances in AI, uses automated facial expression analysis to automatically decode the face and determine the presence of six pain-related facial expressions. PainChek Infant analysis is completed without user subjectivity and in real time, in 3 s. Additionally, PainChek Infant is operated from a mobile device, therefore offering potential benefits in relation to workflow, documentation, and communication. Using these properties together, PainChek Infant presents an opportunity to improve pain assessment in infants, and could lead to changes in clinical practice that help to address the current challenges around making pain visible in the infant population group.[5] When compared with NFCS-R, PainChek Infant shows similar frequencies of recognition of facial actions common to the two scales, supporting the similarity of their construct. Furthermore, changes in total pain scores across the four video segments for PainChek (adaptive and standard), NFCS-R (single and multiple), and ObsVAS followed the same patterns (figure 1), in line with what would be anticipated for infants undergoing immunisation.

Differentiating pain from distress using currently available observational pain scales is difficult.[25,26] Here, we acknowledge the potential for overlap between pain and non-pain related distress on infants' facial expressions. Nonetheless, our findings are encouraging considering a clear change in facial expressions pre-needle and post-needle insertion, starting from the baseline through to the recovery. Our findings confirm the responsiveness of all three scales to pain intervention, as evidenced by clinically and statistically significant effect of the painful procedure. This significance is support by a study from 2012, in which facial expressions similar to those covered by PainChek Infant and NFCS-R were analysed in infants aged 2, 4, 6, and 12 months, for up to a minute post-needle insertion. The study showed that the expressions were associated with pain or pain-related distress, rather than other emotions such as sadness or anger.[27] Interestingly, AU25 (parting lips), as detected using PainChek Infant, and horizontal mouth opening as defined by NFCS-R, were the most common features detected in both pre-vaccination segments (ie, at baseline and preparation). These two segments occur before the painful stimuli (ie, the injection), and so it would be reasonable to assume that AU25 is not representative of pain-related distress. This finding could assist users in differentiating non-pain related facial expressions (neutral, positive, or negative) from those of pain, especially when no source of potential pain is suspected. However, AU25 has been shown to be a prominent feature of pain-related distress in infants undergoing immunisation across a range of ages, occurring together with AU3 or AU4 (brow lowered or pulled together), AU6 or AU7 (cheek raised or lower eyelid tightening), AU9, AU10, or AU11 (lip raise or nasolabial furrow), AU20, AU25, and AU26c, or AU27 (widely open cry mouth), and AU43 (eyes closed).[28] AU4, AU9, AU20, AU25, and AU43 are included in PainChek Infant together with AU15 (lip corner depressor), which was shown to be prominent during pain in the original 20 facial actions used to code the images and used to develop the PainChek Infant algorithms. A 2011 study also reported AU15's association with pain, as adjudged by the NFCS.[29] NFCS-R recognised facial actions also show concordance with these facial actions, specifically AU4, AU9, AU20, AU25, and AU43. Hence, both tools have clinical utility in verifying and quantifying pain when a source of pain is known or suspected.

PainChek Infant showed excellent convergent validity; results of PainChek Infant aligned with both NFCS-R and ObsVAS. PainChek Infant also exhibited moderate to excellent inter-rater reliability across all four video segments. Furthermore, its inter-rater reliability was comparable to, or better than, the comparator scales: NFCS-R and ObsVAS. Strong inter-rater reliability was particularly evident for assessments done at baseline and during the preparation phase. The internal consistency of the two modes of PainChek Infant assessment were also shown to be more than acceptable, with high α values. The internal consistency for NFCS-R after multiple viewings of the video segment were also high and comparable with that reported for the Modified Behavioural Pain Scale.[26] The inter-rater reliability of the three tools used was excellent (with overall ICC values being ≥0·96), as were the α scores.

Despite the strategies used to mitigate against various methodological limitations, some limitations remained. For one, there are challenges associated with evaluating the psychometric properties of scales in which a gold standard does not exist. Assessment is therefore dependent on the results from a range of indirect measures of validity, all of which have limitations. It is not possible to blind the assessors to the circumstances surrounding the infant, therefore potentially biasing assessors' scale application. To help overcome this potential bias, four assessors (of clinical and non-clinical background) were used and, additionally, the EDMS employed in the study automatically muted all videos to ensure that sound did not affect the assessors' scores. Assessors were also broadly aware of the purpose of the study, and although specific details and hypotheses were not revealed, this could have influenced their application of the scales. Also, establishing the validity of one

measure on the basis of correlation with another might rely on circular logic, hence the use of multiple assessment methods to establish scale validity. To further establish the validity of our findings, we used PainChek Infant in both standard and adaptive modes, therefore conducting assessments using a fixed video duration and a minimum number of valid images, and we also used NFCS-R in single view and multiple viewing scenarios.

With respect to examining the responsiveness of the assessments, we found a very low frequency of high pain or distress levels at baseline or preparation at both testing sessions, and the data were insufficient for statistical examination. Although data were pooled to allow responsiveness to be examined, the relatively low sample, combined with a non-significant result, suggests that responsiveness should be further examined with a larger sample to confirm responsiveness of the scale. Future work is also required to evaluate the specificity of the tool, and differences in facial expressions in response to similarly intense painful and non-painful distress stimuli should be compared; this was beyond the scope of the current study.

This study is also limited by the fact that assessors did not face the infants directly, but did their assessments by viewing video recordings. Although this method of assessing pain differs from clinical practice, the use of video recordings to validate pain assessment has been successfully done before.[8,9,25,26] Also, video recording allows multiple viewings of the segments (such as in the case of NFCS-R and ObsVAS), which has been considered a strength.[27] Furthermore, we showed that by using the PainChek Infant simulator, the results obtained by pointing the device at the computer screen were comparable to those obtained by presenting the videos directly to the software development kit (ie, real world processing). Several failures were recorded with the simulator system; however, considering that in real world use the assessor decides on the best time to initiate assessment (eg, at a time when the infant's face is not obstructed) these failures would likely not occur in the real world. Every effort has been made to minimise the risk of failure of the automated facial analysis through innovative design, inbuilt alerts, and user training; however, it is acknowledged that failure still might occur, and, dependent on the results of trials in clinical practice, a manual assessment option might need to be added.

The algorithms incorporated in PainChek Infant have been trained and validated using images of infants aged 1–12 months. As such, use of the tool on younger or older children is yet to be tested. Furthermore, the infants included in the evaluation were all White, and this must be considered when using it with infants of other races. However, it should be noted that much has been written about the universality of facial expressions, and existence of common pain expressions.[30] Additionally, we acknowledge that facial expressions in some infants might be absent due to conditions associated with facial palsy. In these cases, multidimensional scales would be preferred.

In this study we showed that PainChek Infant has good-to-excellent validity, reliability, and internal consistency when compared with the NFCS-R and the ObsVAS. In taking only 3 s to complete, removing observer bias associated with high exposure to facial expressions of pain[29] and automating pain assessment, we believe PainChek Infant represents a meaningful advance in the assessment and monitoring of procedural pain in infants. Still, further research is required to evaluate its clinical utility in clinical practice.

Contributors

JDH and KH conceived the study and contributed to the methodological design and data collection. All authors contributed to data interpretation, writing, and reviewing the manuscript. KH and PTC verified the data and contributed equally to this work. PTC did the data analysis. All authors had access to all the data in the study and had final responsibility for the decision to submit the research for publication.

Declaration of interests

KH and JDH are shareholders in PainChek (formerly known as EPAT Technologies), which is commercialising the PainChek Infant. They are also named as coinventors with Mustafa Atee on the patent entitled a pain assessment method and system (patent granted in Australia [AU2015306075B2], Japan [JP6657216B2], USA [US10398372B2], and China [CN106572820A]. Patent pending from Europe [EP3182893A4], and from the World Intellectual Property Organization [WO2016025989A1]). KH is employed as a consultant by PainChek, while also serving as an associate Professor at the University of Prishtina, Kosovo and university associate at the Curtin Medical School, Curtin University, WA, Australia. JDH is employed as the chief scientific officer of PainChek and holds an adjunct Professor appointment in the Curtin Medical School, Curtin University, WA, Australia. PTC was engaged through DATaR Consulting and was paid as an independent private consultant to undertake the biostatistical analysis for the project by PainChek. PTC also holds adjunct appointments at the Institute for Health Research, the University of Notre Dame Australia, Fremantle, WA, Australia, and School of Medical and Health Sciences, Edith Cowan University, Joondalup, WA, Australia.

Data sharing

After signing a data use agreement, deidentified data that do not violate confidentiality can be made available upon reasonable request from the date of publication. The agreement will specify that the deidentified data can only be used for research purposes, and not for product related work, and cannot be shared with a third party. Additionally, related documents can also be made available (eg, study protocol, statistical analysis plan) on reasonable request; please contact the corresponding author. The software and source code are the property of PainChek and, as such, any enquiry for access to these should be made directly to PainChek directly.

References

1 Czarnecki ML, Turner HN, Collins PM, Doellman D, Wrona S, Reynolds J. Procedural pain management: a position statement with clinical practice recommendations. *Pain Manag Nurs* 2011; **12:** 95–111.

2 Wilson-Smith EM. Procedural pain management in neonates, infants and children. *Rev Pain* 2011; **5:** 4–12.

3 Shave K, Ali S, Scott SD, Hartling L. Procedural pain in children: a qualitative study of caregiver experiences and information needs. *BMC Pediatr* 2018; **18:** 324.

4 Pancekauskaitė G, Jankauskaitė L. Paediatric pain medicine: pain differences, recognition and coping acute procedural pain in paediatric emergency room. *Medicina (Kaunas)* 2018; **54:** 94.

5 Eccleston C, Fisher E, Howard RF, et al. Delivering transformative action in paediatric pain: a Lancet Child & Adolescent Health Commission. *Lancet Child Adolesc Health* 2021; **5:** 47–87.

6 Bradford JY, Stapleton SJ, Horigan A, et al. Needle-related or minor procedural pain in pediatric patients. *J Emerg Nurs* 2019; **45:** 437.e1–32.

7 Herr K, Coyne PJ, Ely E, Gélinas C, Manworren RCB. Pain assessment in the patient unable to self-report: clinical practice. Recommendations in support of the ASPMN 2019 Position Statement. *Pain Manag Nurs* 2019; **20:** 404–17.

8 Chen Z, Ansari R, Wilkie. Automated pain detection from facial expressions using FACS: a review. *arXiv* 2018; published online Nov 13. https://arxiv.org/abs/1811.07988 (preprint).

9 Zamzmi G, Kasturi R, Goldgof D, Zhi R, Ashmeade T, Sun Y. A review of automated pain assessment in infants: features, classification tasks, and databases. *IEEE Rev Biomed Eng* 2018; **11:** 77–96.

10 Grunau RVE, Craig KD. Pain expression in neonates: facial action and cry. *Pain* 1987; **28:** 395–410.

11 Oster H. Baby FACS: facial action coding system for infants and young children. Monograph and coding manual. 2010. New York, NY, USA.

12 Grunau RE, Craig KD. Neonatal facial coding system revised: training manual. 2010. University of British Columbia, Vancouver, Canada.

13 Grunau RVE, Johnston CC, Craig KD. Neonatal facial and cry responses to invasive and non-invasive procedures. *Pain* 1990; **42:** 295–305.

14 Peters JW, Koot HM, Grunau RE, et al. Neonatal Facial Coding System for assessing postoperative pain in infants: item reduction is valid and feasible. *Clin J Pain* 2003; **19:** 353–63.

15 Johnston CC, Stevens B, Yang F, Horton L. Developmental changes in response to heelstick in preterm infants: a prospective cohort study. *Dev Med Child Neurol* 1996; **38:** 438–45.

16 Craig KD, Whitfield MF, Grunau RVE, Linton J, Hadjistavropoulos HD. Pain in the preterm neonate: behavioural and physiological indices. *Pain* 1993; **52:** 287–99.

17 Grunau RE, Oberlander T, Holsti L, Whitfield MF. Bedside application of the Neonatal Facial Coding System in pain assessment of premature neonates. *Pain* 1998; **76:** 277–86.

18 Craig KD, Hadjistavropoulos HD, Grunau RVE, Whitfield MF. A comparison of two measures of facial activity during pain in the newborn child. *J Pediatr Psychol* 1994; **19:** 305–18.

19 Lilley CM, Craig KD, Grunau RE. The expression of pain in infants and toddlers: developmental changes in facial action. *Pain* 1997; **72:** 161–70.

20 Taddio A, O'Brien L, Ipp M, Stephens D, Goldbach M, Koren G. Reliability and validity of observer ratings of pain using the visual analog scale (VAS) in infants undergoing immunization injections. *Pain* 2009; **147:** 141–46.

21 Herr K, Bjoro K, Decker S. Tools for assessment of pain in nonverbal older adults with dementia: a state-of-the-science review. *J Pain Symptom Manage* 2006; **31:** 170–92.

22 Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat* 2007; **17:** 571–82.

23 Zou GY. Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. *Stat Methods Med Res* 2013; **22:** 630–42.

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

25 Crellin DJ, Harrison D, Santamaria N, Huque H, Babl FE. The psychometric properties of the FLACC scale used to assess procedural pain. *J Pain* 2018; **19:** 862–72.

26 Crellin DJ, Babl FE, Santamaria N, Harrison D. The psychometric properties of the MBPS scale used to assess procedural pain. *J Pain* 2018; **19:** 660–69.

27 Kohut SA, Riddell RP, Flora DB, Oster H. A longitudinal analysis of the development of infant facial expressions in response to acute pain: immediate and regulatory expressions. *Pain* 2012; **153:** 2458–65.

28 O'Neill MC, Ahola Kohut S, Pillai Riddell R, Oster H. Age-related differences in the acute pain facial expression during infancy. *Eur J Pain* 2019; **23:** 1596–607.

29 Prkachin KM. Facial pain expression. *Pain Manag (Lond)* 2011; **1:** 367–76.

30 Kappesser J. The facial expression of pain in humans considered from a social perspective. *Philos Trans R Soc Lond B Biol Sci* 2019; **374:** 20190284.