BMJ Open

Current state of science in machine learning methods for automatic infant pain evaluation using facial expression information: study protocol of a systematic review and meta-analysis

Dan Cheng,[1,2] Dianbo Liu,[3] Lisa Liang Philpotts,[4] Dana P Turner,[2] Timothy T Houle,[2] Lucy Chen,[2] Miaomiao Zhang,[5] Jianjun Yang,[1] Wei Zhang,[1] Hao Deng [2,6]

To cite: Cheng D, Liu D, Philpotts LL, *et al*. Current state of science in machine learning methods for automatic infant pain evaluation using facial expression information: study protocol of a systematic review and meta-analysis. *BMJ Open* 2019;**9**:e030482. doi:10.1136/bmjopen-2019-030482

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-030482).

Received 16 March 2019
Revised 18 November 2019
Accepted 18 November 2019

For numbered affiliations see end of article.

Correspondence to
Dr Hao Deng;
hdeng1@mgh.harvard.edu

Professor Wei Zhang;
zhangw571012@126.com

ABSTRACT

Introduction Infants can experience pain similar to adults, and improperly controlled pain stimuli could have a long-term adverse impact on their cognitive and neurological function development. The biggest challenge of achieving good infant pain control is obtaining objective pain assessment when direct communication is lacking. For years, computer scientists have developed many different facial expression-centred machine learning (ML) methods for automatic infant pain assessment. Many of these ML algorithms showed rather satisfactory performance and have demonstrated good potential to be further enhanced for implementation in real-world clinical settings. To date, there is no prior research that has systematically summarised and compared the performance of these ML algorithms. Our proposed meta-analysis will provide the first comprehensive evidence on this topic to guide further ML algorithm development and clinical implementation.

Methods and analysis We will search four major public electronic medical and computer science databases including Web of Science, PubMed, Embase and IEEE Xplore Digital Library from January 2008 to present. All the articles will be imported into the Covidence platform for study eligibility screening and inclusion. Study-level extracted data will be stored in the Systematic Review Data Repository online platform. The primary outcome will be the prediction accuracy of the ML model. The secondary outcomes will be model utility measures including generalisability, interpretability and computational efficiency. All extracted outcome data will be imported into RevMan V.5.2.1 software and R V3.3.2 for analysis. Risk of bias will be summarised using the latest Prediction Model Study Risk of Bias Assessment Tool.

Ethics and dissemination This systematic review and meta-analysis will only use study-level data from public databases, thus formal ethical approval is not required. The results will be disseminated in the form of an official publication in a peer-reviewed journal and/or presentation at relevant conferences.

PROSPERO registration number CRD42019118784.

Strengths and limitations of this study

- This research will be the first systematic review and meta-analysis quantitatively comparing accuracy among current facial expression-based machine learning neonatal and infant pain prediction algorithms.
- This research will be the first study to comprehensively assess accuracy, generalisability, interpretability and computational efficiency of these machine learning algorithms by both model characteristics (eg, algorithm type, task type, data input format, etc) and patient/procedure characteristics (eg, age, disease type, procedure type, etc).
- The major limitation of this study will be the potential clustering effects due to the limited number of infant facial expression databases collected in real-world clinical environments or controlled trials available for modelling research.
- Many of these machine learning prediction studies will use the same database for modelling, and we will use mixed-effects meta-regression to address this clustering effect.

INTRODUCTION

Neonates and infants feel and experience pain much like or even more sensitively than adults, but their pain management is underemphasised even for today.[1 2] Historically, some clinicians and scientists believed that babies' brains are not fully developed enough to experience pain as adults do. Combined with the fear of usage of pain medications in infants due to potential side effects, pain relief treatments are inadequately provided to infants and neonates receiving painful procedures despite the fact that these treatments are provided to elder children and adults every day.[3–7] However, previous research has found that individuals' pain experience during infancy can have both short-term

and long-term adverse influences on their cognitive and neurological function development.[8–11] Therefore, appropriate pain relief for infants is a necessity, and there is a need for development of a systematic guideline for infant and neonatal pain management.

The biggest challenge to developing an infant and neonatal pain management guideline is the accurate measurement of pain. Infant pain is known to be difficult to evaluate because infants cannot accurately assess and verbally communicate their pain experience. The current evaluation system heavily relies on observer-based measurements obtained by trained clinicians and nurses. These observer-based assessment tools commonly integrate both physiological variations (eg, breathing pattern, heart rate, oxygen saturation, etc) and behavioural changes (eg, facial expression, crying, body activity, sleeping state, etc), and a summarised score is calculated based on clinical observations and physiological monitoring results.[12–14] In the last decades, several indicator-based pain scale systems have been developed and validated for infant and neonatal pain assessment.[15–20] Nevertheless, these pain scale-based systems have two major limitations: subjectivity and human resource expense. Because pain measures are obtained by human observers, the pain assessment process is inevitably subjective and suffers from observer bias.[21 22] Depending on the practitioner's clinical experience, training level and even spontaneous or personalised factors, their assessment results can be vastly different and inaccurate, which could be either overly sensitive or fail to discriminate between different levels of pain or pain from normal physiological activities (eg, discomfort, hunger, stress, etc). In addition, these tools are human-resource intensive and require a considerable amount of training to master. Under scenarios when continuous frequent pain monitoring is needed (eg, postoperation, necrotising enterocolitis, abdominal colic, epidermolysis bullosa, etc), the traditional human observer-based assessment method is inefficient and resource-intensive.

In the emerging field of computer vision and emotion research, human facial expression, a promising behavioural indicator for emotion and pain recognition, has been studied and used for automated infant pain assessment for years.[23] Modern data scientists can use state-of-the-art machine learning (ML) algorithms to extract, track and analyse human facial expressions from recorded images and video data to predict pain and emotions. Compared with traditional observer-based pain scale systems, the ML algorithm has a unique opportunity to provide objective and continuous pain detection and assessments for neonates and infants. The first published ML-based infant pain evaluation study, known as the Classification of Pain Expressions (COPE) project, collected facial image data in four routine clinical procedures and used three different face classifiers (principal component analysis, linear discriminant analysis and support vector machine) to recognise infant painful expressions.[24] Since then, researchers have developed several different ML methods (eg, Probabilistic Neural Network, Gaussian, Nearest Mean, Convolutional Neural Networks, Boosted Gabor Features, etc) to classify pain/no-pain images using COPE or similar databases, and many have shown satisfactory performance.[25–28] In recent years, there have also been several modern projects directly analysing infant facial video sequences using deep learning methods to recognise and assess pain, which have also presented considerable high accuracy.[29–31] ML-based automated pain assessment algorithms have the potential to control observer/recording bias from human observations, help reduce the training costs and relieve human resource burden when continuous or high-frequency clinical pain measurements are required. However, no research has quantitatively analysed and compared the performance of these pain prediction methods. In addition, these computational algorithms are rarely published in general or professional medical journals, therefore their clinical utility and interpretability are limited. In our proposed project, we will conduct a systematic review and meta-analysis to provide summarised and quantitative evidence on the performance of facial expression-based ML methods for infant pain assessment.

Objectives

The primary objective of our study is to assess the accuracy (outcome) of automatic facial expression ML algorithms (test measurement method) compared with various indicator-based pain assessments (gold-standard) in assessing pain intensity for infants experiencing pain (population). Additionally, we plan to perform subgroup analyses to compare model accuracy, generalisability, interpretability and computational efficiency by both model and study population characteristics.

Methods and analysis

Our study team prepared this study protocol following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols checklist.[32]

Research question development (PICO/PECO)

Our research question was developed based on the PICO/PECO research framework.[33] Details are reported in table 1.

Eligibility criteria

Our proposed project will systematically search and include all eligible ML methodological and application studies that use facial information to automatically assess pain among infants. The population of our study will be infants experiencing pain. Infants will be defined as young children no more than 12 months, including newborn or neonate, full term, premature and postmature infants. Infant pain can be heel stick, arterial puncture, intravenous cannula, finger stick, nasal aspiration or postoperation pain. We intend to include computer science algorithms papers (methodology and performance evaluation), clinical research (application studies), systematic reviews and meta-analysis research for analysis. Regular

Table 1 PICO/PECO research question development

Name	Description
Population	Infants experiencing pain.
Intervention/ Exposure	The intervention/exposure will be pain assessment using automatic facial recognition ML algorithms.
Control	The study control will be indicator-based pain assessment gold-standard (eg, pain scale, pain score and category).
Outcome	Primary outcome: Model accuracy by predicted assessment measures type: 1. Numeric score: mean SE or equivalence; 2. Categorical pain degree (yes/no; no/moderate/severe): Concordance statistic (AUC ROC) or equivalence. Secondary outcomes: 1. Generalisability; 2. Interpretability; 3. Computational efficiency and related costs.

Definition of infants: infants will be defined as young children no more than 12 months, including newborn or neonate, full term, premature and postmature infants.
AUC ROC, area under the curve for receiver operating characteristic curve.

reviews and qualitative studies will be excluded from the analysis, but their reference lists will be screened to identify potential eligible studies. Systematic review and meta-analysis will be used for the extraction of citations for reducing publication bias. The study exclusion criteria will include: (1) algorithm not for automatic infant pain assessment; (2) algorithm not using facial expression information; (3) not computer science algorithms paper (methodology and performance evaluation), clinical research (application studies) or systematic reviews and meta-analysis concerning ML methods; (4) no measurement of algorithm accuracy (primary outcome); (5) facial expression data not in image or video format; (6) children more than >12 months and adults; (7) research not written in English language.

Search strategy

An experienced medical librarian (LLP) with systematic review expertise will conduct searches in four major public electronic medical and computer science databases including Web of Science, PubMed, Embase and IEEE Xplore Digital Library from January 2008 to present. We will search from 2008 onwards because to the best of our knowledge, major advances in ML techniques, especially for deep learning methods (eg, Convolutional Neural Network), started to rapidly evolve and be widely applied within the most recent 5–6 years due to the unprecedented functional improvement of hardware (eg, GPU for computing) and parallel-computing capacity (eg, Hadoop).[34 35] Thus, algorithms developed 10 or 20 years ago may not be applicable or clinically meaningful because new modern algorithms would easily outperform them. In our study, we intend to include a balanced collection of both classic and modern algorithms, so we have decided to search for studies starting from January 2008 to assure an extended 10-year long search period for good coverage of ML studies. To account for publication bias, we will include qualitative review articles and systematic review and meta-analysis articles to identify missing unpublished literature through their reference lists. We will perform forward and backward citation screening through citations and reference lists of systematic reviews to find more relevant papers. We will also search related professional meeting abstracts and preprints (eg, IEEE conferences, Conference on Computer Vision and Pattern Recognition, Conference on Neural Information Processing Systems, Topics and Advances in Pediatrics, Florida Academy of Pain Medicine, 2018 and 2019 Annual Scientific Meeting, arXiv.org). We will use keywords and subject headings related to three concepts: infants, pain and ML. Multiple synonyms for each concept will be incorporated into the search. Details of search strategies are provided in table 2.

Study selection

Two authors (DL and DC) will independently review and screen searched article records to identify eligible studies according to our inclusion and exclusion criteria using the *Covidence* digital platform. A third investigator (HD or WZ) will resolve the disagreements between these two evaluators. In this step, two authors will screen titles and abstracts of searched articles for primary exclusion on the *Covidence* platform. We will pilot the process by screening the first 10 studies under the principal investigator (PI)'s supervision. Since our study topic is uniquely focused on ML algorithms on infant pain prediction, it will require expertise in two fields including computer science and medicine. We believe that it will be preferable and time-efficient to have both the computer scientist and the physician work together to review the full-text to avoid selection bias (ie, only read the sections they are familiar with) in the full-text screening process. It would be ideal to have two independent groups of specialists (two computer scientist-physician pairs as review teams), but our research team has only one data scientist (DL). Therefore, in our study, two authors DL (computer scientist) and DC (physician) will screen the full-text together to decide the eligibility of included studies after the title and abstract screening round. The PI (HD) will resolve the conflicts or answer questions when needed. Moreover, because many health-related computer science studies are reported with limited details,[36] it is difficult for a single field specialist (either a computer scientist or a physician) to understand and extract the necessary information alone. We will further discuss this issue in the 'Discussion' section. Once eligible studies are identified, we will extract study information for data synthesis. The measures in the identified studies contain both quantitative and qualitative data, thus we will perform quantitative data synthesis or qualitative summary as appropriate. Excluded studies will be listed in the PRISMA flow chart with specific reasons for exclusion in figure 1.

Table 2 Search strategy

PubMed	
#1	'Infant'(Mesh)
#2	(infant*(Title/Abstract)OR neonat*(Title/Abstract)OR baby(Title/Abstract)OR babies(Title/Abstract)OR newborn*(Title/Abstract))
#3	#1 OR #2
#4	'Pain'(Mesh) OR 'Pain Measurement'(Mesh)
#5	(pain*(Title/Abstract)OR hurt*(Title/Abstract)OR agony(Title/Abstract)OR agonising(Title/Abstract)OR agonising(Title/Abstract)OR suffer*(Title/Abstract)OR distress*(Title/Abstract))
#6	#4 OR #5
#7	('Machine Learning'(Mesh) OR 'Algorithms'(Mesh:NoExp)OR 'Expert Systems'(Mesh) OR 'Limit of Detection'(Mesh) OR 'Artificial Intelligence"(Mesh:NoExp)OR 'Neural Networks (Computer)"(Mesh) OR 'Facial Recognition"(Mesh) OR 'Biometric Identification"(Mesh:NoExp)OR 'Facial Expression"(Mesh:NoExp)OR 'pattern recognition, automated"(mesh))
#8	('facial recognition'(title/abstract)OR 'pain recognition'(title/abstract)OR 'pain detection'(Title/Abstract)OR 'detecting pain'(title/abstract) OR automated(Title/Abstract)OR automatic(title/abstract)OR 'recognising pain'(title/abstract)OR 'machine learning'(Title/Abstract)OR 'deep learning'(Title/Abstract)OR algorithm*(Title/Abstract)OR 'neural network'(Title/Abstract)OR 'neural networks'(Title/Abstract)OR SVM(Title/Abstract)OR 'support vector machine'(Title/Abstract)OR 'support vector machines'(Title/Abstract)OR 'computer vision'(Title/Abstract)OR 'artificial intelligence'(Title/Abstract)OR RVM(Title/Abstract)OR 'relevance vector machine'(Title/Abstract)OR 'relevance vector machines'(Title/Abstract)OR AAM(Title/Abstract)OR 'active appearance model'(Title/Abstract)OR 'active appearance models'(Title/Abstract)OR 'K NN'(Title/Abstract)OR 'k nearest neighbour'(Title/Abstract)OR 'random forest trees'(Title/Abstract)OR 'random forest tree'(Title/Abstract)OR PNN(Title/Abstract)OR 'gaussian classifier'(Title/Abstract)OR 'gaussian classifiers'(Title/Abstract)OR 'nearest mean classifier'(Title/Abstract)OR 'nearest mean classifiers'(Title/Abstract))
#9	#7 OR #8
#10	#3 AND #6 AND #9 AND ("2008/01/01"(PDAT) : '3000/12/31'(PDAT))
Embase	
#1	'infant'/exp
#2	infant*:ab,ti OR neonat*:ab,ti OR baby:ab,ti OR babies:ab,ti OR newborn*:ab,ti
#3	#1 OR #2
#4	'pain'/exp OR 'pain measurement'/de
#5	pain*:ab,ti OR hurt*:ab,ti OR agony:ab,ti OR agonising:ab,ti OR agonising:ab,ti OR suffer*:ab,ti OR distress*:ab,ti
#6	#4 OR #5
#7	'algorithm'/de OR 'machine learning'/exp OR 'expert system'/de OR 'limit of detection'/exp OR 'artificial intelligence'/exp OR 'pattern recognition'/exp
#8	'facial recognition':ab,ti OR 'pain recognition':ab,ti OR 'pain detection':ab,ti OR 'detecting pain':ab,ti OR automated:ab,ti OR automatic:ab,ti OR 'recognising pain':ab,ti OR 'machine learning':ab,ti OR 'deep learning':ab,ti OR algorithm*:ab,ti OR 'neural network':ab,ti OR 'neural networks':ab,ti OR svm:ab,ti OR 'support vector machine':ab,ti OR 'support vector machines':ab,ti OR 'computer vision':ab,ti OR 'artificial intelligence':ab,ti OR rvm:ab,ti OR 'relevance vector machine':ab,ti OR 'relevance vector machines':ab,ti OR aam:ab,ti OR 'active appearance model':ab,ti OR 'active appearance models':ab,ti OR 'k nn':ab,ti OR 'k nearest neighbour':ab,ti OR 'random forest trees':ab,ti OR 'random forest tree':ab,ti OR pnn:ab,ti OR 'gaussian classifier':ab,ti OR 'gaussian classifiers':ab,ti OR 'nearest mean classifier':ab,ti OR 'nearest mean classifiers':ab,ti
#9	#7 OR #8
#10	#3 AND #6 AND #9 AND(2008–2019)/py
IEEE	2008 to present (infant* OR neonat* OR baby OR babies OR newborn*) AND (pain* OR hurt OR hurts OR hurting OR agony OR agonising OR agonising OR suffer OR suffering OR suffers OR suffered OR distress*)
Web of Science	
#1	TOPIC: (infant* or neonat* or baby or babies or newborn*)
#2	TOPIC: ('facial recognition' OR 'pain recognition' OR 'pain detection' OR 'detecting pain' OR automated OR automatic OR 'recognising pain' OR 'machine learning' OR 'deep learning' OR algorithm* OR 'neural network' OR 'neural networks' OR SVM OR 'support vector machine' OR 'support vector machines' OR 'computer vision' OR 'artificial intelligence' OR RVM OR 'relevance vector machine' OR 'relevance vector machines' OR AAM OR 'active appearance model' OR 'active appearance models' OR 'K NN' OR 'k nearest neighbour' OR 'random forest trees' OR 'random forest tree' OR PNN OR 'gaussian classifier' OR 'gaussian classifiers' OR 'nearest mean classifier' OR 'nearest mean classifiers')
#3	TOPIC: (pain* or hurt* or agony or agonising or agonising or suffer* or distress*)
#4	#3 AND #2 AND #1 Refined by: PUBLICATION YEARS: (2019 OR 2010 OR 2018 OR 2009 OR 2017 OR 2008 OR 2016 OR 2015 OR 2014 OR 2013 OR 2012 OR 2011)Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC.

AAM, active appearance model; PNN, probabilistic neural network; RVM, relevance vector machine; SVM, support vector machine.

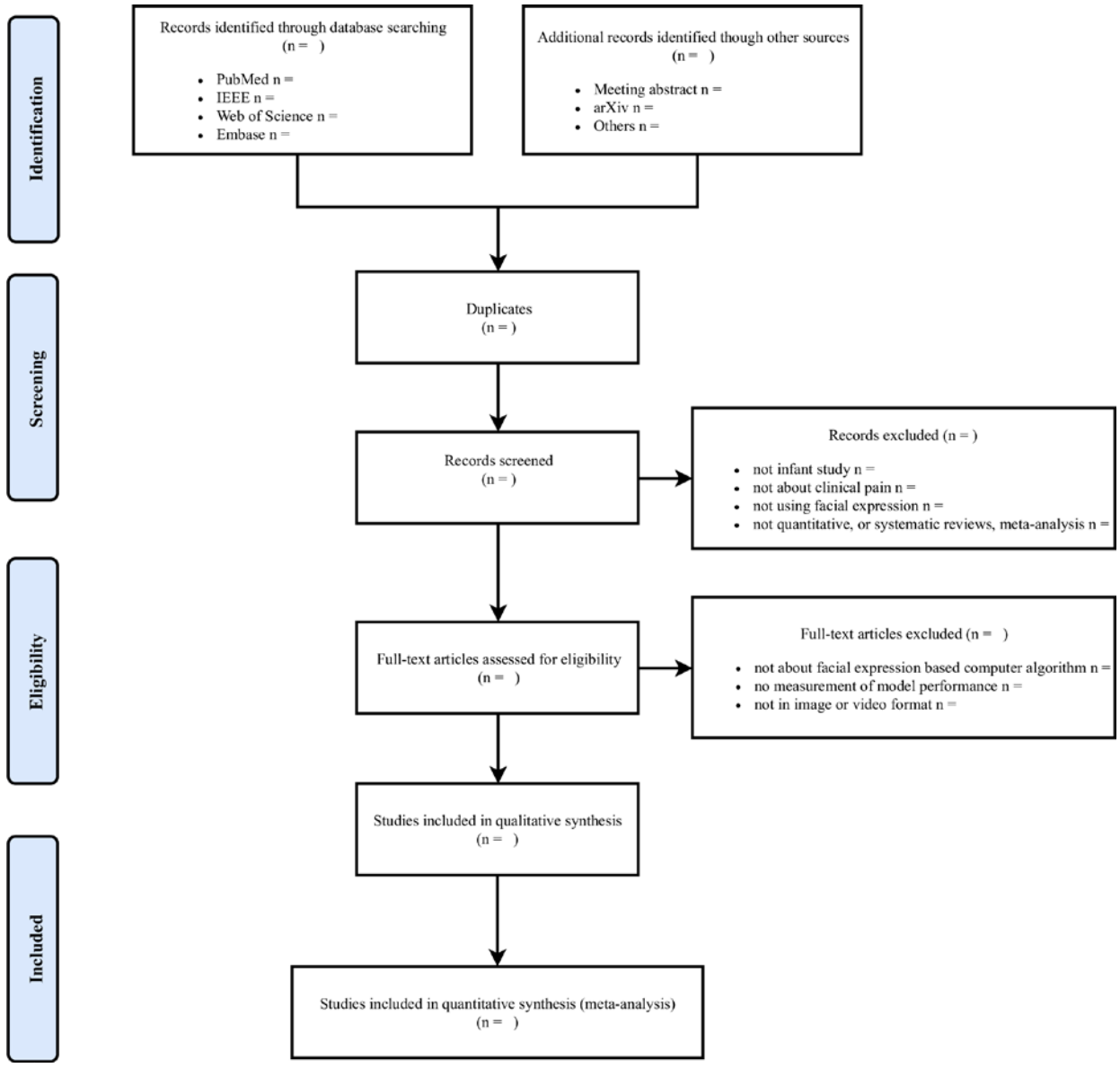

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2009 flow diagram.

Data management

We will import all eligible study records information including author, title and abstract into the Covidence platform. This online tool will facilitate the literature review process by integrating all reviewing activities on one platform with logged records. We will design a digital data extraction form to store study information using the Systematic Review Data Repository (SRDR) online repository website. Moreover, we will upload related data for each step and will update our protocols and analytic plans on the Open Science Framework website for transparency and version control purposes (https://osf.io/f9t4v/).

Data extraction

We will extract the study-level data from all eligible studies using a prepared data extraction form created by the *SRDR* online tool. We will divide items within the data collection form into four blocks: (1) study information including publication year, author information, funding or sponsorship information, type of study, journal name and PICO elements; (2) database information including name, subject size, image or video size; (3) patient demographic information including gender, age, race, disease diagnosis and types of pain; (4) ML methodological information including ML model name, type, format of input feature, optimisation algorithm, objective function, feature extraction methods, type of extraction feature and computational efficiency and cost, etc. An example of the data extraction form is presented in table 3.

Machine learning methods

The most commonly used ML methods for this particular question could include multiple general categories of models such as support vector machine, relevance vector machine, active appearance model and k-nearest neighbour, random forest trees, probabilistic neural

Table 3 An example of variables collected in data extraction table

Study information	
Study year	Year of the study published
Author information	Last name of author, whether clinical practitioners participated in the study
Type of study	Prospective cohort study or study that used a published database
Journal name	Journal name
PICO/PECO elements	PICO/PECO elements in summary
Database information	
Database name	Name of the database used for modelling
Host organisation	Name of the hosting organisation of the database
Sample size	Sample size of the database (image or video)
Sponsorship	The funding or sponsorship information
Patient demographic information	
Gender	Gender of infants (both, only boy, only girl)
Age	Age distribution
Race	Race/country of participants
Disease diagnosis	Disease diagnosis
Medical procedures	Procedure categories
Machine learning method information	
Model name	The name of the model
Model type	Machine learning model type
Model task	Classification, regression or both
Objective function	The objective function for modelling
Optimisation algorithm	The optimisation method for modelling
Format of input feature	Frame, sequence or image
Positive/negative size input	The size of positive and negative for modelling
Feature extraction method	The methods of feature extraction
Type of extracted feature	Pixel feature, AU, landmark or transformed feature
Model performance	Performance metrics and score of performance
Computational efficiency and cost	Computational efficiency (speed, cloud space, etc) and cost related to the algorithm (eg, require GPU resources, large cluster, etc)

AU, action unit; PICO/PECO, population, intervention (exposure), control, outcome.

network (PNN), Gaussian and nearest mean classifier, etc. In order to summarise and analyse these ML methods systematically, each ML algorithm will be presented and partitioned into a set of technical properties including: (1) feature: which features the algorithm uses (eg, pixel feature, action units, landmarks or transformed feature); (2) model: the underlying mathematical model (eg, artificial neural networks, random forests); (3) optimisation: the computational algorithm to find the optimum solution (eg, stochastic gradient descent, Bayesian variational inference); (4) performance: the measures for assessing the performance of the model (eg, our primary outcome accuracy, computational efficiency, etc). In this study, we will first collect the qualitative and descriptive information of these properties and then explore whether we can manually categorise them into fixed and organised classes. If enough data points can be obtained for each class, we will further perform an intraclass correlation analysis for subgroup analysis.

Study outcomes

Our primary outcome will be the measurement of model prediction accuracy (eg, area under the curve for receiver operating characteristic curve (AUC ROC), F1 score and proper score function such as Brier score if available). The secondary outcomes will be model utility measures including generalisability, interpretability, computational efficiency and cost.

Primary outcome: standardised measurement of model prediction accuracy

Based on our previous experience,[23 36] it is likely that the experimental setting, patient population, methods and results reporting styles in these eligible studies could be heterogeneous in presentation but clustered by used infant pain databases. To describe the model prediction accuracy, we will extract both qualitative descriptions and quantitative measures from original texts in these studies. The accuracy of the model prediction usually comprises two pieces of information: calibration and accuracy. Considering computer science studies usually use a different system for model reporting (eg, SEs and correlation coefficients) compared with statistical learning (eg, discrimination, calibration, Brier score, etc), our study will mainly focus on the discrimination domain of the model. After collecting the qualitative description of these accuracy measures, we will perform standardisation procedures on them for further quantitative meta-analysis. For regression algorithms, error measurements (eg, mean absolute error) will be standardised and converted to mean square error (MSE) for parallel comparisons. All correlation measurements (eg, Pearson's correlation) will be converted to ranked correlation (eg, Spearman's correlation). For classification algorithms, all accuracy measures will be converted to AUC ROC and F1 score for comparisons. The measures that cannot be standardised will be qualitatively reported as original values and excluded from the analysis. If diagnostic test accuracy measures such as sensitivity and specificity are reported, this information will also be collected.

Secondary outcomes: generalisability, interpretability and computational efficiency

We will conduct qualitative assessment of model generalisability and interpretability as study secondary outcomes.

These qualitative assessments will be performed by our study staff (DL or DC) in the form of judgement ranks. The levels of judgement rank for secondary outcomes will include high, moderate, low and very low. Computational efficiency and cost data of models will be analysed if data are available for quantitative analysis. Otherwise, they will be assessed qualitatively in judgement ranks.

Incomplete information and missing data

We will try to collect missing study information by contacting the authors if we cannot find the information through public channels. If we cannot obtain sufficient data, then these missing data will be omitted from the data synthesis and analyses.

Risk of bias assessment

The risk of bias assessment will be performed by two authors (DC and DL) using the Prediction model study Risk of Bias Assessment Tool (online supplementary appendix 1).[37] It includes 20 signalling questions across four domains (participants, predictors, outcome and analysis). Based on the ratings of signalling questions, risk of bias for each domain will be ranked as low risk, high risk or too unclear for judgement.

Assessment of reporting quality

Liu *et al* found that the reporting of studies evaluating ML models was often incomplete and non-standardised, which increasingly became a barrier to robust evaluation of artificial intelligence-based models.[38] Collins *et al* highlighted that complete and transparent reporting of the key aspects of ML prediction model studies is vital to ensure the quality and interpretability of the studies in medical and technical fields.[39] In order to assess the reporting quality of eligible studies, we will use Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (online supplementary appendix 2) checklist.[40] Each item will be ranked as no report, inadequate report and adequate report, scored as 0, 1 and 2, respectively. To account for specific methodological aspects of ML algorithms studies, we developed a customised ML-specific reporting checklist for this study as the supplementary ML-method reporting assessment tool (online supplementary appendix 3). This assessment checklist includes description of database, research team, data preprocessing, method and approach, objective function, optimisation technique and computational efficiency. Each item will be ranked as no report, inadequate report and adequate report, scored as 0, 1 and 2, respectively.

Statistical analysis and data synthesis

Extracted outcome data stored in the *SRDR* website will be imported into RevMan V.5.2.1 software and R V.3.3.2 for analysis. Because reported model outputs may be different across individual studies, we will use standardised measures to synthesise model accuracy and calibration. In order to assess accuracy, we will use the C-statistic (AUC ROC) for classification models and MSE for regression models along with their 95% CIs. To assess calibration, we will use the Hosmer-Lemeshow χ^2 test and ranked correlation, if applicable. To measure heterogeneity, we plan to use a Galbraith plot and Higgins and Thompson's I^2. A fixed-effects model will be used only if there is no evidence of statistical heterogeneity. We will pool the summary measures across the studies using a random effects model optimised using DerSimonian and Laird's method, if considerable heterogeneity is indicated (I^2>50%). Furthermore, we will also explore the possible sources of heterogeneity from both clinical and methodological perspectives to provide an explanation or conduct subgroup analysis. Meta-regression will be conducted if applicable.

Subgroup analysis

We intend to conduct subgroup analyses by ML model types (eg, regression vs classification, neural networks vs traditional ML), facial data input format (eg, images vs landmarks) and medical procedure type (eg, acute procedural pain vs postoperative pain). More post hoc exploratory subgroup analyses will be decided during the process of data extraction and analysis.

Publication bias

We will search related professional meeting abstracts, technical preprints, reference lists of qualitative review articles, systematic reviews and meta-analysis articles to identify missing unpublished literature (details in 'Search strategy' section). We will use contour-enhanced funnel plots to assess publication bias.

Confidence in cumulative evidence

Confidence in cumulative evidence will be conducted in accordance with the Grading of Recommendations, Assessment, Development and Evaluations guideline.[41] Inconsistency will be assessed using the I^2 test and Galbraith plot as described in previous sections. Indirectness will be assessed by examining the collected PICO elements of eligible studies and comparing generalisability (one of our secondary outcomes). Imprecision will be assessed by examining the study sample sizes and CIs of outcomes of interest.

DISCUSSION

Neonatal care has developed rapidly within the last decade and has greatly improved survival outcomes of premature and sick neonates. However, when large numbers of painful and stressful treatment procedures were performed on infant patients, only very few were accompanied by adequate analgesia.[3] Several researchers found that neonatal intensive care unit (NICU) admission rates increased steadily yearly[42] and large numbers of neonates could be exposed to acute pain from invasive procedures or prolonged pain from surgery or inflammation.[43] Traditional observer-based infant pain assessment is difficult to train, time-consuming, labour-intensive,

subjective and a source of conflict in NICU care. Developing an automatic pain assessment instrument will be resource-saving and improve the quality of care through fast and accurate pain assessment.

Infants' responses to pain and stress are non-specific and can be easily misinterpreted. The Newborn Individualised Developmental Care and Assessment Programme (NIDCAP) is widely used for infant stress assessment.[44] Holsti *et al* also developed multidimensional assessments, including the full NIDCAP to distinguish pain and stress.[45] However, these human-based instruments are by definition labour-intensive and expensive to implement, and understanding how pain and stress can affect infant development and properly evaluating them is a highly specialised practice area requiring substantial training.[46] Mansor *et al* developed a PNN classifier to distinguish pain from other non-pain tasks (rest/cry, air puff, friction) in the COPE database, and results were remarkable with the classification accuracy >90%.[26] This indicates that ML methods have the potential to distinguish pain and stress when combined with other clinical indicators.

The age of artificial intelligence has come. Computer scientists are now equipped with tools and algorithms to train machines to identify behavioural and physiological indicators associated with pain. Facial expression-based ML algorithms have the potential to overcome observational bias, reduce training costs and provide possibilities for continuous infant pain monitoring without clinician involvement. It is promising that these ML algorithms will provide fast, accurate and continuous pain assessment in both routine practice and NICU with better training data and more advanced ML methods. Our proposed systematic review and meta-analysis will provide the first systematic and quantitative report of model accuracy, generalisability and interpretation capability of different available ML algorithms for neonatal and infant pain assessment. Also, it is not common for computer scientists to clearly disclose costs related to their algorithms.[35 47] We hope computer scientists can disclose both training and running costs of their algorithms to provide better references for cost-effectiveness research for future real-world implementation.

Author affiliations
[1]Department of Anesthesiology, Pain and Perioperative Medicine, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
[2]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4]Treadwell Library, Massachusetts General Hospital, Boston, Massachusetts, USA
[5]Department of Engineering, University of Virginia, Charlottesville, Virginia, USA
[6]DRPH Program, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA

Contributors HD, DC and DL contributed to the conception and design of the study. The manuscript protocol was drafted by DC and was revised by DPT, TTH, LC, MZ, JY, WZ and HD. The search strategy was developed by LLP, HD, DC, and revised by the other authors. Search strategy will be performed by DC and DL, who will also independently screen the potential studies, extract data from the included studies, assess the risk of bias and complete the data synthesis. HD and WZ will arbitrate in cases of disagreement and ensure the absence of errors. All authors approved the publication of the protocol.

Funding This work is supported by the National Natural Science Foundation of China, with the funding reference number of 81571082. This project is also supported by the Youth Creative Fund of The First Affiliated Hospital of Zhengzhou University.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This systematic review and meta-analysis will only use study-level data from public databases, thus formal ethical approval is not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
Hao Deng http://orcid.org/0000-0002-0331-2427

REFERENCES

1 Goksan S, Hartley C, Emery F, *et al*. fMRI reveals neural activity overlap between adult and infant pain. *Elife* 2015;4.
2 Porter FL, Wolf CM, Miller JP. Procedural pain in newborn infants: the influence of intensity and development. *Pediatrics* 1999;104:e13.
3 Carbajal R. Epidemiology and treatment of painful procedures in neonates in intensive care units. *JAMA* 2008;300.
4 Johnston CC, Fernandes AM, Campbell-Yeo M. Pain in neonates is different. *Pain* 2011;152:S65–73.
5 Byrd PJ, Gonzales I, Parsons V. Exploring barriers to pain management in newborn intensive care units: a pilot survey of NICU nurses. *Adv Neonatal Care* 2009;9:299–306.
6 Cong X, Delaney C, Vazquez V. Neonatal nurses' perceptions of pain assessment and management in NICUs: a national survey. *Adv Neonatal Care* 2013;13:353–60.
7 Pillai Riddell RR, Stevens BJ, McKeever P, *et al*. Chronic pain in hospitalized infants: health professionals' perspectives. *J Pain* 2009;10:1217–25.
8 Beggs S. Long-Term consequences of neonatal injury. *Can J Psychiatry* 2015;60:176–80.
9 Donia AE-S, Tolba OA. Effect of early procedural pain experience on subsequent pain responses among premature infants. *Egyptian Pediatric Association Gazette* 2016;64:74–80.
10 Peters JWB, Schouw R, Anand KJS, *et al*. Does neonatal surgery lead to increased pain sensitivity in later childhood? *Pain* 2005;114:444–54.
11 Buonocore G, Bellieni CV. Neonatal Pain: Suffering, Pain, and Risk of Brain Damage in the Fetus and Newborn. Springer Science & Business Media, 2008. Available: https://market.android.com/details?id=book-2kv-8WsGCqEC
12 Craig KD, Whitfield MF, Grunau RVE, *et al*. Pain in the preterm neonate: behavioural and physiological indices. *Pain* 1993;52:287–99.
13 Perlman J, Thach B. Respiratory origin of fluctuations in arterial blood pressure in premature infants with respiratory distress syndrome. *Pediatrics* 1988;81:399–403.
14 Anand KJ, Hickey PR. Pain and its effects in the human neonate and fetus. *N Engl J Med* 1987;317:1321–9.
15 Merkel SI, Voepel-Lewis T, Shayevitz JR, *et al*. The FLACC: a behavioral scale for scoring postoperative pain in young children. *Pediatr Nurs* 1997;23:293–7.
16 Hummel P, Puchalski M, Creech SD, *et al*. Clinical reliability and validity of the N-PASS: neonatal pain, agitation and sedation scale with prolonged pain. *J Perinatol* 2008;28:55–60.
17 Debillon T. Development and initial validation of the EDIN scale, a new tool for assessing prolonged pain in preterm infants. *Arch Dis Child Fetal Neonatal Ed* 2001;85:36F–41.
18 Lawrence J, Alcock D, McGrath P, *et al*. The development of a tool to assess neonatal pain. *Neonatal Netw* 1993;12:59–66.
19 Krechel SW, Bildner J. Cries: a new neonatal postoperative pain measurement score. initial testing of validity and reliability. *Pediatric Anesthesia* 1995;5:53–61.

20 Grunau RE, Oberlander T, Holsti L, *et al*. Bedside application of the neonatal facial coding system in pain assessment of premature infants. *Pain* 1998;76:277–86.
21 Prkachin KM, Berzins S, Mercer SR. Encoding and decoding of pain expressions: a judgement study. *Pain* 1994;58:253–9.
22 Pillai Riddell RR, Horton RE, Hillgrove J, *et al*. Understanding caregiver judgments of infant pain: contrasts of parents, nurses and pediatricians. *Pain Res Manag* 2008;13:489–96.
23 Zamzmi G, Kasturi R, Goldgof D, *et al*. A review of automated pain assessment in infants: features, classification tasks, and databases. *IEEE Rev Biomed Eng* 2018;11:77–96.
24 Brahnam S, Chuang C-F, Shih FY, *et al*. Machine recognition and representation of neonatal facial displays of acute pain. *Artif Intell Med* 2006;36:211–22.
25 Naufal Mansor M, Mansor MN, Rejab MN. A computational model of the infant pain impressions with Gaussian and nearest mean classifier. in: 2013 IEEE International Conference on control system, computing and engineering 2013.
26 Mansor MN, Junoh AK, Ahmed A, *et al*. *Single scale self quotient image and PNN for infant pain detection. 2014 IEEE International Conference on control system, computing and engineering (ICCSCE 2014)*, 2014.
27 Celona L, Manoni L. Neonatal facial pain assessment combining Hand-Crafted and deep features. New trends in image analysis and processing – ICIAP 2017 2017:197–204.
28 Yuan L, Bao FS, Lu G. Recognition of neonatal facial expressions of acute pain using boosted Gabor features. 2008 20th IEEE International Conference on tools with artificial intelligence 2008.
29 Zamzmi G, Pai C-Y, Goldgof D, *et al*. An approach for automated multimodal analysis of infants' pain. 2016 23rd International Conference on Pattern Recognition (ICPR) 2016.
30 Gholami B, Haddad WM, Tannenbaum AR. Agitation and pain assessment using digital imaging. *Conf Proc IEEE Eng Med Biol Soc* 2009;2009:2176–9.
31 Gholami B, Haddad WM, Tannenbaum AR. Relevance vector machine learning for neonate pain intensity assessment using digital imaging. *IEEE Trans Biomed Eng* 2010;57:1457–66.
32 Moher D, Shamseer L, Clarke M, *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015;4.
33 Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. *AMIA Annu Symp Proc* 2006:359–63.
34 Valstar MF, Almaev T, Girard JM, *et al*. FERA 2015 - second Facial Expression Recognition and Analysis challenge. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 2015.
35 Chen Z, Ansari R, Wilkie D. Automated pain detection from facial expressions using FACS: a review. arXiv preprint 2018: arXiv:1811.07988.
36 Liu D, Cheng D, Houle TT, *et al*. Machine learning methods for automatic pain assessment using facial expression information: protocol for a systematic review and meta-analysis. *Medicine* 2018;97:e13421.
37 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51–8.
38 Liu X, Faes L, Kale AU, *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 2019;1:e271–97.
39 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *The Lancet* 2019;393:1577–9.
40 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015;102:148–58.
41 Langer G, Meerpohl JJ, Perleth M, *et al*. GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables]. *Z Evid Fortbild Qual Gesundhwes* 2012;106:357–68.
42 Harrison W, Goodman D. Epidemiologic trends in neonatal intensive care, 2007-2012. *JAMA Pediatr* 2015;169:855–62.
43 Hall RW, Anand KJS. Pain management in newborns. *Clin Perinatol* 2014;41:895–924.
44 Als H. A Synactive Model of Neonatal Behavioral Organization: Physical & Occupational Therapy In Pediatrics 1986;6:3–53.
45 Holsti L, Grunau RE, Oberlander TF, *et al*. Body movements: an important additional factor in discriminating pain from stress in preterm infants. *Clin J Pain* 2005;21:491–8.
46 Holsti L, Grunau RE. Extremity movements help occupational therapists identify stress responses in preterm infants in the neonatal intensive care unit: a systematic review. *Can J Occupation Ther* 2007;74:183–94.
47 Sikka K, Sharma G, Bartlett M. LOMo: latent ordinal model for facial analysis in Videos. 2016 IEEE conference on computer vision and pattern recognition (CVPR) 2016.