BMJ Health & Care Informatics

Data pipeline quality: development and validation of a quality assessment tool for data-driven algorithms and artificial intelligence in healthcare

Eris van Twist ,[1] Brian van Winden,[1] Rogier de Jonge,[1] H Rob Taal,[1] Matthijs de Hoog,[1] Alfred Schouten,[2] David Tax,[3] Jan Willem Kuiper[1]

To cite: van Twist E, van Winden B, de Jonge R, *et al*. Data pipeline quality: development and validation of a quality assessment tool for data-driven algorithms and artificial intelligence in healthcare. *BMJ Health Care Inform* 2026;**33**:e101608. doi:10.1136/bmjhci-2025-101608

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2025-101608).

Received 27 May 2025
Accepted 23 January 2026

Check for updates

[1]Department of Neonatal and Pediatric Intensive Care, Division of Pediatric Intensive Care, Erasmus MC, Rotterdam, The Netherlands
[2]Department of Biomechanical Engineering, Faculty of Mechanical Engineering, Delft University of Technology, Delft, The Netherlands
[3]Pattern Recognition Laboratory, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Correspondence to
Eris van Twist;
e.vantwist@erasmusmc.nl

ABSTRACT

Objectives To develop and validate a tool for standardised quality assessment of data-driven algorithms in healthcare, focusing on the underlying data pipeline.

Methods Data Assessment Tool for Algorithm Critical Appraisal and Robust Evidence (DATA-CARE) was iteratively developed from the established Quality In Prognosis Studies framework, selected after reviewing 10 existing quality assessment tools for observational and artificial intelligence studies. DATA-CARE evaluates five quality domains of the data pipeline: study population, data, algorithm, outcome and report transparency. Each domain comprises three to five quality criteria. With a total score of 75 points, study quality is categorised as low (<45), moderate (45–59) or high (≥60). DATA-CARE was validated during a systematic review on data-driven algorithms using continuous physiological monitoring data within the paediatric intensive care unit. Two independent reviewers performed quality assessment using DATA-CARE of included studies. Tool validation was evaluated using inter-rater agreement and intraclass correlation coefficient (ICC).

Results DATA-CARE demonstrated robust inter-rater agreement (93.5%) with ICC 0.98 (95% CI 0.96 to 0.99). Of 3858 screened studies, 31 were reviewed in the use case, describing diverse algorithms. Studies were predominantly low (32.3%) to moderate (41.9%) and sporadically (25.8%) high quality.

Discussion Predominance of low-to-moderate quality studies reveals critical barriers to clinical implementation of data-driven algorithms, including low quality data capture and processing, lacking validation strategies and non-transparent reporting of findings.

Conclusions DATA-CARE allows standardised and reliable critical appraisal for a wide variety of algorithms, addressing current gaps in standardised and reproducible algorithm development.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Algorithm quality hinges on its underlying data pipeline, specifically source data, processing and analytical methodology and overall reproducibility. Existing quality assessment tools often neglect this, limiting the ability to review and reproduce algorithms and thus hindering their clinical implementation.

WHAT THIS STUDY ADDS

⇒ This study introduces and validates Data Assessment Tool for Algorithm Critical Appraisal and Robust Evidence (DATA-CARE), a quality assessment tool that evaluates five key domains of the data pipeline. It demonstrates high inter-rater reliability and reveals that most reviewed studies in a paediatric intensive care use case are of low-to-moderate quality.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ DATA-CARE provides a standardised and reliable framework for evaluating algorithms, supporting reproducible research and transparent reporting. Its adoption could guide researchers, reviewers and policymakers in improving the quality and clinical readiness of data-driven algorithms in healthcare.

INTRODUCTION

Data-driven healthcare, powered by artificial intelligence (AI) and big data analytics, has emerged as a transformative force in modern healthcare.[1] The ability to harness data for bedside monitoring and decision support via actionable algorithms holds the promise of improved and personalised care.[1] Despite increasing research on this topic, a critical gap persists between algorithm development and clinical implementation, often attributed to lack of standardised methodology.[2–4]

In data-driven healthcare, continuously measured data are used to guide clinical decision making.[5] Patients constitute the source from which data follows a pipeline where raw digitalised signals from bedside monitors and devices are collected, processed and used in algorithmic analysis to produce actionable insights (figure 1).[6] The quality of the data and integrity of each step in the data pipeline affects derived insights and their validity. A robust and reproducible data pipeline is

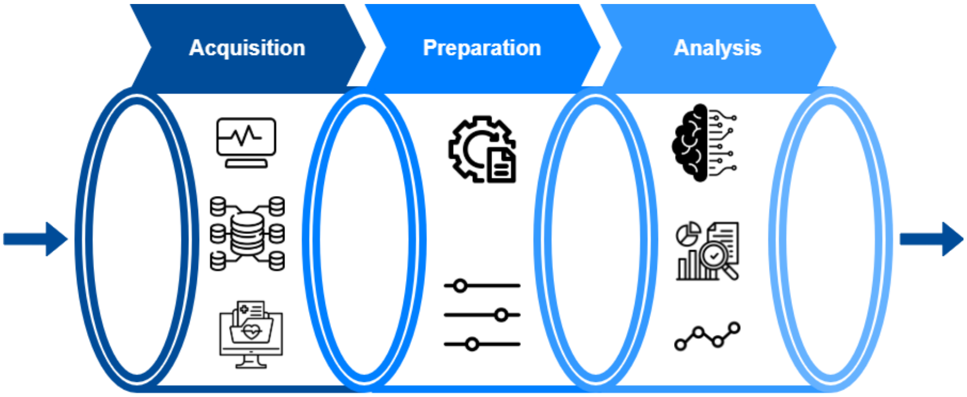

Figure 1 Schematic overview of the data pipeline. Raw data are acquired from various sources and ingested (in batches from files, via streaming of cloud databases) into the data pipeline where it undergoes systematic processing (such as noise cleaning) to prepare it for analysis (machine learning, statistical modelling, signal analytics, dashboarding). This is a simplified overview; there may be more details to the data pipeline (eg, aggregation of preprocessed data) or feedback loops (eg, following analysis, preprocessing is enhanced).

therefore of critical importance to ensure data-driven healthcare is effective, reliable and generalisable. However, to our knowledge, there is no tool for quality assessment of the data pipeline.

Quality assessment tools are often developed for a specific study design or objective, focusing on individual components rather than providing a comprehensive assessment of the data pipeline.[7 8] As such, available tools tend to be fragmented and limited in scope, fail to capture critical domains of the data pipeline or are too focused on specific algorithm types.[7 8] Algorithm quality and returned output are largely dependent on source data and how this is ingested and processed in the pipeline.[9] Perhaps most important of all is the ability to reproduce and validate the data pipeline. This requires transparency in research, in particular on study population selection (data source), data quality and processing, algorithm development and validation and (desired) outcomes. Available tools that address these domains may pose a suitable basis for quality assessment of data-driven healthcare, but need to be adjusted to become widely applicable to studies on data-driven healthcare.

A quality assessment tool for data-driven healthcare enables critical appraisal of existing research and guides towards standardised and reproducible data pipelines for actionable clinical algorithms. Therefore, the aim of this study was to develop and validate a quality assessment tool for data-driven healthcare, based on the domains study population, data, algorithm, outcome and report transparency adjusted from the Quality In Prognosis Studies (QUIPS) framework.

METHODS

Tool development

Tool development occurred iteratively in a six-member working group, including an epidemiologist (RdJ), clinicians (RdJ, JWK), data scientists (EvT, BvW) and engineers (AS, DT) at Erasmus MC Sophia Children's Hospital and Delft University of Technology (figure 2). The tool was developed through refinement and expansion of the QUIPS, which were most suitable among ten existing tools identified for observational and AI studies (online supplemental table S1).[10–24] The QUIPS tool was chosen as it guides systematic and comprehensive critical appraisal in a user-friendly and widely applicable format, with domains that adhere to the data pipeline. Original domains study participation, prognostic factor measurement, outcome measurement, study confounding and statistical analysis and reporting were translated to data-driven healthcare. The domains were divided into criteria that determine the quality per domain, covering the data pipeline from input to output. Quality domains and criteria were ranked based on applicability and

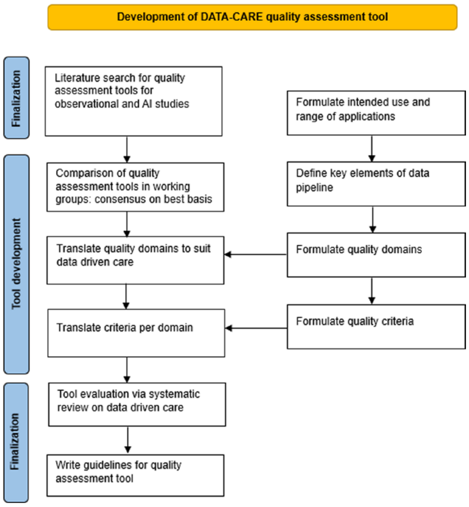
Development of DATA-CARE quality assessment tool
Finalization
Tool development
Finalization
Literature search for quality assessment tools for observational and AI studies
Comparison of quality assessment tools in working groups: consensus on best basis
Translate quality domains to suit data driven care
Translate criteria per domain
Tool evaluation via systematic review on data driven care
Write guidelines for quality assessment tool
Formulate intended use and range of applications
Define key elements of data pipeline
Formulate quality domains
Formulate quality criteria

Figure 2 Schematic overview of stepwise development of DATA-CARE in working group. Criteria for quality assessment were formulated and reformulated iteratively during consensus meetings. AI, artificial intelligence; DATA-CARE, Data Assessment Tool for Algorithm Critical Appraisal and Robust Evidence.

clinical relevance for selection. Guidelines were written with signalling questions and a scoring system was adopted from de Jonge *et al.*[25 26] The score per domain reflects robustness of that domain and the overall score reflects overall quality and risk of bias in a study with regard to the data pipeline. Per domain, 15 points can be allocated, distributed over 3–5 criteria, yielding 17 criteria in total with a maximum score of 75 points. Each criterion is scored out of 3 or 5 points depending on the number of criteria within the domain, with 50% penalty for incomplete or methodologically flawed information. A score of 80% (≥60 points) constituted high quality, between 60% and 80% (45–59 points) moderate quality and below 60% (<45 points) low quality.[26]

Quality domains

The study population domain guides readers to assess whether the target population (data source) was captured and whether risk of bias was introduced when certain participants were (not) included in the study, for example, due to selection bias, ascertainment bias or loss to follow-up.[12 27] It is based on the three criteria: recruitment, inclusion and exclusion criteria and baseline study population.

The data domain assesses whether data capture was adequate and of sufficient quality. It is the only domain containing five criteria: data acquisition, data set size and balance, missing data, preprocessing and feature derivation. The criteria adhere to the data quality dimensions of accuracy, completeness, redundancy, readability, accessibility, consistency, usefulness and trust.[28] During data acquisition, sample frequencies determine data resolution and may cause aliasing if not appropriate to signal bandwidth.[29] Characteristics like size, balance and completeness significantly impact algorithm performance and stability. Imbalanced data lead to overrepresentation of the majority class and may bias the algorithm to better distinguish this class.[30] Missing data can introduce bias similar to loss to follow-up, as eligible participants have incomplete data or data of insufficient quality to be included in analysis.[31] Missing data can be missing completely at random, missing at random or missing not at random, where the latter two introduce high risk of bias if not accounted for during analysis.[31] Preprocessing and feature derivation are important as medical data cannot be mistaken for ground truth, affecting algorithm generalisability and computational efficiency while also posing a risk of bias and/or overfitting.[32–34]

The algorithm domain assesses whether the computational approach to derive an outcome of interest was standardised, robust and valid across participants and its generalisability to the target population, based on algorithm architecture, development and evaluation. An algorithm should be appropriate to the intended use and requires systematic data partitioning, configuration and validation.[35 36] Data partitioning refers to the train–test split, ideally on a participant level to prevent data leakage which may cause overfitting and reduce algorithm generalisability.[35] If algorithms are patient-tailored, this should be explicitly stated. To advance to population inference, algorithms require internal (unseen test set) and external validation (newly sampled data).[37] Complementary performance metrics are vital to clinical interpretation, including metrics of significance (eg, p values) and uncertainty (eg, CIs) or of discrimination (eg, balanced accuracy) and calibration (eg, R^2 curves).[37 38] Discriminative metrics allow the reader to interpret how well the algorithm can identify positive and negative instances, and calibration metrics allow interpretation of how reliable this identification is.[38]

The outcome domain assesses whether the outcome was standardised and measured reliably across participants and judges the risk of bias due to mislabelling, as algorithms can only be implemented in clinical practice if they can reflect on an outcome of interest. To ascertain clinical implications, the outcome and its labelling must represent a ground truth, that is, the objective reality based on (reliable) measurement or observation.[15] Outcomes need clear definitions and standardised assessment, ideally via the reference standard. Labelling maps outcomes to individual data points is especially essential for supervised algorithms which must learn to reproduce outcome labels, while unsupervised algorithms create custom labels.[34] Standardisation of outcome and labelling may be limited by inter-rater variability.[39] If a ground truth is not available (eg, clinical deterioration), labels may be engineered (unsupervised) or derived from clinically relevant endpoints (eg, therapeutic intervention), with implications considered in the discussion.

The report transparency domain assesses the risk of bias due to incomplete reporting or inappropriate statistical methodology, based on presentation of data and findings, reporting of results and statistical analysis. Data and findings should reflect the study objective and methods but are not overstated and limitations are discussed. Selective reporting is avoided by presenting all results, including algorithm subtypes and subgroups where applicable. On indication, sensitivity analysis and/or post hoc analysis is reported. Statistical evaluation was specified, statistical assumptions have been met and results are consistently presented throughout the study (eg, OR as positive decimal). To minimise bias and contribute to fairness, adequate measures of significance or uncertainty are provided with correction for multiple testing where applicable.[40] This allows interpretation of findings, which is dependent on objective, sample sizes and assumptions.[40]

Tool validation

Data Assessment Tool for Algorithm Critical Appraisal and Robust Evidence (DATA-CARE) was validated during a use-case on data-driven healthcare in the paediatric intensive care unit (PICU). Studies were identified using a systematic review conducted in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.[41] The search strategy is detailed in online supplemental A. Study selection was based on study design, setting, data and sampling frequency as described in online supplemental B. Data synthesis included information on study methodology and characteristics of the data pipeline (online supplemental C). Included articles were qualitatively assessed by two reviewers (EvT, BvW). After quality assessment, scoring was compared, with disagreements resolved during consensus meetings (EvT, BvW). Tool validation was based on inter-rater agreement (%) and intraclass correlation coefficient (ICC) for total quality score and quality category.

RESULTS

The DATA-CARE tool

We present DATA-CARE, with quality domains and criteria summarised in table 1. The full tool and guidelines are available in online supplemental D.

Data-driven healthcare in the PICU

Out of 3858 studies identified, 31 were included for quality assessment after duplicates removal, screening and full-text retrieval (online supplemental figure S1). Main reasons for exclusion were non-monitoring objectives and discontinuous data. The included studies and characteristics of the data pipeline are presented in online supplemental tables S2 and S3. Studies were generally retrospective (23 (74.2%)) with median sample size 90 (28–215) participants. Algorithms were mainly AI (16 (51.6%)), specifically machine learning, followed by signal analysis (12 (38.7%)) and were mainly intended for prediction (13 (41.9%)) or monitoring (10 (32.3%)). Studies on AI typically evaluated multiple classifiers, most commonly neural networks and random forest (8 (25.6%)). None of the included studies developed dashboards or provided decision support. Data sampling and window sizes varied widely. 19 studies (61.3%) used separate data sets for algorithm development and validation and specified data partitioning, 9 of which introduced data leakage. Cross-validation was reported by 12 studies (38.7%), optimisation by 16 studies (51.6%) and handling of imbalanced data by 6 studies (19.4%), all mostly AI. Additional validation post train–test procedure was mentioned by three studies (9.7%). One study (3.2%) performed external validation using prospectively collected data. Performance metrics were mainly discriminatory, including area under the receiver-operator curve (AUROC), accuracy, sensitivity and specificity.

Table 1 Overview of DATA-CARE and point allocation for scoring

	Score*		
Criteria	+	±	–
Study population			
Recruitment	5	2.5	0
Inclusion and exclusion criteria	5	2.5	0
Baseline study population	5	2.5	0
Data			
Data acquisition	3	1.5	0
Data set size and balance	3	1.5	0
Missing data	3	1.5	0
Data preprocessing	3	1.5	0
Feature derivation	3	1.5	0
Algorithm			
Algorithm architecture	5	2.5	0
Algorithm development	5	2.5	0
Algorithm evaluation	5	2.5	0
Outcome			
Definition of outcome	5	2.5	0
Method and setting of outcome assessment	5	2.5	0
Outcome labelling	5	2.5	0
Report transparency			
Presentation of data and findings	5	2.5	0
Reporting of results	5	2.5	0
Statistical analysis	5	2.5	0

Adapted from Hayden *et al* and de Jonge *et al* with permission.[12 25 26]

*Scoring symbols refer to maximum score (+), average score (±) and minimum score (–).

DATA-CARE, Data Assessment Tool for Algorithm Critical Appraisal and Robust Evidence.

Tool validation

Mean quality score was 50.6 (12.6) points, with 10 (32.3%) low quality studies (<45 points), 13 (41.9%) moderate quality studies (45–59 points) and 8 (25.8%) high quality studies (≥60 points). Quality scores per study are available in table 2. Most points were withheld in data and report transparency domains as studies neglected to report on data set size and balance, missing data and/or lacked statistical rigour. Most points were allocated in the outcome and algorithm domains. Inter-rater agreement was 63.6% for total score and 93.5% for quality category, with ICC 0.98 (95% CI 0.96 to 0.99) and 0.95 (95% CI 0.90 to 0.98). Across quality domains, inter-rater agreement for domain scores was 64.5% for study population, 67.7% for data, 70.1% for algorithm, 70.1% for outcome and 67.7% for report transparency.

Table 2 Quality assessment of included studies using DATA-CARE

Study	Study population	Data	Algorithm development	Outcome	Report transparency	Score
Singh *et al* [32]	15	12	10	15	12.5	64.5*
Azriel *et al*[49]	7.5	10.5	12.5	12.5	10	53
Rooney *et al*[50]	15	4.5	10	15	10	54.5
Badke *et al*[51]	12.5	10.5	15	15	12.5	65.5*
Joram *et al*[52]	15	10.5	10	10	10	55.5
Amiri *et al*[53]	5	6	10	15	2.5	38.5
Castineira *et al*[54]	7.5	13.5	7.5	12.5	2.5	43.5
Sorensen *et al*[55]	12.5	7.5	10	10	12.5	52.5
Bose *et al*[56]	5	10.5	10	12.5	12.5	50.5
Marsillio *et al*[57]	15	6	15	5	15	56
Messinger *et al*[58]	15	9	7.5	5	7.5	44
Matam *et al* (2019)[59]	7.5	7.5	12.5	10	7.5	45
Kamaleswaran *et al*[60]	7.5	6	12.5	12.5	7.5	46
Rusin *et al*[61]	12.5	6	10	10	12.5	51
Zhang *et al*[62]	5	7.5	12.5	0	5	30
Biswas *et al*[63]	12.5	7.5	12.5	5	12.5	50
Si *et al*[64]	2.5	10.5	5	7.5	2.5	28
Martin *et al*[65]	15	9	7.5	15	10	56.5
Kirschen *et al*[66]	15	7.5	12.5	12.5	15	62.5*
Matam *et al* (2014)[67]	6.5	9	7.5	2.5	2.5	28
Izquierdo *et al*[68]	0	9	10	5	0	24
Zoodsma *et al*[69]	15	9	10	12.5	10	56.5
Tabassum *et al*[70]	5	10.5	7.5	7.5	7.5	38
Liu *et al*[71]	15	13.5	15	15	15	73.5*
van Twist *et al*† (EEG)[72]	12.5	12	12.5	15	12.5	64.5*
Macabiau *et al*[73]	7.5	12	12.5	10	2.5	44.5
Le *et al*[74]	7.5	12	12.5	10	2.5	44.5
Kwon *et al*[75]	12.5	7.5	10	12.5	12.5	55
Hunfeld *et al*[76]	15	9	12.5	12.5	12.5	61.5*
van Twist *et al*† (ECG)[77]	12.5	17	12.5	15	15	72*
Silva *et al*[78]	15	7.5	12.5	12.5	12.5	60*
Mean (SD)	10.5 (4.5)	9.4 (2.7)	10.9 (4.1)	10.6 (4.1)	9.2 (4.6)	50.6 (12.6)

Note there are two pairs of studies with a similar author, where additional information is provided in brackets for distinction.
*Studies with a high quality (≥60 points).
†Study by same author as the present study.
DATA-CARE, Data Assessment Tool for Algorithm Critical Appraisal and Robust Evidence.

DISCUSSION

We have developed DATA-CARE, a quality assessment tool for systematic critical appraisal of data-driven algorithms in healthcare. This tool, based on the widely recognised QUIPS, addresses five quality domains of the data pipeline, including study population, data, algorithm, outcome and report transparency. Validation of DATA-CARE during a use-case on data-driven healthcare in the PICU showed the tool can be applied to a wide variety of algorithms, obtaining robust consensus in our working group with 93.5% agreement and 0.98 correlation. As such, DATA-CARE supports reproducible and transparent research through structured critical appraisal of data-driven algorithms.

To our knowledge, DATA-CARE is the first quality assessment tool suited to the diverse and fast-growing field of data-driven healthcare. While available quality assessment tools for observational studies were relevant for epidemiological aspects (eg, study population), they lacked domains that directly address the data pipeline (eg, data processing).[10 11 13 14] Quality assessment tools addressing the data pipeline were mainly intended for AI and typically occurred as reporting checklists, spanning between 4 and 27 domains with variable numbers of items per domain.[15–21] Such checklists, however, require fundamental knowledge on data science and may provoke inter-rater variation.[42] Checklists also omit the issue that computerised algorithms lack human judgement and can therefore not identify inherent bias in the data.[35] For example, all checklists included data partitioning, but partitioning on a subpatient level (eg, event level) introduces data leakage as patients can occur in both train-set and test-set, hampering generalisability.[43] Partitioning should also be done at the beginning of the pipeline, as preprocessing techniques such as scaling on the entire data set cause similar data leakage. Hence, critical appraisal of information is just as important as ascertaining its presence. Non-checklist quality assessment tools included APPRAISE-AI and Prediction model Risk of Bias Assessment Tool (PROBAST-AI), intended for clinical decision support and predictive AI, respectively.[22 23] APPRAISE-AI reported ICC between 0.71 and 1.00 for criteria scores, 0.89 and 0.99 for domain scores and 0.98 for overall scores.[22] A similar agreement was obtained with DATA-CARE, though the agreement varied across quality domains. Development of Quality Assessment of Diagnostic Accuracy Studies (QUADAS-AI) and Standards for Reporting of Diagnostic Accuracy Study (STARD-AI) was ongoing at the time of publication, but all were specifically intended for AI studies.[16 24] DATA-CARE uniquely shifts critical appraisal to the data pipeline and uses key principles of transparent research reporting. As such, DATA-CARE is widely applicable and practical, without compromising on high reliability.

Progression of data-driven healthcare critically hinges on study quality, common barriers being low quality data, lack of external validation and incomplete or non-transparent reporting of findings.[44–46] These barriers were also encountered during validation of DATA-CARE. None of the reviewed algorithms were implemented, DATA-CARE quality scores varied widely and the majority of studies were regarded as low-to-moderate quality. While the lack of progression and low study quality may reinforce one another, research has shown that qualitative issues persist even among algorithms approved as medical devices.[47] Predominant low scores in the data and report transparency domain, contrary to higher scores in the algorithm domain, imply that the current bottleneck of data-driven healthcare is poor quality data or studies simply neglect to reproducibly report their data pipeline. Among reviewed studies, common issues included heterogeneity in design, small and imbalanced data sets, inconsistent data processing and partitioning and lacking validation strategies with only singular metrics (eg, AUROC). While specific train–test sets may be less relevant in non-AI and/or non-prediction studies, alternatives such as stratification were rarely reported. Altogether, these inconsistencies in the data pipeline hamper reproducibility. However, they also extend as significant barriers on a regulatory level when it comes to implementation of data-driven healthcare, in particular under international bodies such as the Medical Device Regulation (MDR).[48] Despite stringent demands with regard to validation and transparency, such regulations lack guidelines on how to achieve this. International regulations such as the MDR could therefore benefit from tools like DATA-CARE to establish guidelines for standardised and reproducible algorithm development.

The strengths of the present study are that it was conducted in a transdisciplinary working group with experts from medical, engineering and research methodology fields. While clinicians are familiar with algorithm output (ie, a clinical outcome), engineers are familiar with the input (ie, data and underlying measurement principles). DATA-CARE comes with comprehensive guidance, including examples, signalling questions and a scoring system. Moreover, DATA-CARE is practical and can be applied to a wide variety of studies on data-driven studies in healthcare. Nevertheless, this study is not without limitations. Because DATA-CARE is intended to be widely applicable to data-driven healthcare, some criteria may be open to interpretation. This is especially the case for criteria in the algorithm development domain, as precise configurations of algorithms (eg, classifier type, intended objective) may vary. However, regardless of the algorithm type, it still requires an architecture with a dedicated input and output or objective, which must be developed and validated. As shown here, agreement between raters using DATA-CARE was overall high, but varied across quality domains. Furthermore, quality assessment was only performed within our own working group, and the possibility of a learning curve within the process was not considered. We still recommend users of DATA-CARE to always carry out quality assessment with two independent reviewers and reach consensus in interpretation of criteria prior to scoring. The compelling need for a widely applicable quality assessment tool for data-driven healthcare supports the present approach.

We encourage further refinement of DATA-CARE. By using the tool prospectively, criteria can be further specified and/or novel criteria can be formulated. Potentially, instead of equal points per domain, some domains may need to be prioritised and receive more points than others. While DATA-CARE is a quality assessment tool, its use highlights recurring methodological issues and reporting issues that could inform the development of future guidelines or checklists for standardised and reproducible data-driven healthcare. Although such guidelines exist, they rarely address the full data pipeline. DATA-CARE's focus on this aspect represents its key novelty and potential contribution to improving study quality. Ultimately, this will contribute to bridging the gap between algorithm development and implementation in clinical care.

Conclusion

DATA-CARE, a quality assessment tool based on the QUIPS, allows reliable critical appraisal for a wide variety of algorithms within data-driven healthcare. The tool is widely applicable, spanning five quality domains that adhere to the data pipeline, addressing current gaps in standardised and reproducible algorithm development.

Acknowledgements The authors wish to thank M F M Engel, W Bramer and E Krabbendam from the Erasmus MC Medical Library for developing and updating the search strategies.

Contributors Conceived and designed the study (including tool development): EvT, BvW, RdJ, JWK, AS and DT. Collected the data (including literature search): EvT. Performed the analysis: EvT and BvW. Wrote the manuscript: EvT. Revised the manuscript: EvT, BvW, RdJ, JWK, AS, HT, DT and MdH. Guarantor: EvT.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing is not applicable as no data sets were generated and/or analysed for this study. All data relevant to the study are included in the article or uploaded as supplementary information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Author note The lead author and guarantor EvT affirms that the manuscript is an honest, accurate and transparent account of the study being reported; no important aspects of the study have been omitted; any discrepancies from the study as planned (and, if relevant, registered) have been explained.

ORCID iD

Eris van Twist https://orcid.org/0000-0002-0968-5400

REFERENCES

1 Bates DW, Saria S, Ohno-Machado L, *et al*. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)* 2014;33:1123–31.

2 Subrahmanya SVG, Shetty DK, Patil V, *et al*. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Ir J Med Sci* 2022;191:1473–83.

3 Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. *Ann Transl Med* 2020;8:501.

4 King AJ, Kahn JM. The Role of Data Science in Closing the Implementation Gap. *Crit Care Clin* 2023;39:701–16.

5 Freitas AT. Data-driven approaches in healthcare: challenges and emerging trends. In: Antunes HS, Sousa Antunes H, Freitas PM, *et al*., eds. *Multidisciplinary perspectives on artificial intelligence and the law*. Cham: Springer International Publishing, 2024: 65–80.

6 Wu W-T, Li Y-J, Feng A-Z, *et al*. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil Med Res* 2021;8:44.

7 Hosseinzadeh E, Afkanpour M, Momeni M, *et al*. Data quality assessment in healthcare, dimensions, methods and tools: a systematic review. *BMC Med Inform Decis Mak* 2025;25:296.

8 Lewis AE, Weiskopf N, Abrams ZB, *et al*. Electronic health record data quality assessment and tools: a systematic review. *J Am Med Inform Assoc* 2023;30:1730–40.

9 Batini C, Scannapieco M. Introduction to information quality. In: *Data and information quality: dimensions, principles and techniques*. Cham: Springer International Publishing, 2016: 1–19.

10 Wells GA, Shea BJ, O'Connell D, *et al*. The newcastle-ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2000.

11 Elm E von, Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.

12 Hayden JA, van der Windt DA, Cartwright JL, *et al*. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158:280–6.

13 Cichy C, Rass S. An Overview of Data Quality Frameworks. *IEEE Access* 2019;7:24634–48.

14 Batini C, Rula A, Scannapieco M, *et al*. From Data Quality to Big Data Quality. *Journal of Database Management* 2015;26:60–82.

15 Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* 2020;2:e200029.

16 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.

17 Collins GS, Dhiman P, Andaur Navarro CL, *et al*. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.

18 Lekadir K, Feragen A, Fofanah AJ, *et al*. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *arXiv* 2023.

19 Norgeot B, Quer G, Beaulieu-Jones BK, *et al*. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020;26:1320–4.

20 Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, *et al*. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc* 2020;27:2011–5.

21 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.

22 Kwong JCC, Khondker A, Lajkosz K, *et al*. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. *JAMA Netw Open* 2023;6:e2335377.

23 Moons KGM, Damen JAA, Kaul T, *et al*. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 2025;388:e082505.

24 Sounderajah V, Ashrafian H, Rose S, *et al*. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med* 2021;27:1663–5.

25 de Jonge RCJ, van Furth AM, Wassenaar M, *et al*. Predicting sequelae and death after bacterial meningitis in childhood: a systematic review of prognostic studies. *BMC Infect Dis* 2010;10:232.

26 Samuels N, van de Graaf RA, de Jonge RCJ, *et al*. Risk factors for necrotizing enterocolitis in neonates: a systematic review of prognostic studies. *BMC Pediatr* 2017;17:105.

27 Dias S, Sutton AJ, Welton NJ, *et al*. Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. London NICE Decision Support Unit Technical Support Documents. National Institute for Health and Care Excellence (NICE); 2012.

28 Batini C, Scannapieco M. Data quality dimensions. In: Batini C, Scannapieco M, eds. *Data and information quality: dimensions, principles and techniques*. Cham: Springer International Publishing, 2016: 21–51.

29 Nyquist H. Certain Topics in Telegraph Transmission Theory. *Trans Am Inst Electr Eng* 1928;47:617–44.

30 Haixiang G, Yijing L, Shang J, *et al*. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst Appl* 2017;73:220–39.

31 Graham JW. Missing Data Analysis: Making It Work in the Real World. *Annu Rev Psychol* 2009;60:549–76.

32 Singh D, Singh B. Investigating the impact of data normalization on classification performance. *Appl Soft Comput* 2020;97:105524.

33 Jundong L, Cheng K, Wang S, *et al*. Feature Selection: A Data Perspective. *ACM Comput Surv* 2017;50:94.

34 Rehman A, Naz S, Razzak I. Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems* 2022;28:1339–71.

35 Hunter DJ, Holmes C. Where Medical Statistics Meets Artificial Intelligence. *N Engl J Med* 2023;389:1211–9.

36 Bourke C, Deng K, Scott SD, *et al*. On reoptimizing multi-class classifiers. *Mach Learn* 2008;71:219–42.

37 Vollmer S, Mateen BA, Bohner G, *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927.

38 Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.

39 Tuijn S, Janssens F, Robben P, *et al*. Reducing interrater variability and improving health care: a meta-analytical review. *J Eval Clin Pract* 2012;18:887–95.

40 Greenland S, Senn SJ, Rothman KJ, *et al*. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31:337–50.

41 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.

42 Catchpole K, Russ S. The problem with checklists. *BMJ Qual Saf* 2015;24:545–9.

43 Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 2023;4:100804.

44 Heneghan JA, Walker SB, Fawcett A, *et al*. The Pediatric Data Science and Analytics Subgroup of the Pediatric Acute Lung Injury and Sepsis Investigators Network: Use of Supervised Machine Learning Applications in Pediatric Critical Care Medicine Research. *Pediatr Crit Care Med* 2024;25:364–74.

45 van de Sande D, van Genderen ME, Huiskens J, *et al*. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med* 2021;47:750–60.

46 Fu L-H, Schwartz J, Moy A, *et al*. Development and validation of early warning score system: A systematic literature review. *J Biomed Inform* 2020;105:103410.

47 Muralidharan V, Adewale BA, Huang CJ, *et al*. A scoping review of reporting gaps in FDA-approved AI medical devices. *NPJ Digit Med* 2024;7:273.

48 Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). 2017.1–175.

49 Azriel R, Hahn CD, De Cooman T, *et al*. Machine learning to support triage of children at risk for epileptic seizures in the pediatric intensive care unit. *Physiol Meas* 2022;43.

50 Rooney SR, Reynolds EL, Banerjee M, *et al*. Prediction of extubation failure in the paediatric cardiac ICU using machine learning and high-frequency physiologic data. *Cardiol Young* 2022;32:1649–56.

51 Badke CM, Marsillio LE, Carroll MS, *et al*. Development of a Heart Rate Variability Risk Score to Predict Organ Dysfunction and Death in Critically Ill Children. *Pediatr Crit Care Med* 2021;22:e437–47.

52 Joram N, Beqiri E, Pezzato S, *et al*. Continuous Monitoring of Cerebral Autoregulation in Children Supported by Extracorporeal Membrane Oxygenation: A Pilot Study. *Neurocrit Care* 2021;34:935–45.

53 Amiri P, Abbasi H, Derakhshan A, *et al*. Potential prognostic markers in the heart rate variability features for early diagnosis of sepsis in the pediatric intensive care unit using convolutional neural network classifiers. 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) in Conjunction with the 43rd Annual Conference of the Canadian Medical and Biological Engineering Society. 10.1109/EMBC44109.2020.9175481 Available: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9167168

54 Castiñeira D, Schlosser KR, Geva A, *et al*. Adding Continuous Vital Sign Information to Static Clinical Data Improves the Prediction of Length of Stay After Intubation: A Data-Driven Machine Learning Approach. *Respir Care* 2020;65:1367–77.

55 Sorensen MW, Sadiq I, Clifford GD, *et al*. Using pulse oximetry waveforms to detect coarctation of the aorta. *Biomed Eng Online* 2020;19:31.

56 Bose SN, Verigan A, Hanson J, *et al*. Early identification of impending cardiac arrest in neonates and infants in the cardiovascular ICU: a statistical modelling approach using physiologic monitoring data. *Cardiol Young* 2019;29:1340–8.

57 Marsillio LE, Manghi T, Carroll MS, *et al*. Heart rate variability as a marker of recovery from critical illness in children. *PLoS ONE* 2019;14:e0215930.

58 Messinger AI, Bui N, Wagner BD, *et al*. Novel pediatric-automated respiratory score using physiologic data and machine learning in asthma. *Pediatr Pulmonol* 2019;54:1149–55.

59 Matam BR, Duncan H, Lowe D. Machine learning based framework to predict cardiac arrests in a paediatric intensive care unit: Prediction of cardiac arrests. *J Clin Monit Comput* 2019;33:713–24.

60 Kamaleswaran R, Akbilgic O, Hallman MA, *et al*. Applying Artificial Intelligence to Identify Physiomarkers Predicting Severe Sepsis in the PICU. *Pediatr Crit Care Med* 2018;19:e495–503.

61 Rusin CG, Acosta SI, Shekerdemian LS, *et al*. Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data. *J Thorac Cardiovasc Surg* 2016;152:S0022-5223(16)30106-4:171–7:.

62 Zhang Y. Real-time development of patient-specific alarm algorithms for critical care. Annu Int Conf IEEE Eng Med Biol Soc; 2007:4351–4.

63 Biswas AK, Scott WA, Sommerauer JF, *et al*. Heart rate variability after acute traumatic brain injury in children. *Crit Care Med* 2000;28:3907–12.

64 Si Y, Gotman J, Pasupathy A, *et al*. An expert system for EEG monitoring in the pediatric intensive care unit. *Electroencephalogr Clin Neurophysiol* 1998;106:488–500.

65 Martin S, Du Pont-Thibodeau G, Seely AJE, *et al*. n.d. Heart Rate Variability in Children with Moderate and Severe Traumatic Brain Injury: A Prospective Observational Study. *J Pediatr Intensive Care*.

66 Kirschen MP, Majmudar T, Beaulieu F, *et al*. Deviations from NIRS-derived optimal blood pressure are associated with worse outcomes after pediatric cardiac arrest. *Resuscitation* 2021;168:S0300-9572(21)00377-4:110–8:.

67 Matam BR, Fule BK, Duncan HP. Predictability of unplanned extubations. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); Valencia, Spain, 2014

68 Izquierdo LM, Nino LF, Prieto Rojas J. Modeling the vital sign space to detect the deterioration of patients in a pediatric intensive care unit. 16th International Symposium on Medical Information Processing and Analysis; Lima, Peru, 2020

69 Zoodsma RS, Bosch R, Alderliesten T, *et al*. Continuous Data-Driven Monitoring in Critical Congenital Heart Disease: Clinical Deterioration Model Development. *JMIR Cardio* 2023;7:e45190.

70 Tabassum S, Ruesch A, Acharya D, *et al*. Clinical translation of noninvasive intracranial pressure sensing with diffuse correlation spectroscopy. *J Neurosurg* 2023;139:184–93.

71 Liu R, Majumdar T, Gardner MM, *et al*. Association of Postarrest Hypotension Burden With Unfavorable Neurologic Outcome After Pediatric Cardiac Arrest. *Crit Care Med* 2024;52:1402–13.

72 van Twist E, Hiemstra FW, Cramer ABG, *et al*. An electroencephalography-based sleep index and supervised machine learning as a suitable tool for automated sleep classification in children. *J Clin Sleep Med* 2024;20:389–97.

73 Macabiau C, Le T-D, Albert K, *et al*. n.d. Label Propagation Techniques for Artifact Detection in Imbalanced Classes Using Photoplethysmogram Signals. *IEEE Access*12:81221–35.

74 Le T-D, Macabiau C, Albert K, *et al*. A Novel Transformer-Based Self-Supervised Learning Method to Enhance Photoplethysmogram Signal Artifact Detection. *IEEE Access* 2024;12:159860–74.

75 Kwon SB, Weinerman B, Nametz D, *et al*. Non-invasive pulse arrival time is associated with cardiac index in pediatric heart transplant patients with normal ejection fraction. *Physiol Meas* 2024;45:07NT01.

76 Hunfeld M, Verboom M, Josemans S, *et al*. Prediction of Survival After Pediatric Cardiac Arrest Using Quantitative EEG and Machine Learning Techniques. *Neurology (ECronicon)* 2024;103.

77 van Twist E, Meester AM, Cramer ABG, *et al*. Supervised machine learning on electrocardiography features to classify sleep in noncritically ill children. *J Clin Sleep Med* 2025;21:261–8.

78 Silva MJ, Gonçalves H, Almeida R, *et al*. Cardiovascular responses as predictors of mortality in children with acute brain injury. *Pediatr Res* 2025;97:2347–53.