JUPITER-MARS: a reporting framework and preliminary assessment score for pathophysiology-driven clinical digital twin studies.

M. Ester Canepa, Pasquale Striano, Francesco Vinci, Andrea Calandrino, Davide Cangelosi, Luca Ramenghi

Submitted to: JMIR Formative Research
on: May 29, 2026

Table of Contents

JUPITER-MARS: a reporting framework and preliminary assessment score for pathophysiology-driven clinical digital twin studies.

M. Ester Canepa[1] PhD; Pasquale Striano[1, 2] MD, PhD; Francesco Vinci[1, 2] MD, PhD; Andrea Calandrino[1, 2] MD, PhD; Davide Cangelosi[2] MD, PhD; Luca Ramenghi[1, 2] MD, PhD

[1]Department of Neuroscience, Rehabilitation, Genetics, Ophthalmology and Mother-Child Health University of Genoa Genoa IT
[2] Istituto Giannina Gaslini Genoa IT

Corresponding Author:
Andrea Calandrino MD, PhD
Department of Neuroscience, Rehabilitation, Genetics, Ophthalmology and Mother-Child Health
University of Genoa
Largo Daneo, 3
Genoa
IT

Abstract

Background: Clinical digital twins are emerging as promising tools to support personalized medicine by integrating multimodal biological, clinical and computational data into virtual representations of patients or organs. However, despite increasing methodological development in digital twin modelling, standardized reporting frameworks specifically addressing pathophysiology-driven clinical digital twin studies remain limited. Existing validation approaches, such as the Verification, Validation and Uncertainty Quantification (VVUQ) framework, primarily focus on model evaluation rather than structured reporting and methodological transparency.

Objective: This study aims to introduce JUPITER-MARS, a methodological framework designed to support transparent reporting and preliminary assessment of pathophysiology-driven clinical digital twin studies. Specifically, it is a credibility-oriented framework integrating reporting (JUPITER), assessment (MARS), and validation (VVUQ) to address the current methodological gap in clinical digital twin research.

Methods: We developed JUPITER, a structured reporting framework aligned with conventional scientific manuscript architecture, covering seven key domains: justification, underlying pathophysiology, parameter selection, integration modelling, testing (VVUQ), expected outputs and reproducibility. JUPITER is designed not only as a reporting checklist but as a design-oriented framework to guide the development of pathophysiology-driven digital twins. In parallel, we propose MARS, a complementary domain-based preliminary assessment score intended to support structured evaluation of reporting completeness in digital twin studies. To facilitate literature screening and methodological auditing, we additionally developed a prototype computational tool implemented in Python within a Google Colab environment. The tool uses heuristic keyword-based text mining and biomedical language models to identify key methodological reporting elements in digital twin publications. The JUPITER-MARS-VVUQ tool was benchmarked with a corpus of papers (digital twin and computational models using VVUQ) extrapolated from the review by Tudor et al., 2025. Given the limited formal adoption of VVUQ in current clinical digital twin literature, the benchmarking corpus was intentionally small and was used to support an initial methodological validation rather than definitive performance estimation.

Results: Our findings suggest that the lack of VVUQ implementation represents a systemic limitation across both mechanistic and AI-driven digital twin paradigms. This clustering represents a novel methodological taxonomy of clinical digital twins based on physiological grounding and VVUQ maturity. JUPITER provides a structured checklist to guide the design and reporting of pathophysiology-driven digital twin protocols, while MARS enables a preliminary evaluation of reporting transparency across seven domains including data description, multimodal integration, outcome reporting and model performance. The associated prototype computational tool demonstrates the feasibility of supporting structured literature screening and methodological signal detection within the emerging digital twin literature.

Conclusions: JUPITER-MARS offers a structured methodological framework to improve transparency, reproducibility and comparability in clinical digital twin research. By complementing existing validation approaches such as VVUQ, the framework

may support both prospective digital twin development and systematic evaluation of the growing digital twin literature. Future studies may adopt JUPITER-MARS to standardize reporting practices and facilitate methodological rigor in digital twin research across medical domains. Clinical Trial: Not applicable.

(JMIR Preprints 29/05/2026:102850)
DOI: https://doi.org/10.2196/preprints.102850

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Yes – Make my manuscript publicly available as a preprint immediately.**
No – I do not wish to publish a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

JUPITER-MARS: a reporting framework and preliminary assessment score for pathophysiology-driven clinical digital twin studies.

ME Canepa (PhD)[1], P Striano (MD/PhD)[1,2], F Vinci (MD/PhD)[1,2], A Calandrino (MD/PhD)[1,2] , D Cangelosi (MD/PhD)[2], LA Ramenghi (MD/PhD)[1,2]
1 Department of Neuroscience, Ophthalmology, Genetics and Mother-Child Health, University of Genoa, Genoa, Italy
2 IRCCS Giannina Gaslini, Genoa, Italy.
Corresponding author: dr. Andrea Calandrino (andreacalandrino@gaslini.org).
Keywords: digital twin1; pathophysiology2; score3; reporting4;

Abstract

Background:
Clinical digital twins are emerging as promising tools to support personalized medicine by integrating multimodal biological, clinical and computational data into virtual representations of patients or organs. However, despite increasing methodological development in digital twin modelling, standardized reporting frameworks specifically addressing pathophysiology-driven clinical digital twin studies remain limited. Existing validation approaches, such as the Verification, Validation and Uncertainty Quantification (VVUQ) framework, primarily focus on model evaluation rather than structured reporting and methodological transparency.

Objective:
This study aims to introduce JUPITER-MARS, a methodological framework designed to support transparent reporting and preliminary assessment of pathophysiology-driven clinical digital twin studies. Specifically, it is a credibility-oriented framework integrating reporting (JUPITER), assessment (MARS), and validation (VVUQ) to address the current methodological gap in clinical digital twin research.

Methods:
We developed JUPITER, a structured reporting framework aligned with conventional scientific manuscript architecture, covering seven key domains: justification, underlying pathophysiology, parameter selection, integration modelling, testing (VVUQ), expected outputs and reproducibility. JUPITER is designed not only as a reporting checklist but as a design-oriented framework to guide the development of pathophysiology-driven digital twins. In parallel, we propose MARS, a complementary domain-based preliminary assessment score intended to support structured evaluation of reporting completeness in digital twin studies. To facilitate literature screening and methodological auditing, we additionally developed a prototype computational tool implemented in Python within a Google Colab environment. The tool uses heuristic keyword-based text mining and biomedical language models to identify key methodological reporting elements in digital twin publications. The JUPITER-MARS-VVUQ tool was benchmarked with a corpus of papers (digital twin and computational models using VVUQ) extrapolated from the review by Tudor et al., 2025. Given the limited formal adoption of VVUQ in current clinical digital twin literature, the benchmarking corpus was intentionally small and was used to support an initial methodological validation rather than definitive performance estimation.

Results:
Our findings suggest that the lack of VVUQ implementation represents a systemic limitation across both mechanistic and AI-driven digital twin paradigms. This clustering represents a

novel methodological taxonomy of clinical digital twins based on physiological grounding and VVUQ maturity. JUPITER provides a structured checklist to guide the design and reporting of pathophysiology-driven digital twin protocols, while MARS enables a preliminary evaluation of reporting transparency across seven domains including data description, multimodal integration, outcome reporting and model performance. The associated prototype computational tool demonstrates the feasibility of supporting structured literature screening and methodological signal detection within the emerging digital twin literature.

Conclusions:
JUPITER-MARS offers a structured methodological framework to improve transparency, reproducibility and comparability in clinical digital twin research. By complementing existing validation approaches such as VVUQ, the framework may support both prospective digital twin development and systematic evaluation of the growing digital twin literature. Future studies may adopt JUPITER-MARS to standardize reporting practices and facilitate methodological rigor in digital twin research across medical domains

1. Introduction

Digital twins are defined as viewable, virtual replicas of a patient, organ or biological system that contain specific multidimensional information used to inform medical decision-making (1). They have the potential to foster the personalization of care (2).
To achieve these aims of care tailoring, digital twins must be built by balancing pathophysiological knowledge, available data and computational cost, as well as selecting the most appropriate parameters required to address the research question modelled by the digital twin (3).
Several initiatives have been proposed to regulate digital twins in science and medicine, such as the Digital Twin Manifesto for the Pathology Laboratory (4) and the Virtual Human Twin Manifesto proposed by the European Commission in 2023. Nowadays, Equator Network checklists are available for checking diagnostic accuracy and scientific reliability (TRIPOD, STARD, CONSORT, SPIRIT) with AI as well as for providing deployable, trustworthy and explainable medical AI (FUTURE-AI checklist) (5). Noteworthy, the group by Bignami et al. proposed in 2025 a methodology checklist to implement medical AI in policy documents (6) in accordance with AI acts. However, the items “Privacy by Design and Privacy by Default” and “VII. Identify regulatory requirements” are challenging as reported in a review published in 2024 on PLOS Digital Health (7), with social, ethical and legal concerns.
Recent advances in schema-constrained extraction pipelines have improved large-scale evidence synthesis (8)**;** however, these approaches primarily address data extraction rather than the methodological design, validation, and reporting of clinical digital twin models. To the best of our knowledge, the literature still lacks dedicated reporting frameworks specifically designed for pathophysiology-driven clinical digital twin studies, beyond existing validation approaches such as the VVUQ framework (9,10) for clinical digital twins.
To address this gap, we introduce JUPITER, a structured reporting framework designed to guide the design and reporting of pathophysiology-based digital twin protocols suitable for ethical committee submissions and scientific publications. In addition, we present MARS, a complementary domain-based preliminary assessment score intended to support structured evaluation of reporting completeness in clinical digital twin studies. JUPITER and MARS are not only mythological names referring to the king and the warrior among the Roman gods. Additionally, they stand for the key purposes of these frameworks:

J – Justification
U – Underlying pathophysiology
P – Parameters selection
I – Integration modelling
T – Testing (VVUQ)
E – Expected outputs
R – Reproducibility and reporting

M - Multimodal
A – Assessment (of)
R – Reporting
S – Score

2. Methods

Created in collaboration with medical and technical experts, JUPITER and MARS are designed to be flexible across medical specializations and compatible with structured evidence synthesis frameworks such as ATLAS (11 - in submission), a methodological framework for systematic reviews in quantitative imaging research enhanced with radiomics and artificial intelligence (AI). JUPITER is structured according to a conventional scientific manuscript architecture and is intended to function both as a reporting guideline and as a design-oriented framework (12 - in submission) to enhance standardization.

2.1. Development of JUPITER framework

In table 1, JUPITER sections are described in detail.

Tab. 1 – JUPITER framework with guide explanation per item

Item	Explanation
Rationale	Provide the rationale underlying the model, specifying the type of digital twin according to Drummond and Gonsard (1) as reported in Table 2 and the research question.
Background	Illustrate the state of the art of pathophysiology or pathological anatomy to be modularized with your digital twin. If available and VVUQ-reliable, describe shortly the previous digital twins model on your topic.
Method	Describe how to acquire medical data (imaging: indicate plan, type, number of sessions; monitoring devices data; clinical biomarkers, etc) with specifying time intervals windows and instruments to collect data. Furthermore, it is essential to motivate clinical parameters choice as illustrated in tab. 3. Thereafter, provide a figure of Swiss Cheese Model to assign a layer to each kind of data. The Swiss Cheese model may be used as a conceptual visualization framework to represent multimodal pathophysiological

	layers within the digital twin architecture, as described in Cell for glioblastomas (13) or prescribing drugs barriers (14) in computational modules according to the framework by Masison et al. (15) as well as integrating spatial and temporal aspects in a digital twin model guaranteeing information reliability (16). More specifically, the Swiss Cheese model could be applied to represent multimodal data integration, where complementary datasets mitigate individual limitations, enhancing pathophysiological fidelity of the digital twin. For example, in a neonatal respiratory digital twin, ventilatory parameters, imaging data, and blood gas measurements may be represented as distinct layers within the Swiss Cheese model. Their integration enables compensation of individual data limitations (e.g., imaging sparsity or monitoring noise), thereby improving the physiological fidelity and robustness of the digital twin. In this case, Hot Cheese model (17) may additionally serve as a conceptual analogy. It is represented by primary organ injuries and pathophysiology disruption as forces, concomitant latent conditions as loopholes (pressure, glycemia disruptions, etc.) and drips as the new risks for the neonate (secondary heart, lung, brain injury) contributing to systemic instability over time. If AI is used to analyze data, a structured AI hazard results analysis should be included with the TAIHA Taxonomy (18) as described by Cummings in 2024. Describe how to analyze and cluster data. If artificial intelligence is used, please indicate it. If virtual organ modelling is used, please mention it with detailed method plan (i.e. The Virtual Brain; The Virtual Epileptic Patient, etc.). The Hot Cheese model (17) may be used to represent dynamic risk accumulation across layers of the digital twin architecture. In this

	adaptation, forces represent primary lesions or injuries, loopholes represent latent pathological conditions, and drips represent emerging risk factors contributing to systemic instability. The VVUQ framework should be included in the methods section as a standard approach for model validation.
Expected results	Provide information on expected outputs and their impacts on clinical practice by your digital twins.
Results	In the results section, VVUQ framework results should be reported. Moreover, the number of identified pathophysiological clusters should be reported. The contribution of the model to previous pathophysiological knowledge should be described.
Discussion	Comment the results obtained based on the previous research. Furthermore, specify the following: the digital twin has provided results in accordance with clinical reality? If yes/no, say why.
Conclusions and further application	These sections should summarize the results obtained and discuss the strengths, limitations, and possible future applications of the model.

Tab. 2 – Glossary of kind of digital twins as provided by Drummond et Gonsard, 2024 (1)

Type	Explanation
Simulation	The model is designed to simulate a condition/outcome based on patients' data
Monitoring	The model is designed to be constantly upgraded with current smart devices data (digital stethoscope, wearable devices like smartwatches, etc.) in a bidirectional way from patient data to the digital twin system.
Research-oriented	The model is designed to foster knowledge on a research realm without clinical practice.

Tab. 3 – Template for motivating clinical parameters in the Method Section

Parameter	Method of acquiring and analysis	Motivation
Indicate a parameter per line in this column	Describe how to acquire and analyse data, indicating the data analysis tools/software. If a data collection technique could be of ethical concern, please provide strategies to mitigate/avoid them as much as possible.	Explain the reason why of choice based on the current evidence

Tab. 4 – Synthetic description per section and item of JUPITER.

Section	Item	Description
Justification	Research question	Clearly state the clinical question addressed by the digital twin
Underlying pathophysiology	Biological rationale	Describe the pathophysiological mechanisms represented
Parameters	Data selection	Specify clinical/imaging/biomarker parameters included
Integration modelling	Model architecture	Describe modelling approach and integration of modalities
Testing	Validation	Report VVUQ or equivalent validation framework
Expected outputs	Clinical relevance	Specify predicted outputs and potential clinical impact
Reproducibility	Reporting transparency	Provide sufficient methodological details for reproducibility

2.2. Development of MARS framework: preliminary reporting assessment

To facilitate structured evaluation of reporting completeness in digital twin studies, we introduce MARS, a domain-based preliminary assessment score designed to complement the JUPITER reporting framework.

MARS is intended primarily for literature review and methodological auditing purposes. It does not replace formal methodological evaluation but provides a structured overview of key reporting components in clinical digital twin studies.

The MARS framework includes seven domains reflecting essential elements of transparent digital twin reporting with preliminary weights:

- Dataset Overview → 5%
- Systematic Overview & Log → 15%
- Standardized Twin Reporting Architecture → 20%
- Multimodal Data & Integration → 20%
- Outcome & Biomarker Reporting → 15%
- Feasibility & Clinical Utility → 10%
- Model Performance → 15%

These domains reflect core components necessary to evaluate reporting transparency, methodological clarity and reproducibility in digital twin research. Furthermore, preliminary domain weights were assigned on the basis of expert-informed prioritization of reporting

components considered most relevant for transparency and interpretability in clinical digital twin studies. The proposed domain weights should be considered preliminary and hypothesis-generating rather than definitive, pending future external validation and international expert consensus studies.

2.3. JUPITER-MARS computational tool, benchmark corpus and validation procedure

A prototype screening aid was implemented to demonstrate feasibility of structured methodological auditing using Python and implemented in a Google Colab environment. It is intended to facilitate structured literature screening,
The JUPITER-MARS tool uses heuristic keyword-based text mining to extract relevant methodological signals from digital twin publications in PDF format. These signals are mapped to JUPITER reporting items and MARS assessment domains.
The tool can support literature reviews by:

- identifying mentions of common validation metrics (e.g., AUC, ROC, sensitivity and specificity)
- estimating a preliminary VVUQ fulfillment score
- detecting references to internal or external validation procedures
- assisting preliminary classification of documents using biomedical language models such as SciBERT or BioBERT
- generating structured matrices to assist evidence synthesis workflows.

The outputs generated by the tool are intended as preliminary screening aids and should always be interpreted and verified by human reviewers. To provide an initial methodological validation, the JUPITER-MARS tool was benchmarked on a curated corpus of digital twin studies with explicit or implicit VVUQ-related features, selected from the review by Tudor et al., 2025 (19). They were digital twins (20–23) and computational models **(24).** PDFs were downloaded from PubMed and thereafter, they were submitted to MARS-VVUQ and JUPITER tools. Due to the limited formal adoption of VVUQ in clinical digital twin literature (19). Studies were categorized based on explicit, implicit, or absent adherence to VVUQ principles. The outputs provided by the tool were checked manually with an AI expert (fifth author).

Tab. 5 – Outputs generated with JUPITER-MARS-VVUQ Tool compared with Tudor et al., 2025 review's snippets. (19)

Study	Tudor et al., 2025 review's snippets	VVUQ-MARS Score	JUPITER Score (Quality)	Disease
Bethéncourt et al., 2021	"Explicit mention of VVUQ"	Partial VVUQ; 65 points MARS Score	12/14 (85.7%)	Lymphedema monitoring
Salvador et al., 2024	"(this study) addressed many of the VVUQ concerns without explicitly mentioning the term"	VVUQ-compliant; 74 points MARS Score	9/14 (64.3%)	Pediatric cardiology (HLHS)
Santiago et al., 2022	"(this study) showcase detailed	VVUQ compliant. 87 points of MARS Score.	12/14 (85.7)	Computational model of left ventricular flow

	strategies toward the design and execution of robust VVUQ plans, exemplifying how standardized methodologies can enhance model reliability and stakeholder confidence".			after Left Ventricular Assist Device (LVAD) implementation
Silfvergren et al., 2022	"(this study) addressed many of the VVUQ concerns without explicitly mentioning the term"	Weak/insufficient VVUQ. 71 points of MARS Score.	12/14 (85.7%)	Metabolism (resting/diet)
Thamotaran et al., 2023	(This study) Met the NASEM standard digital twin criteria; no mention of VVUQ	Weak/ insufficient VVUQ. 85 points of MARS Score	11/14 (78.6%)	Diabetes 2 monitoring (geriatrics)

3. Results

3.1. Benchmark evaluation of Digital Twin studies.

Our findings suggest that the lack of VVUQ implementation represents a systemic limitation across both mechanistic and AI-driven digital twin paradigms. Greater compliance with VVUQ principles and JUPITER items was associated with higher MARS-DT scores. These findings, summarized in Table 5 and Figures 1–2, support a structured methodological evaluation of clinical digital twins integrating reporting quality (MARS), conceptual completeness (JUPITER), and credibility assessment through Verification, Validation and Uncertainty Quantification (VVUQ, fig. 3). A VVUQ standardized pathway is represented by Santiago et al.'s study showing a step-by-step procedure table.

Bethencourt et al. (20) demonstrated high methodological completeness according to JUPITER criteria, particularly in parameterization and integration modelling. Specifically, mechanistic models, such as those exemplified by Silfvergren et al. (22), demonstrate strong physiological interpretability and alignment with known biological processes. However, these models consistently lack formal VVUQ frameworks, limiting their credibility and clinical translatability despite their theoretical robustness. Conversely, AI-driven and control-based models, as observed in Thamotaran et al. (23), exhibit advanced personalization and optimization capabilities through techniques such as model predictive control and machine learning integration. Despite high reporting completeness and technical sophistication, these approaches often lack both explicit physiological grounding and rigorous VVUQ components, resulting in biologically opaque systems with limited interpretability and uncertain reliability. By contrast, Santiago et al. (24) represents an

engineering-grade benchmark for VVUQ implementation, achieving full compliance through an explicitly designed verification, validation, and uncertainty quantification plan, demonstrating that maximum model credibility can be attained even outside the digital twin paradigm. However, the absence of explicit verification processes and limited validation strategies resulted in only partial VVUQ compliance, highlighting a critical gap between model construction and model credibility assessment. Salvador et al. (21) represents a fully VVUQ-compliant digital twin, integrating physics-based modelling, surrogate learning, and Bayesian uncertainty quantification with direct validation against patient-specific clinical data, thereby achieving high model credibility despite only moderate reporting completeness according to JUPITER criteria.

3.2. Methodological clustering of clinical digital twins

Accordingly, we introduce a novel conceptual taxonomy of clinical digital twins based on physiological grounding and validation maturity. Indeed, a key finding is the emergence of distinct digital twin archetypes, characterized by differing balances between physiological grounding referring to the extent to which a model explicitly represents biological or pathophysiological mechanisms, rather than relying predominantly on data-driven optimization alone, and computational sophistication. Based on differences in physiological grounding, reporting completeness, and VVUQ maturity, five methodological clusters were identified in Fig. 4.

- Cluster 1 (20): Low physiological grounding; moderate reporting completeness (MARS/JUPITER), low VVUQ maturity, low clinical integration, low overall maturity.
- Cluster 2 (22): High physiological grounding; high reporting completeness (MARS/JUPITER), low VVUQ maturity, moderate clinical integration, intermediate overall maturity.
- Cluster 3 (23). Moderate physiological grounding; high reporting completeness; low VVUQ maturity; moderate clinical integration; intermediate overall maturity
- Cluster 4 (21). Moderate-to-high physiological grounding; high reporting completeness, VVUQ maturity, clinical integration and overall maturity.
- Cluster 5 (24). Moderate physiological grounding; high reporting completeness, VVUQ maturity and overall maturity; moderate clinical integration.

Fig. 1 – JUPITER domains heatmap per paper generated with MARS-VVUQ and JUPITER tools. Picture generated with MARS Tool.

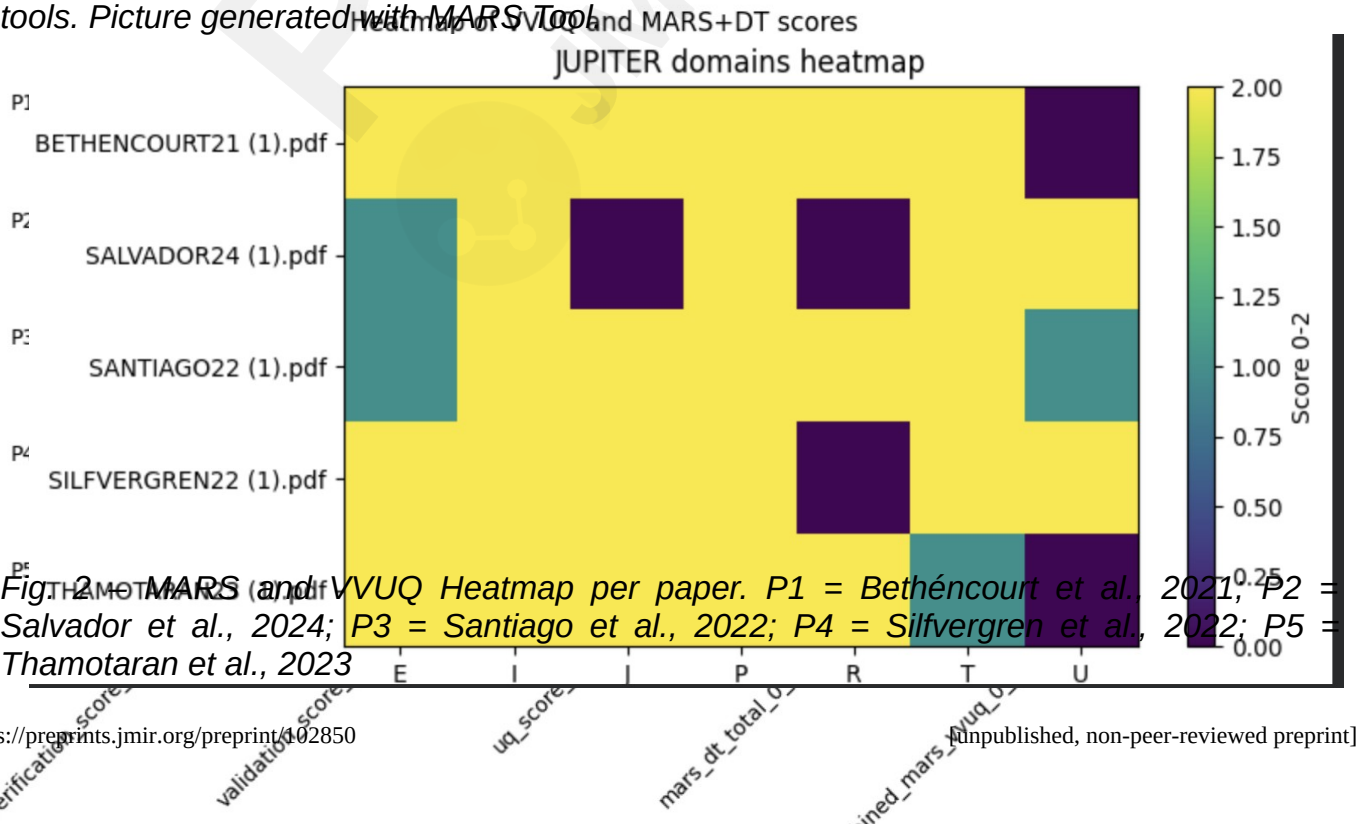

Fig. 2 – MARS and VVUQ Heatmap per paper. P1 = Bethéncourt et al., 2021; P2 = Salvador et al., 2024; P3 = Santiago et al., 2022; P4 = Silfvergren et al., 2022; P5 = Thamotaran et al., 2023

Aspect				Evaluation		
				Max.	Goal	Obt.
1. Verification (Sec. *Verification*)	1.1. Code	1.1.1. software quality assurance (SQA)		C	B [SQA procedures are specified and documented.]	C
		1.1.2. Numerical code verification		D	C [The numerical solution is compared to an exact solution.]	D
	1.2. Calculation	1.2.1. Discretisation error LightGray		C	B [Convergence analysis are performed obtaining stable behaviours.]	C
		1.2.2. Numerical solver error		C	B [Solver parameters are based on values from a previously verified model.]	B
		1.2.3. User error		D	B [Key inputs and outputs were verified by [redacted].]	C
2. Validation (Secs. *Sensitivity analysis* and *Validation with uncertainty quantification*)	2.1 Computational model	2.1.1. Model form		C	B [Influence of some assumptions is explored.]	B
		2.1.2 Model inputs	2.1.2.1. Quantification of sensitivities	C	B [A SA of the expected key parameters is performed.]	B
			2.1.2.2. Quantification of uncertainties	D	B [UQ is executed on expected key inputs but not propagated to the QoIs.]	C
	2.2. Comparator	2.2.1. Test samples	2.2.1.1. Quantity of test samples	C	A [A single sample is used.]	A
			2.2.1.2. Range of characteristics of test samples	D	A [A single test condition is examined.]	A
			2.2.1.3. Measurements of test samples	C	B [One or more key characteristic are measured.]	C
			2.2.1.4. Uncertainty of test samples measurements	C	A [Characteristics uncertainty is not addressed.]	A
		2.2.2. Test conditions	2.2.2.1. Quantity of test conditions	B	B [Multiple test conditions.]	B
			2.2.2.2. Range of test conditions	D	B [Test conditions representing a range of conditions near nominal range are examined.]	C
			2.2.2.3. Measurements of test conditions	C	B [One or more key test conditions are measured.]	B
			2.2.2.4. Uncertainty of test conditions measurements	C	B [UQ of the test conditions incorporated instrument accuracy only.]	B
	2.3. Assessment	2.3.1. Equivalence		C	B [The types of all inputs are similar, but ranges are not equivalent.]	C
		2.3.2. Output Comparison	2.3.2.1. Quantity	B	B [Multiple outputs were compared.]	B
			2.3.2.2. Equivalency of output parameters	C	B [Most types of outputs are similar.]	C
			2.3.2.3. Rigour of output comparison	C	B [Comparison was performed by arithmetic difference.]	C
			2.3.2.4. Agreement of output comparison	C	B [The level of agreement is satisfactory for some key comparisons.]	B
3. Applicabilty (Sec. *Discussion on the V&V40 credibility factors: Achieved score*)	3.1. Relevance of the Quantity of interest			C	B [A subset of the QoIs are identical to those for the CoU.]	C
	3.2. Relevance of the validation activities to the CoU			D	B [There is partial overlap between the ranges of the validation points and the CoU.]	C

Fig. 3 – VVUQ Standardized Pathway (by Santiago et al., 2022) based on ASME criteria (25). UQ = Uncertainty Quantification; SQA = Software Quality Assurance

Fig. 4. Heatmap of methodological maturity across clinical digital twin clusters. Legend (color scale): Low / insufficient (red); Moderate / intermediate (yellow); High / strong (green).

Cluster	Study	Physiological grounding	Reporting completeness (MARS/JUPITER)	VVUQ maturity	Clinical integration	Overall maturity
C1	Bethencourt 2021	Low	Moderate	Low	Low	Low
C2	Silfvergren 2022	High	High	Low	Moderate	Intermediate
C3	Thamotaran 2023	Moderate	High	Low	Moderate	Intermediate
C4	Salvador 2024	Moderate–High	High	High	High	High
C5	Santiago 2022	Moderate	High	High	Moderate	High

Tab. 6 - Methodological clustering of clinical digital twin studies based on physiological grounding, reporting completeness, and VVUQ maturity. Legend: Physiological grounding: extent to which the model reflects biological/pathophysiological mechanisms; Reporting completeness: evaluated through JUPITER and MARS frameworks; VVUQ maturity: degree of implementation of verification, validation and uncertainty quantification (VVUQ)

Clus	Label	Physiolo	Reporting	VVU	Represen	Key	Main

ter		gical groundi ng	completen ess (MARS/JU PITER)	Q matu rity	tative studies	characteris tics	limitation s
C1	Low-structur e digital twins	Low	Moderate	Low	Bethenco urt et al., 2021 (20)	Early-stage implementat ion; basic parameter modeling; limited integration	Lack of validation , minimal physiologi cal modelling , weak reproduci bility
C2	Mechan istic non-validate d twins	High	High	Low	Silfvergre n et al., 2022 (22)	Strong pathophysio logical modeling; interpretabl e systems; biologically grounded	Absence of formal VVUQ; limited credibility and clinical trustworth iness
C3	AI/ control-driven opaque twins	Low–Moderate	High	Low	Thamotar An et al., 2023 (23)	Advanced optimization (MPC, ML); personalizat ion capabilities; high technical sophisticati on	Limited physiologi cal interpreta bility; weak VVUQ; potential black-box behavior
C4	Clinicall y-oriented validate d twins	Moderate –High	Moderate–High	High	Salvador et al., 2024 (21)	Integration of clinical data and validation pipelines; emerging reproducibili ty	Partial reporting gaps; heteroge neity in methodol ogical standards
C5	Enginee ring-grade validate d twins	Moderate	High	High	Santiago et al., 2022 (24)	Strong VVUQ implementat ion; rigorous validation; reproducibl e computation al pipelines	Limited clinical integratio n; reduced physiologi cal interpreta bility in some

contexts

3.3. Neonatal proof-of-concept

Thereafter, two neonatal digital twin studies by Forster et al. and Sizemore et al. (26,27) were tested using MARS-JUPITER tool to preliminarily assess the applicability of the JUPITER–MARS framework. The obtained score and original PDFs' snippets are illustrated in fig. 5 and tables 7-8. Forster et al., 2021 is a mechanistic/physiological digital twin with a more "in-silico model" nature as this study used the clinical and ventilatory parameters from another study – that's why JUPITER scored 71/100 (good; 10/14 score per domain); moreover, it showed the 3D representation of airway. Particularly, 27 weeks of gestation patient suffering from pulmonary bronchodysplasia. Hereafter, we reported the snippets detected from original PDF text-mining by JUPITER-MARS tool for JUPITER evaluation:

Fig. 5 – Heatmap of JUPITER (above) and MARS-VVUQ scores (below) for Forster et al., 2021 (P1) and Sizemore et al., 2024 (P2).

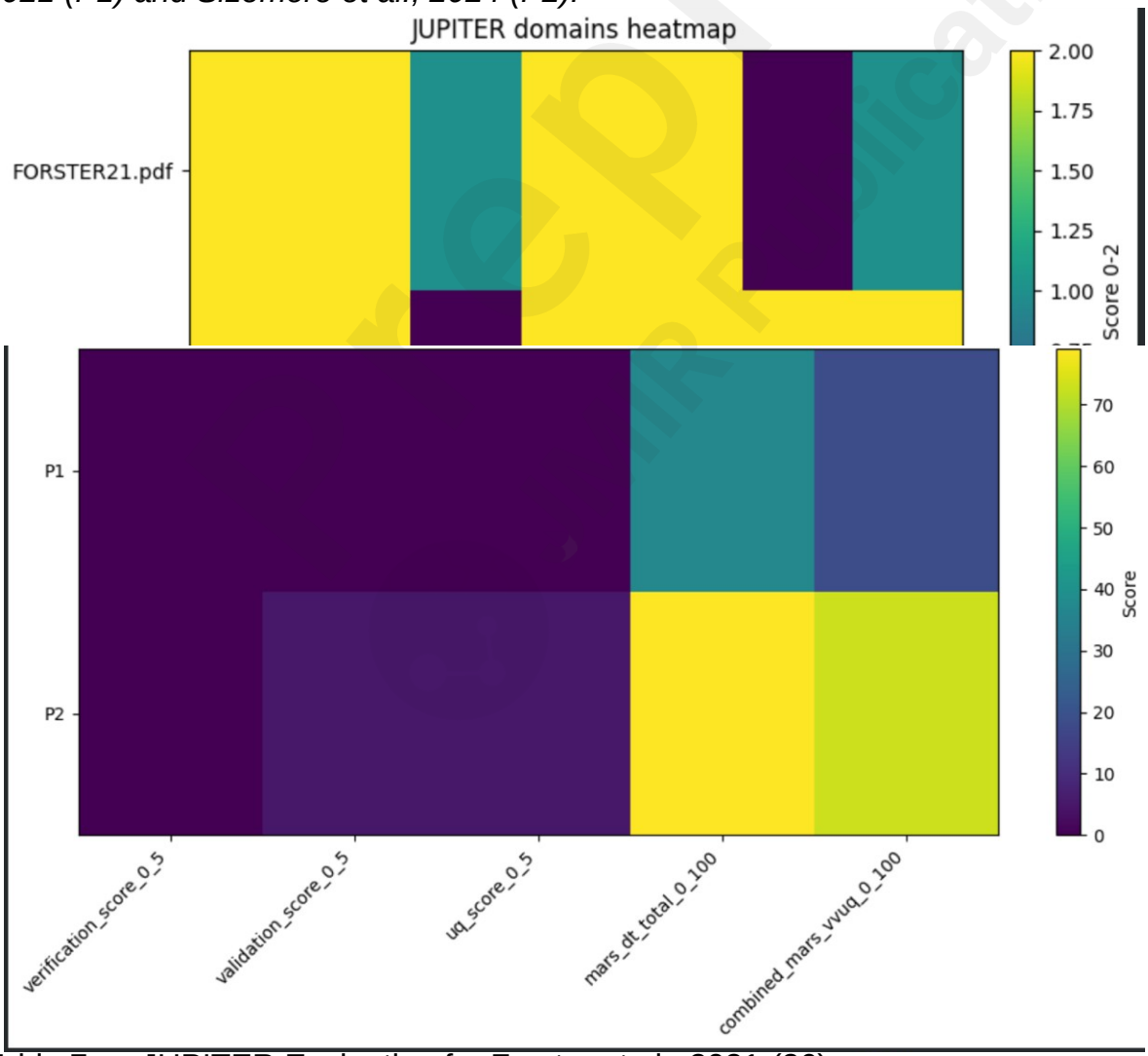

Table 7a – JUPITER Evaluation for Forster et al., 2021 (26)

Justification Score: 1/2 Hits: 1 Weighted hits: 1.5	Underlying Pathophysiology Score: 1/2 Hits: 2 Weighted hits: 2.0	Parameter selection Score: 2/2 Hits: 3 Weighted hits: 3.6	Integration modelling Score: 2/2 Hits: 3 Weighted hits: 3.4	Testing Score: 0/2 Hits: 0 Weighted hits: 0.0	Expected outputs Score: 2/2 Hits: 4 Weighted hits: 5.0	Reproducibility reporting Score: 2/2 Hits: 3 Weighted hits: 3.2
Airway morphology and LM and thus creating a virtual “digital twin.” Slutsky et al. proposed alternative mechanisms of gas exchange such as molecular diffusion, Taylor dispersion, turbulence, asymmetric ventilation	During conventional mechanical ventilation pulmonary gas exchange is maintained by direct alveolar aeration, applying tidal volume to the anatomical dead space, the large airways, and the alveolar compartment. Tidal volumes at or below the anatomical dead space during CV lead to dete (trunked) - (a)eration, applying tidal volume to the anatomical dead space, the large airways, and the alveolar compartment. Tidal volumes at or below the anatomical dead space during CV lead to deterioration of ventilation through CO2-accumulation and may subsequently even impair oxygenation by at (trunked words)	- in the evaluation of OD, our future goal is to investigate the entire gas exchange, including CO2 elimination. This is dependent on several parameters, most notably the perfusion. We plan to address this topic by modeling echocardiography data in the next step. The clinical applicability - (trunked word) ional oxygenation during HFOV. There are limitations to our study that need to be considered when interpreting the data. First, the medical imaging data used in this study were obtained from a preterm infant at the age of 3 months. This might underestimate the flow situation in an ext (trunked words) - (trunked words) ave adequate plausibility checks in our modeling process.12 Beyond that, the difficulties of invasive measurements and sequential medical imaging, especially in the preterm infant, are over- come by means of our numerical approach using an advanced computational lung model based on th (trunked words)	- notably the perfusion. We plan to address this topic by modeling echocardiography data in the next step. The clinical applicability of our model is currently curtailed as we are dealing with a heterogeneous patient population with respect to the disease pro- cess and the range - medical imaging, especially in the preterm infant, are overcome by means of our numerical approach using an advanced computational lung model based on the real anatomy and physiology of the infant's lung. This detailed insight into HFOV in terms of supplied oxygen, pressure, and h - (mod)eling, we simulated for the first time the effects of different ventilator settings during HFOV and CV for a preterm infant by generating a model based on a patient's individual FIGURE 2 Regional distribution of oxygen in comparison with the regional lung compliance. The compliance is (trunked word)	*JUPITER-MARS tool did not detect snippets for this domain*	first clinical trials comparing invasive ventilation strategies that used conventional respiration rates and HFOV yielded no difference in outcome. Nonetheless, higher rates of intraventricular hemorrhage (IVH) were suspected in the HFOV-treated infants,19 contradicted by studies - In line with this, the follow-up of the UKOS trial showed superior lung function at 11–14 years, with no evi- dence of poorer functional outcomes in those infants who had undergone HFOV, as compared to those who had received CV.6 Primarily based on these data, HFOV is currently sugges - “rectangular block” picture, indicating that all regions, in-dependent of the compliance, are ventilated uniformly. This would indicate an outcome in ventilator therapy that could be very bene- ficial in	- n is visualized oxygenation “heat map.” with high oxygenatio visualized in red and tho low oxygenation are sh transparent. In the conv ventilation (CV) scenari are “hotspots” with oxygenation correspond the regions with compliance - 3 visualizes the regi specific risk for overinflat again outlines the homo regional oxygenation HFOV. There are limitat our study that need considered when inte the data. First, the imaging data used in thi were obtained from a - i M. Förster prepared and tables. Kai M. Christian J. Roth, Wolfg Wall, and Andreas W. F wrote the manuscript. AVAILABILITY STATEMEN are available on request f authors. ORCID Kai M. https://orcid.org/0000-000 1400

					the management of neonates with ventilation–perfusion mismatch. Overinflation, h (trunked)	

Table 7b– JUPITER Evaluation for Sizemore et al., 2024. (27)

Justification **Score: 0/2** **Hits: 0** **Weighted hits: 0.0**	**Underlying Pathophysiology** **Score: 2/2** **Hits: 2** **Weighted hits: 3.0**	**Parameter selection** **Score: 2/2** **Hits: 16** **Weighted hits: 22.8**	**Integration modelling** **Score: 2/2** **Hits: 9** **Weighted hits: 11.4**	**Testing** **Score: 2/2** **Hits: 5** **Weighted hits: 6.9**	**Expected outputs** **Score: 2/2** **Hits: 15** **Weighted hits: 18.3**	**Reproducibility reporting** **Score: 2/2** **Hits: 5** **Weighted hits: 5.6**
JUPITER-MARS tool did not detect snippets for this domain	**- (trunked) and Verrucomicrobia as hub nodes, and in general, the SHCG model has more interactions. Role of clinical factors and diet in modulating HCG phenotypic outcome To assess the importance of clinical factors in modulating microbiome maturation trajectories, we carried out a SHAP analysis o** **- consistent impacts as defined above to deduce patient-¬ specific interventions. Using this notion, we end up identifying three intervention phenotypes: (i) Early Bacteroidia supplantation, (ii) Early Actinobacteria Supplantation, and (iii) Null, or patients for whom no such time-indepen (trunked)**	**- method of computing feature importances where the impact of each feature on the model is uncovered using the game-¬theoretic Shapley values. The SHAP analysis** **- method of computing feature importances where the impact of each feature on the model is uncovered using the game-¬theoretic Shapley values. The SHAP analysis (Fig. 3, G and H) found that the top influen** **- random forest regression model, with the Mδ risk as the dependent variable, and all available clinical factors as explanatory variables or features (Fig. 5). We observe that factors such as longer time until total enteral feeding is achieved, sex being male, the total number of morbidit(trunked)**	**- method of computing feature importances where the impact of each feature on the model is uncovered using the game-¬theoretic Shapley values (50, 51). The SHAP analysis (Fig. 3, G and H) found that the top influencers (exertin** **- jor differences between the digital twins inferred for the optimal and suboptimal cohorts. After computing the set of LOMAR values for each model (see the "LOMAR of microbial relative abundance" section for details and tables S5 to S8 for the computed LOMAR values), we plot the asymme** **- l models are quite distinct, with different roles being played by Actinobacteria and Verrucomicrobia as hub nodes, and in general, the SHCG model has more interactions. Role of clinical factors**	***- igital twins inferred for the optimal and suboptimal cohorts lays the foundation for designing early A B C D E F G H Fig. 3. Classification performance out-¬of-¬sample. (A) Classification performance to recognize infants with eventually suboptimal HCG. The AUC is maximum at 32 weeks PMA rea*** ***- imal cohorts lays the foundation for designing early A B C D E F G H Fig. 3. Classification performance out-¬of-¬sample. (A) Classification performance to recognize infants with eventually suboptimal HCG. The AUC is maximum at***	**- microbia as hub nodes, and in general, the SHCG model has more interactions. Role of clinical factors and diet in modulating HCG phenotypic outcome To assess the importance of clinical factors in modulating microbi- ome maturation trajectories, we carried out a SHAP analysis of a random** **- (trunked)ral feeds, which highly increased risk of SHCG in a small subset of individuals. Nevertheless, male sex emerges as a key factor driving HCG outcomes. Designing personalized interventions to reduce risk of poor HCG outcome The SHAP analysis of the digital twins inferred for the optimal an** **- duals. Nevertheless, male sex**	**- (trunked letter) Regulation (L coefficient(s) for each observed interacting b classes. The coefficients represent ¬¬regulatory/down--regul influence of a taxonomic class on a class, where regulation are caus- ally localized (futur (trunked)** **- n (LOMAR) coefficien each pair of ob interacting bacterial c The LOMAR coef represent up-¬¬regulatory/down- ¬¬regulatory influence source taxonomic clas target class, where reg effects are caus- ally lo in time (future cannot af** **- s diverse microbia obfuscate identificati actionable colo patterns that might fore poor clini- cal outcomes limitations hinder the de clinical inter- ventio vulnerable populations as preterm infants, tha otherwise prove effectiv**

			and diet in modulating HCG phenotypic outcome To assess the importance of clinical factors.	***32 weeks PMA reaching 87.6% for the UChicago cohort (in-¬sample - re we focus on microbial classes that have a consistent aver- age negative SHAP value (i.e., risk-¬decreasing impact) over time to Table 2. Performance measures for classification at 30 weeks PMA. UChicago fpr tpr ppv acc npv LR+ LR− 0.02 0.466 ± 0.015 0.961 ± 0.009 0.714 ± 0.008 0.633 ± 0.***	**emerges as a key factor driving HCG outcomes. Designing personalized interventions to reduce risk of poor HCG outcome The SHAP analysis of the digital twins inferred for the optimal and suboptimal cohorts lays the foundation for designing early A B C D E F**	

Noteworthy, Forster et al. (26) scored 0 of VVUQ meaning that it was an in-silico numerical study without wider validation and patient population, although authors underscored the further integration of echocardiography data. Accordingly, the total MARS integrated score was 18.5 (Weak) with a reporting score of 37 (intermediate reporting).

Instead, Sizemore et al. (27) represented an almost ideal example of a strongly-grounded pathophysiology study (JUPITER score: 85.7/100; synthetic evaluation: very good) in integration with a data-driven, digital twin modelling and high-quality reporting for MARS items (12/14 per domain score; 72.83 intermediate reporting per integrated MARS score). Indeed, it contains detailed information on modelling, testing metrics for model performance and AI precision (AUC, SHAP), and clinical outputs (microbiome to predict neurodevelopmental outcomes in preterm neonates). Moreover, it was more feasible (3 points of feasibility score as data were already available) and partially in accordance with VVUQ framework with 5 points (5/5) for validation and uncertainty quantification score compared with 0 points in Forster et al. in (26), and 8 points in terms of clinical utility compared with 0 points in Forster et al. in (26).

The key aspect of these proof-of-concepts cohorts is that JUPITER-MARS Tool could discriminate well levels of maturity of digital twins in the same medical field (neonatology), with the same analogous score in terms of dataset quality (5/5) and similarly low feasibility scores (0 for Forster, 3 for Sizemore). We think that these feasibility scores reflected the difficulty of retrieving and collecting ex novo the data used by these authors (complete ventilation parameters for conventional and high-frequency ventilation for one patient in (26); microbiome, genome and nutritional data in Sizemore et al., 2024) (27) thereby limiting the reproducibility of these models in real-world NICU setting (Neonatal Intensive Care Unit) where balancing healthcare workforce demands, workflow variability, and

technical AI constraints remains challenging (28–31).

Table 8 – MARS-VVUQ Score for Sizemore et al. (P2) and Forster et al. (P1)

Savvidou et al. (32) described in a textbook manner the current limitations of feasibility for digital twins in the NICU. “First, neonatal organ systems are strongly coupled, but coupling mechanisms are incompletely parameterized in preterm infants (e.g., cardiorespiratory–cerebral perfusion interactions, inflammation–microbiome–neurodevelopment links). Second, multimodal NICU data are frequently asynchronous (different sampling rates and time stamps across monitors, ventilators, labs, imaging), making reliable data fusion difficult. Third, integrated DTs must reconcile scale mismatch (molecular or microbiome dynamics vs. second-to-second physiologic waveforms) and propagate uncertainty across modules, which can amplify error when components are weakly validated. Finally, computational and operational barriers persist: bedside-compatible inference latency, interoperability constraints, and lack of standardized benchmarking datasets across NICUs. These barriers suggest that near-term progress will likely occur through modular DTs (validated organ-specific components) connected via well-defined interfaces rather than “whole-infant” models”. Further, they pointed out the neonatal-specific data and infrastructure challenges of implementing neonatal digital twins, as follows:

- Data Scarcity and Small Cohorts. Neonatal intensive care populations are relatively small compared to adult ICU datasets. Rare outcomes such as necrotizing enterocolitis or severe bronchopulmonary dysplasia further limit sample sizes. This increases the risk of overfitting and reduces model generalizability.

- Rapid Physiological Change. Neonatal physiology evolves over hours to days. Models trained on fixed temporal assumptions may fail to account for dynamic developmental trajectories, especially in extremely preterm infants.

- Data Quality and Signal Noise. NICU monitoring data are prone to artifacts, motion-related signal distortions, missing values, and inconsistent sampling frequencies. High-frequency waveform data require advanced preprocessing pipelines before reliable modeling. Interoperability and Fragmented Systems. NICU data are often distributed across multiple platforms (ventilators, monitors, laboratory systems, imaging systems, electronic health records), which may not be fully interoperable. Lack of standardized data schemas complicates real-time digital twin updating.

- Cohort Bias and Representativeness. Many neonatal modeling studies originate from high-resource academic centers. Limited representation of low-resource settings and diverse populations may introduce bias and restrict external validity.

4. Discussion

4.1. Main results.

Importantly, the absence of VVUQ emerges as a cross-cutting limitation across both

paradigms. This suggests that the primary bottleneck in digital twin maturity is not the modeling strategy itself—whether mechanistic or data-driven—but rather the lack of standardized credibility assessment frameworks.

These findings support the need for a paradigm shift in digital twin development, moving from performance-oriented modeling toward credibility-oriented modeling. Future frameworks should systematically integrate:

(i) physiological interpretability,

(ii) formal verification and validation pipelines,

and (iii) uncertainty quantification mechanisms.

We propose a conceptual taxonomy positioning digital twins along two axes: physiological grounding and VVUQ maturity. This framework enables a more nuanced classification of current models and provides a roadmap for future methodological development.

Ultimately, advancing clinical digital twins toward real-world implementation will require harmonizing computational innovation with robust validation standards, ensuring that model outputs are not only accurate but also trustworthy and clinically actionable

Accordingly, we introduced JUPITER-MARS + VVUQ framework as an ensemble, quality triangle able to orient clinicians (including pediatricians and neonatologists) to evaluate previous models in order to comprehend how they are built and whether it is feasible and methodologically justified to design them. While JUPITER represents how a digital twin is structured clinically, MARS evaluates the reporting in integration with clinics and operational feasibility. Furthermore, VVUQ strives to inform about how much the concerned digital twins outputs are reliable.

4.2. Comparison with other frameworks

Current digital twin research risks overestimating model robustness when evaluation is limited to reporting completeness without formal credibility assessment, underscoring the need for integrated frameworks combining reporting, validation, and uncertainty assessment. Moreover, JUPITER–MARS is capable of capturing meaningful differences in modelling completeness, validation strategies, and clinical applicability across digital twin studies. The framework proved applicable across both mechanistic and data-driven digital twin paradigms.

4.3. Potential applications

To best translate pathophysiology-informed, digital twins models into clinical practice, we provide clinical use examples from studies with different design (RCT, prospective, retrospective) in child epilepsy (33,34) and neonatology (35–41) using data derived from these studies to support the development of DT-like computational models. First, a hypothetical drug-response digital twin (cannabidiol) is composed by ECG-EEG tracking and lipidomics within epileptic pediatric patients. The scientific aim is to assess their efficacy against seizures.

As a further hypothetical application scenario, a prospective pilot digital twin integrating bowel ultrasound and abdominal radiography could be designed to stratify surgical intervention risk in neonates with necrotising enterocolitis (NEC). Relevant studies may include pilot cohorts **(37,39)**, monocentric studies **(38)**, systematic reviews and scoping reviews (40,41) to detect corresponding data from other studies to cluster evidences (through PubMed’s similar/cited articles’ function as recommended by John Hopkins Medical School).

Third, a neonatal Patent Ductus Arteriosus-oriented digital twin could integrate pain therapy

(ibuprofen; paracetamol/acetaminophen) in these suffering neonates submitted to surgical intervention. To achieve diverse grade of evidence, we select a monocentric retrospective study (36) and a multicentric prospective randomized trial (35), combining local institutional datasets with multicentric pediatric cohorts. Better statistical power could be obtained through larger sample size collected through chosen, low risk of bias studies. In these example cases, JUPITER guides structured reporting of model architecture and biological assumptions, while MARS supports quantitative assessment of model reliability. VVUQ principles are applied to validate simulation outputs against clinical EEG and imaging findings. When digital twin modelling is unfeasible due to logistic reasons (data collection, analysis or integration; unavailability/unaffordability of qualified medical engineers), JUPITER alone could be useful even for reporting what we called “analogical pathophysiological models” (APMs) insofar as clinical studies themselves may behave as ‘digital twin-like’ systems as far as they are defining systemic pathophysiology with different instruments/criteria that are corresponding to the current evidence (clinical ground truth) provided from systematic reviews and meta-analysis and umbrella reviews (overview of reviews), better when they are assessed with AMSTAR-2 risk of bias tool for systematic reviews and meta-analysis (42) by the respective authors. Specifically, APMs are intended as structured multimodal clinical models mimicking digital twin logic without requiring full computational twin implementation.

To improve clinical and regulatory transferability, we propose in fig. 6 a step-by-step operational roadmap for JUPITER-MARS implementation.

Fig. 6 - Operational Roadmap for JUPITER-MARS. This figure was created with Microsoft Word (SmartArt)

Future studies developing clinical digital twins may benefit from adopting the JUPITER reporting framework and the complementary MARS assessment score to improve transparency, reproducibility and comparability across studies, in integration with existing validation standards such as VVUQ. For example, further research may implement JUPITER-MARS prospectively when developing clinical digital twin models to ensure transparent model design, parameter justification and validation reporting.

JUPITER-MARS proposes a credibility-oriented methodological framework integrating reporting guidance, preliminary methodological assessment, and validation principles for clinical digital twin research. By combining structured reporting (JUPITER), domain-based assessment (MARS), and compatibility with VVUQ frameworks, the proposed approach may improve transparency, reproducibility, and methodological comparability across heterogeneous digital twin paradigms. Although preliminary and requiring future external validation, JUPITER-MARS provides an initial operational roadmap toward more trustworthy and physiologically grounded clinical digital twins. Moreover, the accompanying prototype screening tool provides an initial computational tool for identifying key reporting elements in the digital twin literature. However, JUPITER-MARS is intended as a preliminary methodological framework and should not yet be considered a formal consensus reporting guideline.

4.4. Limitations

This study has several limitations. First, the MARS score and its domain weights remain preliminary and have not yet undergone formal external validation or Delphi-based expert consensus procedures. Second, the benchmarking corpus was intentionally small due to the limited number of clinical digital twin studies explicitly implementing VVUQ principles. Third, the current computational tool relies partially on heuristic keyword extraction and therefore may not fully capture semantic methodological nuances within heterogeneous publications. Finally, the proposed taxonomy should be considered conceptual and hypothesis-generating rather than definitive.

4.5. Conclusions

Although the present work focuses on methodological development, the JUPITER-MARS framework can be applied to existing and emerging clinical digital twin studies to support structured reporting and evaluation. The framework may also support the reporting of domain-specific digital twin models, such as neurological digital twins described in previous methodological work (12).

Acknowledgements
We are grateful of international research whose most precious findings have been applicated for developing JUPITER-MARS framework.
Funding statement.
None.
Conflicts of interest.
None.
Data Availability Statement.
Full reports by MARS/JUPITER Tool are available in Supplementary Material. The MARS Tool is accessible from https://colab.research.google.com/drive/13CGYRbozv6zS7aW08txUhZrvCQHrYwk1?uspssharing, whereas JUPITER is available at the following link: https://colab.research.google.com/drive/1-BRuwmsDPz8foPDzzPnSl9sZuakyRfQf?usp=sharing.
Authors' contribution:
MEC: Concept; Writing – Draft; Writing – Review & Editing; Data curation; Methodology; Reference retrieval; PS: Writing – Review & Editing; Supervision; Validation; FV: Validation; AC: Validation; DC: Writing – Review & Editing; Supervision; Validation; LAR: Writing – Review & Editing; Supervision; Validation.
Abbreviations.
JUPITER = Justification, Underlying pathophysiology, Parameters, Integration, Expected results and Reproducibility framework for clinical digital twins.
MARS = Multimodal Assessment of Reporting Score
AI = Artificial Intelligence
VVUQ = Verification, Validation and Uncertainty Quantification
AUC = Area Under the Curve
ROC = Receiver Operator Curve.
SciBERT = Scientific Bidirectional Encoder Representations from Transformers.
BioBERT = Bidirectional Encoder Representations from Transformers for Biomedical Text Mining.
ASME = American Society of Mechanical Engineers.
SQA = Software Quality Assurance.
References

1. Drummond D, Gonsard A. Definitions and Characteristics of Patient Digital Twins Being

Developed for Clinical Use: Scoping Review. J Med Internet Res. 2024 Nov 13;26:e58504. doi:10.2196/58504

2. De Oliveira El-Warrak L, Miceli De Farias C. Could digital twins be the next revolution in healthcare? European Journal of Public Health. 2025 Feb 1;35(1):19–25. doi:10.1093/eurpub/ckae191

3. Gonsard A, Genet M, Drummond D. Digital twins for chronic lung diseases. Eur Respir Rev. 2024 Oct;33(174):240159. doi:10.1183/16000617.0159-2024

4. Eccher A, Pagni F, Dominici M, Bonetti LR, Marletta S, Munari E, et al. Digital twin manifesto for the pathology laboratory. Diagn Pathol. 2025 Jul 17;20(1):84. doi:10.1186/s13000-025-01679-2

5. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025 Feb 5;388:e081554. doi:10.1136/bmj-2024-081554

6. Bignami E, Darhour LJ, Franco G, Guarnieri M, Bellini V. AI policy in healthcare: a checklist-based methodology for structured implementation. J Anesth Analg Crit Care. 2025 Sep 25;5(1):56, s44158-025-00278–3. doi:10.1186/s44158-025-00278-3

7. Kameyama J, Kodera S, Inoue Y. Ethical, legal, and social issues (ELSI) and reporting guidelines of AI research in healthcare. Al-Jumeily O. B. E. D, editor. PLOS Digit Health. 2024 Sep 19;3(9):e0000607. doi:10.1371/journal.pdig.0000607

8. Mortezaagha P, Shaw J, Sun B, Rahgozar A. From chaos to clarity: schema-constrained AI for auditable biomedical evidence extraction from full-text PDFs. BMC Med Res Methodol. 2026 Apr 14;26(1):119. doi:10.1186/s12874-026-02847-8

9. Sel K, Hawkins-Daarud A, Chaudhuri A, Osman D, Bahai A, Paydarfar D, et al. Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine. npj Digit Med. 2025 Jan 17;8(1):40. doi:10.1038/s41746-025-01447-y

10. Zaid NRR, Hardiansyah D, Yusufaly T, Rahmim A. Verification, Validation, and Uncertainty Quantification (VVUQ) of Physiologically Based Pharmacokinetic Models for Theranostic Digital Twins. PET Clinics. 2026 Jan;21(1):101–15. doi:10.1016/j.cpet.2025.08.004

11. Canepa ME, Striano P, Cangelosi D, Rossi A, Tortora D, Calandrino A, Ramenghi LA. The ATLAS Method: A Framework for Systematic Reviews in Quantitative Radiology Integrating AI, Radiomics, Clinical Biomarkers and Imaging Data Resources. Eur Radiol Exp. 2026.

12. Canepa ME, Villano G, Verrotti A, Ramenghi L, P Striano Canepa ME, Villano G, Verrotti A, Ramenghi L, P Striano. Digital Twins in Neuroscience, Advanced Technology in Neuroscience. Advanced Technology in Neuroscience. 2026.

13. Teran Pumar OY, Lathia JD, Watson DC, Bayik D. ‘Slicing’ glioblastoma drivers with the Swiss cheese model. Trends in Cancer. 2024 Jan;10(1):15–27. doi:10.1016/j.trecan.2023.08.002

14. Vernaz N, Simona A, Samer CF. The Swiss Cheese Prescribing Model for Precision Medicine. The American Journal of Medicine. 2020 Nov;133(11):1249–51.

doi:10.1016/j.amjmed.2020.06.001

15. Masison J, Beezley J, Mei Y, Ribeiro H, Knapp AC, Sordo Vieira L, et al. A modular computational framework for medical digital twins. Proc Natl Acad Sci USA. 2021 May 18;118(20):e2024287118. doi:10.1073/pnas.2024287118

16. Kuznetsova NM, Karlova TV. Application of Digital Twin Technology for Modelling of Information Security Level of Industrial Plant. Nadykto A, Aleksic N, Lima P, Pivkin P, Uvarova L, Jiang X, et al., editors. EPJ Web Conf. 2021;248:02007. doi:10.1051/epjconf/202124802007

17. Li Y, Thimbleby H. Hot Cheese: A Processed Swiss Cheese Model. Journal of the Royal College of Physicians of Edinburgh. 2014 Jun;44(2):116–21. doi:10.4997/JRCPE.2014.205

18. Cummings ML. A Taxonomy for AI Hazard Analysis. Journal of Cognitive Engineering and Decision Making. 2024 Dec;18(4):327–32. doi:10.1177/15553434231224096

19. Tudor BH, Shargo R, Gray GM, Fierstein JL, Kuo FH, Burton R, et al. A scoping review of human digital twins in healthcare applications and usage patterns. npj Digit Med. 2025 Sep 30;8(1):587. doi:10.1038/s41746-025-01910-w

20. Béthencourt L, Dabachine W, Dejouy V, Lalmiche Z, Neuberger K, Ibnouhsein I, et al. Guiding Measurement Protocols of Connected Medical Devices Using Digital Twins: A Statistical Methodology Applied to Detecting and Monitoring Lymphedema. IEEE Access. 2021;9:39444–65. doi:10.1109/ACCESS.2021.3063786

21. Salvador M, Kong F, Peirlinck M, Parker DW, Chubb H, Dubin AM, et al. Digital twinning of cardiac electrophysiology for congenital heart disease. J R Soc Interface. 2024 Jun;21(215):20230729. doi:10.1098/rsif.2023.0729

22. Silfvergren O, Simonsson C, Ekstedt M, Lundberg P, Gennemark P, Cedersund G. Digital twin predicting diet response before and after long-term fasting. Hatzimanikatis V, editor. PLoS Comput Biol. 2022 Sep 12;18(9):e1010469. doi:10.1371/journal.pcbi.1010469

23. Thamotharan P, Srinivasan S, Kesavadev J, Krishnan G, Mohan V, Seshadhri S, et al. Human Digital Twin for Personalized Elderly Type 2 Diabetes Management. JCM. 2023 Mar 7;12(6):2094. doi:10.3390/jcm12062094

24. Santiago A, Butakoff C, Eguzkitza B, Gray RA, May-Newman K, Pathmanathan P, et al. Design and execution of a verification, validation, and uncertainty quantification plan for a numerical model of left ventricular flow after LVAD implantation. McCulloch AD, editor. PLoS Comput Biol. 2022 Jun 13;18(6):e1010141. doi:10.1371/journal.pcbi.1010141

25. ASME. (ASME) TASoME. Assessing Credibility of Computational Modeling Through Verification and Validation: Application to Medical Devices,. ASME V&V. 2018;40.

26. Förster KM, Roth CJ, Hilgendorff A, Ertl-Wagner B, Flemmer AW, Wall WA. In silico numerical simulation of ventilator settings during high-frequency ventilation in preterm infants. Pediatric Pulmonology. 2021 Dec;56(12):3839–46. doi:10.1002/ppul.25626

27. Sizemore N, Oliphant K, Zheng R, Martin CR, Claud EC, Chattopadhyay I. A digital twin of the infant microbiome to predict neurodevelopmental deficits. Sci Adv. 2024 Apr

12;10(15):eadj0400. doi:10.1126/sciadv.adj0400

28. Shahroor M, Lehtonen L, Lee SK, Håkansson S, Vento M, Darlow BA, et al. Unit-Level Variations in Healthcare Professionals' Availability for Preterm Neonates <29 Weeks' Gestation: An International Survey. Neonatology. 2019;116(4):347–55. doi:10.1159/000501801

29. Barrett R, Lawler B, Liu S, Park WY, Davoodi M, Martin B, et al. Transforming neonatal care through informatics: A review of artificial intelligence, data, and implementation considerations. Seminars in Perinatology. 2025 Nov;49(7):152144. doi:10.1016/j.semperi.2025.152144

30. Schulman J, Braun D, Lee HC, Profit J, Duenas G, Bennett MV, et al. Association Between Neonatal Intensive Care Unit Admission Rates and Illness Acuity. JAMA Pediatr. 2018 Jan 1;172(1):17. doi:10.1001/jamapediatrics.2017.3913

31. Goodman DC, Little GA. Data Deficiency in an Era of Expanding Neonatal Intensive Care Unit Care. JAMA Pediatr. 2018 Jan 1;172(1):11. doi:10.1001/jamapediatrics.2017.4042

32. Savvidou D, Dermitzaki N, Baltogianni M, Nikolaou A, Giapros V. Digital Twins in Neonatology: Current Applications and Future Directions: A Narrative Review. Applied Sciences. 2026 Feb 25;16(5):2198. doi:10.3390/app16052198

33. Hopkins S, Weber A, MacDougall J, Evanski J, Marusak HA, Luat A. Preliminary Prospective Study of Pharmaceutical-Grade Cannabidiol for Seizure Frequency, Anxiety, and Comorbid Symptoms in Pediatric Epilepsy: Associations With Circulating Endocannabinoids and Lipid Biomarkers. J Child Neurol. 2026 Apr 25;08830738261439213. doi:10.1177/08830738261439213

34. Gigli L, Sala S, Preda A, Okubo K, Peretto G, Frontera A, et al. Electrocardiogram Changes in the Postictal Phase of Epileptic Seizure: Results from a Prospective Study. JCM. 2023 Jun 17;12(12):4098. doi:10.3390/jcm12124098

35. Dani C, Lista G, Bianchi S, Mosca F, Schena F, Ramenghi L, et al. Intravenous paracetamol in comparison with ibuprofen for the treatment of patent ductus arteriosus in preterm infants: a randomized controlled trial. Eur J Pediatr. 2021 Mar;180(3):807–16. doi:10.1007/s00431-020-03780-8

36. Caruggi S, Calandrino A, Cipresso G, Battaglini M, Massirio P, Vinci F, et al. Second Attempt for Patent Ductus Arteriosus (PDA) Closure: Room for Acetaminophen? A Retrospective Single-Center Experience at Gaslini Children's Hospital. Children. 2025 Apr 29;12(5):577. doi:10.3390/children12050577

37. Priyadarshi A, Angiti R, Chabra S, McAdams R, Webb A, Badawi N, et al. Combining abdominal ultrasound and radiography for surgical risk stratification in necrotising enterocolitis: a prospective cohort pilot study. Arch Dis Child Fetal Neonatal Ed. 2026 Mar 5;fetalneonatal-2025-329960. doi:10.1136/archdischild-2025-329960

38. Vallant N, Thakkar H, Jogeesvaran H, Yardley I. Abdominal Ultrasound Scanning for NEC in Babies at the Threshold of Viability: A Single Centre Experience. Journal of Pediatric Surgery. 2024 Feb;59(2):202–5. doi:10.1016/j.jpedsurg.2023.10.011

39. He Y, Zhong Y, Yu J, Cheng C, Wang Z, Li L. Ultrasonography and radiography findings

predicted the need for surgery in patients with necrotising enterocolitis without pneumoperitoneum. Acta Paediatrica. 2016 Apr;105(4). doi:10.1111/apa.13315

40. Cuna AC, Reddy N, Robinson AL, Chan SS. Bowel ultrasound for predicting surgical management of necrotizing enterocolitis: a systematic review and meta-analysis. Pediatr Radiol. 2018 May;48(5):658–66. doi:10.1007/s00247-017-4056-x

41. Bhattacharjee I, Dolinger MT, Singh R, Singh Y. Ultrasound for the Early Detection and Diagnosis of Necrotizing Enterocolitis: A Scoping Review of Emerging Evidence. Diagnostics. 2025 Jul 23;15(15):1852. doi:10.3390/diagnostics15151852

42. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017 Sep 21;j4008. doi:10.1136/bmj.j4008

Supplementary Files

Multimedia Appendixes

First benchmark.
URL: http://asset.jmir.pub/assets/e68ba7b735bb1e9485102b7d2a7a9e8f.docx

Second benchmark cohort.
URL: http://asset.jmir.pub/assets/923dcc5a60d6e9411c09f9ddd1b3ccf7.docx