International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Synthetic Data inRheumatology: A Systematic Literature Review

INPLASY202630024
doi: 10.37766/inplasy2026.3.0024
Received: 7 March 2026
Published: 7 March 2026

Venerito, V; Morrom, M; Lopalco, G; Del Vescovo, S; Bilgin, E; Gupta, L; Iannone, F.

Corresponding author:
Vincenzo Venerito

vincenzo.venerito@gmail.com

Author Affiliation:
Rheumatology Unit -University of Bari "Aldo Moro", Italy.

ADMINISTRATIVE INFORMATION

Support - NA.

Review Stage at time of this submission - Preliminary searches.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202630024

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 7 March 2026 and was last updated on 7 March 2026.

INTRODUCTION

Review question / Objective To systematically identify, categorize, and critically appraise the methods, validation frameworks, healthcare applications, and regulatory landscape of synthetic data generation, with a specific focus on rheumatology and related autoimmune/musculoskeletal diseases.

The review addresses four questions:
(1) What methods exist for generating synthetic healthcare data and how do they compare?
(2) What validation frameworks are used to evaluate synthetic data quality?
(3) How has synthetic data been applied in healthcare generally and rheumatology specifically?
(4) What are the current privacy and regulatory frameworks governing synthetic health data?

Rationale Background and rationale: Rheumatology research faces critical challenges in data scarcity (rare diseases, small cohorts), phenotypic complexity (overlap syndromes, variable organ involvement), multimodal data requirements (clinical, imaging, biomarker, genomic), and privacy constraints on multi-institutional data sharing (GDPR, HIPAA). Synthetic data methods offer potential solutions to all these challenges, yet no systematic review has comprehensively mapped the synthetic data landscape with a rheumatology focus. The field has experienced rapid growth (approximately 40–70% biannual increases since 2020), making a systematic synthesis timely and necessary.

Condition being studied Domain: Synthetic data generation methodology in healthcare, encompassing: statistical methods (Bayesian networks, copula models, Gaussian mixture models), generative adversarial networks (GANs), diffusion models, large language model (LLM)-based approaches, rule-based simulators (e.g., Synthea), hybrid methods, and digital twin/virtual patient frameworks. The clinical domain focus is rheumatology and autoimmune/musculoskeletal diseases.

METHODS

Search strategy Framework (note: traditional PICO is not applicable to this methodological review):
Topic: Synthetic data generation for healthcare/ clinical data
Context: Healthcare broadly; rheumatology specifically
Themes: (1) Generation methods, (2) Validation frameworks (fidelity, utility, privacy), (3) Healthcare applications, (4) Rheumatology-specific applications, (5) Privacy and regulatory considerations
Outcomes: Categorization of methods, identification of validation gaps, mapping of disease-specific applications, regulatory status.

Participant or population
Inclusion criteria:
1. Primary focus on synthetic data generation, evaluation, or validation methodology; OR
2. Primary focus on privacy-preserving data sharing using synthetic data; OR
3. Primary focus on digital twins or virtual patient/ population generation; OR
4. Primary focus on regulatory frameworks for synthetic data
5. Must have healthcare or clinical application context
6. Must be a full paper: original research, systematic/narrative review, or full conference proceedings (≥4 pages)
Exclusion criteria:
E1: Paper merely uses data augmentation as a preprocessing step for a clinical task
E2: Drug or molecular design using generative models (not patient/clinical data synthesis)
E3: Conference abstract, poster, or oral presentation without full paper
E4: Commentary, editorial, or letter without original data
E5: Standard geometric augmentation only (rotation, flipping, cropping)
E6: Not about healthcare or clinical data.

Intervention NA- PICO Not APplicable.

Comparator NA - PICO not Applicable.

Study designs to be included Study designs to be included: Original research articles (computational methodology, validation, and benchmarking studies), systematic reviews, narrative reviews, full conference proceedings (≥4 pages), and clinical trials. No restriction on study design was applied provided the study presented a full paper with primary focus on synthetic data generation, evaluation, privacy-preserving data sharing, digital twins/virtual populations, or regulatory frameworks in a healthcare context. Commentaries, editorials, letters without original data, and conference abstracts/post.

Eligibility criteria Study selection proceeded in two stages per PRISMA 2020:

Stage 1 — Title and abstract screening: After removing 405 duplicate records, 701 records were screened by title and abstract against inclusion and exclusion criteria. Records that were clearly irrelevant or did not meet inclusion criteria were excluded (n=353).

Stage 2 — Full-text eligibility assessment: Full-text retrieval was attempted for the remaining 348 articles via PubMed Central, arXiv, Unpaywall, and institutional access. Of these, 309 full texts were retrieved; 39 were assessed from abstract only. Full-text assessment excluded 120 articles, yielding 228 studies for qualitative synthesis.

Information sources Data items extracted for each included study:
• Bibliographic information (authors, year, journal, DOI/PMID)
• Study design (original research, review, clinical trial, etc.)
• Synthetic data generation method (statistical, GAN, diffusion, LLM, rule-based, hybrid, digital twin)
• Data type (tabular/EHR, imaging, clinical text, biomarker, genomic, multimodal)
• Application domain (general healthcare, specific disease area)
• Validation metrics used (fidelity, utility, privacy — specific metrics noted)
• Privacy evaluation approach (if any)
• Key findings and main contribution
• Limitations noted by authors
• Disease area (for rheumatology-specific studies).

Main outcome(s) Primary outcomes:
1. Categorization and comparative analysis of synthetic data generation methods in healthcare
2. Assessment of validation framework adoption (fidelity, utility, privacy metrics)
3. Mapping of rheumatology-specific synthetic data applications by disease area and method
4. Identification of gaps between general healthcare synthetic data maturity and rheumatology adoption.

Quality assessment / Risk of bias analysis Quality assessment approach:

A systematic evaluation of validated risk of bias tools (RoB 2, ROBINS-I, QUADAS-2, PROBAST, PROBAST+AI, QUADAS-AI) determined that none are applicable to studies whose primary purpose is the development or evaluation of synthetic data generation methods.

For the 13 rheumatology-specific studies (from which disease-specific conclusions are drawn), a purpose-designed quality assessment checklist was developed, informed by PROBAST, TRIPOD+AI, and synthetic data evaluation criteria from the literature. The checklist comprises 10 items across five domains:
1. Study design and data description (Q1–Q2)
2. Methodological rigor (Q3)
3. Evaluation completeness: fidelity, utility, privacy (Q4–Q6)
4. Comparison and clinical relevance (Q7–Q8)
5. Limitations and reproducibility (Q9–Q10)

Each item scored as Yes (fully met), Partial (partially met), or No (not met). Overall quality rated as High (≥7 Yes), Moderate (4–6 Yes), or Low (≤3 Yes).

For the remaining 215 methodology studies, no formal quality assessment was performed, as no validated instrument exists and these studies are included for methodological context rather than to support clinical conclusions. This is acknowledged as a limitation.

Strategy of data synthesis Synthesis approach:
Qualitative narrative synthesis organized by thematic framework:
1. Methods for synthetic data generation (statistical, GAN, diffusion, LLM, rule-based, hybrid, digital twin)
2. Validation frameworks (fidelity metrics, utility metrics, privacy metrics, the triad trade-off)
3. Healthcare applications (data augmentation, privacy-preserving sharing, clinical trial simulation, synthetic imaging, synthetic EHR)
4. Rheumatology-specific applications (by disease: RA, OA, SLE, SSc, SjS, AS)
5. Privacy and regulatory landscape

No meta-analysis was performed due to the methodological heterogeneity of included studies (diverse methods, data types, application domains, and outcome measures).

Subgroup analysis Subgroup categories:
• Generation method: statistical, GAN, diffusion, LLM, rule-based, hybrid, digital twin
• Data modality: tabular/EHR, imaging, clinical text, multimodal
• Application domain: general healthcare vs. rheumatology-specific
• Disease area (rheumatology): RA, OA, SLE, SSc, SjS, AS, other
• Validation approach: fidelity only, utility only, privacy only, two dimensions, full triad
Sensitivity analysis:
Not applicable (qualitative synthesis, no pooled effect estimates).

Sensitivity analysis Not applicable (qualitative synthesis, no pooled effect estimates).

Language restriction English.

Country(ies) involved Italy, Turkiye, UK.

Keywords Synthetic Data, Rheumatology, Artifcial intelligence.

Dissemination plans Submission to a peer-reviewed journal in the rheumatology or digital health domain. Target journals: Annals of the Rheumatic Diseases, Arthritis & Rheumatology, RMD Open, npj Digital Medicine, or Journal of Medical Internet Research.

Contributions of each author
Author 1 - Vincenzo Venerito.
Author 2 - Maria Morrone.
Author 3 - Giuseppe Lopalco.
Author 4 - Sergio Del Vescovo.
Author 5 - Emre Bilgin.
Author 6 - Latika Gupta.
Author 7 - Florenzo Iannone.