Journal of Translational Gastroenterology **2026** | Epub
DOI: 10.14218/JTG.2026.00011

Mini Review

Artificial Intelligence-driven Phenomics in Celiac Disease: Toward Precision Diagnosis

Hakim Rahmoune[1*], Nada Boutrid[1] and Isra Benchoufi[2]

[1]*LIRSSEI Research Lab, Faculty of Medicine, University of Sétif 1, Sétif, Algeria;* [2]*ENSIA, Algiers, Algeria*

Received: April 04, 2026 | **Revised:** June 07, 2026 | **Accepted:** July 28, 2026 | **Published online:** September 01, 2026

Abstract

Celiac disease remains underdiagnosed despite an established diagnostic pathway, reflecting phenotypic heterogeneity, delayed recognition, and dependence on resource-intensive testing. This narrative review synthesizes evidence on artificial intelligence and phenomics in celiac disease (CD) across four operational layers: electronic health record phenotyping, Human Phenotype Ontology-based semantic encoding, machine-learning pre-screening from routine clinical data, and deep-learning-assisted histopathology. A targeted literature search of PubMed/MEDLINE, Embase, and Google Scholar covered publications from January 2010 through December 2025, using combinations of CD/coeliac disease with artificial intelligence, machine learning, deep learning, phenomics, Human Phenotype Ontology, computable phenotype, electronic health records, and natural language processing, supplemented by targeted searches and citation chaining. We present a curated minimum viable CD phenome and discuss clinical actionability, age-specific considerations, and current pediatric validation gaps. Selected models have demonstrated promising CD detection or pre-screening performance in specific datasets, but prospective, multicenter evidence with external validation remains insufficient to establish clinical utility. Artificial intelligence should augment expert clinical care rather than replace it.

Introduction

Celiac disease (CD) is a chronic immune-mediated enteropathy triggered by dietary gluten in genetically susceptible individuals, affecting an estimated 1.4% of the global population by serology and 0.7% by biopsy confirmation.[1] Despite its prevalence, more than 80% of cases remain undiagnosed.[2,3] Diagnostic delays have been reported across healthcare settings and may be prolonged in both high- and low-resource environments.[4-8]

Diagnostic delay has been associated with clinical consequences, including osteoporosis, iron-deficiency anemia, and neurological complications.[5,6] Persistent underdiagnosis reflects the phenotypic heterogeneity of CD, spanning classical malabsorption to silent or extraintestinal presentations, and the resource-intensive nature of diagnostic evaluation.[9,10] The challenge is particularly relevant in pediatric patients, who may present with short stature (HP:0004322), failure to thrive (HP:0001508), and dental enamel defects, while adults may present predominantly with extraintestinal or minimally symptomatic disease.[11] Potential CD is best described as persistently positive CD-specific serology in the presence of normal or non-atrophic small-intestinal mucosa; precise diagnostic and management criteria depend on the applicable guideline and clinical context.[11]

Artificial intelligence (AI) and phenomics may help address these bottlenecks. Machine learning (ML) may identify pre-diagnostic patterns in routine clinical data,[12] while deep learning (DL) has shown promising performance for automated CD detection from digitized biopsy images.[13] Standardized ontologies such as the Human Phenotype Ontology (HPO) enable computable, interoperable phenotype representations that can bridge clinical records, research databases, and AI pipelines.[14,15] This review synthesizes current evidence across these domains and proposes a four-layer implementation framework for AI-assisted CD diagnostics, with explicit attention to clinical actionability, age-specific phenotypic considerations, and equitable applicability.

A targeted narrative literature search was conducted across PubMed/MEDLINE, Embase, and Google Scholar for publications from January 2010 through December 2025 addressing AI, ML, DL, phenomics, or computable phenotyping in CD; earlier publications were considered without date restriction for foundational HPO and EHR-phenotyping methodology. Search terms

Keywords: Celiac disease; Artificial intelligence; Phenomics; Human Phenotype Ontology; Deep learning; Computable phenotype; Electronic health records; Precision medicine.

***Correspondence to:** Hakim Rahmoune, LIRSSEI Research Lab, Faculty of Medicine, University of Sétif 1, Sétif 19000, Algeria. ORCID: https://orcid.org/0000-0002-9604-3675. Tel: +213-550123279, E-mail: rahmounehakim@gmail.com

How to cite this article: Rahmoune H, Boutrid N, Benchoufi I. Artificial Intelligence-driven Phenomics in Celiac Disease: Toward Precision Diagnosis. *J Transl Gastroenterol*. Published online: Aug 31, 2026. DOI: https://doi.org/10.14218/JTG.2026.00011.

included combinations of "CD" or "coeliac disease" with "artificial intelligence," "machine learning," "deep learning," "phenomics," "Human Phenotype Ontology," "computable phenotype," "electronic health records," and "natural language processing." Additional targeted searches addressed extreme gradient boosting (XGBoost), convolutional neural networks, vision transformers, histopathology, potential CD, and pre-diagnostic screening. Citation chaining from high-yield publications supplemented the database searches. Studies were considered when they reported original AI-model performance data relevant to CD or methodological approaches directly applicable to CD phenotyping. Conference abstracts without full-text data, case reports, and editorials without original quantitative findings were excluded. The literature was reviewed by the authors for relevance and applicability; no formal risk-of-bias or methodological quality-assessment tool was applied.

Computable phenotyping of CD

Limitations of ICD-based case ascertainment

The International Classification of Diseases (ICD) has historically been used to identify CD in administrative and research databases. ICD-10 code K90.0 has shown high specificity in some settings but may have limited sensitivity when diagnoses are recorded under related codes or when extraintestinal manifestations predominate.[16] Evidence on coding performance is heterogeneous, and a specific 15–30% misclassification rate for CD cohorts is not sufficiently supported by the cited literature; accordingly, we do not assign a universal misclassification rate.

Phecodes, which aggregate related ICD codes into clinically meaningful phenotype categories, offer a partial solution. The phecode for CD (phecode 557.1) combines K90.0 with related codes for malabsorption and gluten sensitivity.[17,18] However, phecodes remain dependent on the completeness and accuracy of underlying ICD coding and do not capture laboratory, pathological, or narrative data.

Multimodal electronic health record phenotyping

More sophisticated computable phenotypes for CD integrate multiple electronic health record (EHR) data streams: structured diagnosis codes, laboratory values (tissue transglutaminase immunoglobulin A (tTG-IgA), hemoglobin, ferritin, albumin), medication records, pathology reports, and unstructured clinical notes processed by natural language processing (NLP).[19-21] Algorithmic phenotyping frameworks such as PheNorm and PheVis have been applied to autoimmune and other conditions, [22,23] with approaches that may be transferable to CD phenotyping. The CALIBER platform in the United Kingdom has demonstrated high-fidelity EHR phenotyping across numerous conditions using linked primary care, hospital, and registry data.[24] Ludvigsson *et al*.[25] demonstrated that a computerized algorithm applied to electronic medical record data could identify individuals requiring CD testing with high specificity. The integration of pathology-report NLP with structured EHR data may capture histological and longitudinal dimensions that are not available from ICD codes alone.[21,26]

The Human Phenotype Ontology as a semantic framework for CD

HPO architecture and the 2024 update

The Human Phenotype Ontology (HPO) is a structured, hierarchical vocabulary of human phenotypic abnormalities, originally described by Robinson *et al*.[27] Successive updates have expanded its coverage substantially; the 2024 release comprises more than 18,000 terms and over 170,000 disease–phenotype annotations across thousands of diseases.[14] HPO terms are organized as a directed acyclic graph with major subhierarchies covering abnormalities of the digestive, hematological, metabolic, and nervous systems.[28,29] The 2024 update also expanded multilingual and global resources, supporting broader use in diverse populations in which CD epidemiology and human leukocyte antigen (HLA) distributions may differ.[14,30,31]

HPO-based phenotyping pipelines and age-specific considerations

Several computational pipelines exploit HPO for clinical decision support and differential diagnosis. Encoding clinical data with HPO enables similarity-based matching between patient phenotype profiles and disease-specific HPO annotations, a strategy established in rare-disease diagnostics and potentially applicable to CD.[32,33] For CD, HPO can represent a broad phenotypic spectrum, including villous atrophy (HP:0011473), malabsorption (HP:0002024), and extraintestinal manifestations. The proposed CD phenome should therefore be regarded as a curated conceptual feature set rather than a validated disease-specific frequency model.

An important consideration for HPO-based AI models is the distinction between pediatric and adult phenotypic presentations. In children, CD may manifest through growth failure, failure to thrive, and dental enamel defects. In adults, extraintestinal or minimally symptomatic presentations are common, including iron-deficiency anemia, elevated transaminases, and peripheral neuropathy. This phenotypic divergence suggests that HPO feature weighting, training-cohort composition, and risk thresholds may require age-stratified calibration. The cited ML pre-screening study by Dreyfuss *et al*.[12] included adolescents and adults, and its performance has not been independently validated in an exclusively pediatric population.

The PhenoBrain system illustrates the potential of HPO-based AI in rare-disease differential diagnosis.[34] Evaluated across 2,271 cases covering 431 rare diseases, the system reported top-3 recall of 0.613 and top-10 recall of 0.813 in the reported physician-comparison setting.[34] These findings were obtained in a non-celiac rare-disease cohort and therefore provide indirect methodological support only; they do not establish diagnostic performance in CD and should not be interpreted as evidence that HPO-based AI outperforms specialists in CD.

Curated HPO terms for the CD phenome

Table 1 presents a curated set of HPO terms intended as a minimum viable CD phenome across major clinical domains. The terms were selected conceptually to represent gastrointestinal, hematological, metabolic/growth, neurological, dermatological, and reproductive features relevant to CD. Because the cited literature does not provide a validated CD-specific frequency

Table 1. Curated HPO terms for a conceptual minimum viable celiac disease phenome

HPO term	HPO ID	Clinical domain	Clinical relevance
Villous atrophy	HP:0011473	Gastrointestinal	Important histological feature of established CD; not universal because diagnosis may be established without biopsy in selected pediatric pathways
Malabsorption	HP:0002024	Gastrointestinal	Represents impaired intestinal nutrient absorption and may accompany symptomatic disease
Chronic diarrhea	HP:0002028	Gastrointestinal	Common classical presentation but not required and may be absent in atypical disease
Abdominal pain	HP:0002027	Gastrointestinal	Common but nonspecific gastrointestinal symptom
Abdominal distention	HP:0003270	Gastrointestinal	May accompany gastrointestinal symptoms and malabsorption
Iron deficiency anemia	HP:0001891	Hematological	Important extraintestinal manifestation and a clinically relevant trigger for testing
Decreased circulating folate concentration	HP:0100507	Hematological	May occur with nutritional deficiency or impaired intestinal absorption
Thrombocytosis	HP:0001894	Hematological	May occur as a reactive hematological abnormality
Elevated circulating hepatic transaminase concentration	HP:0002910	Metabolic/hepatic	Can accompany CD and may prompt investigation when otherwise unexplained
Osteoporosis	HP:0000939	Metabolic/skeletal	May occur in association with chronic malabsorption and altered bone health
Short stature	HP:0004322	Metabolic/growth	Relevant pediatric growth phenotype that may prompt CD evaluation
Failure to thrive	HP:0001508	Metabolic/growth	Relevant pediatric phenotype, particularly in younger children
Peripheral neuropathy	HP:0009830	Neurological	Recognized extraintestinal neurological manifestation
Ataxia	HP:0001251	Neurological	Uncommon neurological presentation associated with gluten-related disease
Cognitive impairment	HP:0100543	Neurological	Nonspecific neurological feature reported in association with CD
Abnormal blistering of the skin	HP:0008066	Dermatological	Characteristic blistering and vesiculobullous cutaneous manifestation of gluten-sensitive enteropathy; include when clinically documented
Infertility	HP:0000789	Reproductive	Reported reproductive manifestation, particularly relevant in adult presentations
Recurrent spontaneous abortion	HP:0200067	Reproductive	Reported reproductive manifestation associated with untreated CD

The HPO identifiers in the table were checked and verified against current HPO-linked ontologies and cross-referenced with Orphanet, MedGen, and Monarch Initiative disease annotations. Folate deficiency is represented by HP:0100507 and the cutaneous blistering phenotype of dermatitis herpetiformis by HP:0008066. The table is conceptual and is not a validated CD-specific frequency or feature-importance model. CD, celiac disease; HPO, Human Phenotype Ontology.

estimate or machine-learning assessment of feature utility for this exact 18-term set, the table deliberately omits a frequency column and avoids claims that any term is obligatory, universal, or independently validated as a predictive feature. The set should therefore be considered a hypothesis-generating semantic layer requiring formal validation before clinical deployment.

Artificial intelligence in CD histopathology

The Marsh–Oberhuber system and its limitations

The Marsh–Oberhuber classification remains widely used for describing histological changes associated with CD, grading mucosal abnormalities from increased intraepithelial lymphocytes through villous atrophy.[11] However, interobserver variability can affect intermediate-grade assessment. Quantitative digital pathology offers an approach beyond subjective grading. Gruver *et al*.[35] developed pathologist-trained machine-learning classifiers that quantify villous and crypt measurements and intraepithelial

lymphocyte features, generating quantitative scores that correlate with modified Marsh assessment and dietary intervention response. Griffin *et al*.[36] reported the feasibility of automated cell-type and tissue-region classification across multisite whole-slide images (WSIs).

Deep learning models for CD histopathology and potential CD

Jaeckle *et al*.[13] evaluated a machine-learning model for binary detection of CD from hematoxylin and eosin-stained duodenal biopsy WSIs. The model was trained using 3,383 WSIs from four hospitals and evaluated on 644 previously unseen scans from a geographically separate regional National Health Service (NHS) Trust, constituting an independent test dataset. The model achieved accuracy, sensitivity, and specificity exceeding 95%, with an area under the receiver operating characteristic curve (AUC) exceeding 0.99. Model–pathologist agreement was comparable with interpathologist agreement.[13] Importantly, this study addressed binary CD detection rather than automated Marsh grading.

A particularly challenging diagnostic scenario is potential CD, characterized by persistently positive CD-specific serology in the presence of normal or non-atrophic small-intestinal mucosa; precise definition and management depend on guideline and clinical context.[37] Piccialli *et al*.[37] applied ML to predict biopsy-positive conversion in this population, identifying clinical and laboratory features associated with progression. Such work suggests a possible future role for AI-assisted risk stratification, but it does not establish a validated deep-learning method for continuous mucosal risk assessment.

Explainability remains important for clinical adoption. Attention-map visualization and class activation mapping may help highlight image regions contributing to a model's prediction, potentially supporting review by a pathologist. In a clinical workflow, such visualizations should be treated as aids to interpretation rather than proof of biological causality.

Carreras reported convolutional neural network (CNN)-based CD classification with accuracy of 99.7%, precision of 99.6%, recall of 99.3%, and an F1 score of 99.5% on duodenal biopsy images, although single-center validation limits generalizability.[38] Koh *et al*.[39] and Stoleru *et al*.[40] reported the feasibility of automated biopsy and capsule-endoscopy interpretation, respectively. Soylu and Bozkir applied EfficientNet to IgA-class endomysial antibody immunofluorescence pattern recognition.[41] A systematic review by Hartmann Tolić *et al*.[42] summarized promising performance across AI-assisted CD image-analysis tasks, while emphasizing heterogeneity and the need for broader validation. Key studies across these applications are summarized in Table 2.[12,13,34-41]

Machine learning for pre-diagnostic screening

Pre-diagnostic signatures in routine laboratory data

A pre-diagnostic phase may precede CD recognition, during which laboratory abnormalities such as iron-deficiency anemia and elevated transaminases can occur.[9,10] Dreyfuss *et al*.[12] developed five ML models using deidentified electronic medical record data. The model-development dataset included 677 highly seropositive cases and 176,293 controls; a distinct test dataset included 153 highly seropositive cases and 41,087 controls. The study population comprised adolescents and adults, and the outcome was identification of previously unrecognised CD autoimmunity or seropositivity rather than biopsy-confirmed CD. XGBoost achieved an AUC of 0.86, and performance remained above 0.80 at a four-year pre-index interval in the reported analyses. Iron-deficiency anemia, transaminitis, and low high-density lipoprotein levels were among the leading features identified by model interpretation.[12]

Regarding clinical actionability, an ML model could conceptually be used to prioritize patients for established CD serological testing. Any model-generated threshold would require local calibration and should not be treated as a validated universal cut-off. Once CD-specific serology is obtained, subsequent diagnostic evaluation should follow current age-appropriate clinical guidelines rather than a fixed threshold embedded in this framework. HLA-DQ2/DQ8 testing should be considered an adjunctive test in selected situations of diagnostic uncertainty, not a routine confirmatory step. Any proposed monitoring interval should likewise be regarded as conceptual and unvalidated.

B-cell receptor repertoire and multiomics approaches

Beyond conventional laboratory data, emerging ML approaches exploit deeper biological signals. Shemesh *et al*.[43] applied ML to naïve B-cell receptor repertoire sequencing data, showing that B-cell receptor features distinguished patients with CD from controls in the study dataset. Abraham *et al*.[44] reported genomic prediction of CD risk using statistical learning, and Piroozkhah *et al*.[45] reviewed emerging multiomics approaches that combine genomics, proteomics, metabolomics, microbiome data, and AI. These approaches remain primarily investigational and require validation before they can be incorporated into routine screening.

A four-layer integration framework

Framework architecture and clinical feasibility

The convergence of EHR phenotyping, HPO-based semantic encoding, ML pre-screening, and DL-assisted histopathology can be conceptualized as a four-layer implementation framework for AI-assisted CD diagnostics (Fig. 1).[13] The layers are conceptualized as sequential, with outputs potentially informing the next layer. The framework is not proposed as a replacement for current practice but as a modular architecture that could be implemented incrementally according to institutional infrastructure and validation status.

Layers 1 and 2 may be the most readily deployable components in settings with structured EHR systems. Layer 3 could support targeted serological testing where appropriate data are available, whereas Layer 4 requires digitized WSI infrastructure. The principal implementation challenges include data interoperability, representative training and validation datasets, clinical workflow integration, clinician oversight, and appropriate governance. Claims of clinical readiness should remain proportionate to the available evidence.

- *Layer 1 - EHR phenotyping:* Structured and unstructured EHR data can be processed using NLP and rule-based algorithms to extract computable phenotypes. ICD codes can be augmented with laboratory values, medication records, and pathology

Table 2. Summary of key artificial intelligence studies in celiac disease diagnosis and phenotyping

Study	Algorithm	Dataset/task	Key result and interpretation
Jaeckle *et al.* (2025)[13]	ML ensemble	3,383 training/cross-validation WSIs; 644 WSI independent test set from geographically separate NHS Trust. Binary CD vs non-CD detection	Accuracy, sensitivity and specificity > 95%; AUC > 0.99. Promising independent-test performance; not automated Marsh grading
Dreyfuss *et al.* (2024)[12]	XGBoost and comparators	Development: 677 highly seropositive cases/176,293 controls; test: 153 cases/41,087 controls; Maccabi, Israel	XGBoost AUC 0.86. Adolescents/adults; seropositivity/autoimmunity outcome, not biopsy-confirmed CD
Gruver *et al.* (2023)[35]	Pathologist-trained ML	Multisite biopsy data/WSIs; quantitative feature assessment	Scores correlated with modified Marsh assessment and dietary response
Griffin *et al.* (2024)[36]	Automated DL	Multisite WSIs; cell-type and tissue-region classification	Feasibility demonstrated; broader external validation remains needed
Piccialli *et al.* (2021)[37]	ML	Potential CD cohort; biopsy-positive conversion prediction	Potential risk-stratification role; not a validated DL mucosal-risk model
Carreras (2024)[38]	CNN	Single-center duodenal biopsy images; CD vs non-CD	Accuracy 99.7%; broader external validation needed
Koh *et al.* (2021)[39]	CNN/ML	Duodenal biopsy images; automated interpretation	Supports feasibility of automated biopsy-image interpretation
Stoleru *et al.* (2022)[40]	ML	Capsule endoscopy images; CD detection	Accuracy 94.1%; F1 score 94%
Soylu and Bozkir (2025)[41]	EfficientNet	EMA immunofluorescence images	Emerging AI application to serological image interpretation
Mao *et al.*/ PhenoBrain (2025)[34]	HPO-based AI	2,271 cases, 431 rare diseases	Non-celiac rare-disease cohort; indirect methodological support only

AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CD, celiac disease; CNN, convolutional neural network; DL, deep learning; EMA, endomysial antibody; F1, harmonic mean of precision and recall; HPO, Human Phenotype Ontology; ML, machine learning; NHS, National Health Service; WSI, whole-slide image; XGBoost, extreme gradient boosting.

reports to construct multimodal patient representations.[19-21,26]

- *Layer 2 - HPO semantic encoding:* Patient phenotype profiles can be mapped to HPO terms using automated NLP pipelines or structured data extraction.[32,33,46] The resulting HPO-encoded profiles can support similarity-based matching, interoperability, and research data integration. This layer draws on the curated minimum viable CD phenome in Table 1, which is a conceptual feature set requiring formal validation.
- *Layer 3 - ML pre-screening:* HPO-encoded phenotype profiles, augmented with longitudinal laboratory data, can be processed by ML models to generate a CD risk score.[12,25,37] In a validated implementation, model outputs could help prioritize patients for serological testing. Clinician review of model-derived feature contributions may provide additional transparency.
- *Layer 4 - DL histopathology:* For patients undergoing endoscopic biopsy, digitized WSIs can be processed by DL models for automated CD detection or quantitative histological analysis.[13,35,38-40] The Jaeckle *et al.*[13] model performs binary CD detection, and its AUC > 0.99 was obtained on an independent test set of 644 WSIs from a geographically separate NHS Trust. It should not be characterized as an automated Marsh-grading system. Other studies have explored quantitative Marsh-related features, but these approaches should be distinguished from binary CD detection. Discordant or low-confidence cases should remain under expert pathologist review.

HPO-to-phecode-to-ICD translation

A potential enabling component of the framework is cross-mapping among HPO terms, phecodes, and ICD codes. McArthur *et al.*[32] developed a systematic mapping between HPO terms and phecodes, enabling phenotype data generated in research contexts to be related to clinical coding systems. The Open Annotations for Rare Diseases resource provides real-world-data-derived HPO and rare-disease phenotype annotations extracted from EHRs using ontology mapping and NLP.[47] This rare-disease resource is not a CD-specific validation resource, but it illustrates how real-world EHR data can complement expert-curated phenotype annotations.

Limitations

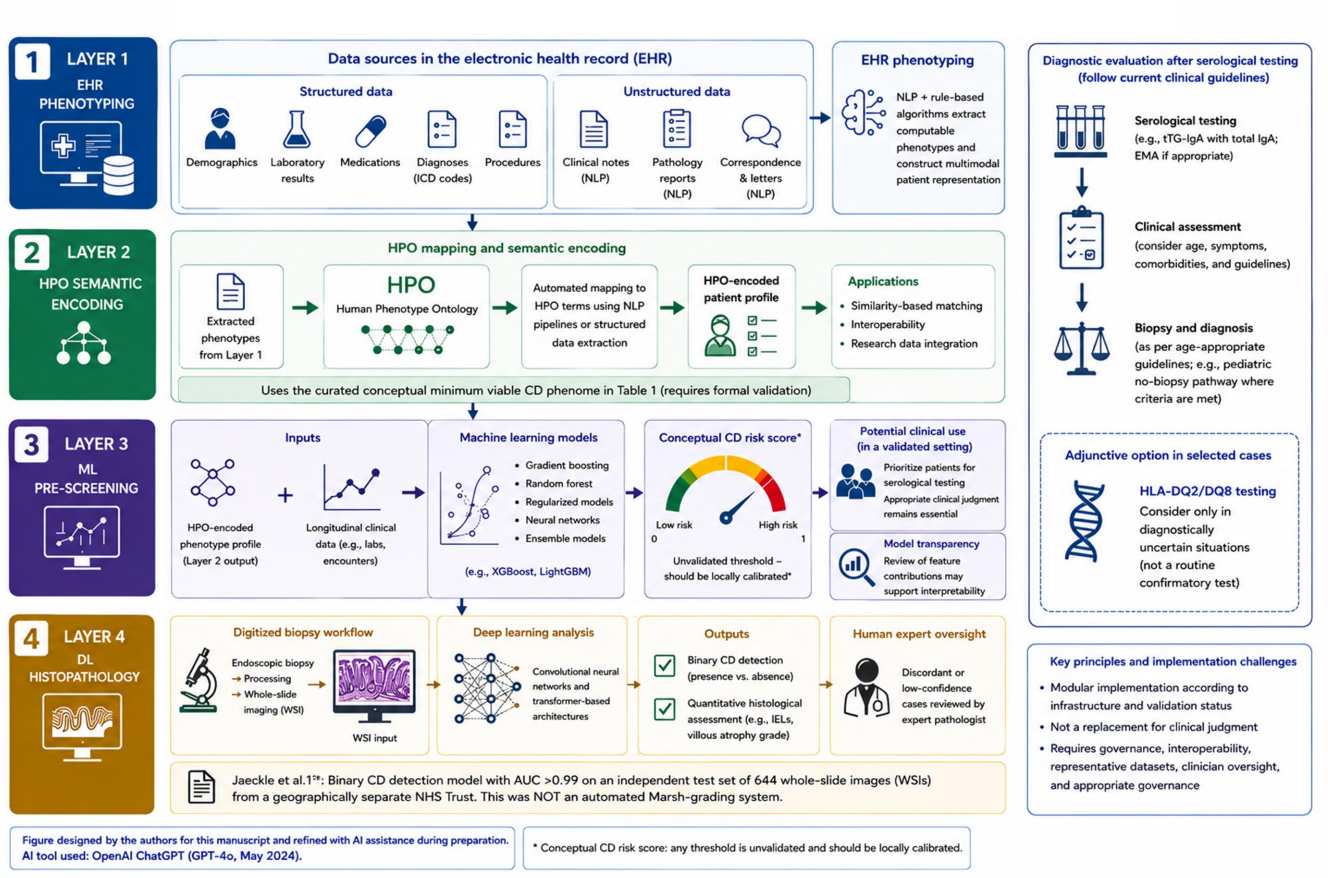

Fig. 1. Four-layer AI implementation framework for celiac disease precision diagnostics. Layer 1 (EHR phenotyping) extracts computable phenotypes from structured and unstructured EHR data. Layer 2 (HPO semantic encoding) maps extracted phenotypes to HPO terms using the curated conceptual minimum viable CD phenome in Table 1. Layer 3 (ML pre-screening) uses phenotype and longitudinal clinical data to generate a conceptual CD risk score and prioritize appropriate serological evaluation; any model threshold is unvalidated and should be locally calibrated. * indicates that any risk threshold is conceptual, unvalidated, and would require local calibration in a future validated implementation. Layer 4 (DL histopathology) represents AI-assisted analysis of digitized biopsy images for binary CD detection or quantitative histological assessment. Jaeckle *et al.*[13] reported an AUC > 0.99 on an independent test set of 644 WSIs from a geographically separate NHS Trust; this was binary CD detection, not automated Marsh grading. Diagnostic evaluation after serological testing should follow current age-appropriate clinical guidelines. HLA-DQ2/DQ8 testing is an adjunctive option only in selected situations of diagnostic uncertainty. The figure was designed by the authors for this manuscript and refined with AI assistance (OpenAI ChatGPT, GPT-4o) during preparation. AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CD, celiac disease; DL, deep learning; EHR, electronic health record; EMA, endomysial antibody; HLA, human leukocyte antigen; HPO, Human Phenotype Ontology; ICD, International Classification of Diseases; IEL, intraepithelial lymphocyte; IgA, immunoglobulin A; ML, machine learning; NHS, National Health Service; NLP, natural language processing; tTG-IgA, tissue transglutaminase IgA; WSI, whole-slide image.

Data quality and standardization

The performance of AI models for CD diagnosis is constrained by the quality, representativeness, and standardization of training data. Histopathological datasets vary in staining protocols, scanner hardware, sampling, and annotation practices, limiting cross-site generalizability.[42] Federated learning and harmonized data standards may facilitate broader validation. The AUC > 0.99 reported by Jaeckle *et al.*[13] was obtained on an independent test set from a geographically separate NHS Trust, which is an important strength; nevertheless, prospective, multicenter studies assessing clinical utility and validation in broader populations remain necessary. Binary CD detection should not be conflated with automated Marsh grading.

Algorithmic bias and health equity

AI models trained predominantly on data from high-income populations may perform differently in underrepresented groups. CD has distinct epidemiological patterns and HLA distributions across populations,[30,31] and AI tools should therefore be evaluated in diverse cohorts before deployment. The current evidence base remains concentrated in selected academic and healthcare systems, underscoring the need for inclusive dataset curation and prospective bias assessment.

Explainability and clinical trust

The adoption of AI in clinical practice requires interpretable and clinically reviewable outputs. Shapley-value-based explanations and attention-map visualizations may assist clinicians in understanding model behavior and identifying potential errors.[12,13] Regulatory requirements vary by jurisdiction and intended use; therefore, developers should address applicable regulatory, validation, documentation, monitoring, and post-market requirements rather than assuming a single universal pathway.

Age-specific validation gaps

As noted in the discussion of HPO-based phenotyping and age-specific considerations, the principal ML pre-screening evidence currently includes adolescents and adults rather than exclusively pediatric cohorts.[12] The lack of prospective validation in dedicated pediatric populations is a critical limitation. Pediatric CD registries capable of providing adequately powered, age-stratified datasets are therefore needed.

Future directions

Emerging technologies

Foundation models pretrained on large pathology image corpora may improve generalization for histopathological tasks when CD-specific training data are limited, but their performance in CD remains to be established prospectively. Large language models integrated with HPO-based pipelines may further automate phenotype extraction from clinical notes.[33,46] Multiomics integration, combining genomics, proteomics, microbiome data, and EHR phenotypes, represents a potential future layer for precision medicine.[45]

Research priorities

Future research should prioritize: (1) construction of large, diverse, multicenter CD histopathology and EHR datasets that include underrepresented populations; (2) prospective real-world validation of ML pre-screening algorithms in primary care and pediatric-specific cohorts; (3) international harmonisation of CD-relevant HPO annotation sets with explicit age-stratified feature weighting; (4) development and external validation of DL models for subtle mucosal changes; (5) prospective bias auditing and inclusive AI development; and (6) implementation and regulatory science studies to support safe translation of AI-based CD tools into clinical workflows.

Conclusions

Celiac disease presents a compelling setting for the application of AI and phenomics to a common, underdiagnosed, and phenotypically heterogeneous condition. The convergence of computable EHR phenotyping, HPO-based semantic encoding, ML pre-screening, and DL-assisted histopathology provides a coherent four-layer framework for future precision diagnostics. Selected models have demonstrated promising CD detection performance, including performance comparable with that of pathologists in specific datasets, while HPO-based AI has shown promise in non-celiac rare-disease differential diagnosis. These findings support further investigation but do not establish clinical superiority or routine clinical utility in CD.

Realizing the potential of this framework requires attention to dataset diversity and standardization, algorithmic bias, age-specific validation, clinical integration, and governance. The four-layer architecture is modular and may be implemented incrementally as individual components achieve adequate validation. Prospective, multicenter studies with external validation are needed before routine clinical deployment can be recommended.

Acknowledgments

The authors thank the open-access scientific community for making the literature and ontological resources cited in this review freely available. In accordance with the journal's policy on AI tool disclosure, the authors declare the use of BIOMNI as an AI-assisted tool to support literature synthesis and language editing during manuscript preparation. Figure 1 was designed by the authors specifically for this manuscript; its visual elements, including icons and schematic components, were created by the authors rather than copied or adapted from third-party copyrighted sources. The figure was subsequently refined with the assistance of artificial intelligence during manuscript preparation. The authors remain responsible for verifying the scientific accuracy and licensing status of all visual components.

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. No grant numbers apply, as no funding was received in support of this research.

Conflict of interest

The authors declare no conflict of interest related to this publication.

Author contributions

Conceptualization (HR, NB, IB); literature review (HR, NB); writing—original draft (HR); writing—review and editing (HR, NB, IB); informatics and technical review (IB); figure and table design (HR). All authors contributed to interpretation, approved the final manuscript, and agree to be accountable for all aspects of the work.

References

[1] Singh P, Arora A, Strand TA, Leffler DA, Catassi C, Green PH, *et al*. Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. Clin Gastroenterol Hepatol 2018;16(6):823–836.e2. DOI: 10.1016/j.cgh.2017.06.037, PMID: 29551598.

[2] Choung RS, Larson SA, Khaleghi S, Rubio-Tapia A, Ovsyannikova IG, King KS, *et al*. Prevalence and Morbidity of Undiagnosed Celiac Disease From a Community-Based Study. Gastroenterology 2017; 152(4):830–839.e5. DOI: 10.1053/j.gastro.2016.11.043, PMID: 27916669.

[3] Whitburn J, Rao SR, Paul SP, Sandhu BK. Diagnosis of celiac disease is being missed in over 80% of children particularly in those from

socioeconomically deprived backgrounds. Eur J Pediatr 2021;180(6): 1941–1946. DOI: 10.1007/s00431-021-03974-8, PMID: 33569662.

[4] Gatti S, Rubio-Tapia A, Makharia G, Catassi C. Patient and Community Health Global Burden in a World With More Celiac Disease. Gastroenterology 2024;167(1):23–33. DOI: 10.1053/j.gastro.2024.01.035, PMID: 38309629.

[5] Mehta S, Agarwal A, Pachisia AV, Singh A, Dang S, Vignesh D, *et al*. Impact of delay in the diagnosis on the severity of celiac disease. J Gastroenterol Hepatol 2024;39(2):256–263. DOI: 10.1111/jgh.16385, PMID: 37963456.

[6] Fuchs V, Kurppa K, Huhtala H, Mäki M, Kekkonen L, Kaukinen K. Delayed celiac disease diagnosis predisposes to reduced quality of life and incremental use of health care services and medicines: A prospective nationwide study. United European Gastroenterol J 2018;6(4):567–575. DOI: 10.1177/2050640617751253, PMID: 29881612.

[7] Zylberberg HM, Miller EBP, Reidy D, Avery K, Newberry C, Ratner A, *et al*. Delay in Celiac Disease Diagnosis Among Patients with High-Risk Screening Conditions: Results from a United States Claims Database. J Clin Med 2025;14(18):6471. DOI: 10.3390/jcm14186471, PMID: 41010675.

[8] Bianchi PI, Lenti MV, Petrucci C, Gambini G, Aronico N, Varallo M, *et al*. Diagnostic Delay of Celiac Disease in Childhood. JAMA Netw Open 2024;7(4):e245671. DOI: 10.1001/jamanetworkopen.2024.5671, PMID: 38592719.

[9] Doyle JB, Silvester J, Ludvigsson JF, Lebwohl B. Advances in the pathophysiology, diagnosis, and management of celiac disease. BMJ 2025;391:e081353. DOI: 10.1136/bmj-2024-081353, PMID: 41093604.

[10] Caio G, Volta U, Sapone A, Leffler DA, De Giorgio R, Catassi C, *et al*. Celiac disease: a comprehensive current review. BMC Med 2019; 17(1):142. DOI: 10.1186/s12916-019-1380-z, PMID: 31331324.

[11] Husby S, Koletzko S, Korponay-Szabó I, Kurppa K, Mearin ML, Ribes-Koninckx C, *et al*. European Society Paediatric Gastroenterology, Hepatology and Nutrition Guidelines for Diagnosing Coeliac Disease 2020. J Pediatr Gastroenterol Nutr 2020;70(1):141–156. DOI: 10.1097/MPG.0000000000002497, PMID: 31568151.

[12] Dreyfuss M, Getz B, Lebwohl B, Ramni O, Underberger D, Ber TI, *et al*. A machine learning tool for early identification of celiac disease autoimmunity. Sci Rep 2024;14(1):30760. DOI: 10.1038/s41598-024-80817-0, PMID: 39730479.

[13] Jaeckle F, Denholm J, Schreiber B, Evans SC, Wicks MN, Chan JYH, *et al*. Machine Learning Achieves Pathologist-Level Coeliac Disease Diagnosis. NEJM AI 2025;2(4):aioa2400738. DOI: 10.1056/AIoa2400738, PMID: 40438269.

[14] Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, *et al*. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res 2024;52(D1):D1333–D1346. DOI: 10.1093/nar/gkad1005, PMID: 37953324.

[15] Rahmoune H, Boutrid N, Benchoufi I. Precision medicine in celiac disease: A step ahead. Artif Intell Gastroenterol 2025;6(1):105682. DOI: 10.35712/aig.v6.i1.105682.

[16] Nelson SJ, Yin Y, Trujillo Rivera EA, Shao Y, Ma P, Tuttle MS, *et al*. Are ICD codes reliable for observational studies? Assessing coding consistency for data quality. Digit Health 2024;10: 20552076241297056. DOI: 10.1177/20552076241297056, PMID: 39493629.

[17] Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu Rev Biomed Data Sci 2021;4:1–19. DOI: 10.1146/annurev-biodatasci-122320-112352, PMID: 34465180.

[18] Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, *et al*. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One 2017;12(7):e0175508. DOI: 10.1371/journal.pone.0175508, PMID: 28686612.

[19] Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: a methodical review. J Am Med Inform Assoc 2023;30(2):367–381. DOI: 10.1093/jamia/ocac216, PMID: 36413056.

[20] Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(e1):e20–e27. DOI: 10.1093/jamia/ocv130, PMID: 26338219.

[21] Chen W, Huang Y, Boyle B, Lin S. The utility of including pathology reports in improving the computational identification of patients. J Pathol Inform 2016;7:46. DOI: 10.4103/2153-3539.194838, PMID: 27994938.

[22] Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, *et al*. Enabling phenotypic big data with PheNorm. J Am Med Inform Assoc 2018;25(1):54–60. DOI: 10.1093/jamia/ocx111, PMID: 29126253.

[23] Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronic health record: PheVis algorithm. J Biomed Inform 2021;117:103746. DOI: 10.1016/j.jbi.2021.103746, PMID: 33746080.

[24] Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, *et al*. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26(12):1545–1559. DOI: 10.1093/jamia/ocz105, PMID: 31329239.

[25] Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, *et al*. Use of computerized algorithm to identify individuals in need of testing for celiac disease. J Am Med Inform Assoc 2013;20(e2): e306–e310. DOI: 10.1136/amiajnl-2013-001924, PMID: 23956016.

[26] Escudié JB, Rance B, Malamut G, Khater S, Burgun A, Cellier C, *et al*. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. BMC Med Inform Decis Mak 2017;17(1):140. DOI: 10.1186/s12911-017-0537-y, PMID: 28962565.

[27] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83(5):610–615. DOI: 10.1016/j.ajhg.2008.09.017, PMID: 18950739.

[28] Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, *et al*. Expansion of the Human Phenotype Ontology knowledge base and resources. Nucleic Acids Res 2019;47(D1):D1018–D1027. DOI: 10.1093/nar/gky1105, PMID: 30476213.

[29] Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, *et al*. The Human Phenotype Ontology in 2021. Nucleic Acids Res 2021;49(D1):D1207–D1217. DOI: 10.1093/nar/gkaa1043, PMID: 33264411.

[30] Catassi C, Gatti S, Lionetti E. World perspective and celiac disease epidemiology. Dig Dis 2015;33(2):141–146. DOI: 10.1159/000369518, PMID: 25925915.

[31] Makharia GK, Chauhan A, Singh P, Ahuja V. Review article: Epidemiology of coeliac disease. Aliment Pharmacol Ther 2022;56 Suppl 1:S3–S17. DOI: 10.1111/apt.16787, PMID: 35815830.

[32] McArthur E, Bastarache L, Capra JA. Linking rare and common disease vocabularies by mapping between the human phenotype ontology and phecodes. JAMIA Open 2023;6(1):ooad007. DOI: 10.1093/jamiaopen/ooad007, PMID: 36875690.

[33] Yang J, Liu C, Deng W, Wu D, Weng C, Zhou Y, *et al*. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. Patterns (N Y) 2024;5(1):100887. DOI: 10.1016/j.patter.2023.100887, PMID: 38264716.

[34] Mao X, Huang Y, Jin Y, Wang L, Chen X, Liu H, *et al*. A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. NPJ Digit Med 2025;8(1):68. DOI: 10.1038/s41746-025-01452-1, PMID: 39875532.

[35] Gruver AM, Lu H, Zhao X, Fulford AD, Soper MD, Ballard D, *et al*. Pathologist-trained machine learning classifiers developed to quantitate celiac disease features differentiate endoscopic biopsies according to modified marsh score and dietary intervention response. Diagn Pathol 2023;18(1):122. DOI: 10.1186/s13000-023-01412-x, PMID: 37951937.

[36] Griffin M, Gruver AM, Shah C, Wani Q, Fahy D, Khosla A, *et al*. A feasibility study using quantitative and interpretable histological analyses of celiac disease for automated cell type and tissue area classification. Sci Rep 2024;14(1):29883. DOI: 10.1038/s41598-024-79570-1, PMID: 39622903.

[37] Piccialli F, Calabrò F, Crisci D, Cuomo S, Prezioso E, Mandile R, *et al*. Precision medicine and machine learning towards the prediction of the outcome of potential celiac disease. Sci Rep 2021;11(1):5683. DOI: 10.1038/s41598-021-84951-x, PMID: 33707543.

[38] Carreras J. Celiac Disease Deep Learning Image Classification Using Convolutional Neural Networks. J Imaging 2024;10(8):200. DOI: 10.3390/jimaging10080200, PMID: 39194989.

[39] Koh JEW, De Michele S, Sudarshan VK, Jahmunah V, Ciaccio EJ, Ooi CP, *et al*. Automated interpretation of biopsy images for the detection of celiac disease using a machine learning approach. Comput Methods Programs Biomed 2021;203:106010. DOI: 10.1016/j.cmpb.2021.106010, PMID: 33831693.

[40] Stoleru CA, Dulf EH, Ciobanu L. Automated detection of celiac disease using Machine Learning Algorithms. Sci Rep 2022;12(1):4071. DOI: 10.1038/s41598-022-07199-z, PMID: 35260574.

[41] Soylu M, Bozkir AS. Recognizing IgA-class endomysial antibody equivalent binding patterns on monkey liver substrate through EfficientNet architectures and deep learning. PeerJ 2025;13:e20191. DOI: 10.7717/peerj.20191, PMID: 41112771.

[42] Hartmann Tolić I, Habijan M, Galić I, Nyarko EK. Advancements in Computer-Aided Diagnosis of Celiac Disease: A Systematic Review. Biomimetics (Basel) 2024;9(8):493. DOI: 10.3390/biomimetics9080493, PMID: 39194472.

[43] Shemesh O, Polak P, Lundin KEA, Sollid LM, Yaari G. Machine Learning Analysis of Naïve B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. Front Immunol 2021;12:627813. DOI: 10.3389/fimmu.2021.627813, PMID: 33790900.

[44] Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, Inouye M. Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet 2014;10(2):e1004137. DOI: 10.1371/journal.pgen.1004137, PMID: 24550740.

[45] Piroozkhah M, Kahangi MF, Asri N, Piroozkhah M, Moradi N, Nazemalhosseini-Mojarad E, *et al*. Toward Precision Medicine in Celiac Disease: Emerging Roles of Multiomics and Artificial Intelligence. Anal Cell Pathol (Amst) 2025;2025:6570955. DOI: 10.1155/ancp/6570955, PMID: 41080452.

[46] Garcia BT, Westerfield L, Yelemali P, Gogate N, Rivera-Munoz EA, Du H, *et al*. Improving automated deep phenotyping through large language models using retrieval-augmented generation. Genome Med 2025;17(1):91. DOI: 10.1186/s13073-025-01521-w, PMID: 40826123.

[47] Liu C, Ta CN, Havrilla JM, Nestor JG, Spotnitz ME, Geneslaw AS, *et al*. OARD: Open annotations for rare diseases and their phenotypes based on real-world data. Am J Hum Genet 2022;109(9):1591–1604. DOI: 10.1016/j.ajhg.2022.08.002, PMID: 35998640.