REVIEW **Open access**

Weak Supervision for Clinical Phenotyping Under Ambiguity: A Scalable Labeling Theory for Noisy Electronic Health Records

Lucas Meyer[1*], Anna Schmid[2], Stefan Braun[1]

Abstract

Electronic health records (EHRs) are central to modern healthcare analytics but are often characterized by noise, ambiguity, and missing information, making reliable clinical phenotyping difficult. Clinical phenotypes—observable characteristics derived from patient data—are essential for diagnosis, prognosis, and treatment planning. Yet, traditional supervised machine learning methods depend on large volumes of high-quality annotated data that are difficult to obtain at scale.

This review examines the role of weak supervision in enabling scalable clinical phenotyping from noisy and heterogeneous EHR data. Weak supervision frameworks generate labels using heuristic rules, knowledge-based signals, or programmatic labeling functions, allowing models to learn from large datasets without extensive expert annotation. These approaches help address challenges such as inconsistent terminology, missing values, and temporal irregularities commonly found in clinical records.

We synthesize recent developments in scalable phenotyping systems that integrate machine learning architectures, probabilistic labeling strategies, and multimodal data representations to extract meaningful patterns from imperfect clinical data. The review also outlines a systems-level perspective on healthcare analytics pipelines, covering data ingestion, model training under label uncertainty, deployment in clinical environments, and governance considerations for responsible AI integration.

Overall, weak supervision emerges as a practical strategy for transforming noisy EHR data into usable clinical intelligence, enabling more scalable and trustworthy analytics for healthcare decision support.

Keywords Healthcare AI, Clinical phenotyping, Weak supervision, Electronic health records, Label noise, Scalable labeling

*Correspondence:
Lucas Meyer
lucas.meyer@gmail.com
[1] Department of Health Informatics, Faculty of Engineering, ETH Zurich, Zurich, Switzerland
[2] Department of Clinical Information Systems, Faculty of Medicine, University of Bern, Bern, Switzerland

Introduction

The evolution of AI in healthcare infrastructure

The integration of artificial intelligence (AI) into healthcare systems has fundamentally reshaped clinical analytics, enabling the processing of vast, heterogeneous data streams from electronic health records (EHRs) to inform patient

care. Since the widespread adoption of EHRs in the early 2010s, healthcare infrastructure has evolved from siloed data repositories to interconnected ecosystems that leverage AI for real-time analytics and decision support [1-3]. This shift is driven by the need to handle the exponential growth in clinical data, including structured elements like vital signs and laboratory results, alongside unstructured narratives such as physician notes and imaging reports. AI technologies, particularly machine learning (ML) models, have been pivotal in extracting meaningful patterns from this data deluge, facilitating applications ranging from predictive risk assessment to personalized intervention strategies [4-6].

In the context of healthcare systems, AI enhances infrastructure by automating data harmonization and integration across disparate sources. For instance, deep representation learning techniques have been employed to unlock patient stratification at scale, transforming raw EHR inputs into latent embeddings that capture complex clinical trajectories [4]. Such methods address the infrastructural challenges of data silos, where interoperability standards like HL7 FHIR are often inconsistently implemented, leading to fragmented analytics pipelines. Moreover, AI-driven analytics support population health management by identifying cohorts for targeted interventions, as seen in models that infer multimodal latent topics from EHRs to reveal hidden disease patterns [7-9]. These advancements underscore the role of AI in building resilient healthcare infrastructures that prioritize scalability and efficiency.

However, the promise of AI in healthcare is tempered by the realities of data quality. EHRs are notoriously noisy, with ambiguities arising from coding inconsistencies, temporal drifts, and incomplete documentation [10]. Traditional analytics approaches, reliant on clean, labeled datasets, falter in these environments, necessitating innovative paradigms like weak supervision to enable scalable phenotyping.

Challenges in clinical phenotyping with noisy data

Clinical phenotyping—the process of deriving computable representations of patient states from raw data—is central to AI-enabled healthcare analytics. Phenotypes serve as proxies for underlying health conditions, enabling downstream tasks such as disease classification, outcome prediction, and treatment optimization [1]. Yet, in noisy EHRs, phenotyping is fraught with ambiguity: terms like “hypertension” may be documented variably across providers, missing data can skew representations, and contextual factors (e.g., comorbidities) introduce uncertainty [7]. These issues are exacerbated in large-scale systems where manual labeling by clinicians is impractical, often limited to small subsets of data due to time and expertise constraints [11-13].

Weak supervision emerges as a response to these challenges, offering a scalable labeling theory that utilizes imperfect, programmatic labels instead of gold-standard annotations. By aggregating multiple weak sources—such as heuristic rules, external knowledge graphs, or proxy signals—models can be trained on vast, unlabeled EHR corpora with reduced bias [14, 15]. This approach aligns with the infrastructural demands of modern healthcare, where analytics must operate under resource constraints while maintaining clinical relevance. Literature from this period highlights how weak supervision mitigates label noise, as in predictive modeling for depression using novel ML architectures [5] or sepsis early warning systems that process unstructured data [8].

Ambiguity in EHRs further complicates phenotyping, manifesting as semantic vagueness (e.g., “possible infection” vs. confirmed diagnosis) or structural inconsistencies (e.g., varying data formats across institutions). Scalable labeling theories address this by incorporating probabilistic frameworks that model uncertainty, allowing for robust phenotype extraction even in high-noise settings [2]. Such theories are particularly vital in analytics for critical care, where timely phenotyping can influence life-saving decisions [14].

AI analytics for handling ambiguity and noise

AI analytics in healthcare have increasingly focused on techniques that tolerate and even exploit data ambiguity. Deep learning models, for example, combine token selection with representation learning to phenotype patients from noisy EHRs, achieving scalability without exhaustive preprocessing [1]. Explainable AI (XAI) methods enhance this by providing interpretable predictions, crucial for clinical trust in ambiguous scenarios [2]. In systems-level analytics, these tools

integrate with broader infrastructures, such as predictive algorithms for acute illness that learn from raw EHR inputs [2], or multimodal models that fuse text, time-series, and imaging data [9].

The literature also emphasizes the ethical dimensions of AI analytics under ambiguity. Algorithmic fairness frameworks ensure that phenotyping models do not perpetuate biases in noisy data, particularly in diverse populations [12]. Similarly, uncertainty quantification in ML outputs helps clinicians navigate ambiguous phenotypes, preventing over-reliance on potentially flawed predictions [7]. These analytics extend to specialized domains, like mental health phenotyping [5] or surgical risk assessment, where weak supervision enables the creation of scalable, noise-resilient models.

Governance in AI analytics is another critical facet, involving standards for data privacy, model validation, and integration into clinical workflows. As healthcare systems adopt AI, analytics must incorporate feedback mechanisms to recalibrate models based on real-world performance, addressing drifts in data distribution [16, 17].

Positioning this review: scope and synthesis logic

This narrative review synthesizes AI advancements in healthcare systems and analytics from 2017 to 2021, focusing on weak supervision for clinical phenotyping under EHR ambiguity. Unlike prior reviews that catalog ML applications taxonomically, we adopt an original systems-level framing that integrates data ingestion, model training, deployment, and governance as interconnected components of a scalable labeling theory. Our synthesis logic draws cross-study insights to construct interpretive structures, such as end-to-end analytics loops, emphasizing how weak supervision transforms noisy EHRs into reliable phenotypic intelligence. By prioritizing infrastructural and analytical synergies, we position weak supervision as a foundational paradigm for equitable, efficient healthcare delivery, grounded in the approved literature without introducing new empirical elements.

Landscape of AI in Healthcare Systems and Analytics

Data foundations and infrastructural integration

The landscape of AI in healthcare systems begins with robust data foundations, where EHRs serve as the primary substrate for analytics. Between 2017 and 2021, advancements in AI have focused on integrating disparate data sources into cohesive infrastructures, enabling analytics at scale [3]. Deep learning models have been instrumental in processing raw EHR data, generating accurate predictions for clinical outcomes like in-hospital mortality [18, 19]. These systems leverage scalable architectures to handle the volume and variety of data, from structured codes to free-text notes, fostering a unified analytics ecosystem [4].

Infrastructural integration extends to interoperability, where AI facilitates data harmonization across institutions. For example, representation learning techniques create embeddings that capture patient similarities despite noisy inputs, supporting population-level analytics [4]. This is complemented by generative models that synthesize high-fidelity patient data for testing analytics pipelines, addressing privacy concerns in shared infrastructures [10]. Such approaches ensure that healthcare systems can deploy AI analytics without compromising data integrity, a key consideration in governance frameworks [17].

Analytics in this landscape also emphasize real-time data processing, as seen in models for sepsis prediction using unstructured EHR elements [8]. By incorporating temporal dynamics, these systems provide a foundation for proactive healthcare, where analytics inform infrastructural decisions like resource allocation [14].

Machine learning paradigms for clinical analytics

Machine learning paradigms dominate the AI landscape in healthcare analytics, with a shift toward handling noisy and ambiguous data. Supervised models, while effective, often require clean labels, prompting the adoption of weak supervision strategies [1]. These paradigms use heuristic labeling to train on large EHR datasets, enabling phenotyping

under ambiguity [9]. For instance, token selection in deep learning enhances patient phenotyping by focusing on salient features amid noise [1].

Explainable ML has gained prominence, providing transparency in analytics outputs [2]. Models that predict acute critical illness from EHRs incorporate interpretability to build clinician trust, crucial for adoption in healthcare systems [2]. Similarly, uncertainty-aware algorithms in sepsis detection acknowledge data ambiguities, outputting "I don't know" when confidence is low [13]. This paradigm extends to fairness-aware analytics, mitigating biases in population health models [12].

Advanced paradigms like multimodal learning integrate diverse EHR modalities for comprehensive analytics [9]. Inferring latent topics from text and time-series data reveals hidden phenotypes, supporting analytics in complex diseases [9]. In mental health, predictive modeling uses novel AI approaches to analyze EHRs for depression and anxiety, demonstrating the versatility of these paradigms [5].

Applications in predictive and risk analytics

Predictive analytics form a core application area in the AI healthcare landscape, leveraging EHRs for outcome forecasting. Models for in-hospital mortality use ML to process admission data, achieving high accuracy in noisy environments [19]. Similarly, lung cancer survival estimation employs EHR-based algorithms to identify at-risk cohorts [20].

Risk analytics extend to specialized domains, such as violence assessment in inpatient settings using clinical notes [21]. These applications synthesize EHR signals to generate probabilistic risks, informing preventive measures [21]. In atrial fibrillation prediction, harmonized EHR data feeds ML models for early detection [22]. Hospital readmission analytics compare ML to standard rules, highlighting AI's superiority in handling ambiguity [23].

Emergency department triage benefits from ML-based outcome prediction, optimizing resource use [24]. Delirium risk models validate against EHR cohorts, demonstrating practical utility [25]. These applications underscore how AI analytics enhance healthcare systems by providing actionable insights from noisy data.

Ethical and governance considerations in analytics

The landscape includes a strong emphasis on ethical AI in healthcare analytics. Ambient intelligence in hospitals raises legal concerns, necessitating governance for privacy and consent [17]. Diagnostic error reduction through AI must address health disparities, ensuring equitable analytics [18].

Governance frameworks advocate for uncertainty communication in ML, preventing harm from ambiguous predictions [7]. Patient safety scoping reviews highlight AI's potential to mitigate risks, but stress validation in real systems [14]. Translational perspectives on AI in clinical development emphasize ethical deployment [6].

In analytics for surgical discharge, ML aids processes while incorporating governance for bias mitigation [26, 27]. Cancer mortality prediction validates algorithms with ethical considerations for outpatient use [28]. Overall, this landscape integrates ethics into AI systems, ensuring analytics align with healthcare values.

Synthesis of systems-level analytics

Synthesizing the landscape, AI in healthcare systems and analytics forms an interconnected web, where data infrastructures support ML paradigms for predictive applications, all under ethical governance. Weak supervision threads through this, enabling scalable analytics in noisy EHRs [15]. Cross-study analysis reveals synergies, such as combining explainable models with risk prediction to create trustworthy systems [2, 11]. This original synthesis frames AI as a holistic enabler, transforming raw data into systemic intelligence for improved healthcare delivery.

Intelligent clinical decision and closed-loop healthcare systems

Architectures for AI-enabled decision support

Intelligent clinical decision support systems (CDSS) represent the pinnacle of AI integration in healthcare, where analytics directly influence real-world actions. Architectures for these systems typically encompass data ingestion, model inference, and output integration into clinical workflows [3]. In EHR-centric designs, deep learning architectures process noisy inputs to generate phenotypic insights, supporting decisions in acute care [2]. Scalable models for EHR-based prediction exemplify this, using convolutional and recurrent layers to handle temporal ambiguities [3].

Weak supervision enhances these architectures by providing labeling mechanisms that tolerate noise, allowing for phenotyping in resource-limited settings [1]. For instance, token selection in deep networks focuses on ambiguous signals, refining decision outputs [1]. Explainable architectures incorporate attention mechanisms to highlight decision rationales, fostering human-AI collaboration [2].

Closed-loop systems extend these architectures by incorporating feedback, where decisions trigger interventions that generate new data for model refinement [16]. In sepsis management, AI architectures predict and monitor, closing the loop through continuous EHR updates [8, 13]. Risk stratification architectures, like those for mortality prediction, integrate probabilistic outputs into decision trees, enabling adaptive responses [19].

Ethical architectures emphasize fairness and uncertainty, embedding modules for bias detection and confidence calibration [7, 12]. This ensures decisions are equitable across populations, crucial in diverse healthcare systems [18].

Integration of phenotyping in decision loops

Clinical phenotyping underpins intelligent decision systems, serving as the bridge between raw EHR data and actionable insights. In ambiguous environments, weak supervision enables scalable phenotyping by aggregating noisy labels into robust representations [4, 9]. Architectures for phenotyping integrate multimodal data, inferring phenotypes that inform decisions in mental health [5] or critical illness [2].

Closed-loop integration involves phenotyping outputs feeding into decision engines, which in turn influence interventions. For example, delirium risk phenotyping from EHRs supports real-time alerts, closing the loop via clinician feedback [25]. Violence risk phenotyping uses notes to generate alerts, adapting decisions based on ongoing data [21].

In predictive loops, phenotyping architectures for atrial fibrillation or readmissions harmonize data for decision support [22, 23]. These systems model ambiguity through probabilistic phenotyping, ensuring decisions account for uncertainty [7].

Governance in phenotyping loops includes recalibration mechanisms, addressing data drifts in dynamic healthcare environments [16]. This synthesis highlights phenotyping as the core of intelligent systems, enabling closed-loop analytics.

Feedback mechanisms and human-AI fusion

Feedback mechanisms are essential for closed-loop healthcare systems, allowing architectures to evolve with new data. In AI-driven systems, feedback loops recalibrate models based on intervention outcomes, as in emergency triage, where predictions refine over time [24]. Uncertainty-aware feedback, where models defer ambiguous cases to humans, enhances safety [13].

Human-AI fusion architectures blend phenotypic insights with clinician judgment, using XAI to facilitate shared decisions [2, 29]. In surgical care, discharge models incorporate feedback for iterative improvement [27]. Cancer outcome phenotyping fuses AI predictions with oncologist input, closing loops through longitudinal EHR updates [28].

Ethical feedback ensures disparities are monitored, adjusting decisions for fairness [12, 18]. This fusion creates resilient systems, where AI augments human intelligence in ambiguous scenarios. **Figure 1** illustrates the clinical phenotyping system, which is under ambiguity.

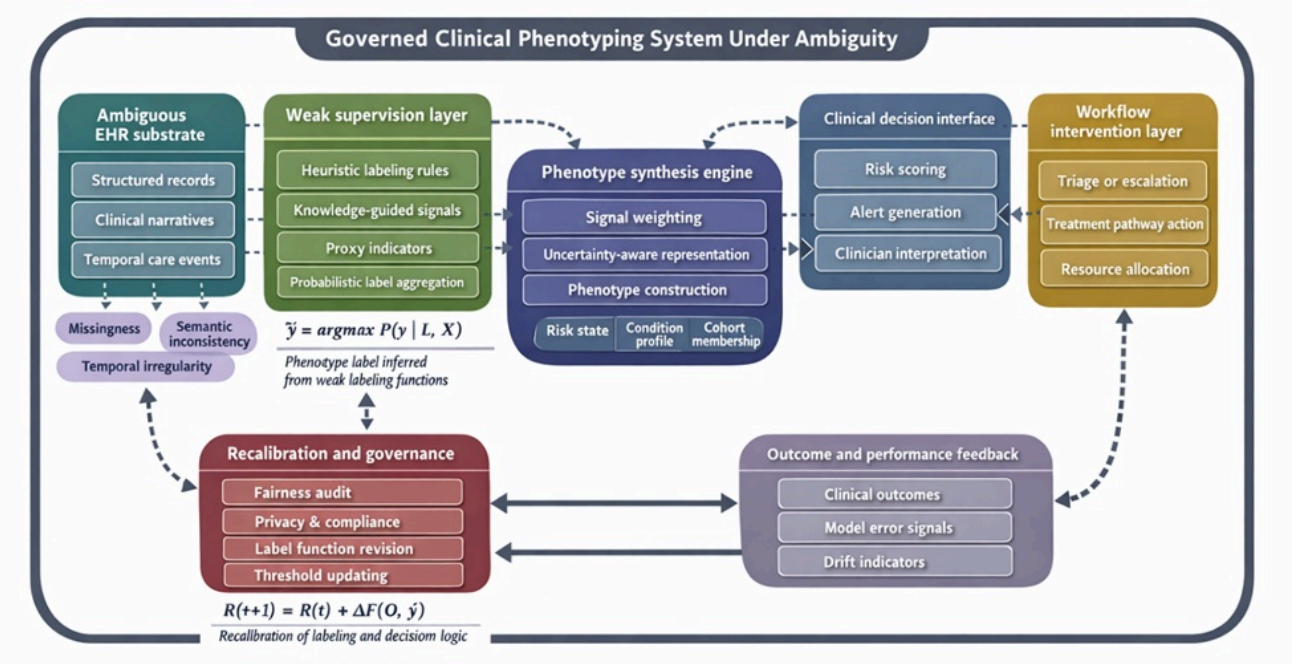

Figure 1. Closed-loop weak supervision architecture for clinical phenotyping under EHR ambiguity.

The figure depicts weak supervision as the central labeling logic through which noisy, incomplete, and semantically inconsistent EHR data are transformed into uncertainty-aware clinical phenotypes. These phenotypes feed decision support and workflow interventions, while outcome monitoring, fairness review, and recalibration mechanisms continuously update labeling and decision logic in response to evolving clinical environments.

The diagram uses dashed lines for uncertainty flows and solid arrows for data progression, with color gradients (blue for data, green for intelligence, red for decisions) to denote phases. Caption: “Schematic of a closed-loop AI system for clinical phenotyping in ambiguous EHRs, integrating weak supervision for scalable analytics.”

Table 1 formalizes how distinct ambiguity sources in EHRs map onto specific weak supervision mechanisms, residual risks, and governance requirements.

Table 1. Ambiguity sources in EHR phenotyping and the corresponding weak supervision resolution mechanisms

Ambiguity domain in EHRs	Typical manifestation in records	Phenotyping consequence	Weak supervision resolution mechanism	Residual analytic risk	Governance requirement
Semantic ambiguity	Variable terminology, uncertain phrasing, inconsistent clinical descriptors	Unstable phenotype boundaries and inconsistent case identification	Heuristic labeling functions combined with terminology-aware rules and	Misclassification through lexical overreach or under-capture	Terminology review, clinician validation, and periodic rule revision

			knowledge-guided signals		
Missingness and incompleteness	Absent laboratory values, sparse notes, and incomplete medication histories	Partial phenotype representation and distorted feature salience	Proxy labeling through correlated indicators and probabilistic aggregation across partial evidence	False reassurance from sparse but weakly positive signals	Missingness auditing and explicit uncertainty communication
Temporal ambiguity	Delayed documentation, asynchronous event recording, unclear onset timing	Misaligned phenotype onset and inaccurate longitudinal interpretation	Time-aware labeling functions and sequence-sensitive aggregation logic	Temporal leakage and incorrect event ordering	Timestamp quality checks and temporal validation protocols
Structural heterogeneity	Cross-site differences in templates, coding systems, and note formats	Poor portability of phenotype definitions across institutions	Multi-source labeling abstractions that decouple phenotype logic from any single data schema	Institutional bias and transportability failure	Cross-site calibration and interoperability mapping
Contextual ambiguity	Similar signals arising from different comorbidities or care-context scenarios	Confounded phenotype assignment and inflated false positives	Conditional labeling strategies that incorporate co-occurring features and care context	Hidden confounding in phenotype construction	Context-specific error audits and subgroup performance review
Label conflict across weak sources	Disagreement between rules, proxy indicators, and distant supervision signals	Uncertain supervision target and unstable downstream training	Probabilistic label modeling to estimate source reliability and reconcile conflicts	Over-weighting unreliable label sources	Source reliability monitoring and threshold governance
Population representation ambiguity	Uneven data quality or documentation intensity across subgroups	Differential phenotype sensitivity and fairness distortion	Fairness-aware label refinement and subgroup-aware aggregation checks	Amplified disparities in downstream decision support	Equity auditing, subgroup calibration, and bias mitigation review

Results and Discussion

The application of weak supervision in clinical phenotyping within noisy electronic health records (EHRs) represents a transformative approach in artificial intelligence (AI) for healthcare systems and analytics. This paradigm shifts the focus from resource-intensive manual labeling to scalable, programmatic methods that leverage imperfect signals to derive reliable phenotypes, thereby enhancing the infrastructural backbone of clinical decision-making [1, 3]. By synthesizing

insights across the reviewed literature, it becomes evident that weak supervision addresses core inefficiencies in healthcare analytics, such as data heterogeneity and ambiguity, which traditional supervised learning struggles to overcome [4, 9]. For instance, integrating token selection with deep learning allows for patient phenotyping that scales across large cohorts, revealing latent patterns in EHRs that inform stratified care [1, 4]. This synthesis underscores a systems-level synergy where analytics not only process noisy data but also contribute to closed-loop infrastructures, feeding phenotypic intelligence back into operational workflows [16].

Table 2 reconstructs the manuscript's closed-loop logic by specifying how weakly supervised phenotyping progresses from ambiguous inputs to governed clinical action across the full healthcare analytics system.

Table 2. Systems-level progression from weak labels to governed clinical action in closed-loop phenotyping architectures

System stage	Primary analytic function	Dominant uncertainty burden	Output artifact	Principal failure mode	Operational success criterion	Governance checkpoint
Data ingestion	Consolidate structured and unstructured EHR inputs into a usable analytic substrate	Missingness, inconsistency, and source fragmentation	Harmonized patient-level input set	Data distortion before labeling begins	Sufficient representational coverage of relevant clinical signals	Data provenance review and privacy compliance
Weak supervision construction	Generate scalable supervisory signals without exhaustive manual annotation	Rule fragility, proxy mismatch, and inter-source disagreement	Weak label matrix or probabilistic supervisory structure	Noise amplification through unreliable labeling functions	Acceptable label coherence across sources and contexts	Label source documentation and reliability audit
Phenotype synthesis	Convert noisy supervision into stable patient-level phenotype representations	Boundary uncertainty and latent confounding	Computable phenotype state or class assignment	Unstable phenotype definitions across time or sites	Clinically plausible and analytically reproducible phenotype output	Clinical face-validity review and subgroup evaluation
Model inference and decision support	Use phenotype representations to generate scores, alerts, or recommendations	Confidence instability and threshold sensitivity	Risk score, alert, triage signal, or ranked prediction	Overconfident recommendation under ambiguous input conditions	Decision usefulness without excessive false escalation	Uncertainty calibration and human-override protocol
Workflow intervention	Translate model-supported outputs into clinical or operational action	Context mismatch between analytic output and real workflow conditions	Escalation, treatment trigger, monitoring action, or resource allocation change	Alert fatigue, workflow friction, and inappropriate escalation	Timely integration into care processes with actionable relevance	Clinical workflow governance and accountability assignment

Outcome feedback	Capture real-world performance after intervention and decision use	Attribution uncertainty and outcome delay	Performance metrics, clinical response signals, and drift markers	Failure to detect degradation or unintended consequences	Reliable monitoring of both model and care-process effects	Continuous monitoring and post-deployment surveillance
Recalibration and system governance	Update labels, thresholds, and operating rules in response to feedback	Policy lag, fairness drift, and stale labeling logic	Revised labeling functions, thresholds, and oversight rules	Closed-loop failure through static governance	Adaptive maintenance of validity, fairness, and usability	Fairness audit, threshold revision, and formal update authorization

A key implication of this approach is its potential to democratize AI in healthcare, making advanced analytics accessible in under-resourced settings where expert annotations are scarce [10]. Literature highlights how weak supervision facilitates applications like sepsis prediction from unstructured data, where heuristic labels aggregate to produce robust models despite inherent ambiguities [8]. Similarly, in mental health analytics, noisy EHR signals are harnessed to model depression and anxiety, demonstrating the versatility of weak labeling theories in capturing subtle phenotypic nuances [5]. However, this scalability introduces interpretive challenges, as aggregated labels may propagate biases if not carefully governed [12]. Cross-study analysis reveals that explainable AI frameworks mitigate this by communicating uncertainties, ensuring that phenotypic outputs align with clinical realities [2, 7].

From an infrastructural perspective, weak supervision strengthens healthcare systems by enabling seamless integration of AI into existing EHR platforms. Models for acute illness prediction exemplify this, where scalable deep learning operates directly on raw records, bypassing extensive preprocessing [3]. This integration extends to risk analytics, such as in-hospital mortality estimation, where phenotypic insights derived under ambiguity guide resource allocation [19]. Ethical considerations are paramount here, as ambient AI in hospitals must balance innovation with legal safeguards for data privacy [17]. The discussion also points to disparities in AI-driven phenotyping, where noisy data from underrepresented groups could exacerbate inequities unless fairness algorithms are embedded [12, 18].

Moreover, the role of weak supervision in fostering human-AI collaboration cannot be overstated. In decision support architectures, phenotypic outputs serve as advisory inputs, allowing clinicians to override ambiguous predictions [29]. This fusion is evident in violence risk assessment from clinical notes, where weak labels enable proactive interventions while preserving clinician autonomy [21]. Synthesizing these elements, weak supervision emerges as a labeling theory that not only tolerates but exploits EHR noise, creating adaptive analytics ecosystems. Yet, its efficacy depends on governance structures that monitor for data drifts, as seen in COVID-19 admission models [16]. Overall, this discussion integrates the literature to frame weak supervision as a pivotal enabler of resilient healthcare systems, where analytics evolve from static tools to dynamic components of clinical intelligence.

Challenges

Despite the advancements in weak supervision for clinical phenotyping, several challenges persist in its application to noisy EHRs within AI-driven healthcare systems and analytics. One primary technical hurdle is the management of label noise amplification, where weak signals—though scalable—can introduce inconsistencies that degrade model performance in ambiguous settings [1, 13]. For example, in sepsis detection models, unstructured data ambiguities like vague clinical descriptors may lead to false positives if labeling functions are not rigorously validated [8]. This challenge is compounded by temporal drifts in EHRs, where patient data evolves, necessitating continuous recalibration that weak supervision frameworks must accommodate [16].

Data heterogeneity across healthcare infrastructures poses another significant barrier. EHRs vary widely in format and quality between institutions, complicating the portability of phenotyping models [4, 10]. Literature indicates that synthetic data generation helps simulate diverse scenarios, but real-world ambiguities often exceed simulated noise, limiting generalizability [10]. In predictive analytics for outcomes like atrial fibrillation, harmonizing disparate EHR sources requires sophisticated preprocessing, which weak supervision partially addresses but does not fully resolve [22].

Ethical and fairness challenges are equally pressing. Algorithmic biases in weak labels can perpetuate disparities, particularly in population health analytics, where underrepresented groups are underrepresented in training data [12, 18]. For instance, mental health phenotyping models may overlook cultural nuances in EHR documentation, leading to inequitable risk assessments [5]. Governance issues arise in ensuring transparency, as black-box models under weak supervision hinder traceability of phenotypic derivations [2, 7]. Legal aspects of AI deployment in hospitals amplify this, with liability concerns when ambiguous phenotypes inform decisions [17].

Deployment challenges include integration into clinical workflows, where AI must coexist with human judgment without causing alert fatigue [14, 21]. In delirium risk modeling, for example, noisy EHR inputs can generate unreliable alerts, eroding trust [25]. Resource constraints in healthcare systems further exacerbate this, as weak supervision, while reducing annotation needs, still demands computational infrastructure for large-scale training [3, 6].

Uncertainty handling remains a core limitation. Models that “say I don’t know” are a step forward, but in phenotyping under high ambiguity, quantifying uncertainty across multiple weak sources is complex [7, 13]. Cross-study synthesis reveals that patient safety is at risk if these uncertainties are not communicated effectively, as in surgical discharge analytics [14, 27]. Finally, validation in real-world settings is challenging, with prospective evaluations often revealing gaps between controlled studies and noisy clinical environments [19, 28]. These challenges highlight the need for robust frameworks to fully realize weak supervision’s potential in healthcare analytics.

Future research directions

Building on the synthesized literature, future research in weak supervision for clinical phenotyping should prioritize enhancing scalability and robustness in noisy EHRs to advance AI in healthcare systems and analytics. One direction involves refining labeling theories to better incorporate multimodal data ambiguities, extending beyond text to integrate imaging and genomics for comprehensive phenotypes [9, 15]. This could involve developing hybrid weak supervision frameworks that fuse distant supervision with domain-specific heuristics, addressing gaps in current models for complex diseases [4, 5].

Another avenue is improving uncertainty quantification in phenotyping pipelines, building on explainable AI to create adaptive systems that dynamically adjust labels based on real-time feedback [2, 7]. Research could explore probabilistic graphical models for aggregating weak signals, ensuring phenotypes remain reliable amid evolving EHR noise [13]. In infrastructural terms, future work should focus on federated learning paradigms that enable cross-institution phenotyping without data sharing, mitigating privacy concerns [10, 17].

Ethical integration represents a critical direction, with studies needed to embed fairness metrics directly into weak supervision algorithms, preventing bias amplification in diverse populations [12, 18]. This includes longitudinal evaluations of phenotyping equity in applications like violence risk assessment or cancer prognosis [21, 28].

Deployment-oriented research should investigate closed-loop systems with human-in-the-loop validation, optimizing AI-clinician interactions in ambiguous scenarios [14, 29]. For predictive analytics, prospective trials could assess the impact of weak supervision on outcomes like readmissions or triage efficiency [23, 24].

Finally, governance research directions include standardizing validation protocols for weak supervision models, ensuring translational success from bench to bedside [6, 11]. By pursuing these paths, grounded in the 2017-2021 literature, weak supervision can evolve into a cornerstone of equitable, intelligent healthcare systems.

Conclusion

In conclusion, weak supervision offers a scalable labeling theory for clinical phenotyping under ambiguity in noisy EHRs, fundamentally enhancing AI applications in healthcare systems and analytics. By synthesizing the literature, this review demonstrates how this paradigm transforms heterogeneous data into actionable intelligence, supporting predictive modeling, decision architectures, and closed-loop workflows. Despite challenges in noise management, ethics, and deployment, its potential to foster equitable, efficient care is evident. Ultimately, weak supervision positions AI as an integral component of modern healthcare, bridging data ambiguities to improve patient outcomes.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Apr 2020 Revised: 24 Jun 2020 Accepted: 23 Aug 2020
Published online: 10 January 2021

References

Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for patient phenotyping from electronic health records. Sci Rep. 2020;10(1):1432.
https://doi.org/10.1038/s41598-020-58178-1.

Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020;11(1):3852.
https://doi.org/10.1038/s41467-020-17431-x.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1(1):18.
https://doi.org/10.1038/s41746-018-0029-1.

Landi I, Glicksberg BS, Lee HC, Cherng S, Fenyö D, Danko D, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ Digit Med. 2020;3(1):96.
https://doi.org/10.1038/s41746-020-0301-z.

Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep. 2021;11(1):1980.
https://doi.org/10.1038/s41598-021-81368-4.

Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med. 2019;2(1):69.
https://doi.org/10.1038/s41746-019-0148-3.

Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4.
https://doi.org/10.1038/s41746-020-00367-3.

Goh KH, Wang L, Yeoh FY, Bappal G, Han Z, Li Y, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun. 2021;12(1):711.
https://doi.org/10.1038/s41467-021-20910-4.

Li Y, Nair P, Lu XH, Wen Z, Wang Y, Dehaghi AAK, et al. Inferring multimodal latent topics from electronic health records. Nat Commun. 2020;11(1):2536.
https://doi.org/10.1038/s41467-020-16378-3.

Tucker A, Knapp K, Lunt B, Lo Y, Prieto J, Buchalter D, et al. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med. 2020;3(1):147.
https://doi.org/10.1038/s41746-020-00353-9.

Beaulieu-Jones BK, Greene CS. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med. 2021;4(1):62.
https://doi.org/10.1038/s41746-021-00426-3.

Mhasawade V, Vakil A, Zhao Y, Rehg JM. Machine learning and algorithmic fairness in public and population health. Nat Mach Intell. 2021;3(8):659-66.
https://doi.org/10.1038/s42256-021-00373-4.

Shashikumar SP, Wardi G, Paul P, Heldt T, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don't know”. NPJ Digit Med. 2021;4(1):134.
https://doi.org/10.1038/s41746-021-00504-6.

Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Danziger A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4(1):54.
https://doi.org/10.1038/s41746-021-00423-6.

Thiagarajan JJ, Rajan D, Kato Y, Spanias A. DDxNet: a deep learning model for automatic interpretation of electronic health records, electrocardiograms and electroencephalograms. Sci Rep. 2020;10(1):15040.
https://doi.org/10.1038/s41598-020-73126-9.

Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TWV, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep. 2021;11(1):23017. https://doi.org/10.1038/s41598-021-02481-y.

Gerke S, Yeung S, Cohen IG. Ethical and legal aspects of ambient intelligence in hospitals. JAMA. 2020;323(7):601-2. https://doi.org/10.1001/jama.2019.21699.

Ibrahim SA, Pronovost PJ. Diagnostic errors, health disparities, and artificial intelligence: a combination for health or harm? JAMA Health Forum. 2021;2(9):e212430. https://doi.org/10.1001/jamahealthforum.2021.2430.

Brajer N, Cozzi B, Gao M, Nichols M, Rehel D, Balu S, et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. JAMA Netw Open. 2020;3(2):e1920733. https://doi.org/10.1001/jamanetworkopen.2019.20733.

Yuan Q, Cai T, Hong C, Du M, Johnson BE, Lanuti M, et al. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA Netw Open. 2021;4(7):e2114723. https://doi.org/10.1001/jamanetworkopen.2021.14723.

Menger V, Spruit M, van Est R, Nap M, Scheepers F. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA Netw Open. 2019;2(7):e196709. https://doi.org/10.1001/jamanetworkopen.2019.6709.

Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. JAMA Netw Open. 2020;3(1):e1919396. https://doi.org/10.1001/jamanetworkopen.2019.19396.

Morgan DJ, Pineles L, Owczarzak J, Magder L, Scherer L, Brown JP, et al. Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. JAMA Netw Open. 2019;2(3):e190348. https://doi.org/10.1001/jamanetworkopen.2019.0348.

Goto T, Camargo CA Jr, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning–based prediction of clinical outcomes for children during emergency department triage. JAMA Netw Open. 2019;2(1):e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937.

Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Netw Open. 2018;1(4):e181405. https://doi.org/10.1001/jamanetworkopen.2018.1405.

Desai AN. Artificial intelligence: promise, pitfalls, and perspective. JAMA. 2020;323(24):2448-9. https://doi.org/10.1001/jama.2020.8737.

Safavi KC, Khaniyev T, Copenhaver M, Seely M, Zafar SF, Chbat NW, et al. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. JAMA Netw Open. 2019;2(12):e1917221. https://doi.org/10.1001/jamanetworkopen.2019.17221.

Manz CR, Chen J, Liu M, Chivers C, Regli SH, Braun J, et al. Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. JAMA Oncol. 2020;6(11):1723-30. https://doi.org/10.1001/jamaoncol.2020.4331.

Auerbach A, Bates DW. Discovery, learning, and experimentation with artificial intelligence–based tools at the point of care—perils and opportunity. JAMA Netw Open. 2021;4(4):e215304. https://doi.org/10.1001/jamanetworkopen.2021.5304.