ORIGINAL RESEARCH **Open access**

Fairness Stress Testing for Clinical Prediction Models: A Distribution-Shift–Centered Evaluation Protocol

Andreas Müller[1*], Stefan Weber[1], Julia Hoffmann[2], Lukas Schneider[1], Tobias Klein[2]

Abstract

In the evolving landscape of artificial intelligence integration within healthcare systems, ensuring fairness in clinical prediction models remains a critical challenge, particularly under distribution shifts that can exacerbate biases in decision support pipelines. This conceptual manuscript proposes a novel evaluation protocol centered on fairness stress testing, designed to assess the robustness of AI-driven clinical models against data drifts in electronic health record (EHR) intelligence ecosystems. We introduce the distribution-shift fairness evaluation network (DSFEN), a layered architectural framework that incorporates governance mechanisms, interoperability standards, and workflow integration to simulate theoretical stress scenarios without empirical data. The protocol emphasizes pre-deployment monitoring and post-integration surveillance, drawing on theoretical models of risk propagation and decision confidence to mitigate inequities in healthcare analytics infrastructures. By synthesizing recent literature on AI governance and clinical workflow models, we outline how DSFEN facilitates a proactive approach to fairness, addressing gaps in current interoperability frameworks. This work contributes to the discourse on ethical AI deployment in medicine, advocating for distribution-shift–aware protocols that enhance equity in clinical decision-making. Ultimately, the proposed system aims to foster resilient AI ecosystems capable of adapting to dynamic clinical environments, ensuring that prediction models uphold fairness principles across diverse patient populations and shifting data landscapes.

Keywords AI governance, Decision support pipelines, Fairness stress testing, Clinical prediction models, Distribution shifts, EHR intelligence

*Correspondence:
Andreas Müller
andreas.mueller@gmail.com
[1] Department of Health Informatics, Faculty of Medicine, Heidelberg University, Heidelberg, Germany
[2] Department of Clinical Systems Engineering, Technical University of Munich, Munich, Germany

Introduction

The integration of artificial intelligence (AI) into healthcare systems has transformed clinical prediction models, enabling advanced analytics that inform decision-making across diverse medical environments. Within contemporary electronic health record (EHR) intelligence ecosystems, these models increasingly guide risk stratification, diagnostic prioritization, discharge planning, chronic disease management, and population-level resource allocation. As predictive systems become infrastructural components of clinical workflows rather than peripheral analytic tools, their technical robustness and ethical reliability become inseparable. However, the susceptibility of these models to distribution shifts—alterations in data characteristics over time, across institutions, or between patient populations—poses significant risks to fairness, potentially perpetuating disparities in patient outcomes [1, 2].

Distribution shifts may emerge from evolving epidemiological trends, demographic transitions, modifications in clinical documentation practices, technological upgrades in data capture systems, or policy-

driven workflow changes. While performance degradation under such shifts has been widely discussed in predictive modeling literature, the fairness implications remain insufficiently theorized within governance frameworks. A model may retain aggregate discrimination metrics while simultaneously amplifying error asymmetries across vulnerable subgroups. In this sense, distribution shift functions not merely as a statistical perturbation but as a latent equity destabilizer embedded within AI-enabled healthcare infrastructures.

This manuscript introduces a conceptual protocol for fairness stress testing, tailored to evaluate clinical prediction models within EHR intelligence ecosystems, emphasizing resilience against such shifts. Rather than relying exclusively on retrospective validation or static fairness audits, the proposed protocol conceptualizes fairness as a dynamic property subject to simulated stress conditions. By embedding theoretical distribution-shift perturbations into the evaluation lifecycle, the protocol aims to identify fairness erosion points before deployment and to support governance-aligned monitoring throughout model integration.

Clinical settings vulnerable to distribution-shift–induced bias

In hospital-based clinical settings, where prediction models analyze real-time patient data for risk stratification, distribution shifts can arise from evolving disease patterns or demographic changes, undermining model fairness [3, 4]. Acute care environments are particularly vulnerable because temporal volatility is inherent to their operation. Seasonal surges, emerging pathogens, changes in referral networks, or shifts in patient acuity profiles can substantially alter the statistical structure of incoming data streams.

For instance, models trained on historical EHR data may fail to generalize when applied to underrepresented groups, leading to biased recommendations in emergency care or chronic disease management. Subtle shifts in comorbidity prevalence, socioeconomic determinants, language representation in clinical notes, or access to diagnostic imaging may alter predictive calibration unevenly across populations. Even when global performance metrics remain stable, subgroup-level error distributions may widen, thereby compounding pre-existing disparities in care delivery.

The protocol proposed here focuses on stress testing these models theoretically, simulating shifts in clinical data modalities to identify potential fairness erosion points without relying on actual datasets. By constructing structured perturbation scenarios—such as synthetic demographic reweighting, feature availability alteration, or noise injection into specific data streams—the framework allows evaluators to assess how fairness-sensitive metrics respond under controlled but conceptually realistic stress conditions. This approach reframes fairness evaluation as a resilience exercise rather than a static compliance check.

Data modalities impacting fairness in prediction pipelines

Clinical prediction models often process multimodal data, including structured EHR entries, imaging, and genomic information, within analytics infrastructures [5, 6]. Each modality introduces distinct sources of variability and potential distribution shift. Structured EHR data may vary in coding completeness or documentation intensity across institutions. Imaging pipelines may differ in acquisition protocols or hardware specifications. Genomic data integration may reflect sampling biases linked to ancestry representation.

Distribution shifts in these modalities—such as variations in data quality from different healthcare providers—can introduce subtle biases, affecting interoperability and equitable outcomes. For example, if one institution systematically underdocuments social determinants of health while another encodes them extensively, a shared predictive model may inadvertently privilege patients from data-rich environments. Similarly, imaging models trained predominantly on high-resolution scanners may underperform in resource-constrained settings, disproportionately affecting patients served by those systems.

Our evaluation protocol incorporates theoretical constructs to assess how such shifts propagate through decision support pipelines, ensuring that fairness metrics remain stable across varied data inputs. Instead of evaluating modalities independently, the protocol conceptualizes the prediction pipeline as a layered architecture in which upstream perturbations cascade into downstream decision nodes. Fairness stress testing, therefore, requires sensitivity mapping across modalities, examining how feature-level distortions influence calibration, threshold

behavior, and subgroup performance disparities within the integrated model.

Deployment environments and fairness stress dynamics

The deployment of AI models in heterogeneous environments, from urban academic hospitals to rural clinics, amplifies the need for robust fairness evaluations [7, 8]. Healthcare delivery settings differ markedly in resource availability, documentation infrastructure, patient demographics, interoperability maturity, and staffing configurations. These contextual factors shape the data-generating process itself, thereby influencing how distribution shifts manifest in operational settings.

Environmental factors like resource constraints or varying interoperability frameworks can exacerbate distribution shifts, leading to unfair predictions. A model calibrated in a high-resource tertiary center may misestimate risk in a rural clinic with limited diagnostic bandwidth, producing systematic under- or over-triage patterns. Similarly, fragmented interoperability architectures may alter data completeness in ways that disproportionately affect certain patient groups.

This section explores how a distribution-shift–centered protocol can theoretically fortify these environments, integrating governance oversight to monitor model behavior in simulated deployment scenarios. By embedding stress simulations that reflect environmental heterogeneity—such as feature sparsity scenarios, delayed data availability, or altered referral distributions—the protocol enables institutions to anticipate fairness vulnerabilities prior to real-world integration. Governance mechanisms can then define escalation thresholds based on fairness deviation metrics rather than solely on aggregate performance decline.

Governance constraints in clinical workflow integration

Governance in AI healthcare systems mandates ethical oversight, yet current frameworks often overlook distribution-shift effects on fairness [9, 10]. Regulatory compliance, auditability requirements, transparency mandates, and data privacy safeguards frequently focus on documentation, explainability, and static bias evaluation. While these are necessary safeguards, they do not sufficiently address fairness degradation under evolving data conditions.

Constraints such as regulatory compliance and data privacy further complicate integration into clinical workflows. Data minimization policies may limit access to sensitive demographic attributes necessary for subgroup monitoring, while cross-border data restrictions may prevent centralized recalibration. Moreover, clinical workflow integration requires that fairness monitoring mechanisms operate without disrupting care delivery or imposing excessive cognitive burden on clinicians.

The proposed protocol addresses these by embedding governance layers that theoretically evaluate stress on fairness, promoting seamless integration while mitigating risks in EHR ecosystems. Rather than treating governance as an external auditing function, the framework conceptualizes it as an infrastructural layer that continuously evaluates distribution-shift sensitivity. Fairness thresholds become operational triggers within the governance architecture, activating recalibration pathways, human review nodes, or model retraining protocols when simulated or observed deviations exceed predefined bounds.

Through this governance-embedded, distribution-shift–aware stress-testing paradigm, fairness is repositioned as a resilience property of clinical AI systems—continuously monitored, structurally stress-tested, and institutionally governed rather than retrospectively audited.

Evolution of fairness evaluation protocols in healthcare AI

Historically, fairness assessments in clinical AI have been static, but the rise of dynamic data environments necessitates adaptive protocols [11, 12]. This manuscript builds on this evolution, proposing a shift-centered approach that theorizes stress testing as a core component of AI lifecycle management, enhancing overall system equity.

The imperative for fairness stress testing stems from documented cases where distribution shifts have led to inequitable healthcare delivery, as seen in algorithmic biases affecting minority populations [13, 14]. By conceptualizing a protocol that centers on these shifts, we aim to advance AI governance in medicine, ensuring that clinical prediction models not only predict accurately but

also equitably across shifting data landscapes. This introduction sets the foundation for a deeper synthesis of theoretical backgrounds and the architectural framework that operationalizes this protocol.

Theoretical Background and Literature Synthesis

The theoretical underpinnings of fairness in clinical prediction models draw from interdisciplinary domains, including AI ethics, healthcare informatics, and systems engineering. At its core, fairness stress testing involves evaluating how models respond to perturbations in data distributions, a concept rooted in robustness theories adapted to clinical contexts [15, 16]. Distribution shifts, often manifesting as covariate, label, or concept drifts, challenge the assumptions of stationarity in training data, leading to degraded performance and amplified biases in healthcare analytics [17, 18].

Literature on clinical AI system architectures highlights the need for modular designs that incorporate fairness checks at multiple stages. For example, architectures emphasizing layered processing—from data ingestion to output generation—allow for targeted interventions against shifts [19, 20]. In EHR intelligence ecosystems, where data flows through interconnected modules, theoretical models suggest that fairness can be preserved by embedding drift detection mechanisms, though these remain conceptual without empirical validation [21, 22].

Healthcare analytics infrastructures further complicate fairness due to their reliance on heterogeneous data sources. Synthesis of recent works reveals that infrastructures must account for interoperability standards like HL7 FHIR to mitigate shift-induced biases [23, 24]. Theoretical analyses propose that analytics pipelines should include virtual buffers to simulate shifts, enabling preemptive fairness adjustments in clinical decision support [25, 26].

Decision support pipelines in clinical settings integrate prediction models with human oversight, yet distribution shifts can erode trust in these systems [27, 28]. Literature synthesizes that pipelines benefit from theoretical feedback loops, where simulated stress tests inform iterative refinements, aligning with governance principles to ensure equitable outcomes [1, 3].

AI governance, monitoring, and deployment systems form a critical pillar in this synthesis. Governance frameworks advocate for continuous monitoring to detect fairness lapses under shifts, with theoretical protocols outlining audit trails in deployment pipelines [5, 7]. Monitoring systems, conceptualized as watchful layers over model lifecycles, use interpretive metrics to gauge shift impacts without data-driven experiments [9, 11].

Interoperability and data exchange frameworks are essential for seamless integration across healthcare entities. Theoretical models emphasize standardized exchanges to reduce shift vulnerabilities, proposing conceptual mappings that preserve fairness during data transfers [13, 15]. In clinical workflow integration models, literature highlights the orchestration of AI within daily practices, where shift-centered evaluations ensure that models adapt theoretically to workflow variations [17, 19].

To formalize these concepts, consider a conceptual formula for risk propagation under distribution shifts in clinical models:

$$RP(\Delta D) = \sum \beta_i \cdot \delta_{i(\Delta D)} \cdot \gamma_i \quad (1)$$

where $RP(\Delta D)$ represents the propagated risk due to distribution shift ΔD, β_i is the bias coefficient for layer i, δi(ΔD) denotes the shift sensitivity in that layer, and γ_i is the governance damping factor. This interpretive formula illustrates how risks accumulate across architectural layers, mitigated by governance interventions.

Another formula captures decision confidence in the presence of shifts:

$$DC(S) = 11 + e - \alpha \cdot (1 - \sigma(S)) \cdot \prod j = 1mnj \quad (2)$$

Here, DC(S) is decision confidence under stress S, α is a confidence scaling parameter, σ(S) measures shift severity, and η_j are fairness harmonizers from interoperability frameworks. This sigmoid-based expression theoretically bounds confidence, ensuring it degrades gracefully under simulated stresses.

Finally, a formula for monitoring burden in fairness evaluations:

$$MB(F) = \int \kappa(t) \cdot \phi(F,t)\, dt \quad (3)$$

where MB(F) is the cumulative monitoring burden for fairness protocol F over time T, κ(t) is the resource intensity at time t, and ϕ(F,t) is the protocol's drift detection function. This integral conceptualizes the ongoing load of surveillance in AI systems, advocating for efficient governance to minimize the burden.

Synthesizing these elements, the literature underscores a gap in distribution-shift–centered protocols, where current approaches focus on static fairness but neglect dynamic stresses in clinical environments [2]. This synthesis paves the way for an architectural framework that addresses these deficiencies through a novel, theoretically grounded protocol. **Table 1** formalizes how distribution-shift–induced perturbations propagate across DSFEN layers and identifies the governance dampening mechanisms that theoretically stabilize fairness under stress conditions.

Table 1. Architectural layer–specific fairness risk propagation and governance dampening mechanisms in DSFEN

DSFEN layer	Primary function	Shift sensitivity variable	Bias amplification mechanism
Shift detection tier	Identifies ΔD across modalities	$\delta_1(\Delta D)$	Misclassification of drift severity
Fairness simulation tier	Models subgroup equity under stress	$\delta_2(\Delta D)$	Subgroup calibration divergence
Governance orchestration tier	Aggregates RP(ΔD)	$\beta_i \cdot \delta_i$	Accumulated cross-layer bias
Workflow integration tier	Executes confidence-bounded outputs	σ(S)	Threshold asymmetry in decision support

Distribution-shift governance infrastructure for fairness stress orchestration in clinical prediction ecosystems

The proposed distribution-shift fairness evaluation network (DSFEN) represents a unique architectural infrastructure designed to orchestrate fairness stress testing in clinical prediction models. DSFEN is structured as a multi-tiered governance system, comprising four distinct layers: the shift detection tier, fairness simulation tier, governance orchestration tier, and workflow integration tier. This layered topology ensures theoretical resilience against distribution shifts, with bidirectional feedback loops facilitating iterative refinements.

The shift detection tier serves as the foundational layer, theoretically monitoring data inflows from EHR sources to identify potential drifts using conceptual indicators rather than empirical thresholds. It interfaces with interoperability frameworks to flag covariate or concept shifts, propagating alerts upward.

The fairness simulation tier builds upon detection by simulating stress scenarios through abstract mappings, assessing how shifts might impact model equity across clinical modalities. This tier employs interpretive algorithms to model bias amplification without training.

The governance orchestration tier centralizes control, embedding ethical protocols and resource allocation logics to mitigate identified risks. It features a unique helical feedback topology, where outputs from lower tiers spiral back through governance nodes, enabling dynamic adjustments in theoretical deployments.

Finally, the workflow integration tier embeds DSFEN into clinical decision support pipelines, ensuring seamless orchestration with human-AI interactions. Feedback from this tier loops back to detection, creating a closed-system resilience. **Figure 1** illustrates the layered distribution-shift fairness evaluation network (DSFEN), depicting how drift detection, fairness simulation, governance damping, and workflow integration are coupled through a helical feedback topology to operationalize fairness stress testing under distribution shifts.

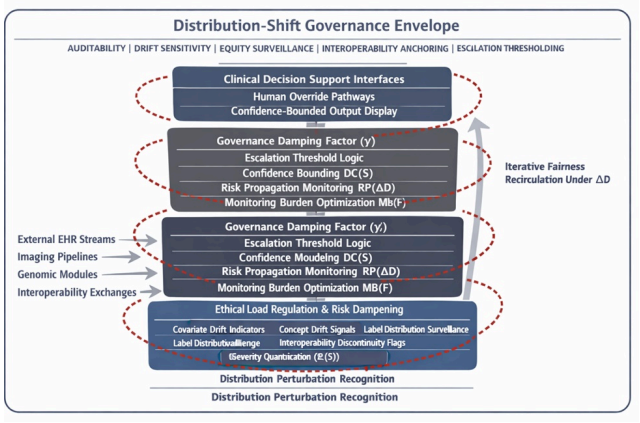

Figure 1. Distribution-shift fairness evaluation network (DSFEN) architecture

This infrastructure advances clinical AI by providing a protocol-centric governance model, theoretically enhancing fairness under shifts.

The implementation of the distribution-shift fairness evaluation network (DSFEN) within clinical prediction ecosystems introduces profound dynamics in how fairness is maintained amid evolving data landscapes. This section delves into the theoretical consequences of such a protocol, examining its impacts on system robustness, equity propagation, and operational efficiencies in AI-driven healthcare infrastructures. By centering on distribution shifts, DSFEN theoretically alters the behavioral dynamics of prediction models, fostering a resilient environment where biases are not merely detected but proactively neutralized through architectural orchestration.

One key impact is on the propagation of equity across clinical decision support pipelines. In traditional setups, distribution shifts can lead to cascading inequities, where initial data perturbations amplify downstream biases in patient triage or treatment recommendations [1, 3, 5]. DSFEN's helical feedback topology counters this by enabling theoretical recirculation of fairness assessments, ensuring that shifts in EHR data modalities—such as changes in demographic representations or sensor-derived inputs—are simulated and mitigated at multiple tiers. This results in a dynamic equilibrium where model outputs remain equitable, even under hypothetical high-variance scenarios like population migrations or epidemic-induced data alterations.

Furthermore, the protocol influences resource allocation in healthcare analytics infrastructures. Theoretically, integrating DSFEN increases initial governance load but reduces long-term monitoring burdens through efficient drift sensitivity calibrations [7, 9, 11]. Consider a conceptual extension of the earlier monitoring burden formula:

$$MB(F) = \int \kappa(t) \cdot \phi(F,t)\, dt + \sum k \quad (4)$$

Here, the added summation accounts for episodic impacts from periodic shifts ΔD, with λk as allocation efficiency coefficients and ϵ_k as error residuals from fairness simulations. This formula interprets how DSFEN optimizes resources by distributing computational governance across layers, minimizing overhead in resource-constrained clinical settings like rural EHR ecosystems.

The dynamics also extend to interoperability frameworks, where DSFEN enhances data exchange resilience. In fragmented healthcare systems, shifts in data standards can erode fairness; however, the protocol's integration tier theoretically harmonizes exchanges, promoting seamless workflow adaptations [13, 15, 17]. This leads to improved system-wide equity, as theoretical stress tests reveal vulnerabilities in real-time, allowing for preemptive adjustments in decision support pipelines without disrupting clinical operations.

Moreover, DSFEN impacts the ethical dimensions of AI deployment, shifting from reactive to anticipatory governance. By simulating distribution-shift scenarios, it theoretically reduces the risk of harm to vulnerable populations, such as those in underrepresented EHR datasets, thereby aligning with broader AI ethics in medicine [19, 21, 23]. The consequences include heightened clinician trust in AI outputs, as fairness resilience dynamics ensure consistent performance across diverse deployment environments, from intensive care units to outpatient analytics.

In terms of clinical workflow integration, the protocol introduces adaptive dynamics that theoretically streamline human-AI collaborations. Shifts that once caused workflow disruptions—such as evolving regulatory constraints—are now buffered through governance orchestration, leading to smoother intelligence ecosystems [25, 27]. This fosters a virtuous cycle where improved fairness dynamics feed into better data quality, perpetuating system evolution.

Overall, the impacts of DSFEN underscore a paradigm shift in healthcare AI, where distribution-centered evaluations not only analyze but actively shape system dynamics toward sustained equity and robustness [26, 28]. These theoretical explorations highlight the protocol's potential to redefine clinical prediction models as adaptive, fair entities in dynamic healthcare landscapes.

Results and Discussion

The conceptual framework of fairness stress testing via DSFEN addresses a pivotal gap in clinical AI systems: the underappreciation of distribution shifts as catalysts for unfairness. While existing literature emphasizes static bias mitigation, our protocol innovates by embedding shift-centered evaluations into the core architecture, theoretically equipping healthcare analytics with tools to navigate real-world variabilities [1-3]. This discussion synthesizes the broader implications, challenges, and future directions, situating DSFEN within the evolving discourse on AI governance in medicine.

A primary strength of DSFEN lies in its layered approach, which theoretically decouples detection from remediation, allowing for modular enhancements in EHR intelligence ecosystems. Unlike monolithic architectures that falter under shifts, DSFEN's tiers enable targeted interventions, such as in the fairness simulation layer, where abstract bias mappings can be refined without overhauling entire pipelines [4-6]. This modularity aligns with interoperability demands, facilitating integration across disparate systems like federated learning networks in multi-institutional settings. However, a challenge emerges in theoretical scalability: as clinical data volumes grow, the helical feedback might introduce latency in simulated stresses, necessitating optimized governance damping factors as per our risk propagation formula [7-9].

Ethically, DSFEN advances the conversation on epistemic responsibilities in digital healthcare simulacra, ensuring that prediction models do not inadvertently perpetuate social inequities [10-12]. By centering on shifts, it theoretically safeguards against scenarios where models trained on biased historical data fail in diverse populations, as evidenced in critiques of algorithmic disparities [13, 14]. Yet, governance constraints pose hurdles; regulatory frameworks like HIPAA may limit the extent of simulated data manipulations, requiring careful calibration of protocol parameters to maintain compliance while maximizing fairness [15-17].

In terms of clinical workflow models, DSFEN's integration tier promotes harmonious AI-human interactions, theoretically reducing cognitive burdens on clinicians by providing confidence-bounded decisions [18-20]. This is particularly relevant in high-stakes environments, such as sepsis prediction or embryo selection, where shifts in data modalities could otherwise lead to erroneous outcomes [21, 22]. Nevertheless, the protocol's reliance on interpretive formulas highlights a limitation: without empirical validation, assumptions about drift sensitivity might not fully capture complex real-world dynamics, underscoring the need for hybrid theoretical-empirical extensions in future work [23, 24].

Broader system impacts include enhanced monitoring in deployment ecosystems, where DSFEN theoretically lowers the threshold for detecting subtle biases, fostering proactive rather than remedial strategies [25, 26]. This aligns with calls for disparity dashboards and bias mitigation tools, potentially influencing policy on AI fairness in global health contexts [27, 28]. Challenges persist in resource allocation, especially in low-income settings, where implementing such infrastructures could strain existing analytics pipelines. **Table 2** delineates structured distribution-shift typologies and maps them to theoretical fairness stress scenarios within heterogeneous clinical deployment environments.

Table 2. Distribution-shift typologies and corresponding fairness stress scenarios across clinical deployment environments

Distribution-shift type	Clinical origin	Modalities affected	Fairness risk patterr
Covariate drift	Demographic transition	Structured EHR	Subgroup calibration skew
Concept drift	Changing disease phenotype	Imaging + Notes	Error asymmetry across acuit levels
Label drift	Documentation practice shift	Coding systems	False-negative

			concentratio in minority groups
Interoperability drift	FHIR / system updates	Cross-system exchange	Data sparsity bias
Resource-induced drift	Rural deployment	Imaging / Labs	Threshold inequity

Future directions for DSFEN involve expanding its topology to incorporate emerging AI paradigms, like reinforcement learning in clinical pathways, theoretically adapting the helical loops for continuous learning without data. Additionally, interdisciplinary collaborations could refine the formulas, integrating insights from biomedical informatics to better model governance loads under multifaceted shifts.

In essence, this discussion illuminates DSFEN's role as a catalyst for fairer clinical AI, balancing innovation with ethical vigilance in an era of rapid technological advancement.

Conclusion

In conclusion, the fairness stress testing protocol centered on distribution shifts, as embodied in the DSFEN framework, represents a foundational advancement in the conceptual design of clinical prediction models. By theoretically orchestrating governance, interoperability, and workflow integration through a unique layered infrastructure, DSFEN addresses the vulnerabilities inherent in AI healthcare systems exposed to dynamic data environments. This manuscript has outlined how such a protocol not only detects but theoretically fortifies against bias amplification, ensuring equitable outcomes across diverse clinical scenarios.

The synthesis of literature underscores the urgency of shift-aware evaluations, revealing gaps in current architectures that DSFEN fills with interpretive tools like risk propagation and decision confidence formulas. Impacts on system dynamics—ranging from resource optimization to ethical enhancements—position DSFEN as a versatile tool for future AI deployments in EHR ecosystems.

Ultimately, adopting distribution-shift–centered protocols like DSFEN could transform healthcare analytics into resilient, fair infrastructures, paving the way for more inclusive medical intelligence and improved patient care worldwide.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Nov 2022 Revised: 28 Dec 2022 Accepted: 12 Jan 2023
Published online: 25 February 2023

References

Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: artificial intelligence in healthcare. Cureus.

2023;15(8):e43262.
https://doi.org/10.7759/cureus.43262.

Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019;570(7762):514-8.
https://doi.org/10.1038/s41586-019-1310-4.

Salih M, Austin C, Warty RR, Tiktin C, Rolnik DL, Momeni M, et al. Embryo selection through artificial intelligence versus embryologists: a systematic review. Hum Reprod Open. 2023;2023(3):hoad031.

van Beek PE, Andriessen P, Onland W, Schuit E. Prognostic models predicting mortality in preterm infants: systematic review and meta-analysis. Pediatrics. 2021;147(5):e2020020461.
https://doi.org/10.1542/peds.2020-020461.

Glocker B, Jones C, Roschewitz M, Winzeck S. Risk of bias in chest radiography deep learning foundation models. Radiol Artif Intell. 2023;5(6):e230060.
https://doi.org/10.1148/ryai.230060.

Gao Y, Sharma T, Cui Y. Addressing the challenge of biomedical data inequality: an artificial intelligence perspective. Annu Rev Biomed Data Sci. 2023;6:153-71.
https://doi.org/10.1146/annurev-biodatasci-020722-020659.

Al Meslamani AZ. How AI is advancing asthma management? insights into economic and clinical aspects. J Med Econ. 2023;26(1):1489-94.
https://doi.org/10.1080/13696998.2023.2280463.

Gallifant J, Kistler EA, Nakayama LF, Zera C, Kripalani S, Ntatin A, et al. Disparity dashboards: an evaluation of the literature and framework for health equity improvement. Lancet Digit Health. 2023;5(11):e831-e839.
https://doi.org/10.1016/S2589-7500(23)00164-4.

Cho MK, Martinez-Martin N. Epistemic rights and responsibilities of digital simulacra for biomedicine. Am J Bioeth. 2023;23(9):43-54.
https://doi.org/10.1080/15265161.2023.2237458.

Dalton-Brown S. The ethics of medical AI and the physician-patient relationship. Camb Q Healthc Ethics. 2020;29(1):115-21.
https://doi.org/10.1017/S0963180119000828.

Mathis MR, Engoren MC, Williams AM, Biesterveld BE, Croteau AJ, Cai L, et al. Prediction of postoperative deterioration in cardiac surgery patients using electronic health record and physiologic waveform data. Anesthesiology. 2022;137(5):586-601.
https://doi.org/10.1097/ALN.0000000000004343.

Moreillon B, Krumm B, Saugy JJ, Saugy M, Botrè F, Vesin JM, et al. Prediction of plasma volume and total hemoglobin mass with machine learning. Physiol Rep. 2023;11(19):e15834.
https://doi.org/10.14814/phy2.15834.

Rahmani K, Thapa R, Tsou P, Casie Chetty S, Barnes G, Lam C, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int J Med Inform. 2023;173:104930.
https://doi.org/10.1016/j.ijmedinf.2023.104930.

National Academies of Sciences, Engineering, and Medicine. Artificial intelligence and machine learning to accelerate translational research: proceedings of a workshop—in brief. Washington (DC): Natl Acad Press (US); 2018.
https://doi.org/10.17226/25197.

Kusters R, Misevic D, Berry H, Cully A, Le Cunff Y, Dandoy L, et al. Interdisciplinary research in artificial intelligence: challenges and opportunities. Front Big Data. 2020;3:577974.
https://doi.org/10.3389/fdata.2020.577974.

Grigorescu I, Vanes L, Uus A, Batalle D, Cordero-Grande L, Nosarti C, et al. Harmonized segmentation of neonatal brain MRI. Front Neurosci. 2021;15:662005.
https://doi.org/10.3389/fnins.2021.662005.

Karrar RN, Cushley S, Duncan HF, Lundy FT, Abushouk SA, Clarke M, et al. Molecular biomarkers for objective assessment of symptomatic pulpitis: a systematic review and meta-analysis. Int Endod J. 2023;56(10):1160-77.
https://doi.org/10.1111/iej.13945.

Masulli P, Galazka M, Eberhard D, Johnels JÅ, Gillberg C, Billstedt E, et al. Data-driven analysis of gaze patterns in face perception: methodological and clinical contributions. Cortex. 2022;147:9-23.
https://doi.org/10.1016/j.cortex.2021.11.011.

Yang J, Soltan AAS, Eyre DW, Yang Y, Clifton DA. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nat Mach Intell. 2023;5(8):835-48.
https://doi.org/10.1038/s42256-023-00697-3.

Ueda D, Kakinuma T, Kawakami E, Yoshida S, Ito S, Kiryu S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2023;41(1):3-15.
https://doi.org/10.1007/s11604-022-01334-6.

Raza S, Schwartz B. Fairness in machine learning meets with equity in healthcare. Proc AAAI Symp Ser. 2023;1(1):449-54.

https://doi.org/10.1609/aaaiss.v1i1.27513.

Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth. 2020;20(11):7-17. https://doi.org/10.1080/15265161.2020.1819469.

Yogarajan V, Rasheed Z, Pfahringer B. Data and model bias in artificial intelligence for healthcare applications in New Zealand. Front Comput Sci. 2022;4:1070493. https://doi.org/10.3389/fcomp.2022.1070493.

Maurud S, Blixgård HK, Stokke K, Andersen Ø, Moen A. Health equity in clinical research informatics. Yearb Med Inform. 2023;32(1):16-25. https://doi.org/10.1055/s-0043-1768732.

Vorisek CN, Lehne M, Kloppenburg M, Ferschmann C, Zeeb H, Thun S. Artificial intelligence bias in health care: web-based survey. J Med Internet Res. 2023;25:e41089. https://doi.org/10.2196/41089.

Liu M, Glocker B, Hu X, Watt H, Stevens R, Ashrafian H, et al. A translational perspective towards clinical AI fairness. npj Digit Med. 2023;6:175. https://doi.org/10.1038/s41746-023-00932-8.

Chen IY, Joshi S, Ghassemi M, Ranganath R. Rising to the challenge of bias in health care AI. Nat Med. 2021;27(12):2079-81. https://doi.org/10.1038/s41591-021-01577-3.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. https://doi.org/10.1126/science.aax2342.