Viewpoint

Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support

Carlos Sáez, PhD; Pablo Ferri, PhD; Juan M García-Gómez, PhD

Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Valencia, Spain

Corresponding Author:
Carlos Sáez, PhD
Biomedical Data Science Lab
Instituto Universitario de Tecnologías de la Información y Comunicaciones
Universitat Politècnica de València
Camino de Vera s/n
Valencia, 46022
Spain
Phone: 34 963877000 ext 85247
Email: carsaesi@upv.es

Abstract

Artificial intelligence (AI)–based clinical decision support systems are gaining momentum by relying on a greater volume and variety of secondary use data. However, the uncertainty, variability, and biases in real-world data environments still pose significant challenges to the development of health AI, its routine clinical use, and its regulatory frameworks. Health AI should be resilient against real-world environments throughout its lifecycle, including the training and prediction phases and maintenance during production, and health AI regulations should evolve accordingly. Data quality issues, variability over time or across sites, information uncertainty, human-computer interaction, and fundamental rights assurance are among the most relevant challenges. If health AI is not designed resiliently with regard to these real-world data effects, potentially biased data-driven medical decisions can risk the safety and fundamental rights of millions of people. In this viewpoint, we review the challenges, requirements, and methods for resilient AI in health and provide a research framework to improve the trustworthiness of next-generation AI-based clinical decision support.

(J Med Internet Res 2024;26:e50295) doi: 10.2196/50295

KEYWORDS

artificial intelligence; clinical decision support; resilience; clinical medicine; machine learning; data quality; fairness; trustworthy AI; regulation; AI regulation; AI Act; EHDS; European Health Data Space; emergency medical dispatch; clinical decision support systems

Introduction

The advent of electronic health record (EHR) data sharing posed expectations for improved, trustworthy development of artificial intelligence (AI) in health through a larger volume and variety of data. However, the development process of health AI and the generalization and fairness of the resulting models face significant challenges due to the inherent biases, uncertainty, variability, and quality levels of real-world data (RWD). These challenges include variable information across different settings and over time, biases affecting underrepresented groups, uncertainty from lacking or overlapping information, or data quality (DQ) issues such as incomplete or implausible information. These issues can be present in training data feeding the AI learning or manifest de novo in the AI routine clinical use (Textbox 1). If health AI is not designed resiliently with regard to these RWD effects, potentially biased data-driven medical decisions can risk the safety and fundamental rights of millions of people.

Textbox 1. Clinical case with potentially suboptimal clinical decision support outcomes in a high-risk artificial intelligence (AI) system according to the European Union (EU) AI Act—Article 6(2)—Annex III.

A medical emergency dispatch center receives a call from a professor informing about an aged 20 years female student showing apparent respiratory distress. After data input, with no known chronic respiratory disease reported, an AI triage support system confirms that the case is not life-threatening. They send a basic life support ambulance. Eventually, the patient died during transport. The autopsy revealed a pulmonary embolism: she had recently started taking oral contraceptives. Clearly, this lack of information affected the AI outcome. Should AI have reacted by warning about potentially high uncertainty or asked for more information?

Alternatively, without previous embolism, a similar case might have occurred in March 2020 as a then-unknown effect of the SARS-CoV-2 infection. Instead of using a static AI decision support system trained on prepandemic data, should AI have automatically adapted to the very recent data patterns to provide better outcomes on new data?

Many recent papers in the highest-impact medical journals warned about the increasing obstacles imposed by RWD challenges for health AI. Cabitza et al [1] warned about the effects of intrinsic uncertainty in medical observations and interpretations on the reliability and accuracy of machine learning (ML). They claimed to develop and validate ML adaptable to the variable nature and actual DQ of medical information. Chen and Asch [2] argued that just relying on past data can have diminishing effects on AI's usefulness and future generalization, as well as the "butterfly effect" of tiny current variations into the future. Gianfrancesco et al [3] outlined the potential contribution to socioeconomic health disparities of ML based on biased data. Rajkomar et al [4] classified the availability of high-quality data and learning from undesirable past practices among the key challenges for medical ML, especially in RWD and nonimaging data from EHR. Google Health and DeepMind teams identified data variability as among the key challenges in delivering clinical impact with AI [5]. Besides, the COVID-19 pandemic highlighted potential limitations in medical AI from inadequate training and evaluation design and DQ and variability issues [6,7]. Not surprisingly, the European Commission has recognized DQ as a risk for the safety and fundamental rights assurance (FRA) in AI and included it along with other RWD challenges within the recently approved AI regulation (Article 10) [8,9] and among the significant issues for the quality and economy of the European Health Data Space (EHDS) [10-13].

We focus on ML as the methodological driver for current well-established health AI and clinical decision support systems (CDSSs) across diverse clinical fields. In health ML, new knowledge is learned computationally from data in a training phase, generally from dedicated data cohorts or the EHR. This knowledge is then used to assist decision-making for new cases in a prediction phase. At prediction, a CDSS can retrieve its inputs from the EHR or manual input. Potential data-related uncertainties, variability, and biases for health AI can arise both at the training and prediction phases (eg, see Table S1 in Multimedia Appendix 1 for some examples).

The definitions we provide in this work apply to multiple traditional and state-of-the-art ML approaches for CDSS, including predictive analytics and the potential upcoming use of conversational AI [14]. The related techniques include, but are not limited to, deep learning, ensemble models, and statistical methods, with knowledge acquisition schemes based on learning from scratch and using pretrained or foundation models [15,16].

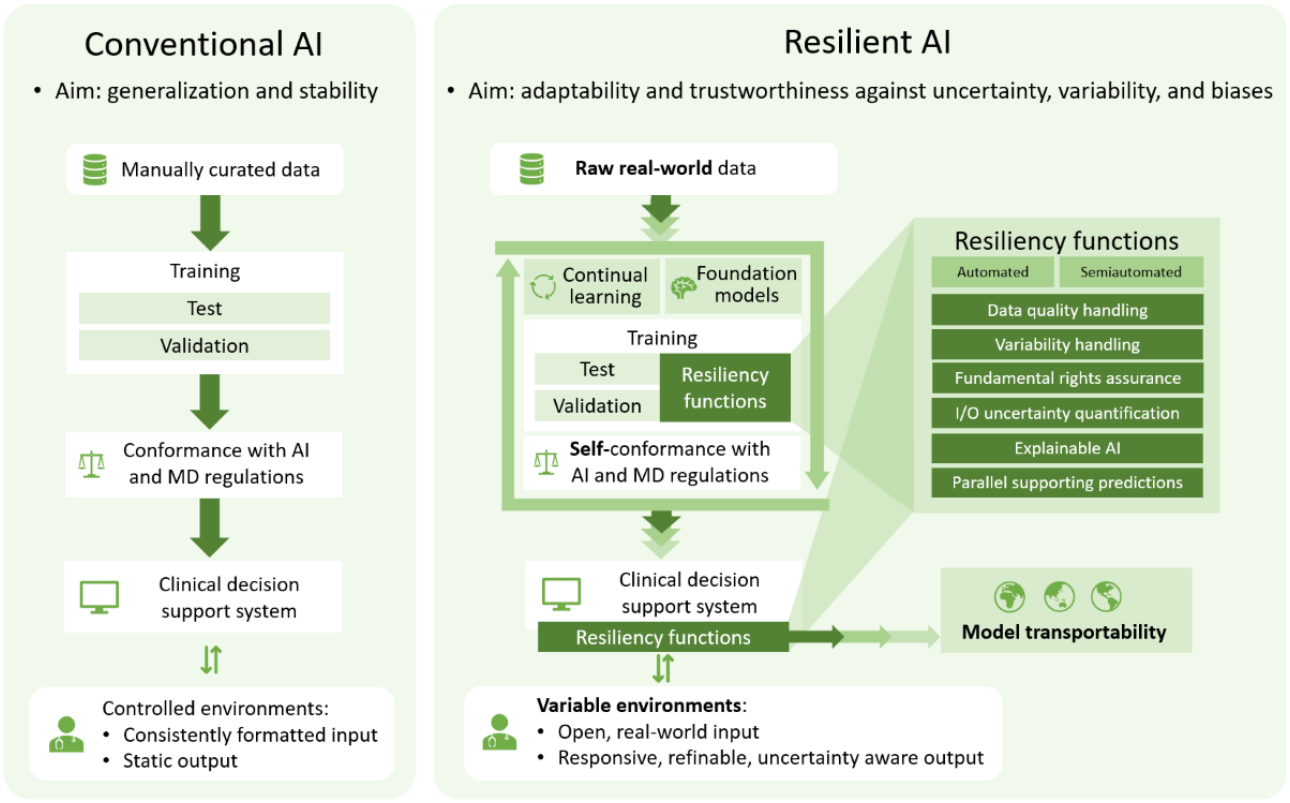
Resilient AI: Definition and Requirements

Definition

We define resilient AI (RAI) as an AI that can automatically or semiautomatically adapt and react to unprecedented, uncertain, or undesired situations in real-world environments during model training and its use.

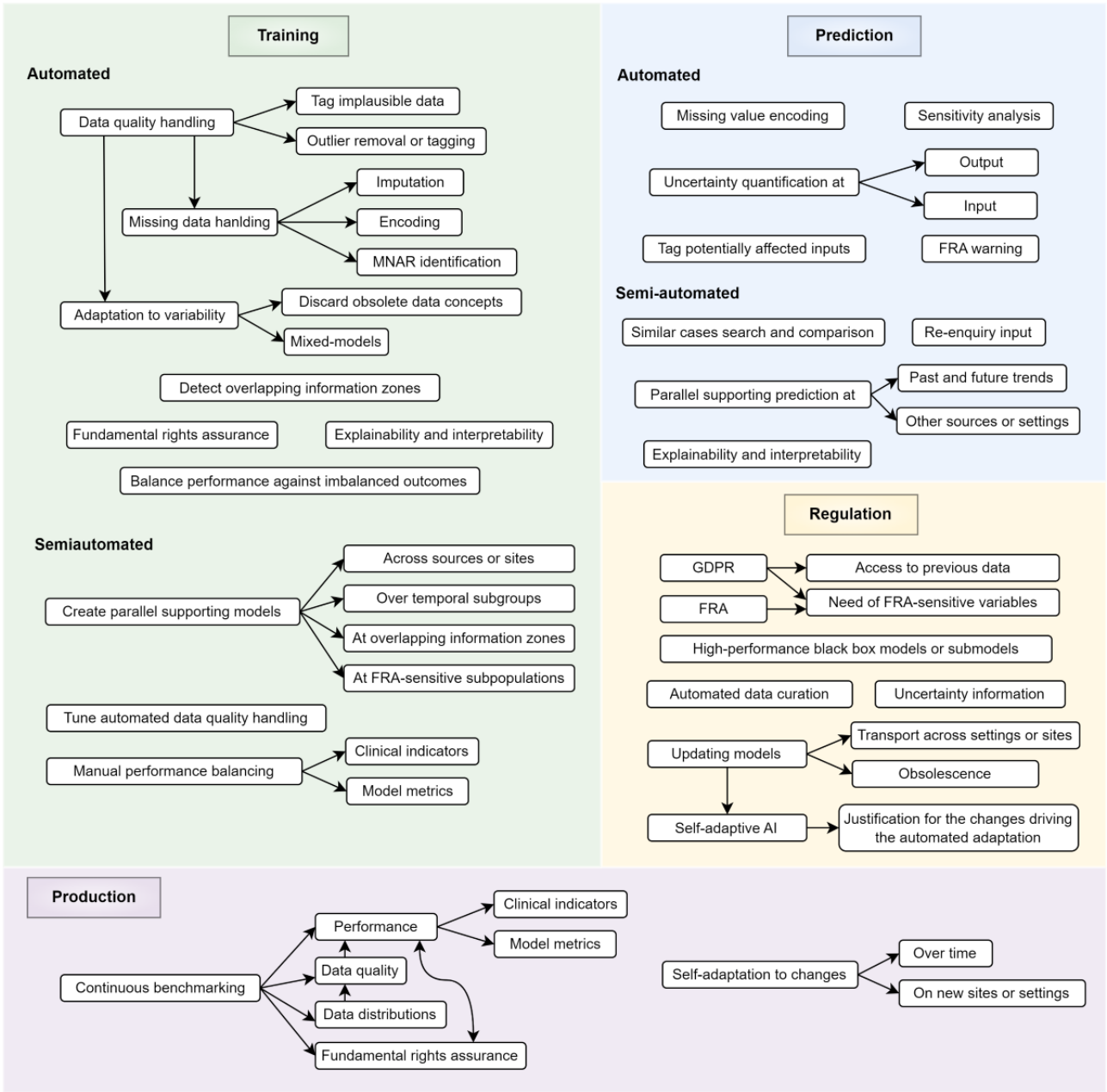
RAI emerges as a new paradigm over conventional AI procedures, generally aimed at model generalization and stability in controlled environments. The conventional AI approach expects learning on preprocessed, curated datasets and then inferring on equivalently consistent input data. The RAI approach, however, aims to learn and predict raw RWD by relying on resiliency-enabling methodologies and functions while improving its generalization and stability in variable environments. In RAI, adaptation would involve using new information to change or update AI knowledge, and reaction would involve providing appropriate information to support decision-making. Overall, RAI aims to enable trustworthy CDSSs for real-world environments, from adapting and explaining against potential biases and variability in the secondary use of EHR to handling uncertainty in decision-making, covering the whole AI lifecycle (Figure 1).

Figure 1. Comparison of the conventional AI process and the proposed resilient AI paradigm. Conventional AI is generally aimed at generalization and stability in controlled environments. That is, typically learning and validating a static model from a standardized, quality-controlled dataset, and applying the inference on equivalently consistent inputs. Resilient AI, however, aims to handle real-world variable environments during all its lifecycle, from adapting and explaining against potential biases and variability in the secondary use of electronic health records, both at model training and during use, to handling uncertainty in decision-making, enabling next-generation, trustworthy clinical decision support systems. AI: artificial intelligence, MD: medical device, I/O: input/output.



The challenges for RAI include DQ issues, data variability over time or across sources, information uncertainty, human-computer interaction, security and privacy, and FRA, among others. By wrapping the desired behavior of health AI against these challenges, we propose a set of requirements for the behavior and functionality of RAI in health (Figure 2). These requirements are organized in the 4 blocks described next.

Figure 2. A classification for the expected behavior and functionality of resilient AI in health, regarding the AI training phase, the prediction in new cases, the routine maintenance of AI during production, and related regulatory aspects. We focus on supervised ML as the methodological driver for current well-established medical AI and clinical decision support systems. AI: artificial intelligence; FRA: fundamental rights assurance; GDPR: General Data Protection Regulation; ML: machine learning; MNAR: missing not-at-random.



Training

RAI should ensure quality, generalizable knowledge through automated or semiautomated data preparation and lifelong learning processes cost-effectively. Manual data preparation generally consumes 80% of the costs of AI projects, and many data issues neither generally addressed nor detected in apparently error-free data can potentially lead to biased or inaccurate models.

Most data-related barriers in training are categorized as DQ issues [17-19], including missing, implausible, or outlier data. RAI should tag DQ issues and allow models to discard or fix them, infer any useful knowledge associated with faulty inputs, or suggest multiple modelling pipelines. For example, repeated missing laboratory tests at emergency admission in patients with COVID-19 due to rapid intensive care unit admissions might lead to a missing not-at-random situation associated with higher severity in the outcome.

RAI training should be resilient against variability over time and across settings or sites. Changing or unexpected information patterns can present as shifts in the covariates—*p(x)*—prior probability shifts in the outcomes—*p(y)*—or as concept shifts in one conditioned to the other—*p(y|x) or p(x|y)*—leading to dataset shifts [20-22]. For example, would prediction models for acute respiratory distress syndrome trained on the first COVID-19 wave EHRs perform on new cases as initially evaluated? Would a model trained on a global European sample perform equally cross-border? These shifts can even relate to FRA-sensitive variables such as gender or race [10]. To delineate and characterize variability is critical, as it is to

optimize model performance according to the population where the model will be used. Therefore, variability-resilient AI training should apply both at the initial dataset learning and during its prospective production cycle, as described in this paper.

Similarly, ML responds suboptimally at overlapping information zones, that is, where repeated similar cases show distinct outcomes, bounding the Bayes error, the minimum and irreducible error rate achievable by a model. RAI should help delineate these zones during training, enabling reviewing, skipping, or fixing these cases or suggesting complementary models trying to optimize prediction at these zones.

RAI training should account for user requirements regarding the expected performance, including cost-benefits for each miss-correct classification, targets for sensitivity, specificity, or positive predictive value, and their trade-offs. These estimations are key in unbalanced outcomes and against FRA-unbalanced subpopulations. Lastly, training should enable prospective explainability and interpretability in the prediction phase.

Prediction

Directly related to clinical decision-making, resilience becomes critical in prediction. Balancing clinical usefulness with time-efficiency—with minimum intrusiveness—RAI should encompass fully automated support with smart human-AI interaction: a fully automated RAI will act without additional user feedback, while a semiautomated RAI will help improve or investigate the results through additional feedback.

Determinant for the user's adherence to the AI decision processes, RAI predictions should be explainable and interpretable [23,24]. Interpretability means understanding in human terms why a model provides an output and how it could vary from changes in the input—for example, through a sensitivity analysis. Explainability means understanding the internal mechanics of a model driving causality in a specific decision-making process.

Uncertainty is intrinsic to decision-making. Therefore, informing and quantifying uncertainty in health AI would significantly increase the system's trustworthiness and confidence in derived decisions. Particularly relevant are input cases showing unprecedented data patterns. RAI outcomes should count, where relevant, with CIs or levels. Likewise, RAI should handle potential input issues by identifying DQ-affected inputs, smartly encoding missing data, and identifying lacking optional information potentially relevant to the prediction.

Toward CDSS transportability, endurance, and fairness, RAI should enable the comparison of predictions obtained at multisite consensus or local level, identify potential changes in the outcome over time, or check for differences at distinct FRA-related subpopulations. Of note is that RAI should compare our case with similar past cases with validated outcomes, particularly beneficial in rare cases, and warn about potential FRA violations.

Production

Once an AI is in routine clinical use, changes can occur in the data acquisition contexts, the target populations of the model, or in clinical knowledge and user requirements. The dataset shifts described before in the training block can still occur in production, either over time or using the AI in different settings from those we initially trained it in. To avoid unexpected biases and obsolescence, we must continuously benchmark AI performance on target populations and enable automated or semiautomated self-adaptation mechanisms to these changes.

Besides benchmarking model performance in clinical indicators—for example, number of hospital readmissions in readmission prediction—or confusion matrix-derived metrics—for example, sensitivity and specificity—an RAI in production should monitor dataset shifts, user-defined DQ rules, user requirements—for example, deprecated or new International Classificafion of Diseases codes—and continuously benchmarking FRA.

On any significant change in AI performance, or even in advance of them, RAI should be able to self-adapt its knowledge over time or when transporting a model to other settings. Further, changes in requirements and clinical knowledge should be easily incorporated into previously built models without requiring a whole reengineering process.

Regulation

Production health AI and CDSS require regulations for medical device products, such as the European Union (EU) Medical Device Regulation [25], and specific regulations for AI, including the European Regulatory Framework for AI or AI Act [8,9,11]. The US Food and Drug Administration provides the AI and ML in Medical Devices white paper and its AI and ML Software as a Medical Device Action Plan [26,27].

The flexibility of current regulations on putting trustworthiness and patient safety as the top priority accommodates most of the described resilient AI needs. However, the wide RAI casuistry, involving more system self-decisions and higher levels of interaction and data access, might deserve more detailed definitions.

To promote AI while addressing potential risks, the EU AI Act includes a list of prohibited AI practices, rules for high-risk AI systems, transparency obligations, liability rules, and an innovation support legal framework, all consistent with the EU General Data Protection Regulation (GDPR). Specifically, the EU AI Act Article 15—Accuracy, Robustness, and Cybersecurity—Section 4 states that high-risk AI systems shall be as resilient as possible regarding errors, faults, or inconsistencies that may occur within the system or the environment in which the system operates, in particular, due to their interaction with natural persons or other systems. How to implement these resilience features is left to the system designers. At the simplest level, RAI could stop or warn users of unexpected situations, such as potential uncertainties or FRA risks at outputs or inputs. How this information is presented and interacted with, since it potentially affects decision-making, might be the subject of more specific regulations. Further,

Article 15, Section 5, also claims resilience against attempts to exploit system vulnerabilities.

However, in addition to during system operation, issues can occur as well at initial model training and in prospective retraining, where alternative resilience features could also take place. In this regard, Article 10—data and data governance—claims that data governance and management practices can be used to mitigate these data-related biases. However, fully resilient AI should, to some extent, make self-decisions for automated data curation or FAIRification procedures, which could be applied both during training and system operation. These may require stricter regulation and a justification supported by the added value of this process.

An innovative regulatory aspect for RAI focuses on self-adaptive AI, which automatically updates its knowledge to maintain performance and avoid obsolescence during production or, similarly, when transporting the AI to other settings. RAI self-updates should avoid, to some extent, new conformance evaluations. Self-adaptive RAI might also need data access regulations for benchmarking and retraining AI while respecting data security and privacy, such as in compliance with the EU GDPR [28]. Besides, accessing the EHRs for similar cases search and comparison, and the need for sensitive, FRA-related information potentially improving the system performance should also be regulated.

Lastly, for black-box models significantly outperforming other interpretable solutions, regulations should consider mitigation through resilient explainable and FRA features rather than limiting the AI use indications [29,30].

Available Methods and Solutions

Overview

Next, we describe a collection of currently available methods and solutions with a feasible while relevant use to address to some extent the RAI objectives and resiliency functions defined in the previous section.

Data Quality

The assurance of DQ can apply to the whole health data and AI lifecycles. Initially, AI is typically trained on secondary use data, of which quality and utility for AI should be labeled appropriately for trustworthy use. The current proposal for a regulation of the EHDS addresses the necessity of labeling the DQ and utility (Article 56) for data uses including research and personalized medicine, as is the case for AI.

DQ assurance can be specified based on DQ dimensions, which characterize data attributes facilitating or impeding its expected use. Dimension definitions vary according to the field of study [17,18,31,32]. Some dimensions are intrinsic to the data contents, including completeness, correctness, plausibility, or stability. Others relate to data access, including availability, accessibility, or security. Since potentially hindering or impeding AI development and use, we set intrinsic and access dimensions as the initial DQ targets in RAI.

Addressing DQ generally involves costly processes, from quality rules checks and exploratory analysis to data curation. Simple rules can find missing data footprints—blank spaces or, in the worst case, negative numbers such as “–9.” Logic rules defined by experts can help find implausible patterns in and between variables. Principal component analysis plots can help find apparently plausible but unlikely data, such as outliers. Further, information variability methods can help delineate temporal and source variability [22,33,34].

Addressing faulty data in training includes skipping faulty cases or recovering faulty values through data imputation methods. DQ flags can occasionally provide information relevant to AI performance, such as in the missing not-at-random case described in section 2. Further, as described in this paper, continual learning and model transportability methods can help address data variability in AI training.

Besides, DQ issues for AI can also occur when data is inputted into the AI systems, either being created de novo or passed through the EHR. At this stage, DQ handling methods can warn or automatically update missing, implausible, or outlying inputs with the most likely values given a context. However, any fully automated modification should be notified, and its effect on the output quantified through sensitivity analysis. This can be achieved through smart human-computer interaction and explainable AI methods.

Uncertainty Management

In health AI, data-related uncertainty can stem from epistemic or aleatoric factors, including lacking or faulty data or from pure statistical randomness in data or information overlap [35-37].

Quantification of uncertainty in AI focuses mainly on prediction accuracy, that is, informing about variability and confidence of the outputs. Available approaches include deterministic, Bayesian, ensemble, and augmentation methods [35,37]. Deterministic methods aim to predict jointly an output and its variability in single models, thus requiring this information at training. Bayesian methods model the statistical distribution of AI parameters—for example, the coefficients of logistic regression—rather than as single values, capturing the training data variability and enabling an output distribution from which to estimate a mean and CI. Bayesian methods include Monte Carlo dropout or Gaussian mixture models. Ensemble methods estimate multiple models—for example, through boosting or bagging approaches—in which combined outputs conform to a distribution. Lastly, augmentation methods apply artificial but realistic perturbations to inputs enabling multiple predictions which then also conform to a distribution.

Further, uncertainty from lacking information is particularly relevant in open inputs, including free text and dialogue systems such as in large language models (LLMs). It can be managed through deep learning network architectures, such as using recurrent neural networks that treat the input as a sequence with order information, recursive neural networks that treat the input as a hierarchical structure, transformers masked models such as BERT, where the different parts of an input sequence can influence each other through attention mechanisms [38], or using input embedding layers, where partial inputs can be expressed as indexes of dense vector representations [39].

Continual Learning and Model Transportability

Continual learning provides AI with mechanisms to self-adapt to information changes over time [40]. The model knowledge, generally represented as hyperparameters—for example, an artificial neural network layout—and parameters—the network weights—updates as new data batches arrive. Continual learning must balance forward and backward knowledge transfer, that is, rapidly adapting to dataset shifts while avoiding a catastrophic forgetting of important predictive past patterns. When using an AI model at settings or sites different from where it was trained, we face the problem of model transportability.

The continual learning and model transportability fields encompass similar methods from diverse ML paradigms, including online, active, transfer, or multitask learning [41]. Given their adaptative flexibility, these methods generally relate to artificial neural networks and deep learning. However, other simpler methods, including logistic regression or random forests, can also use batch or iterative learning. Of note is that these methods can update their knowledge without needing past data, conforming to data access regulations. Continual learning and model transportability can also benefit from infrastructure and organizational practices, including interoperability and domain knowledge management. Further, recent ML methodologies such as ML operations consider continual benchmarking and model updating [42]. Of note is that the use of AI foundation models, pretrained across various domains, can be of potential benefit to optimize resilience in continual learning and model transportability [43].

Fundamental Rights Assurance

Health AI involves potential risks for FRA regarding equality and nondiscrimination, economic and social rights, access to a fair trial and effective remedies, protection of personal data, or good administration [10]. RAI can address some of these rights by attaching to principles including the EU Charter of Fundamental Rights [44], the Ethics Guidelines for Trustworthy AI [45], or the GDPR [28]. In fact, FRA is at the core of medical AI regulations [46]. A successful implementation demands specific computational methods handling potential algorithmic biases.

We can classify FRA computational methods into pretraining, in-training, and posttraining methods [47,48]. Initially, we should identify available or derived variables potentially defining FRA-sensitive subpopulations, such as gender, race, or socioeconomic status. In pretraining methods, we compensate underrepresented subpopulations or imbalanced outcomes via resampling methods. Reweighting mechanisms can mitigate discrimination in outcomes. Further, we can remove or obfuscate FRA-sensitive variables if they do not affect model performance or usability. Besides, in-training methods focus mainly on defining objective functions or constraints in the AI learning process to compensate for unfair performance metrics in sensitive subpopulations—for example, optimizing the false-negative rate for a discriminated subpopulation. In screening and classification tasks, we can set subpopulation-specific cost benefits in confusion matrices or use specific deep learning loss functions—that is, the functions that measure the difference between the model prediction to the true answers to learn the network weights. Lastly, posttraining methods adapt a model's outcome to the sensitive situations of the tested individual—for example, setting custom receiver operating characteristic curves and decision thresholds or readjusting uncertain outcomes to favor sensitive subpopulations.

DQ and variability methods can also help detect unfair differences in data representations, DQ levels, or model performance across subpopulations. Another FRA aspect is decision-making transparency, where the models' interpretability is highly important.

Human-Computer Interaction

Resilient human-computer interaction is key to trustworthy AI and FRA. The methodological drivers for RAI by human-computer interaction include user experience and interface design methods [49], toward developing resilient user-centered CDSSs from the beginning [50], and AI and CDSS engineering aiming for dynamic and explainable AI [23,24].

Dynamic human-computer interaction in AI can use sequential prediction models that generate and update outcomes on partial inputs as new information is introduced, such as in recurrent or transformer neural networks [24]. Sequential models can establish a smart human-computer dialogue, particularly useful for sequential anamnesis inputs and natural language processing, such as using LLM.

Regarding explainable AI, the simpler the model, the more interpretable it is. For example, we can easily reason about a logistic regression behavior based on the feature's coefficients. Explainable AI post hoc methods provide explanations for less interpretable or "black box" models, which typically perform higher in complex tasks, including natural language processing or medical imaging. Post hoc local methods explain specific predictions, including the model-agnostic Shapley values or neural network gradient attribution techniques such as saliency maps. Post hoc global methods estimate each feature's average importance for any prediction. Surrogate interpretable models in points close to the test case can assist in explaining individual predictions from less-interpretable models [51]. Lastly, explainable AI local techniques can provide a sensitivity analysis for the outcomes.

Regulation

Current regulations supporting the development and maintenance of RAI solutions AI include those described in section 2, namely the EU AI Act, the EU Medical Device Regulation, the AI Action Plans by the Food and Drug Administration, and the proposal for the EHDS.

As a first step to facilitate the transportability of production AI systems, the EHDS provides a set of common specifications (Article 23) and a list of interoperability and security requirements (Annex II). These include structural and semantical requirements, and requirements related to DQ, patient safety, and data protection.

Some of the harmonized rules (HRs) that provide concrete details on how to meet the EU AI Act's goals align with our aim for RAI in health: high DQ and proper statistical properties

in training, testing, and validation are essential for the performance of AI and to avoid biases—HR-67; explainability and transparency to avoid opacity is key to use the system appropriately—HR-72; the system performance should be consistent throughout its lifecycle and monitored in real-life settings—HR-74; and the systems should be resilient against errors, faults, inconsistencies, or unexpected situations, as well as malicious actions, to avoid erroneous, potentially harmful decisions or biased outputs—HR-75.

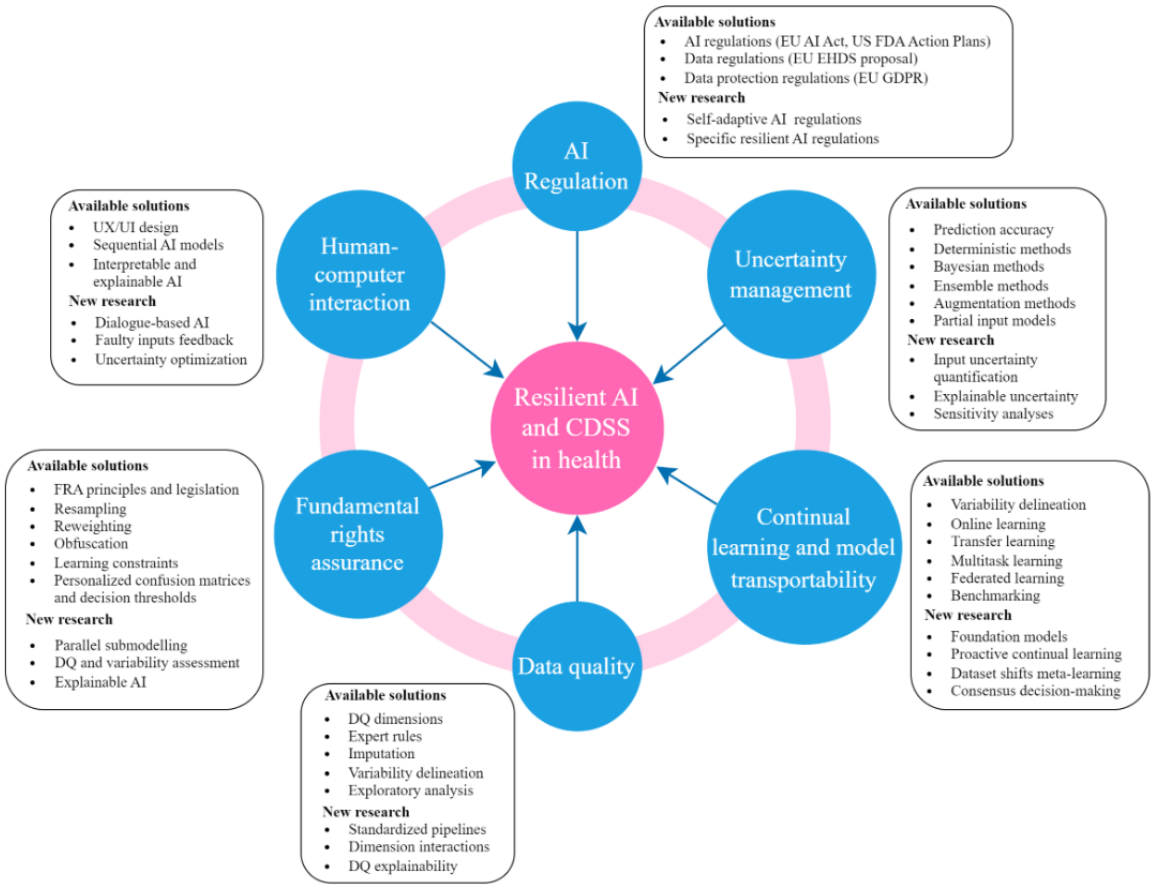
Regarding self-adaptive AI, once a system is put into service, the EU regulatory framework for AI suggests providing rules establishing that changes in the algorithm and its performance, predetermined by the provider and assessed at conformity assessment, should not constitute a substantial modification—HR-128. Similarly, the American Medical Informatics Association suggests a set of recommendations for the safe, effective use of self-adaptive CDSS and their prospective regulation concerning: the design and development based on transparency metrics; implementation using standards for communication and planning retraining criteria; in-situ evaluations and testing; and on-going monitoring including system maintenance and user training [52].

Others

Further transversal approaches can benefit RAI. Federated learning provides algorithms and interoperability infrastructure for multisite AI learning, enabling secure model transportability and continual learning addressing potential data access regulations. In the prediction phase, fast nearest-neighbor search and phenotyping algorithms can proactively display relevant comparative past cases [53-55]. Lastly, we bring attention to the proposed commandments of ethical medical AI by Muller et al [30], which can help design RAI-enabled CDSS while encompassing human-AI decision roles.

Figure 3 summarizes the main solutions described in this section, linking to potential new research aims, as described in this paper in more detail.

Figure 3. Summary of available solutions and new proposed research for resilient AI in health and CDSS. AI: artificial intelligence; CDSS: clinical decision support systems; DQ: data quality; EHDS: European Health Data Space; EU: European Union; FDA: Food and Drug Administration; GDPR: General Data Protection Regulation; UX/UI: user experience and interface.



Research Agenda

Putting the methods and solutions described into practice would be already a huge step toward RAI and trustworthy CDSS (see Textbox 2 for an example of a desired use of an RAI-enabled CDSS). However, this is not free of challenges. The development complexity and costs can be considerable, regulations must be met or extended, and additional validation of the methods is needed. Fortunately, user acceptance and organizational culture regarding RAI may benefit from conventional AI, given its increased, user-centered flexibility. Next, we propose several research aims to address current and future challenges in RAI.

Textbox 2. Example desired use of a resilient artificial intelligence–enabled clinical decision support system following the example from emergency medical dispatch in Textbox 1.

Context

- Suppose a European emergency medical dispatch center develops a clinical decision support system based on resilient artificial intelligence (RAI) for triage support. The aim is to predict the life-threatening situation of incidents based on the live transcription of the during-call conversations to support resource allocation.

Training

- The initial RAI learning is fed from a raw database containing transcriptions, in situ maneuvers and diagnoses, transport, and hospitalization—as the final gold standard—data of the last 10 years. The system vendor and its triage tree changed during this period, leading to different conversation types. Further, the International Classification of Diseases codes at hospitalization were updated from version 9 to version 10. The RAI training is based on a continual, deep learning methodology; the neural network is learnt in temporal batches through a continual fine-tuning strategy. Therefore, the final model will retain past knowledge, but recent patterns will be more relevant. The training also relied on resiliency functions that permitted, for example, alerting about outlying operators with training data biased toward overclassification, prospective uncertainty quantification of inputs based on a Monte Carlo dropout function, and suggesting a parallel model for female cases with poor information in cardiac disease incidents with positive outcome to avoid false negatives.

Prediction

- The case from Textbox 1 is inputted into the system, which classifies it as a high-uncertainty case using the uncertainty quantification function. The system automatically asks for potentially relevant information to reduce uncertainty, such as the use of drugs or recent anxiety attack episodes. In light of neither changes in the input nor interoperability with the patient EHR to retrieve this data, the system uses a reject option-based classification function to classify the case as life-threatening to avoid potentially biased decisions.

Production

- A continuous artificial intelligence (AI) benchmarking system is used, counting with secure interoperability with gold-standard feedback data and automated inspection of causes of changes. The system eventually detects reduced performance in older adults and newborns, finding multidrug-resistant bacterial infections associated with these cases, and suggests retraining the model with these new data.

Regulation

- The system is a high-risk AI system in compliance with the European Union (EU) AI Act, ensuring its safe and trustworthy use. It conforms with the European Health Data Space proposal's common specifications, interoperability, and security requirements. An updated regulation framework considered the digital document with clinically explained motivation for the retraining described before as a valid source for self-conformance of the system to continue its use after its self-update.

We must standardize pipelines for fully automated and semiautomated DQ processing during training. The implications of DQ dimensions interactions, for example, across-site missingness patterns, should be studied. Excessive curation might lead to an unreal dataset; therefore, combining curation with using affected informative patterns—for example, informative missingness [56]—should be further studied. In prediction, we must investigate user-centered human-computer interaction approaches against faulty inputs, for example, using automated DQ corrections and DQ-related sensitivity analysis. Legal aspects for data curation, especially if automated, should also be realized technically [57].

Uncertainty quantification can benefit from addressing input uncertainty besides outcome uncertainty. Input uncertainty quantification could combine imputation and data augmentation methods. This approach can also help analyze the effect of DQ on output uncertainty and lead to input-outcome sensitivity analysis. Additionally, explainability might improve from uncertainty quantification at specific decision stages, for example, visualizing uncertainty at neural network branches.

Conversational, dialogue-based RAI interactions, smartly asking for new parameters anticipating the user needs, while minimizing input costs following an uncertainty-reduction targeted dialogue, are key research aims. This could be achieved by combining recurrent AI with dynamic human-computer interaction, for example, through voice or text-based LLMs [58,59], enabling AI to work with partial information, the natural case in medicine. Besides, user-centered design with a focus on explainability deserves specific research toward its acceptance [60].

Human-AI interaction is key to FRA since it influences users' decision-making. Previous methods for uncertainty quantification, DQ assessment, or RAI interactions shall be put in common with AI and health data regulations, for example, through specific HRs and methods enabling FRA as a priority layer in RAI-based CDSS interactions.

The use of foundation models in CDSS development should be further validated for safe and trustworthy clinical use. They can enable rapid development and model transportability, and show promising behavior to fill knowledge gaps in the training data, increasing generalization [15,16,43]. However, risks and uncertainty must still be thoroughly quantified. Besides, foundation models will still need to be updated as data changes through continual learning methods.

Continual learning can evolve from reactive to proactive. Current resilience against dataset shifts is constrained by the need for new labeled data to update the model's parameters, which is always one step behind the actual context. Further than adapting periodically or after significant shifts, proactive

continual learning can aim to anticipate changes. Domain adaptation methods—for example, transductive learning—can help update knowledge from cases before their gold-standard outcomes are available [61,62]. Besides, comprehensive variability characterizations in historical datasets can provide an extensive knowledge base of variability patterns, potentially enabling the forecasting of changes [22,63]. Overall, proper benchmarking for continual learning in health is still challenging, where we motivate the availability of publicly available patient-level datasets including the cases' date of acquisition.

Predictions based on the "consensus" of parallel models may support human decision-making resiliency—not generally based on a single mental process—through cooperative knowledge from different perspectives. Parallel models can apply to FRA-sensitive subpopulations, missingness patterns, multisite and temporal variability, and high uncertainty, overlapping-information zones. However, this can risk the CDSS user understanding; for example, what is the trustworthiness balance between local and multisite outcomes? Transparent, distributed, and continuous AI benchmarking is key. Complementarily, consensus decision-making can benefit from research in context and logic-based argumentation algorithms [64].

Specific resilience and self-adaptiveness features must be included in health AI regulations. The adaptability of AI to different settings and the temporal evolution of medical practice are recognized regulatory challenges [52,65]. Any change in an AI production system currently requires human intervention and conformance evaluation. Therefore, automatically providing a clinically explainable justification for why and how self-adaptive AI is updated without human intervention—besides planned retraining—is significant future work. Further, we must standardize AI documentation and record-keeping practices to define the rules firing a self-adaptation process and keep track of the latter—for example, based on performance changes or dataset shifts—and, similarly, justify automated DQ curation processes both at training and prediction.

Self-adaptiveness and federated learning infrastructures should be validated and regulated for CDSSs' retraining, production use, and benchmarking with data access limitations. In cases when protected information is required—for example, for training or similarity searches—synthetic data fabrication can be a solution.

Lastly, future work in RAI can address unsupervised and reinforcement learning. Unsupervised learning aims to discover natural subgroups in data. Generally applied at the population level—for example, to find potential immune response patterns in blood tests—it can also apply at the patient level—for example, for tissue segmentation in medical imaging. Besides, reinforcement learning aims to learn optimal decision-making choices in nonstationary environments—for example, to optimize intensive care unit procedures [66]. Improving the resilience of both will potentially improve their contribution to actual clinical practice.

Conclusions

Resilience is a significant factor in the success of health AI. Health data is imperfect, incomplete, and permeated with variable representations. However, this is not wrong. Instead of artificially modifying data for AI, AI should be resilient to the data nature. Methods enabling RAI involve disciplines including DQ, uncertainty management, continual learning, model transportability, foundation models, conversational AI, human-computer interaction, and regulatory aspects. Their implementation and specific research in RAI will define new-generation CDSSs and maximize trustworthiness in AI-enabled health care.

Acknowledgments

This work was cofunded (PID2022-138636OA-I00; KINEMAI) by Agencia Estatal de Investigación—Proyectos de Generación de Conocimiento 2022.

Authors' Contributions

CS conceived this work and wrote this paper. PF and JMG-G reviewed this paper and provided critical revision. All authors approved the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Potential data-related uncertainties, variability, and biases for health artificial intelligence at the training and prediction phases. [PDF File (Adobe PDF File), 211 KB-Multimedia Appendix 1]

References

1. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
2. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507-2509. [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]

3. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544-1547. [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
5. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
6. Wynants L, van Calster B, Collins G, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
7. Sáez C, Romero N, Conejero J, García-Gómez JM. Potential limitations in COVID-19 machine learning due to data source variability: a case study in the nCov2019 dataset. J Am Med Inform Assoc. 2021;28(2):360-364. [FREE Full text] [doi: 10.1093/jamia/ocaa258] [Medline: 33027509]
8. European Commission. Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 [accessed 2024-06-12]
9. European Commission. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Outcome of the European Parliament's first reading (Strasbourg, 11 to 14 March 2024). 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CONSIL:ST_7536_2024_INIT [accessed 2024-06-12]
10. European Union Agency for Fundamental Rights. Data quality and artificial intelligence: mitigating bias and error to protect fundamental rights. LU: Publications Office. 2019. URL: https://data.europa.eu/doi/10.2811/546219 [accessed 2022-06-01]
11. European Commission. White paper on artificial intelligence—a european approach to excellence and trust. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065 [accessed 2024-06-12]
12. European Commission. A european strategy for data. Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 [accessed 2024-06-12]
13. European Commission. Regulation of the European Parliament and of the Council on the European Health Data Space. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197 [accessed 2024-06-12]
14. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res. 2020;22(10):e20346. [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
15. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
16. Rosenbloom L, Hudson DA, Adeli E. On the opportunities and risks of foundation models. arXiv. 2022. [doi: 10.48550/arXiv.2108.07258]
17. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-151. [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
18. Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform. 2012;180:721-725. [Medline: 22874286]
19. Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: organizing data for trustworthy artificial intelligence. Gov Inf Q. 2020;37(3):101493. [doi: 10.1016/j.giq.2020.101493]
20. Quiñonero-Candela J. Dataset shift in machine learning. In: Mass: MIT Press. London, England. MIT Press; 2009.
21. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit. 2012;45(1):521-530. [doi: 10.1016/j.patcog.2011.06.019]
22. Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. Gigascience. 2020;9(8):giaa079. [FREE Full text] [doi: 10.1093/gigascience/giaa079] [Medline: 32729900]
23. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
24. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
25. Regulation (EU) 2017/745 of the European Parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745 [accessed 2024-06-12]

26. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food & Drug Administration. US. U.S. Food & Drug; 2021. URL: https://www.fda.gov/media/145022/download?attachment [accessed 2024-06-12]
27. Artificial Intelligence and Machine Learning in Software as a Medical Device. U.S. Food & Drug Administration. US. U.S. Food & Drug; 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [accessed 2024-06-12]
28. European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation). 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2024-06-12]
29. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA. 2018;320(11):1107-1108. [doi: 10.1001/jama.2018.11029] [Medline: 30178025]
30. Muller H, Mayrhofer MT, van Veen E, Holzinger A. The ten commandments of ethical medical AI. Computer. 2021;54(7):119-123. [doi: 10.1109/mc.2021.3074263]
31. Madnick SE, Wang RY, Lee YW, Zhu H. Overview and framework for data and information quality research. J Data Inf Qual. 2009;1(1):1-22. [doi: 10.1145/1515693.1516680]
32. Aerts H, Kalra D, Sáez C, Ramírez-Anguita JM, Mayer M, Garcia-Gomez JM, et al. Quality of hospital electronic health record (EHR) data based on the international consortium for health outcomes measurement (ICHOM) in heart failure: pilot data quality assessment study. JMIR Med Inform. 2021;9(8):e27842. [FREE Full text] [doi: 10.2196/27842] [Medline: 34346902]
33. Kahn MG, Raebel M, Glanz J, Riedlinger K, Steiner J. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50 Suppl:S21-S29. [FREE Full text] [doi: 10.1097/MLR.0b013e318257dd67] [Medline: 22692254]
34. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. J Am Med Inform Assoc. 2016;23(6):1085-1095. [FREE Full text] [doi: 10.1093/jamia/ocw010] [Medline: 27107447]
35. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion. 2021;76:243-297. [doi: 10.1016/j.inffus.2021.05.008]
36. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn. 2021;110(3):457-506. [doi: 10.1007/s10994-021-05946-3]
37. Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev. 2023;56(S1):1513-1589. [doi: 10.1007/s10462-023-10562-9]
38. Goodfellow I, Bengio Y, Courville A. Deep learning. USA. Cambridge, MA. USA: MIT Press; 2016.
39. Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. Advances in Neural Information Processing Systems. 2000. URL: https://papers.nips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html [accessed 2024-04-16]
40. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. Neural Netw. 2019;113:54-71. [FREE Full text] [doi: 10.1016/j.neunet.2019.01.012] [Medline: 30780045]
41. Mundt M, Lang S, Delfosse Q, Kersting K. CLEVA-compass: a continual learning evaluation assessment compass to promote research transparency and comparability. arXiv. 2022. [doi: 10.48550/arXiv.2110.03331]
42. Dreyfus-Schmidt L. Introducing MLOps. URL: https://learning.oreilly.com/library/view/~/9781492083283/?ar?orpq&email=^u [accessed 2022-06-01]
43. Guo LL, Steinberg E, Fleming SL, Posada J, Lemmon J, Pfohl SR, et al. EHR foundation models improve robustness in the presence of temporal distribution shift. Sci Rep. 2023;13(1):3767. [FREE Full text] [doi: 10.1038/s41598-023-30820-8] [Medline: 36882576]
44. European Commission. Charter of Fundamental Rights of the European Union 2012/C 326/02. 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:C2012/326/02 [accessed 2024-06-12]
45. European Commission. Ethics guidelines for trustworthy AI. Directorate-General for Communications Networks, Content and Technology. 2019. URL: https://data.europa.eu/doi/10.2759/346720 [accessed 2024-06-12]
46. Stöger K, Schneeberger D, Holzinger A. Medical artificial intelligence: the European legal perspective. Commun ACM. 2021;64(11):34-36. [doi: 10.1145/3458652]
47. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv. 2022;55(3):1-44. [doi: 10.1145/3494672]
48. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI Ethics. 2023;3(1):223-240. [doi: 10.1007/s43681-022-00195-z]
49. Margetis G, Ntoa S, Antona M, Stephanidis C. Human-centered design of artificial inteligence. In: Handbook of human factors and ergonomics. USA. John Wiley & Sons; 2021:1085-1086.

50. Melnick ER, Hess EP, Guo G, Breslin M, Lopez K, Pavlo AJ, et al. Patient-centered decision support: formative usability evaluation of integrated clinical decision support with a patient decision aid for minor head injury in the emergency department. J Med Internet Res. 2017;19(5):e174. [FREE Full text] [doi: 10.2196/jmir.7846] [Medline: 28526667]
51. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You": explaining the predictions of any classifier. arXiv. 2016. [doi: 10.48550/arXiv.1602.04938]
52. Petersen C, Smith J, Freimuth R, Goodman KW, Jackson GP, Kannry J, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc. 2021;28(4):677-684. [FREE Full text] [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
53. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21(2):221-230. [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
54. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol. 2015;33(5):462-463. [doi: 10.1038/nbt.3223] [Medline: 25965751]
55. Li W, Zhang Y, Sun Y, Wang W, Li M, Zhang W, et al. Approximate nearest neighbor search on high dimensional data — experiments, analyses, and improvement. IEEE Trans Knowl Data Eng. 2020;32(8):1475-1488. [doi: 10.1109/tkde.2019.2909204]
56. Ferri P, Romero-Garcia N, Badenes R, Lora-Pablos D, Morales TG, Gómez de la Cámara A, et al. Extremely missing numerical data in electronic health records for machine learning can be managed through simple imputation methods considering informative missingness: a comparative of solutions in a COVID-19 mortality case study. Comput Methods Programs Biomed. 2023;242:107803. [FREE Full text] [doi: 10.1016/j.cmpb.2023.107803] [Medline: 37703700]
57. Stöger K, Schneeberger D, Kieseberg P, Holzinger A. Legal aspects of data cleansing in medical AI. Comput Law Secur Rev. 2021;42:105587. [doi: 10.1016/j.clsr.2021.105587]
58. Moore MR, Arar R. Studies in Conversational UX Design. In: Moore RJ, Szymanski MH, Arar R, Ren GJ, editors. Conversational UX design: an introduction. US. Cham: Springer International Publishing; 2018:1-16.
59. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine. 2023;90:104512. [FREE Full text] [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
60. Cabitza F, Campagner A, Malgieri G, Natali C, Schneeberger D, Stoeger K, et al. Quod erat demonstrandum?—towards a typology of the concept of explanation for the design of explainable AI. Expert Syst Appl. 2023;213:118888. [doi: 10.1016/j.eswa.2022.118888]
61. Vladimir V. Transductive inference semi-supervised learning. Semi-Supervised Learning. London, UK. The MIT Press; 2006. URL: https://axon.cs.byu.edu/~martinez/classes/778/Papers/transductive.pdf [accessed 2024-06-12]
62. Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B. Covariate shift by kernel mean matching. Dataset Shift in Machine Learning. UK. The MIT Press; 2008. URL: https://www.gatsby.ucl.ac.uk/~gretton/papers/covariateShiftChapter.pdf [accessed 2024-06-12]
63. Sáez C, García-Gómez JM. Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: functional data analysis of data temporal evolution over non-parametric statistical manifolds. Int J Med Inform. 2018;119:109-124. [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.09.015] [Medline: 30342679]
64. Vassiliades A, Bassiliades N, Patkos T. Argumentation and explainable artificial intelligence: a survey. Knowl Eng Rev. 2021;36:e5. [doi: 10.1017/s0269888921000011]
65. Cohen IG, Evgeniou T, Gerke S, Minssen T. The European artificial intelligence strategy: implications and challenges for digital health. Lancet Digit Health. 2020;2(7):e376-e379. [FREE Full text] [doi: 10.1016/S2589-7500(20)30112-6] [Medline: 33328096]
66. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):16-18. [FREE Full text] [doi: 10.1038/s41591-018-0310-5] [Medline: 30617332]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
DQ: data quality
EHDS: European Health Data Space
EHR: electronic health record
EU: European Union
FRA: fundamental rights assurance
GDPR: General Data Protection Regulation
HR: harmonized rule
LLM: large language model
ML: machine learning
RAI: resilient artificial intelligence

RWD: real-world data

Edited by G Tsafnat; submitted 26.06.23; peer-reviewed by Y Tolias, Z Su, C Manlhiot, A Holzinger; comments to author 27.01.24; revised version received 16.04.24; accepted 18.05.24; published 28.06.24

Please cite as:
Sáez C, Ferri P, García-Gómez JM
Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support
J Med Internet Res 2024;26:e50295
URL: https://www.jmir.org/2024/1/e50295
doi: 10.2196/50295
PMID: 38941134