Issues and Limitations on the Road to Fair and Inclusive AI Solutions for Biomedical Challenges

sensors

MDPI

Perspective

Issues and Limitations on the Road to Fair and Inclusive AI Solutions for Biomedical Challenges

Oliver Faust [1,*], Massimo Salvi [2], Prabal Datta Barua [3,4,5,6,7,8,9,10,11], Subrata Chakraborty [7,12,13], Filippo Molinari [2] and U. Rajendra Acharya [14,15]

1 School of Computing and Information Science, Anglia Ruskin University, Cambridge Campus, Cambridge CB1 1PT, UK
2 PoliToBIOMed Lab, Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca Degli Abruzzi 24, 10129 Turin, Italy; massimo.salvi@polito.it (M.S.); filippo.molinari@polito.it (F.M.)
3 Cogninet Australia, Sydney, NSW 2010, Australia; prabal.barua@unisq.edu.au
4 School of Business (Information Systems), University of Southern Queensland, Toowoomba, QLD 4350, Australia
5 Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
6 Australian International Institute of Higher Education, Sydney, NSW 2000, Australia
7 School of Science and Technology, University of New England, Armidale, NSW 2351, Australia; subrata.chakraborty@une.edu.au
8 School of Biosciences, Taylor's University, Subang Jaya 47500, Malaysia
9 School of Computing, SRM Institute of Science and Technology, Kattankulathur 603203, India
10 School of Science and Technology, Kumamoto University, Kumamoto 860-8555, Japan
11 Sydney School of Education and Social Work, University of Sydney, Camperdown, NSW 2050, Australia
12 Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia
13 Griffith Business School, Griffith University, Brisbane, QLD 4111, Australia
14 School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD 4300, Australia; rajendra.acharya@unisq.edu.au
15 Centre for Health Research, University of Southern Queensland, Ipswich, QLD 4305, Australia
* Correspondence: oliver.faust@gmail.com

Abstract: Objective: In this paper, we explore the correlation between performance reporting and the development of inclusive AI solutions for biomedical problems. Our study examines the critical aspects of bias and noise in the context of medical decision support, aiming to provide actionable solutions. Contributions: A key contribution of our work is the recognition that measurement processes introduce noise and bias arising from human data interpretation and selection. We introduce the concept of "noise-bias cascade" to explain their interconnected nature. While current AI models handle noise well, bias remains a significant obstacle in achieving practical performance in these models. Our analysis spans the entire AI development lifecycle, from data collection to model deployment. Recommendations: To effectively mitigate bias, we assert the need to implement additional measures such as rigorous study design; appropriate statistical analysis; transparent reporting; and diverse research representation. Furthermore, we strongly recommend the integration of uncertainty measures during model deployment to ensure the utmost fairness and inclusivity. These comprehensive recommendations aim to minimize both bias and noise, thereby improving the performance of future medical decision support systems.

Keywords: bias; noise; inclusive AI; trust; explainability; system design

Academic Editors: Cataldo Guaragnella and Maria Rizzi

Received: 5 November 2024
Revised: 14 December 2024
Accepted: 20 December 2024
Published: 2 January 2025

Citation: Faust, O.; Salvi, M.; Barua, P.D.; Chakraborty, S.; Molinari, F.; Acharya, U.R. Issues and Limitations on the Road to Fair and Inclusive AI Solutions for Biomedical Challenges. *Sensors* **2025**, *25*, 205. https://doi.org/10.3390/s25010205

1. Introduction

While artificial intelligence (AI) continues to make remarkable advances in medical decision support, we argue that current approaches to ensure fairness and inclusivity fall short when addressing the unique challenges of the medical domain. This position paper argues that a fundamental rethinking of how we approach fairness in medical AI is necessary due to the high stakes involved in medical judgements, the complexity of multi-modal medical data, and the deeply embedded biases in current healthcare systems.

Achieving fairness in medical AI not only requires addressing biases but also confronting the inherent complexities of real-world medical data. Engineers and computer scientists strive to maintain control over the operating environment for their systems and software programs. In an ideal scenario, these solutions exhibit perfect repeatability, with no variations between operators or conditions. However, practical problem-solving inevitably widens the gap between a system's performance in controlled laboratory settings and its real-world deployment performance [1]. This performance gap is particularly significant in the context of medical decision support systems [2–4]. The extent of the performance gap is directly linked to the level of control over the data acquisition process in the medical environment [5]. Unfortunately, the absence of objective measures hampers the assessment of the quality of medical data acquisition and control [6,7]. The presence of unavoidable noise during measurements introduces uncertainty. In healthcare applications, noise can manifest as variations in disease symptoms, which subsequently affect the measurement data collected from patients at specific points in time.

AI models aim to emulate human decision-making processes in providing medical decision support [8]. This approach is philosophically justified by analogies between human thinking and machine decision-making. Medical professionals navigate uncertain environments with limited control, their expertise serving as invaluable knowledge sources [9,10]. Capturing this knowledge involves human experts labeling data, inevitably introducing subjectivity and bias, which limits the transferability of results from medical decision support systems [11,12]. Data and their interpretation stem from individual or group decisions, influenced by intentions about what to measure, when to measure, and how to interpret [13,14].

Bias, a systematic deviation from objectivity and fairness, can arise from various sources in medical measurements, including selection bias [15,16], measurement bias [17,18], confounding bias [19], and cultural or demographic bias [20]. AI systems can amplify these biases if trained on biased datasets or if algorithms have inherent biases [21]. This can perpetuate existing biases, introduce new ones, and lead to unfair outcomes in diagnostic pathways. The challenge for engineers is developing medical decision support systems that effectively replicate professional decision-making while mitigating biases. This requires effective methods for assessing decision support quality [22,23], with algorithm design and quality measures being fair, inclusive, and comprehensive [24].

In this paper, we adopt the position that quality design and reporting for AI solutions is a multifaceted undertaking. It is necessary to incorporate mitigation measures against bias and noise during AI development and include uncertainty measures during model deployment to ensure fairness and inclusivity. These measures should be used to build trust in the model so that there is a good level of confidence before the clinical validation.

The structure of this document is as follows. The next section outlines the methods used for developing medical decision support systems, covering the basics of AI, with a specific emphasis on state-of-the-art quality measures. Finally, we discuss the implications of our position and provide recommendations for future research and development efforts.

2. Methods

In this section, we provide a critical analysis of various medical decision support technologies, and the methods used to assess their performance. While these technologies and methods are well established, we argue that their application in medical contexts presents unique challenges that are often overlooked in standard performance assessments.

Figure 1 presents an abstract block diagram of a medical decision support system, encompassing crucial elements such as measurement, human-led data interpretation, AI model creation, and model deployment. This holistic view allows us to examine how noise and bias propagate through the entire system, affecting overall performance and fairness.

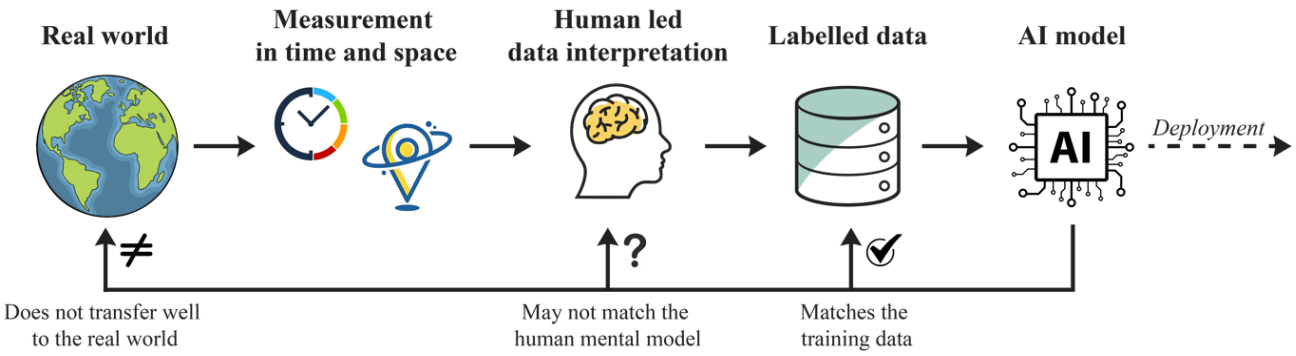

Figure 1. Block diagram illustrating the components and steps involved in the development of a medical decision support system. The diagram highlights four interconnected stages: measurement (introducing system noise), human-led data interpretation (incorporating expert knowledge and potential biases), AI model creation (development and validation), and model deployment (clinical implementation). The connections between components emphasize how noise and bias propagate throughout the system, influencing both the development process and final performance.

We pay particular attention to the measurement block, which introduces noise that impacts AI model performance [23], and the human data interpretation block, which introduces bias into decision support systems. These elements are often treated as separate issues in traditional AI development, but they are intrinsically linked in medical contexts and must be addressed simultaneously. While we discuss AI model creation and associated performance measures [24], our analysis goes beyond standard metrics. We critically examine how these measures may fail to capture the full complexity of medical decision-making and propose alternative approaches that better account for the unique challenges in healthcare settings.

A unique aspect of our analysis is the recognition of the cyclical relationship between measurement noise and interpretation bias in medical contexts. While traditional approaches treat these as independent challenges, we propose that they form a feedback loop: measurement noise influences human interpretation strategies, leading to compensatory biases, while biased interpretation affects future measurement protocols, introducing systematic noise. This interdependence creates what we term a "noise-bias cascade" that conventional AI evaluation frameworks fail to address. This perspective shifts the focus from treating noise and bias as separate technical challenges to understanding them as interlinked components of the medical decision-making ecosystem.

2.1. Measurement

Medical data measurement is a fundamental process that involves systematically and quantitatively assessing various aspects of health, diseases, treatments, or other relevant variables within the field of medicine. It plays a crucial role in evidence-based medicine, clinical research, healthcare quality improvement, and personalized patient care. The

accurate and reliable collection, recording, and analysis of data using specific measurement tools, techniques, or instruments are vital for ensuring the integrity and usefulness of medical information.

The diversity of medical data types (ranging from physiological measurements and clinical observations to patient-reported outcomes, imaging data, and genetic information [25–27]) creates a multifaceted landscape that AI models must navigate. This complexity exacerbates the impact of noise and bias in ways that are distinct from other domains of AI application. Bias in medical data collection often stems from unrepresentative or exclusionary practices. For example, wearable devices calibrated primarily for lighter skin tones may produce inaccurate readings for individuals with darker skin, introducing systematic bias in the data [28]. Similarly, geographic disparities lead to uneven data representation, as rural areas often lack diagnostic infrastructure, resulting in lower-quality or incomplete datasets. To address these challenges, targeted data collection strategies, such as oversampling underrepresented groups and standardizing measurement protocols, are essential. Federated learning can help mitigate biases by training AI models across decentralized datasets without compromising patient privacy. For example, a federated model, trained on patient records from hospitals across diverse regions, can account for demographic and geographic variations, reducing biases in AI predictions. Privacy-preserving techniques, such as differential privacy and secure multiparty computation, further enhance this approach by safeguarding sensitive information during the training process. These advancements ensure that AI models remain inclusive while maintaining compliance with ethical and legal standards.

While strategies such as sensor fusion and uncertainty quantification [29,30] can improve AI model robustness against noise, we argue that these approaches must be tailored specifically for medical applications. Standard validation techniques for measurement tools and data collection procedures [31–33] may not fully capture the nuanced ways in which bias can manifest in medical contexts.

There is a need for a more comprehensive approach to ensuring the accuracy, reliability, and validity of medical data measurement for AI applications. Upholding rigorous standards and practices in medical data measurement is not just about improving data quality, but also about ensuring equitable and safe healthcare outcomes for all populations [34].

2.2. Human-Led Data Interpretation

The application of clinical knowledge and expertise is a crucial aspect of human-based medical data interpretation. By drawing upon their clinical experience and understanding of medical concepts, pathophysiology, and evidence-based practices, medical experts provide valuable insights that complement automated approaches [35,36]. We argue that the integration of this human expertise with AI systems in healthcare settings introduces complexities that go beyond those seen in other domains.

While human expertise is invaluable, it also introduces bias due to subjective judgments and cognitive limitations. For example, clinicians may consciously or unconsciously interpret diagnostic data differently based on a patient's demographic characteristics (age, gender, ethnicity), potentially influencing subsequent AI predictions. This interpretation bias can skew the performance of AI models if these subjective decisions are incorporated into training datasets. Bias quantification methods, such as comparing model performance across subgroups (e.g., by demographic), can help detect these biases. Statistical tests or fairness metrics (such as group-specific precision, recall, or AUC) can be employed to measure how well the model performs for different populations, highlighting potential biases introduced by human-led interpretation.

Another significant concern arises from the labeling of training data. Human error or inconsistency during data labeling, especially in large-scale datasets, can create label noise, further diminishing the model's capacity to generalize. Therefore, it is recommended to apply robustness measures in AI models that account for noisy labels, such as active learning or semi-supervised learning approaches.

2.3. Model Creation

Model creation is a multi-step process that involves several key stages: preprocessing, training, and testing. The first step in creating a model is preprocessing the data to prepare them for analysis by filtering, cleaning, and transforming them into a suitable format for the model. Next, the model is trained on the preprocessed data, with its parameters adjusted iteratively to optimize its performance. Once the model has been trained, it is evaluated on a separate set of data to test its accuracy and generalizability. This testing phase is essential for ensuring that the model is not overfitting to the training data. Performance measures play a crucial role in providing feedback to guide the model creation process [1,37].

2.3.1. Preprocessing

During preprocessing, careful attention must be paid to addressing the potential for bias arising from unrepresentative or incomplete data. It is important to employ techniques that can detect and mitigate biases in AI models [38]. One approach involves utilizing fairness metrics and evaluation methods to identify any disparate impacts that may exist within the system. Once biases are identified, modifications can be made to the algorithms to reduce or eliminate the bias. This can involve adjusting the training data, developing algorithms that are more sensitive to fairness considerations, or implementing post-processing techniques to ensure fairness in decision-making [39]. Furthermore, it is crucial to involve diverse and representative stakeholders throughout the development and evaluation process. This includes considering the perspectives and expertise of individuals from different demographic groups to ensure that potential biases are identified and addressed comprehensively [40].

2.3.2. Training and Testing

At an abstract level, AI algorithms in the field of healthcare are typically based on either supervised or unsupervised learning approaches [41]. Unsupervised learning methods discover hidden data structures and patterns within medical datasets [42], providing insights into disease hotspots, interactions between multiple diseases in multimorbid patients, and the socioeconomic implications. This exploration through unsupervised learning enables knowledge generation, which has the potential to drive medical progress. However, the interpretation of these discovered patterns requires careful consideration, based on medical knowledge, to avoid spurious correlations or clinically irrelevant findings.

Supervised learning aims to tap into this existing knowledge by training and testing models using labeled data [43]. These labels result from human-led data interpretation, as discussed in Section 2.2. On a technical level, the first step in conducting supervised learning is to divide the available data into training and test sets. The training set is utilized by the algorithm to learn and build a model, while the test set is employed to evaluate the model's performance. By splitting the data in this manner, we can assess how well the trained model generalizes to new, unseen data. However, this process tends to preserve and sometimes exacerbate existing biases. For example, training on imbalanced datasets without bias correction leads to models that systematically underperform for minority groups. Similarly, validation bias—when test sets fail to represent the full diversity of the population—can result in misleading performance metrics that may not reflect real-world clinical performance. Recent advancements, such as federated learning [44] and

privacy-preserving AI [45], allow the training of models across decentralized datasets while preserving privacy. These approaches mitigate demographic biases and increase inclusivity by incorporating data from diverse, distributed sources.

To ensure fair and inclusive model development, we advocate for the pervasive use of fairness metrics during both training and testing. Techniques such as group-specific accuracy, equalized odds, or demographic parity should be incorporated into the evaluation process to identify disparities in model outcomes across subgroups. Additionally, the use of cross-validation with demographically stratified data can ensure that the model is tested on a broad and inclusive sample. This might result in a more accurate reflection of its performance in diverse real-world applications.

2.3.3. Performance Measures

Performance measures play a critical role in assessing the effectiveness and efficiency of medical AI. These measures provide quantitative evaluations of how well an AI algorithm performs its intended task, enabling comparisons between different algorithms or variations of the same algorithm [46,47]. The following list introduces commonly used AI performance measures for medical applications:

1. Accuracy is a widely employed performance measure, particularly in classification tasks. It calculates the percentage of correctly classified instances out of the total number of instances. While accuracy is essential, it may not offer a comprehensive view of the algorithm's performance, especially when dealing with imbalanced datasets.
2. Precision and recall are often utilized in binary classification problems. Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances out of all actual positive instances [48]. Precision and recall are commonly combined into a single measure called the F1 score, which provides a balanced evaluation of both precision and recall.
3. Area under the curve (AUC) is commonly employed in binary classification problems to assess the performance of a classifier's receiver operating characteristic (ROC) curve. It quantifies the classifier's ability to rank positive instances higher than negative instances across different classification thresholds [49]. AUC values range from 0.5 (random guessing) to 1.0 (perfect classification).
4. Mean absolute error (MAE) and root mean squared error (RMSE): These measures are frequently used in regression tasks to evaluate the prediction accuracy of continuous variables. MAE calculates the average absolute difference between predicted and actual values, while RMSE calculates the square root of the average squared difference [50]. Lower values indicate better performance.
5. Computational efficiency: In addition to accuracy measures, assessing computational efficiency is crucial. It evaluates the algorithm's speed and usage of resources, such as memory and processing power [51]. Performance measures like training time, prediction time, and memory consumption can be employed to evaluate the efficiency of AI algorithms.

These standard performance measures require careful interpretation based on domain-specific factors. For instance, in many medical applications, false negatives (missed diagnoses) may have more severe consequences than false positives, making recall potentially more critical than precision in certain contexts. The choice of performance measures depends on the specific medical problem, characteristics of the dataset, and desired outcomes. We recommend using a combination of performance measures to establish a comprehensive evaluation of an AI algorithm's performance, particularly in medical applications where multiple aspects of performance may be crucial [52].

In the context of inclusive medical AI, it is necessary to extend performance measures to account for fairness and bias. Metrics such as equalized opportunity and disparity impact can measure model performance across different demographic groups, ensuring that the decision support model does not disproportionately benefit or harm any group. Shapley values and counterfactual fairness tests are useful in assessing whether certain input factors (e.g., race or gender) unduly influence model predictions, providing a mechanism for bias detection and mitigation.

When evaluating the model performance based on noisy data, traditional error metrics such as mean absolute error (MAE) or root mean squared error (RMSE) should be complemented with uncertainty quantification and noise-tolerant measures. Continuous evaluation and refinement of these performance measures is necessary to ensure accurate and meaningful assessments of AI algorithms, considering both potential biases and real-world impact on patient outcomes.

2.4. Deployment

Medical decision support systems are created by deploying a trained AI model into a production environment within a healthcare system. This process involves making the AI model operational and accessible to healthcare professionals for use in diagnosing diseases, predicting outcomes, assisting in treatment decisions, or other relevant applications. By providing a second diagnostic opinion or automating screening processes, these systems have the potential to free up resources and make decision processes more coherent. The deployment of an AI model in the medical domain typically follows these steps:

1. Model selection: The performance measures, discussed in Section 2.3.3, provide an objective basis for selecting a model for deployment.
2. Infrastructure setup: The necessary infrastructure is established to support the deployment of the AI model in the medical setting. This includes ensuring compliance with privacy regulations [53], implementing data security measures, and addressing any specific requirements for handling sensitive patient information [54].
3. Integration with healthcare systems: The AI model is integrated into existing healthcare systems, such as electronic health record (EHR) systems, medical imaging platforms, or clinical decision support tools. This integration enables seamless data exchange and interaction between the AI model and healthcare professionals [55].
4. Data access and pre-processing: The AI model is connected to relevant data sources, such as patient records, medical imaging archives, or real-time monitoring devices. Data preprocessing steps may be implemented to standardize, clean, and anonymize the data while preserving their integrity and privacy [56–58].
5. Testing and validation in the real-world setting: The deployed AI model undergoes extensive testing and validation in real-world medical scenarios. Its performance, accuracy, and safety are evaluated, and necessary adjustments are made to ensure optimal performance and patient safety.
6. Regulatory compliance and ethical considerations: Compliance with regulatory requirements, such as those set by medical authorities or data protection regulations, is addressed to ensure the responsible and ethical deployment of the AI model. Patient consent, privacy, and ethical considerations are given utmost importance [59].
7. Monitoring and maintenance: Deployed AI models are continuously monitored to assess their performance, detect any anomalies, and identify opportunities for improvement [60,61]. Regular maintenance activities, including updating the model with new data, retraining, or refining its algorithms, are carried out to keep the model up-to-date and effective.

8. Collaboration and feedback: Collaboration between AI experts, healthcare professionals, and stakeholders is encouraged to gather feedback, address concerns, and optimize the AI model's performance for better patient outcomes and clinical decision-making [62].

The deployment strategy must ensure the transparency, interpretability, and safety of AI systems in healthcare while also maintaining the expertise and judgment of healthcare professionals in the decision-making process. Continuous bias monitoring is essential as deployment conditions differ from training environments, where demographic and socioeconomic factors can reintroduce biases. We propose post-deployment fairness audits and feedback loops incorporating healthcare professionals and patients to identify and correct discrepancies between model predictions and real-world outcomes, particularly for underrepresented populations.

3. Discussion

AI solutions for biomedical problems are developed within scientific, commercial, and ethical frameworks, which influence their design, deployment, and outcomes [63]. These systems are shaped not only by scientific interest, but also by practical constraints, including data availability and ethical considerations such as fairness and privacy. For example, models that perpetuate biases risk harming marginalized populations and undermining trust in AI-driven healthcare. Ethical AI design mandates participatory approaches, involving diverse stakeholders—patients, clinicians, and policymakers—in every stage of development. Transparency is equally crucial, with open reporting on model limitations and potential biases.

Simply stating that biased data affect model outcomes is insufficient for understanding the full impact on medical decision support. Biases in data often mirror societal inequities, such as underrepresentation of minority groups in clinical trials, geographic disparities between urban and rural healthcare facilities, and socioeconomic barriers to accessing medical care. These variations in data collection methodologies and healthcare access create systemic biases that affect different regions and socioeconomic groups disproportionately. These biases can lead to algorithmic discrimination, where AI models perpetuate existing inequities by providing inaccurate or less effective recommendations for marginalized populations. Therefore, addressing bias requires a more holistic approach that goes beyond technical fixes in data preprocessing or performance metrics. It involves rethinking how data is collected, curated, and applied to ensure more equitable AI outcomes.

3.1. Intelligent Decision Support Systems: Beyond Model Deployment

While our initial recommendations focused on trustworthy AI model evaluation and deployment, we acknowledge that this approach does not fully encompass the complexity of intelligent decision support systems (IDSSs) in healthcare. An IDSS integrates multiple components beyond a deployed model, including data preprocessing pipelines, human–computer interfaces, knowledge bases, and reasoning engines, creating an adaptive, context-aware system for complex decision-making environments. A well-designed IDSS must incorporate mechanisms for integrating clinical expertise, patient data, and real-time feedback from healthcare professionals. It should facilitate collaborative decision-making by providing interpretable results, confidence scores, and clear reasoning pathways. Moreover, an IDSS should account for the dynamic nature of medical practice, where new data, evolving guidelines, and individual patient preferences.

To address these requirements, we propose expanding our design methodology to include the following components:

- Knowledge management systems: To store and update clinical guidelines, medical literature, and historical patient outcomes.
- Clinical workflow integration: To ensure that the AI's outputs are seamlessly integrated into the medical professionals' workflow, providing recommendations at the point of care.
- Interactive user interfaces: Allow clinicians to understand the model's reasoning and adjust parameters or provide feedback, enhancing trust and interpretability.
- Patient-centric components: Such as decision aids that provide patients with understandable explanations of their options, fostering shared decision-making between patients and healthcare providers.

While the traditional steps for deploying an AI model are necessary, they are just one part of the broader ecosystem required for a truly intelligent medical decision support system. By refining our design methodology to include these additional elements, we can create systems that not only improve decision accuracy, but also ensure that medical AI is both transparent and adaptable to the complex realities of healthcare.

3.2. Expanding Bias Mitigation Beyond Model Performance

Addressing structural biases inherent in the healthcare system is essential when deploying AI systems. These systems interact with biased policies, organizational practices, and resource disparities, necessitating a multi-dimensional approach to bias mitigation:

- Auditing data sources for systemic biases (e.g., underrepresentation of certain populations).
- Tracking performance across subgroups to ensure that models perform equitably for all patients.
- Contextualizing AI outputs within the broader healthcare environment to ensure that recommendations align with ethical and clinical best practices.
- Engaging diverse stakeholders, including patients and healthcare providers from underrepresented groups, in the development and evaluation of AI systems.
- Implementing ongoing monitoring and feedback mechanisms to identify and address emergent biases in real-world settings.

Furthermore, we propose that bias mitigation should be integrated into every stage of the AI lifecycle, from problem formulation and data collection to model deployment and ongoing evaluation. This holistic approach requires collaboration between AI developers, healthcare professionals, policymakers, and patient advocates to ensure that AI systems contribute to fairer and more inclusive medical decision support.

3.3. Algorithm Sharing and Technology Reuse

The sharing of algorithms and trained models has gained significant traction in machine learning and deep learning, offering potential for accelerated progress and efficient resource utilization [64]. However, this practice presents ethical and intellectual property (IP) concerns that require careful consideration, particularly in the context of medical decision support systems [65].

Sharing algorithms can have profound consequences for privacy, bias, and discrimination [66]. In medical contexts, where decisions can directly impact patient outcomes, these concerns are particularly acute. An algorithm trained on biased medical data could perpetuate health disparities if shared and implemented widely [67]. Moreover, the use of shared algorithms in sensitive medical decision-making raises issues of transparency and accountability.

Algorithms, like any other software, can be protected by copyright, patents, and trade secrets, potentially leading to conflicts regarding ownership and licensing. Open-source

licenses offer one effective approach, enabling algorithm sharing while protecting creators' rights [68]. For example, the GPL license requires that any derivative works of the original code must be licensed under the same terms, ensuring that the code remains free and open. Several organizations, such as the IEEE, ACM, and Partnership on AI, have formulated guidelines for ethical AI development and deployment [69]. These guidelines cover aspects such as transparency, accountability, fairness, and privacy, providing a framework for the responsible sharing of algorithms.

In healthcare, these approaches must be tailored to address the unique challenges of medical data, including patient privacy, regulatory compliance, and the potential impact on human life. Striking the right balance between openness and protection in medical algorithm sharing is crucial for advancing healthcare innovation while safeguarding patient interests [70].

3.4. Recommendations

Our discussion on the model creation process has highlighted that traditional methods alone are inadequate in addressing challenges stemming from bias and noise in medical AI. Model creation, performance measurements, and benchmarking heavily rely on data availability, which can perpetuate existing biases. We propose moving beyond purely data-driven approaches to incorporate a problem-solution design methodology for fairer and more inclusive AI solutions. We formulate this as a design problem, aiming to reduce biases arising from data, algorithms, and user interaction. The innovation is that data-driven AI model creation is just the first of three steps in a more comprehensive process, requiring continuous refinement toward a practical, trustworthy, and clinically validated solution.

Figure 2 depicts our proposed design methodology. The "Model design and testing" block represents the data-driven approach outlined in the previous section. Empirical performance evaluation and benchmarking will help us to determine if the model is promising and if we can progress in the design methodology. If this is not the case, it is necessary to go back and refine the data-driven AI model. The "Trust building" step transitions into social and clinical science through explainability analysis, testing if the AI model aligns with human mental models. Only trustworthy models progress to clinical validation, where we establish if the model provides efficient medical decision support. Passing these tests indicates sufficient uncertainty reduction for practical deployment.

After outlining the general steps to design fairer and more inclusive medical decision support systems, we now focus on specific design patterns and best practices that should be followed during the design process. Good documentation is essential for successful system design because it enables reproducibility and helps to assess noise and biases, which builds trust in the system. The creation of the AI model should be described in detail, including the technical implementation and performance testing. During the deployment phase, it is crucial to validate the performance of the model in a practical setting. The processes underlying the block descriptions should be clearly explained, specifying the methods employed. Failure to provide detailed descriptions of the methods utilized constitutes a distinct limitation of any study on medical decision support, and it should be explicitly acknowledged.

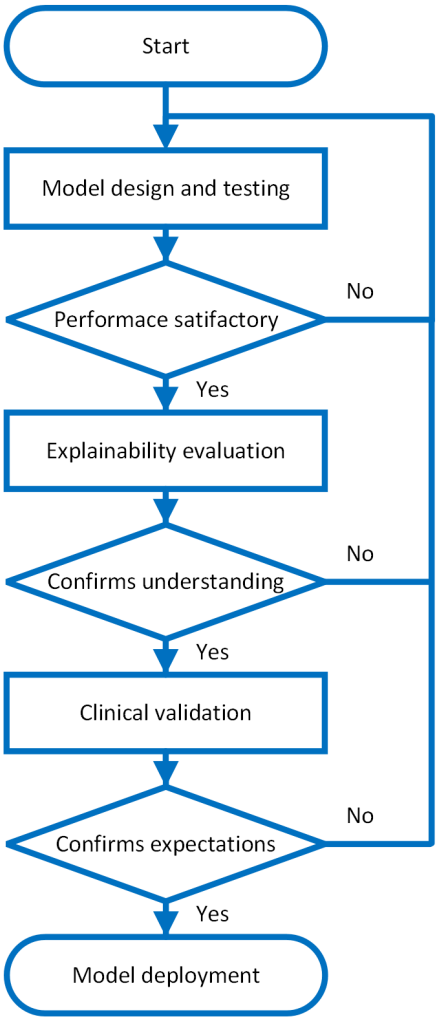

Figure 2. Flow diagram of the proposed AI model design methodology for medical decision support systems. The process consists of three major stages: (1) Model Design and Testing, which involves iterative data-driven development and empirical performance evaluation; (2) Trust Building, where explainability analysis is used to verify alignment between AI decisions and human mental models; and (3) Clinical Validation, which assesses the model's effectiveness in providing medical decision support. The feedback loops indicate that failing to meet criteria at any stage requires returning to previous stages for refinement.

We recommend incorporating the following features into the training and testing processes:

- Multicenter data: Utilizing data from multiple centers can enhance the model's generalization and make it more robust to variations in data collection protocols and equipment. Applying this technique can reduce measurement bias [71].
- Standardization of preprocessing reporting: Documenting the preprocessing steps employed in preparing the data for model training is important. Standardization of preprocessing reporting ensures the reproducibility of experiments. This can help to explain and subsequently address data-related bias.
- Annotated data from multiple operators: Multiple experts or operators in the process of annotating or labeling medical data for training machine learning models or decision support systems. This approach is utilized to mitigate bias, increase accuracy, and ensure diverse perspectives in the annotation process [72].
- Performance reporting standards: Using standard performance metrics relevant to the specific task is crucial for assessing the level of noise in a given dataset and reporting the performance accurately [73].

- Reproducibility: Sharing the source code alongside the dataset, if feasible, can enhance AI methods [74]. By openly providing the source code, researchers and practitioners can replicate and validate the results, ensuring transparency and promoting scientific rigor. Furthermore, sharing the dataset enables other researchers to evaluate and compare different algorithms on the same data, facilitating a comprehensive understanding of the methods' performance and potential biases.
- Explainability: To foster trust and confidence in the model, it is essential to provide explanations for its predictions. Methods such as LIME, SHAP, and Grad-CAM [75] can be employed for explainability [76]. Explainable AI models aid in identifying biases and understanding the decision-making process, enabling stakeholders to effectively address potential biases.
- Uncertainty quantification: Estimating uncertainty in the model's predictions is crucial, especially in medical applications where incorrect predictions can have serious consequences. Recent papers addressing uncertainty quantification [77,78] offer valuable insights into this domain.
- Continuous monitoring and evaluation: Regular monitoring of the system's performance and evaluation for potential biases is crucial [79]. Ongoing assessment can help identify and rectify any bias that emerges over time, enabling the system to adapt and improve its fairness and accuracy.

By incorporating these features, the model development process can be enhanced, resulting in more reliable and accountable AI systems in the medical domain. Figure 2 displays the development diagram of an AI model incorporating all the mitigation measures proposed in this work.

3.5. Limitations and Future Work

Some paragraphs in our position paper may appear vague due to the necessary abstraction required to address broad, systemic issues. Introducing specific technical or methodological details would lead to incompleteness and inconsistency. For instance, discussing CT-based lesion detection would necessitate focusing on a particular medical field—such as oncology, neurology, or pulmonology—each requiring detailed exploration which exceeds the scope of this manuscript. The abstraction enables us to highlight important challenges of bias and noise in biomedical AI. Historically, progress in medicine relied on standardization and education, both subject to systemic noise and bias. AI models, however, allow us to simulate "what-if" scenarios, providing some insights that might help to quantify and mitigate these challenges in contexts like medical decision support. By maintaining this level of abstraction, we aim to provide a clearer reflection on the limitations and opportunities of AI in fostering fair and inclusive biomedical solutions.

In future we anticipate that understanding data sources in terms of noise and bias will become increasingly crucial. Different measurement environments introduce varying levels of noise and bias. For example, clinical measurements of the electrical activity of the human heart using a 12-lead ECG exhibit less noise compared to a 1-lead pickup system commonly utilized for ECG measurements in home environments. However, clinical measurements are often shorter, leading to the introduction of selection bias. Therefore, future studies should prioritize quantifying the disparities between measurement environments with standardized measures for bias and noise.

The development of new AI algorithms is an ongoing process, encompassing both domain-specific and general-purpose approaches. Moving forward, it is crucial to continue exploring a wide range of AI algorithms rather than confining ourselves to dedicated solutions solely for medical decision support. For instance, it may be feasible in the future to provide medical decision support in home environments by executing lightweight AI

models on edge devices. This approach offers potential benefits, such as reducing selection bias through prolonged observation durations. However, if the edge device operates on battery power, the runtime will be influenced by the computational complexity of the AI model. Opting for a lightweight model would result in a longer runtime, thereby enhancing the usability of such solutions. An illustrative example is the development of an AI model for predicting cardiovascular risk. Incorporating federated learning enabled the use of data from multiple regions while preserving privacy. Stratified sampling helped address geographic disparities, ensuring that the model performed equitably across diverse populations.

From a medical perspective, disease-specific decision support systems within diagnostic pathways require tailored solutions with significant implications for noise and bias. Unfortunately, research exploring disease-specific bias and noise remains limited. Different performance measures can impact the trustworthiness and deployability of AI models in diverse ways, necessitating detailed examination of these aspects.

Considering individual diseases as isolated occurrences is a simplistic approximation of medical scenarios. Many clinical scenarios involve co-morbidities, which makes disease diagnosis more complex. Therefore, future studies should incorporate co-morbid data in the testing and evaluation of AI models, such that the training and testing regime reflects complex clinical scenarios. This enables more comprehensive evaluation of model capabilities and potential biases while acknowledging that multiple diseases often coexist. This framework leads to a crucial extension for medical decision support systems: the ability to identify multiple conditions simultaneously. For example, it might be feasible to identify cardiovascular diseases and sleep disorders by observing the electrical activity of the human heart. The task of multi-disease classification could be facilitated by using multi-modal data [80]. Such data can be captured with a wireless body area network that integrates a wide range of physiological sensors.

4. Conclusions

The development of fair and inclusive AI for medical decision support systems is a complex challenge that requires a multifaceted approach. Throughout this paper, we have explored the various sources of noise and bias in medical data and AI models, and proposed strategies to mitigate these issues.

We conclude that the validity and usefulness of medical data can be significantly enhanced by implementing a comprehensive set of measures such as study design, appropriate statistical analysis, transparent reporting, and promoting diverse representation in research. Moreover, our proposed design methodology, which extends beyond traditional data-driven approaches, offers a framework for creating more robust and reliable AI models for medical decision support.

Author Contributions: O.F., Conceptualization, Writing—Review and Editing, Writing—Original Draft; M.S., Conceptualization, Writing—Review and Editing, Writing—Original Draft; P.D.B., Writing—Review and Editing; S.C., Writing—Review and Editing; F.M., Conceptualization; U.R.A., Conceptualization, Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: Prabal Datta Barua was employed by Cogninet Australia. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Injadat, M.; Moubayed, A.; Nassif, A.B.; Shami, A. Machine learning towards intelligent systems: Applications, challenges, and opportunities. *Artif. Intell. Rev.* **2021**, *54*, 3299–3348. [CrossRef]
2. Bal, M.; Amasyali, M.F.; Sever, H.; Kose, G.; Demirhan, A. Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system. *Sci. World J.* **2014**, *2014*, 137896. [CrossRef] [PubMed]
3. Kilsdonk, E.; Peute, L.W.; Jaspers, M.W. Factors influencing implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis. *Int. J. Med. Inform.* **2017**, *98*, 56–64. [CrossRef] [PubMed]
4. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
5. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals: Components of a new Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, E215–E220. [CrossRef]
6. Pezoulas, V.C.; Kourou, K.D.; Kalatzis, F.; Exarchos, T.P.; Venetsanopoulou, A.; Zampeli, E.; Gandolfo, S.; Skopouli, F.; De Vita, S.; Tzioufas, A.G.; et al. Medical data quality assessment: On the development of an automated framework for medical data curation. *Comput. Biol. Med.* **2019**, *107*, 270–283. [CrossRef]
7. Warwick, W.; Johnson, S.; Bond, J.; Fletcher, G.; Kanellakis, P. A framework to assess healthcare data quality. *Eur. J. Soc. Behav. Sci.* **2015**, *13*, 92–98. [CrossRef]
8. Varachiu, N.; Karanicolas, C.; Ulieru, M. Computational intelligence for medical knowledge acquisition with application to glaucoma. In Proceedings of the First IEEE International Conference on Cognitive Informatics, Calgary, AB, Canada, 19–20 August 2003.
9. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2022**, *54*, 1–35. [CrossRef]
10. Ahmad, M.A.; Patel, A.; Eckert, C.; Kumar, V.; Teredesai, A. Fairness in Machine Learning for Healthcare. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 23–27 August 2020.
11. Johnson-Laird, P.N. Mental models and human reasoning. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18243–18250. [CrossRef]
12. Sarker, A.; Mollá, D.; Paris, C. Automatic evidence quality prediction to support evidence-based decision making. *Artif. Intell. Med.* **2015**, *64*, 89–103. [CrossRef]
13. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med.* **2015**, *13*, 1. [CrossRef] [PubMed]
14. Elston, D.M. Participation bias, self-selection bias, and response bias. *J. Am. Acad. Dermatol.* 2021, *in press*. [CrossRef] [PubMed]
15. Bland, J.M.; Altman, D.G. Statistics Notes: Measurement error and correlation coefficients. *BMJ* **1996**, *313*, 41–42. [CrossRef] [PubMed]
16. Pepe, M.S.; Janes, H.; Li, C.I.; Bossuyt, P.M. Comparative accuracy research: Methods, measures, and implications. Annual Review of Statistics and Its Application. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 291–316.
17. Stürmer, T.; Joshi, M.; Glynn, R.J.; Avorn, J.; Rothman, K.J.; Schneeweiss, S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J. Clin. Epidemiol.* **2006**, *59*, 437.e1–437.e24. [CrossRef]
18. Terry, R.; Posadzki, P.; Watson, L.K.; Ernst, E. The use of ginger (*Zingiber officinale*) for the treatment of pain: A systematic review of clinical trials. *Pain Med.* **2011**, *12*, 1808–1818. [CrossRef]
19. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **2019**, *1*, e271–e297. [CrossRef]
20. John Lu, Z.Q. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2010**, *173*, 693–694. [CrossRef]
21. Tucker, A.; Kannampallil, T.; Fodeh, S.J.; Peleg, M. New JBI policy emphasizes clinically-meaningful novel machine learning methods. *J. Biomed. Inform.* **2022**, *127*, 104003. [CrossRef]
22. Horry, M.J.; Chakraborty, S.; Pradhan, B.; Paul, M.; Zhu, J.; Loh, H.W.; Barua, P.D.; Acharya, U.R. Development of Debiasing Technique for Lung Nodule Chest X-ray Datasets to Generalize Deep Learning Models. *Sensors* **2023**, *23*, 6585. [CrossRef]
23. Gupta, S.; Gupta, A. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Comput. Sci.* **2019**, *161*, 466–474. [CrossRef]

24. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
25. Billig, J.I.; Sears, E.D.; Travis, B.N.; Waljee, J.F. Patient-reported outcomes: Understanding surgical efficacy and quality from the patient's perspective. *Ann. Surg. Oncol.* **2020**, *27*, 56–64. [CrossRef] [PubMed]
26. Wiklund, I. Assessment of patient-reported outcomes in clinical trials: The example of health-related quality of life. *Fundam. Clin. Pharmacol.* **2004**, *18*, 351–363. [CrossRef]
27. Guan, H.; Liu, M. Domain adaptation for medical image analysis: A survey. *IEEE Trans. Biomed. Eng.* **2021**, *69*, 1173–1185. [CrossRef]
28. Koerber, D.; Khan, S.; Shamsheri, T.; Kirubarajan, A.; Mehta, S. Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: A systematic review. *J. Racial Ethn. Health Disparities* **2023**, *10*, 2676–2684. [CrossRef]
29. Schulz, K.F.; Chalmers, I.; Hayes, R.J.; Altman, D.G. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* **1995**, *273*, 408–412. [CrossRef]
30. Bajpai, R.; Bajpai, S.B.R.C. Goodness of measurement: Reliability and validity. *Int. J. Med. Sci. Public Health* **2014**, *3*, 112–115. [CrossRef]
31. Rothman, K.J.; Lash, T.L.; Greenland, S. *Modern Epidemiology*, 3rd ed.; Lippincott Williams and Wilkins: Philadelphia, PA, USA, 2012.
32. Donaldson, M.S.; Corrigan, J.M.; Kohn, L.T. (Eds.) *To Err Is Human: Building a Safer Health System*; National Academies Press: Washington, DC, USA, 2000.
33. Balayn, A.; Lofi, C.; Houben, G.J. Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *VLDB J.* **2021**, *30*, 739–768. [CrossRef]
34. Snyder, C.F.; Aaronson, N.K. Use of patient-reported outcomes in clinical practice. *Lancet* **2009**, *374*, 369–370. [CrossRef]
35. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **2020**, *32*, 18069–18083. [CrossRef]
36. Nauck, D.; Kruse, R. Obtaining interpretable fuzzy classification rules from medical data. *Artif. Intell. Med.* **1999**, *16*, 149–169. [CrossRef] [PubMed]
37. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef]
38. Seker, E.; Talburt, J.R.; Greer, M.L. Preprocessing to address bias in healthcare data. In *Challenges of Trustable AI and Added-Value on Health*; IOS Press: Amsterdam, The Netherlands, 2022; p. 327.
39. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [CrossRef]
40. Marcelin, J.R.; Siraj, D.S.; Victor, R.; Kotadia, S.; Maldonado, Y.A. The impact of unconscious bias in healthcare: How to recognize and mitigate it. *J. Infect. Dis.* **2019**, *220* (Suppl. 2), S62–S73. [CrossRef]
41. Alanazi, A. Using machine learning for healthcare challenges and opportunities. *Inform. Med. Unlocked* **2022**, *30*, 100924. [CrossRef]
42. Krishnan, R.; Rajpurkar, P.; Topol, E.J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **2022**, *6*, 1346–1352. [CrossRef]
43. Jiang, T.; Gradus, J.L.; Rosellini, A.J. Supervised machine learning: A brief primer. *Behav. Ther.* **2020**, *51*, 675–687. [CrossRef]
44. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *NPJ Digit. Med.* **2020**, *3*, 119. [CrossRef]
45. Khalid, N.; Qayyum, A.; Bilal, M.; Al-Fuqaha, A.; Qadir, J. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Comput. Biol. Med.* **2023**, *158*, 106848. [CrossRef]
46. Flach, P. Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. *Proc. Conf. AAAI Artif. Intell.* **2019**, *33*, 9808–9814. [CrossRef]
47. Botchkarev, A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv* **2018**, arXiv:1809.03006.
48. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
49. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef]
50. Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.* **2022**, *15*, 5481–5487. [CrossRef]
51. Hager, G.; Wellein, G. *Introduction to High Performance Computing for Scientists and Engineers*; CRC Press: Boca Raton, FL, USA, 2010.

52. Reinke, A.; Tizabi, M.D.; Sudre, C.H.; Eisenmann, M.; Rädsch, T.; Baumgartner, M.; Acion, L.; Antonelli, M.; Arbel, T.; Bakas, S.; et al. Common limitations of image processing metrics: A picture story. *arXiv* **2021**, arXiv:2104.05642.
53. Panch, T.; Mattie, H.; Celi, L.A. The 'inconvenient truth' about AI in healthcare. *NPJ Digit. Med.* **2019**, *2*, 77. [CrossRef]
54. Yue, X.; Wang, H.; Jin, D.; Li, M.; Jiang, W. Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control. *J. Med. Syst.* **2016**, *40*, 218. [CrossRef]
55. Scheurwegs, E.; Luyckx, K.; Luyten, L.; Daelemans, W.; Van den Bulcke, T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *J. Am. Med. Inform. Assoc.* **2016**, *23*, e11–e19. [CrossRef]
56. Lin, J.-H.; Haug, P.J. Data preparation framework for preprocessing clinical data in data mining. *AMIA Annu. Symp. Proc.* **2006**, *2006*, 489–493.
57. Kashina, M.; Lenivtceva, I.D.; Kopanitsa, G.D. Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification. *Procedia Comput. Sci.* **2020**, *178*, 284–290. [CrossRef]
58. Wang, S.; Celebi, M.E.; Zhang, Y.D.; Yu, X.; Lu, S.; Yao, X.; Zhou, Q.; Miguel, M.G.; Tian, Y.; Gorriz, J.M.; et al. Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects. *Inf. Fusion* **2021**, *76*, 376–421. [CrossRef]
59. Gerke, S.; Minssen, T.; Cohen, G. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 295–336.
60. Wright, A.; Sittig, D.F.; Ash, J.S.; Bates, D.W.; Feblowitz, J.; Fraser, G.; Maviglia, S.M.; McMullen, C.; Nichol, W.P.; Pang, J.E.; et al. Governance for clinical decision support: Case studies and recommended practices from leading institutions. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 187–194. [CrossRef] [PubMed]
61. Wu, Y.; Dobriban, E.; Davidson, S. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*; PMLR: Baltimore, MD, USA, 2020; pp. 10355–10366.
62. Sutton, R.T.; Pincock, D.; Baumgart, D.C.; Sadowski, D.C.; Fedorak, R.N.; Kroeker, K.I. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digit. Med.* **2020**, *3*, 17. [CrossRef]
63. Beam, A.L.; Kompa, B.; Schmaltz, A.; Fried, I.; Weber, G.; Palmer, N.; Shi, X.; Cai, T.; Kohane, I.S. Clinical concept embeddings learned from massive sources of multimodal medical data. *Pac. Symp. Biocomput.* **2020**, *25*, 295–306.
64. Faust, O.; Hagiwara, Y.; Hong, T.J.; Lih, O.S.; Acharya, U.R. Deep learning for healthcare applications based on physiological signals: A review. *Comput. Methods Programs Biomed.* **2018**, *161*, 1–13. [CrossRef]
65. Gaonkar, B.; Cook, K.; Macyszyn, L. Ethical issues arising due to bias in training AI algorithms in healthcare and data sharing as a potential solution. *AI Ethics J.* **2020**, *1*, 1–11. [CrossRef]
66. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012.
67. Verma, S.; Ernst, M.; Just, R. Removing biased data to improve fairness and accuracy. *arXiv* **2021**, arXiv:2102.03054.
68. Muñoz Ferrandis, C.; Duque Lizarralde, M. Open Sourcing AI: Intellectual Property at the Service of Platform Leadership. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4018413 (accessed on 20 November 2024).
69. Shahriari, K.; Shahriari, M. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In Proceedings of the 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, Canada, 21–22 July 2017; pp. 197–201.
70. Wickramasinghe, C.S.; Marino, D.L.; Grandio, J.; Manic, M. Trustworthy AI development guidelines for human system interaction. In Proceedings of the 2020 13th International Conference on Human System Interaction (HSI), Tokyo, Japan, 6–8 June 2020.
71. Fallahpoor, M.; Chakraborty, S.; Heshejin, M.T.; Chegeni, H.; Horry, M.J.; Pradhan, B. Generalizability assessment of COVID-19 3D CT data for deep learning-based disease detection. *Comput. Biol. Med.* **2022**, *145*, 105464. [CrossRef]
72. Bernal, J.; Tajkbaksh, N.; Sanchez, F.J.; Matuszewski, B.J.; Chen, H.; Yu, L.; Angermann, Q.; Romain, O.; Rustad, B.; Balasingham, I.; et al. Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans. Med. Imaging* **2017**, *36*, 1231–1249. [CrossRef]
73. Hicks, S.A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M.A.; Halvorsen, P.; Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **2022**, *12*, 5979. [CrossRef] [PubMed]
74. McDermott, M.B.; Wang, S.; Marinsek, N.; Ranganath, R.; Foschini, L.; Ghassemi, M. Reproducibility in machine learning for health research: Still a ways to go. *Sci. Transl. Med.* **2021**, *13*, eabb1655. [CrossRef] [PubMed]
75. Loh, H.W.; Ooi, C.P.; Seoni, S.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Programs Biomed.* **2022**, *226*, 107161. [CrossRef] [PubMed]
76. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
77. Seoni, S.; Jahmunah, V.; Salvi, M.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Comput. Biol. Med.* **2023**, *165*, 107441. [CrossRef]

78. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **2021**, *76*, 243–297. [CrossRef]
79. Weng, S.F.; Reps, J.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **2017**, *12*, e0174944. [CrossRef]
80. Salvi, M.; Loh, H.W.; Seoni, S.; Barua, P.D.; García, S.; Molinari, F.; Acharya, U.R. Multi-modality approaches for medical support systems: A systematic review of the last decade. *Inf. Fusion* **2023**, *103*, 102134. [CrossRef]