# Building Trustworthy AI Models for Medicine: From Theory to Applications

Soumyadeep Roy
IIT Kharagpur
Kharagpur, India
soumyadeep.roy9@iitkgp.ac.in

Sowmya S. Sundaram
Stanford University
California, United States
sowmyasm@stanford.edu

Dominik Wolff
Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School
Hannover, Germany
Wolff.Dominik@mh-hannover.de

Niloy Ganguly
IIT Kharagpur
Kharagpur, India
niloy@cse.iitkgp.ac.in

## Abstract

AI is emerging as an efficient companion in medicine. While AI holds promise for reducing the cognitive load of researchers and practitioners, its adoption is often hindered by a lack of trust in new AI advancements. We present sophisticated techniques for developing trustworthy artificial intelligence (AI) models in medicine, bridging breakthroughs in AI research with practical healthcare applications. We will discuss in-depth the four stages (*Design, Development, Implementation*, and *Evaluation*) involved in the process of building trustworthy AI models customized for the medical domain. We present various techniques for incorporating important Trustworthy AI principles like data privacy, robustness, explainability, interpretability, medical experts-in-the-loop, and risk assessment while developing AI models for medicine. In contrast to prior tutorials, we make the following two key contributions: (i) While explaining the *'Implementation'* stage, we cover various real-world healthcare applications developed as part of research projects in academia in collaboration with medical schools in India and Germany. (ii) By including a health informatics professional as one of the tutorial organizers, we provide a fresh and much-needed perspective on the research challenges and mitigation strategies in building AI models for medicine.

## CCS Concepts

• **Applied computing** → *Health informatics*; • **Computing methodologies** → **Natural language processing**; *Machine learning*.

## Keywords

Trustworthy AI, Medical NLP, Knowledge Integration in Healthcare

**ACM Reference Format:**
Soumyadeep Roy, Sowmya S. Sundaram, Dominik Wolff, and Niloy Ganguly. 2025. Building Trustworthy AI Models for Medicine: From Theory to Applications. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701551.3703477



*WSDM '25, March 10–14, 2025, Hannover, Germany*

ACM ISBN 979-8-4007-1329-3/25/03
https://doi.org/10.1145/3701551.3703477

## 1 Information

**Suggested duration** Half day (3 hrs)
**Type of tutorial:** Lecture-Style
**Intended Audience:** Intermediate
**Main Contact Person:** Soumyadeep Roy

## 2 Presenters

**Prof. Niloy Ganguly** is a Professor in the Dept. of Computer Science and Engineering at IIT Kharagpur and a Fellow of the Indian Academy of Engineering. He is currently the Head of the Department of Artificial Intelligence and the Project Director of AI4ICPS at IIT Kharagpur. He is the founding member of the Complex Networks Research Group (CNeRG) of IIT Kharagpur. His research interests lie primarily in Natural Language Processing, Machine Learning, Social Computing, and Network Science. He has published in 80 journals and 200 conferences in reputed international venues related to AI and NLP. He has guided 22 Ph.D. and 9 M.S. students during this tenure.

**Dr. Sowmya S. Sundaram** is a postdoctoral scholar at Stanford University, advised by Prof. Mark Musen on improving metadata of medical datasets using NLP techniques. Previously, she was a postdoctoral scholar at L3S Research Center under the mentorship of Prof. Wolfgang Nejdl, where she worked with physicians, clinical informaticians, and NLP researchers to work on interdisciplinary research. She received her PhD from IIT Madras where she worked on mathematical question answering. Her overarching interest has been in NLP applications and alignment in the domains of mathematical reasoning, medical reasoning, fairness, and alignment with publication footprints in AMIA, ECAI, AAAI, EDBT, DMKD journal, ESWC, and so on. She delivered a tutorial on 'ML for Automatic Word Problem Solving' at ECML-PKDD 2019.

**Dr. Dominik Wolff** leads the junior research group *iXplain_CDS* at Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Germany. Dominik holds a PhD from TU Braunschweig, where he focused on recommendation systems, the incorporation of tacit knowledge, and the hybridization of knowledge and data-driven models. His research focuses

on explainability and interoperability in clinical decision support and biomarker identification. He focuses on synergies that arise from the combined expertise of artificial and human experts and the evaluation of AI-based systems in medicine to close the implementation gap. He is the author of over 20 scientific publications. He has extensive teaching experience and has been awarded five teaching prizes in the field of *Artificial Intelligence in Biomedicine*.

**Soumyadeep Roy** is a fifth-year Ph.D student in the Department of Computer Science and Engineering at IIT Kharagpur. His research interests lie in Natural Language Processing and Generative AI, with a focus on medical NLP applications. He has spent 2.5 years during his Ph.D working as a Research Associate with Prof. Wolfgang Nejdl at the Leibniz AI Future Lab, L3S Research Center, Germany, focused on applying artificial intelligence to personalized medicine applications, in collaboration with the Hannover Medical School, Germany. His work has been published in prestigious venues like EMNLP, IJCAI, SIGIR, ECAI, CIKM, IEEE ICDH, WebSci, and ACM Transactions of the Web. He has been a Teaching Assistant for various courses at IIT Kharagpur and Leibniz University Hannover, such as Natural Language Processing and Information Retrieval.

## 3 Motivation

Building trustworthy AI models for medicine is imperative as it directly impacts the quality and safety of healthcare delivery. Trustworthy AI models are essential for ensuring that AI applications in healthcare are reliable, transparent, and ethically sound.

**Relevance to WSDM.** The Web research community and AI in medicine share challenges, such as managing large datasets, improving searchability, and ensuring interpretability, but the contexts differ. Web research prioritizes optimizing search and information retrieval, while healthcare AI ensures accurate, trustworthy outcomes for critical medical decisions. Both areas rely on interdisciplinary collaboration and face concerns around bias and ethics, though healthcare operates under more stringent regulations due to its direct influence on patient care. The deployment of AI models in clinical settings must consider the medico-legal implications and economic factors, as well as the potential for automation bias, which can affect the vigilance of healthcare providers [16]. Developing high-performing, interpretable, transparent AI models is essential to building trust among clinicians and patients [1, 7]. Furthermore, the use of AI in healthcare is expanding across various specialties, such as radiology and dermatology, necessitating a comprehensive understanding of AI methodologies and their ethical implications [7]. Overall, the journey from theory to applications in building trustworthy AI models in medicine involves addressing key challenges related to data quality and availability, model validation, acceptance, and the integration of AI into existing clinical workflows while ensuring that these models are accountable and inclusive.

**Expected Outcomes.** Our proposed tutorial aims to bridge the gap between theory and practice, offering actionable insights and strategies for leveraging AI models (not just LLMs, in contrast to most prior tutorials) to improve healthcare outcomes. We present real-world case studies, mostly from university projects related to developing trustworthy AI models for healthcare, of which the tutorial organizers were a part, from IIT Kharagpur, India, and L3S Research Center and Hannover Medical School, Germany. The tutorial organizers have almost five years of experience in this space. We will present our findings in the form of checklists and project frameworks that may be used by other researchers working on similar research projects. Additionally, we will present five case studies of real-world AI in medicine projects and describe the research challenges, design choices, and solutions. These projects involve extensive collaboration with medical centers and health informatics professionals. Therefore, from a health informatics background, Dr. Dominik Wolff is part of this tutorial and has the necessary background to provide a much-needed perspective on this emerging, highly interdisciplinary topic.

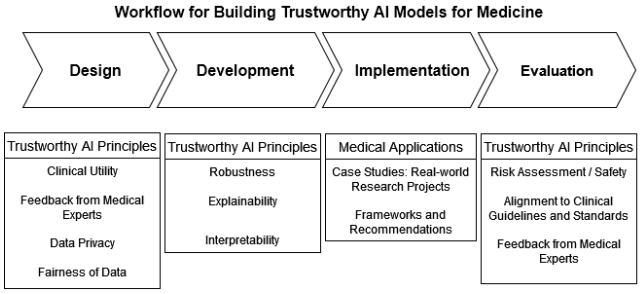


**Figure 1: Tutorial Outline**

**Target Audience and Prerequisites.** This interdisciplinary tutorial aims to engage computer scientists, artificial intelligence researchers, and health informatics professionals who are interested in learning recent developments in advanced, Trustworthy AI techniques in healthcare settings. Although participants possessing expertise in these domains would have the most significant advantage, the tutorial is structured to deliver an exhaustive overview of AI within healthcare, ensuring no particular foundational or background knowledge is required.

## 4 Tutorial Details

**Tutorial Schedule**

- Introduction to Trustworthy AI in Medicine (10 mins)
- Building Trustworthy AI Models for Medicine
  - Design Stage (40 minutes)
  - Development Stage (40 minutes)
  - Implementation Stage (40 minutes)
  - Evaluation Stage (40 minutes)
- Conclusion and Open Questions (10 mins)

**Resources for the Audience:** We will provide a reading list of important papers to the audience in the area of developing trustworthy AI models in medicine, which are organized into sub-topics aligning with the topics covered in the tutorial. We will provide a list of recommendations in the form of checklists and project frameworks that the research community may use.

### 4.1 Tutorial Outline

The tutorial will cover the four stages of building trustworthy AI models for medicine - *Design, Development, Implementation,* and *Evaluation*. Figure 1 further mentions the Trustworthy AI principles

associated with each stage of the pipeline, specifically tailored to the healthcare setting.

*4.1.1 Design Stage.* Here, we will cover the trustworthy AI principles involving the following project activities such as (i) Problem formulation and feasibility analysis in a **medical experts-in-the-loop** manner, (ii) **Data privacy** and usage issues, specifically sensitive patient data (clinical and genomic), while maintaining **fairness of data** at the same time, (iii) Supplementary resources available in open-domain such as MetaMap, SemRep, bibliographic databases such as Pubmed and its associated citation network.

*4.1.2 Development Stage.* Here, we will cover various modeling techniques to improve the **robustness** of open-domain AI models, including LLMs, in downstream medical NLP tasks. For example, we will discuss supervised and unsupervised domain adaptation techniques. Domain adaptation techniques focus on the challenges of adapting AI models like large language models (LLMs) to expert domains such as medicine, where issues like high vocabulary mismatch, limited labeled data, and the need to integrate structured domain knowledge into models arise. Standard domain adaptation techniques often fall short in this context; necessitating novel approaches customized for the medical domain.

The 'black box' nature of many deep learning and generative AI models are particularly problematic in medical contexts, where understanding the reasoning behind a recommendation is often as important as the recommendation itself. **Explainability** and **interpretability** of the calculated recommendations, including comprehensibility of the underlying algorithmic logic, are essential for trust in and acceptance of such decision support. We explore state-of-the-art methods for creating explainable AI systems, focusing on techniques that can provide interpretable insights to healthcare professionals.

*4.1.3 Implementation Stage.* We will present real-world research projects from related university projects as **case studies** where the organizers of this tutorial were personally involved and can thus provide an in-depth perspective. The projects range from Parkinson's Disease patient subtyping to cochlear implant data analysis, children's brain maturity estimation, and dental implants. We demonstrate how these theoretical advancements translate into practical medical care and research improvements. Some common themes are the incompatibility of the latest deep learning models with low dataset size, the lack of interpretability, and the cost of utilizing multiple biomarkers. This section describes the best practices in eliciting requirements and designing AI applications with trustworthiness in mind.

*4.1.4 Evaluation Stage.* In the tutorial, state-of-the-art evaluation protocols that go beyond standard automated metrics like accuracy and F1 score are presented, to show how the **risk assessment** or **safety** principle of trustworthy AI is implemented. If implemented correctly, they improve the trust of medical staff and patients and are vital for implementing AI in the clinical routine. **Alignment** to clinical guidelines or standards explores the importance of ensuring that AI outputs align with established medical guidelines and best practices, given the potential for LLMs to generate hallucinations. This is crucial for building AI systems that can be trusted in clinical decision-making.

## 4.2 Selected Topics

Here, we provide a detailed description of a subtopic from each of the four stages involved in building Trustworthy AI models for medicine - *Design, Development, Implementation*, and *Evaluation*, as shown in Figure 1.

*4.2.1 Design Stage: Designing Trustworthy Clinical Decision Support Systems.* Clinical decision support systems can assist medical professionals in specific decision-making tasks, such as diagnosis or treatment planning, as they can rapidly process large amounts of clinical (and non-clinical, e.g. patient-generated) data. The explainability and interpretability of the calculated recommendations, including comprehensibility of the underlying algorithmic logic, are essential for the trust in and acceptance of such decision support. Conversely, widespread soft computing models are typical "black boxes" with opaque underlying prediction reasoning. Applying post-hoc explainable AI (XAI) methods to these models aims to make the reasoning interpretable for humans.

Here, we will highlight stages in the implementation cycle where XAI can increase trustworthiness and depict state-of-the-art medical examples. On the other hand, we will discuss how XAI holds the risk of strengthening trust in false reasoning, and still, humans' mental work is needed to comprehend the explanation of the reasoning. Another critical aspect of trustworthy clinical decision support lies in evaluating the systems developed. These are strategies that, if implemented correctly, can improve the trust of medical staff and patients in the AI systems. We will present and discuss state-of-the-art evaluation protocols, which go beyond commonly used metrics and are a vital factor for implementing AI in the clinical routine. In this context, we will discuss in which scenarios a human in the loop is beneficial and which synergies can arise from combined artificial and human expertise. For example, the necessity of high-quality evaluations will be discussed in light of the vast implementation gap of AI in medicine centering on the patient's well-being.

*4.2.2 Development Stage: Improving Robustness based on Domain Adaptation Techniques.* Domain adaptation techniques help to address the 'domain shift' issue because the training data distribution of open-domain models differs significantly from the downstream datasets of the medical domain. Supervised domain adaptation techniques leverage available labeled data from the target domain or task, while unsupervised domain adaptation methods do not require any labeled data from the target domain or task. The thesis investigates two primary target domains: (i) the medical NLP domain, focusing on English-based downstream tasks, and (ii) the genomic domain, centered on DNA sequences composed of nucleotides. Within these domains, we will explore two broad categories of downstream tasks: (i) query and response understanding, encompassing medical forum question classification [10], medical question-answering [12], medical text summarization [2] and clinical trial search [11], and (ii) gene sequence classification [13] tasks, including splice sites prediction and the identification of gene regulatory elements in both human and non-human species such as Yeast and Mouse.

*4.2.3 Implementation Stage: Case Studies of Medical Applications.* We will present a list of relevant university projects that the tutorial organizers were part of, spanning around five years — (i) Decision

tree-based robust patient subtyping of Parkinson's Disease, (ii) Classifying the natural language notes presented by the physician for Cochlear implants, (iii) CDS for brain maturity estimation in small children based on MRI, and (iv) Patient education via chatbot that learns from user feedback. We will first describe the problem statement for each case study or healthcare application and then explain the process of building trustworthy AI models to solve the research problem. The study emphasizes the importance of domain adaptation, robust evaluation, interpretability, and alignment in creating reliable AI models.

*4.2.4 Evaluation Stage: Enforcing Alignment to Clinical Standards and Guidelines.* Studies show that LLMs [14, 18] struggle with medical fact-checking as LLMs are trained on text that provides contradictory pieces of information. One of the endeavors in this space is to train LLMs to develop fact-checking mechanisms from pieces of text and provide evidence for the same[17]. Our initial analysis suggests LLMs may inject some knowledge outside the purview of the provided text for inference. Initial efforts in this space [4] include fashioning a detailed rubric for evaluating fact-checking based on human experts. In another related application, analyzing the metadata of medical datasets reveals a gap in alignment[3]. Metadata for datasets needs to be designed ideally to enhance search. Data scientists spend significant time and effort to comb through datasets that have been assigned poorly designed metadata [8] to build their cohort of interest. We encountered alignment issues while deploying LLMs for metadata correction to conform with the metadata guidelines. The language models can generate linguistically appropriate metadata but cannot conform to restrictions to ontologies and other structural requirements. With the help of a detailed template in the prompt, we could nudge the LLM to bring in closer alignment [15].

## 5 Related Works

In recent years, integrating large language models (LLMs) into healthcare has garnered significant attention, as evidenced by several notable tutorials and research efforts. Poon et al. [9] provided a comprehensive overview of the application of foundation models in precision health and focused on how these models can enhance healthcare delivery by integrating novel information seamlessly despite the challenges posed by extensive unstructured data and manual processing. Zang et al. [19] underscored the practical applications of NLP in extracting valuable insights from electronic health records (EHRs), demonstrating the potential of LLMs to improve patient care through efficient data processing. Kim et al. [5] discussed the latest trends and methodologies in applying LLMs to healthcare data, focusing on the significant impact these models have on medical text analysis and the extraction of actionable insights. Koyejo and Li [6] delved into the theoretical aspects of building trustworthy LLMs. Unlike our proposed tutorial, which emphasizes practical applications and guidelines, this tutorial focused on the capabilities of LLMs and the theoretical foundations of trustworthiness. It examines the challenges and solutions to ensuring the reliability and ethical use of LLMs in real-world applications. These tutorials underscore the importance of practical guidelines and real-world applications in deploying LLMs in healthcare.

## References


[1] Ahmed Shihab Albahri, Ali Faisal Mohammed, Mohammed A. Fadhel, et al. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* 96 (2023), 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

[2] Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. MEDVOC: Vocabulary Adaptation for Fine-tuning Pre-trained Language Models on Medical Text Summarization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, 6180–6188. https://doi.org/10.24963/ijcai.2024/683 Main Track.

[3] Rafael S Gonçalves and Mark A Musen. 2019. The variable quality of metadata about biological samples used in biomedical experiments. *Scientific data* 6, 1 (2019), 1–15.

[4] Sebastian Joseph, Lily Chen, Jan Trienes, et al. 2024. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 8437–8464. https://aclanthology.org/2024.acl-long.459

[5] Yunsoo Kim, Jinge Wu, and Honghan Wu. 2024. Healthcare Text Analytics in the Era of Large Language Models. Tutorial presented at the 7th Healthcare Text Analytics Conference. Institute of Health Informatics, University College London.

[6] Sanmi Koyejo and Bo Li. 2024. Towards Trustworthy Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) *(WSDM '24)*. 1126–1127. https://doi.org/10.1145/3616855.3636454

[7] Pranjal Kumar, Siddhartha Chauhan, and Lalit Kumar Awasthi. 2023. Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions. *Engineering Applications of Artificial Intelligence* 120 (2023), 105894–105894. https://doi.org/10.1016/j.engappai.2023.105894

[8] Mark A Musen. 2022. Without appropriate metadata, data-sharing mandates are pointless. *Nature* 609, 7926 (2022), 222–222.

[9] Hoifung Poon, Tristan Naumann, Sheng Zhang, and Javier González Hernández. 2023. Precision Health in the Age of Large Language Models. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. 5825–5826. https://doi.org/10.1145/3580305.3599568

[10] Soumyadeep Roy, Sudip Chakraborty, Aishik Mandal, et al. 2021. Knowledge-Aware Neural Networks for Medical Forum Question Classification. In *Proceedings of the 30th ACM International Conference on Information &; Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. 3398–3402. https://doi.org/10.1145/3459637.3482128

[11] Soumyadeep Roy, Niloy Ganguly, Shamik Sural, and Koustav Rudra. 2023. Interpretable Clinical Trial Search using Pubmed Citation Network. In *2023 IEEE International Conference on Digital Health (ICDH)*. 328–338. https://doi.org/10.1109/ICDH60066.2023.00056

[12] Soumyadeep Roy, Aparup Khatua, Fatemeh Ghoochani, et al. 2024. Beyond Accuracy: Investigating Error Types in GPT-4 Responses to USMLE Questions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1073–1082. https://doi.org/10.1145/3626772.3657882

[13] Soumyadeep Roy, Jonas Wallat, Sowmya S. Sundaram, et al. 2023. GENEMASK: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland (Frontiers in Artificial Intelligence and Applications, Vol. 372)*. 2002–2009. https://doi.org/10.3233/FAIA230492

[14] Lichao Sun, Yue Huang, Haoran Wang, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* (2024).

[15] Sowmya S Sundaram, Benjamin Solomon, Avani Khatri, et al. 2024. Use of a Structured Knowledge Base Enhances Metadata Curation by Large Language Models. *arXiv preprint arXiv:2404.05893* (2024).

[16] Hari Trivedi and Judy Gichoya. 2024. Breathing Life Into Artificial Intelligence. *Critical Care Medicine* 52, 2 (2024), 345–348. https://doi.org/10.1097/ccm.0000000000006124

[17] Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, 8095–8107. https://aclanthology.org/2024.lrec-main.709

[18] Kevin Wu, Eric Wu, Ally Cassasola, et al. 2024. How well do LLMs cite relevant medical references? An evaluation framework and analyses. arXiv:2402.02008 [cs.CL] https://arxiv.org/abs/2402.02008

[19] Chengxi Zang, Weishen Pan, and Fei Wang. 2023. Mining Electronic Health Records for Real-World Evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. 5837–5838. https://doi.org/10.1145/3580305.3599566