University of Dundee

# A historical perspective of biomedical explainable AI research

Malinverno, Luca; Barros, Vesna; Ghisoni, Francesco; Visonà, Giovanni; Kern, Roman; Nickel, Philip J.

# Patterns

## Perspective

# A historical perspective
# of biomedical explainable AI research

Luca Malinverno,[1,14] Vesna Barros,[2,3,14,*] Francesco Ghisoni,[1,14] Giovanni Visonà,[4] Roman Kern,[5,6] Philip J. Nickel,[7] Barbara Elvira Ventura,[1] Ilija Šimić,[6] Sarah Stryeck,[8] Francesca Manni,[9] Cesar Ferri,[10] Claire Jean-Quartier,[11] Laura Genga,[7] Gabriele Schweikert,[12] Mario Lovrić,[6,13] and Michal Rosen-Zvi[2,3]

[1]Porini SRL, Via Cavour, 222074 Lomazzo, Italy
[2]AI for Accelerated Healthcare & Life Sciences Discovery, IBM R&D Laboratories, University of Haifa Campus, Mount Carmel, Haifa 3498825, Israel
[3]The Hebrew University of Jerusalem, Ein Kerem Campus, 9112102, Jerusalem, Israel
[4]Empirical Inference, Max-Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
[5]Institute of Interactive Systems and Data Science, Graz University of Technology, Sandgasse 36/III, 8010 Graz, Austria
[6]Know-Center GmbH, Sandgasse 36/4A 8010, Graz, Austria
[7]Eindhoven University of Technology, 5135600 MB Eindhoven, The Netherlands
[8]Research Center Pharmaceutical Engineering GmbH, Inffeldgasse 138010 Graz, Austria
[9]Philips Research, HTC 4, 5656 AE Eindhoven, The Netherlands
[10]VRAIN, Universitat Politècnica de València, Camino de Vera, s/n 46022 Valencia, Spain
[11]Research Data Management, Graz University of Technology, Brockmanngasse 84, 8010 Graz, Austria
[12]School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK
[13]Centre for Applied Bioanthropology, Institute for Anthropological Research, 10000 Zagreb, Croatia
[14]These authors contributed equally
*Correspondence: vesna.barros@ibm.com
https://doi.org/10.1016/j.patter.2023.100830

---

**THE BIGGER PICTURE** Understanding the inner working of machine-learning models has become a crucial point of discussion in fairness and reliability of artificial intelligence (AI). In this perspective, we reveal insights from recently published scientific works on explainable AI (XAI) within the biomedical sciences. Specifically, we speculate that the COVID-19 pandemic is associated with the rate of publications in the field. Current research efforts seem to be directed more toward explaining black-box machine-learning models than designing novel interpretable architecture. Notably, an inflection period in the publication rate was observed in October 2020, when the quantity of XAI research in biomedical sciences surged upward significantly. While a universally accepted definition of explainability is unlikely, ongoing research efforts are pushing the biomedical field toward improving the robustness and reliability of applied machine learning, which we consider a positive trend.

**1 2 3 4 5**   **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

---

## SUMMARY

The black-box nature of most artificial intelligence (AI) models encourages the development of explainability methods to engender trust into the AI decision-making process. Such methods can be broadly categorized into two main types: post hoc explanations and inherently interpretable algorithms. We aimed at analyzing the possible associations between COVID-19 and the push of explainable AI (XAI) to the forefront of biomedical research. We automatically extracted from the PubMed database biomedical XAI studies related to concepts of causality or explainability and manually labeled 1,603 papers with respect to XAI categories. To compare the trends pre- and post-COVID-19, we fit a change point detection model and evaluated significant changes in publication rates. We show that the advent of COVID-19 in the beginning of 2020 could be the driving factor behind an increased focus concerning XAI, playing a crucial role in accelerating an already evolving trend. Finally, we present a discussion with future societal use and impact of XAI technologies and potential future directions for those who pursue fostering clinical trust with interpretable machine learning models.

## INTRODUCTION

The COVID-19 pandemic has accelerated medical research and the way medical care is provided. Vaccines were developed and approved in record time,[1] and novel drugs were rapidly designed for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection.[2] Simultaneously, COVID-19 has caused major disruptions and backlogs in health systems, leaving many millions of people without care in the EU[3] and around the world.[4] As a direct response to the mounting pressures, primary and community care underwent a significant transformation, accelerating the use of remote consultation,[5] automatic triage,[6] and virtual monitoring and care.[7,8] In order to decrease the backlog and restore and improve equitable healthcare, a combination of strategies have been identified, including the efficient adaptation of new technologies and digital solutions, additional funding, sufficient workforce, and infrastructure improvements.[3] Even pre-COVID-19, information technology was suggested to be a key driver for higher quality, better effectiveness, and more efficiency in health systems,[9] and several studies have found that computational models can be on par with or outperform human experts in certain diagnostic and prognostic tasks.[10] However, while artificial intelligence (AI) technologies have become a growth engine in many other industries, development and adoption of such technologies across the health ecosystem have been much slower.[11] During the pandemic, many AI research teams have stepped up their efforts in this application area, and hundreds of predictive tools to combat COVID-19 have been devised and, in some cases, applied in the clinic. Their potential and overall clinical relevancy are currently under critical discussion in several recent reviews.[12–14] The opacity and black-box nature of many state-of-the-art AI systems additionally raise concerns in clinicians, care providers, and regulators and prevent their rapid integration into the high-stakes clinical decision-making process.[15]

It has been argued that explainable AI can engender transparency and increase adoption; however, the term "explainable AI" (XAI) is still not a well-defined concept. As discussed in Miller,[16] explainability or interpretability is how human comprehensible the decisions of an AI system are, which is commonly referred to as XAI. The concept of XAI dates back to the mid-1970s, while the term itself is newer. It was coined in 2004, but only since 2016 has XAI received significant attention,[17] including discussions of what its underlying concept is,[18] the role of XAI in trustworthiness,[17,19,20] and its importance in the biomedical domain.[21,22]

In recent years, a considerable number of review and perspective papers have been published with the goal of organizing the field of XAI (see, e.g., Adadi and Berrada,[17] Barredo Arrieta et al.,[23] Das and Rad,[24] and Vilone and Longo,[25] and references therein). Yet, only a few studies have focused on XAI in the field of medicine and health (see, e.g., Jiménez-Luna et al.,[26] Tjoa and Guan,[27] and Loh et al.[28]). Conversely, we have not found a study that focuses on how the field of biomedical XAI has changed in recent years. In this perspective, we provide a historical and conceptual perspective of the evolution of XAI research within biomedical sciences. We aim at giving an overview of the state-of-the-art XAI, the trends observed in the last few years, and the current explainability techniques. We hypothesize that the interest of the community is not primarily in the development of new XAI methods but rather in applying existent techniques to explain their black-box models. We hope that this work benefits researchers from various fields, serving as a reference point for those who want to get a deeper understanding of the direction in which XAI is going. The perspective is organized as follows: it starts with a discussion of the scope of XAI in the context of biomedicine and health. This is followed by a historical perspective, a review of the change in the extent of research performed over the years, and a conceptual perspective, a mapping of the research into five different types of studies. Finally, the perspective ends with emerging themes and open research areas in the field.

## BIOMEDICAL XAI: OVERVIEW OF MEANING AND SCOPE

The ultimate aim of XAI is to explain why a model produced a specific result. Such a goal can be achieved in a variety of manners, ranging from post hoc explanations to inductive biases that constrain model architectures. A particularly compelling framework to consider within this setting is the study of causality, through which we can move from observational explanations to a more in-depth analysis of the model in question, in the form of interventional and counterfactual interrogation. When a causal model is available, not only can we attempt to diagnose the working of the model, but we can also ask questions such as "how would this prediction change if a specific variable had a different value?" As a stricter framework, causality offers compelling tools to explain the workings of machine-learning models, and as such, we opted to include publications relevant to causality in biomedicine in the present meta-analysis. Undoubtedly, counterfactuals have become an essential part of XAI, enabling researchers to explain past outcomes and predict future events by identifying causal relations in the data.[29,30] Thus, causal machine learning plays a central role in the biomedical domain.[31,32] For instance, learning the optimal medication or therapy for a patient requires carefully curated data or causal effect estimation, which classic machine-learning models do not necessarily identify.[33] Building causal models of the world can support the understanding of why a model selects a certain intervention as an optimal one, which in turn can help the AI developer in improving the technology, ultimately leading to an expansion of the medical knowledge. Notably, while XAI helps build user trust in the AI systems and provides useful insights for AI developers, trustworthy AI is a different concept. Trustworthiness encompasses not only technologies but also practices, and it has been institutionalized in policy documents guiding AI innovation, such as the EU's High Level Expert Group on AI. To develop a historical view and a perspective on biomedical XAI, we started by reviewing citations from PubMed that fit the above scope definition.

PubMed is a free database of citations of publications in the biomedical domain that was developed by the National Center for Biotechnology Information (NCBI). It is a major source of literature analysis for biomedical researchers due to its comprehensiveness and accessibility. We performed a systematic review of PubMed papers following the recommendations of the PRISMA statement.[34] We applied manual and automatic methods as
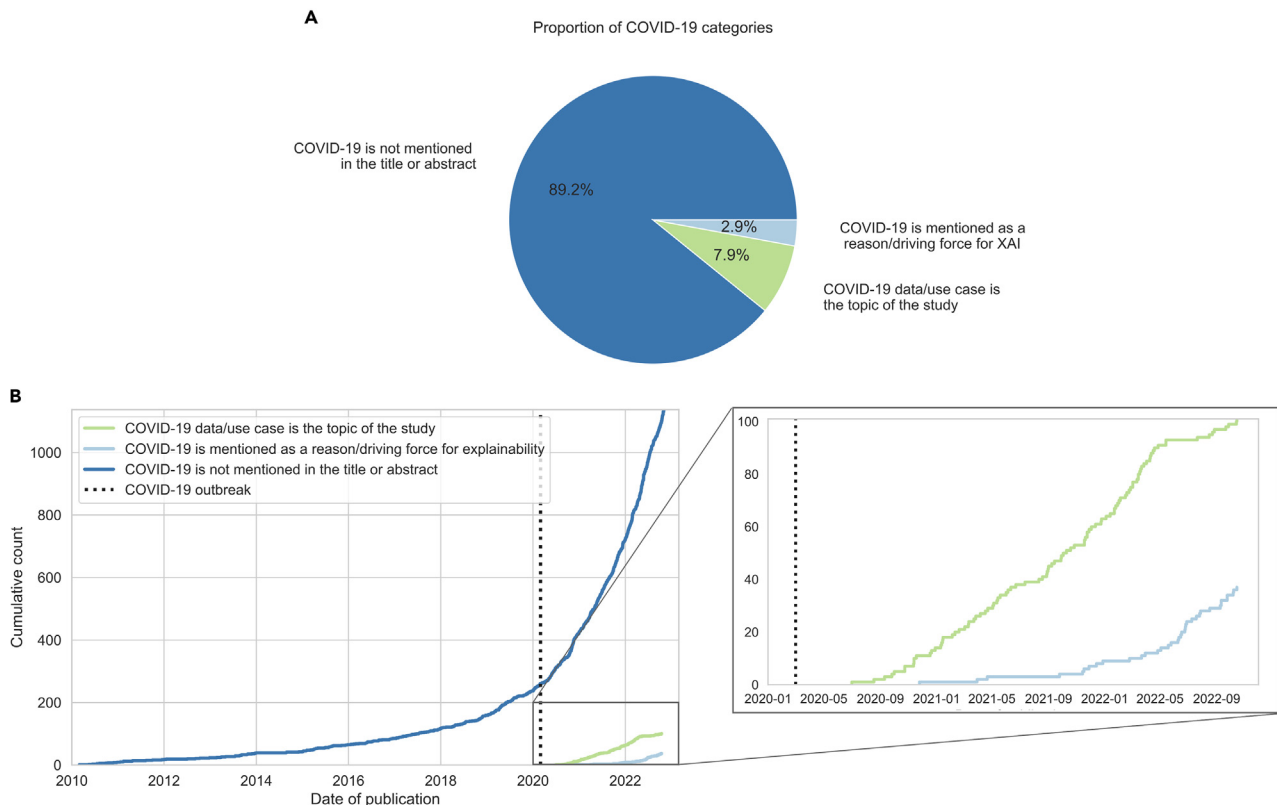
**Figure 1. Trends in XAI papers related to COVID-19**
(A) Proportions of COVID-19 categories in the dataset.
(B) Cumulative number of XAI papers for each category throughout time. The increase in COVID-19-related papers, represented by the light blue and light green curves, set off around 7 months after the pandemic onset.

illustrated in Figure S1; that is, we searched in the PubMed database for all citations until November 3, 2022, with no restrictions on regions, languages, or publication types. Citations were added to our cohort if they contained biomedical concepts in the abstracts or titles, which were automatically extracted based on a prespecified set of keywords (Table S1; Data S1). We excluded studies that were not related to AI and limited the search space with a start date of January 1, 2010 (defined based on the step change of significant attention to the advances of deep-learning technologies). Lastly, we included only papers that were related to XAI, considered as the wide term that includes concepts of the two broad and overlapping subjects discussed above: explainability and causal analysis. The complete information regarding the citation collection criteria used in the data collection process is illustrated in Figure S2. Manual review of the abstracts complemented the automatic extraction. The result is a cohort of 1,276 peer-reviewed papers that study biomedical XAI. In the next section, we discuss research trends that we found in our analysis.

## BIOMEDICAL XAI: A HISTORICAL PERSPECTIVE THROUGH THE LENS OF COVID-19

To investigate the trend of biomedical XAI and the possible role of the COVID-19 pandemic in bringing this field to the forefront of AI research, we used the publication date and constructed a time series defined by the number of biomedical XAI PubMed papers published per month. We focused on the date of COVID-19 outbreak, which was considered March 1, 2020. 261 abstracts on biomedical XAI were published before the outbreak and 1,017 after. After careful selection of studies and iterative manual revision of abstracts, we investigated whether publications related to causality and explainability showed the same trends pre- and post-COVID-19 (Figure S3). There was a much stronger growth in the rate of explainability publications than in causality, which could possibly be due to the difficulties in performing causal inference versus how convenient explaining an existing model is.

We then manually annotated the role of COVID-19 in each paper, where each annotator had three options to label a study: (1) studies that do not mention the pandemic in their titles or abstracts, (2) studies that used COVID-19 data (e.g., chest radiographs or computed tomography [CT] scans, number of SARS-CoV-2 infections, or administrated vaccines) to predict clinical outcomes, or (3) studies that mentioned COVID-19 as a reason or driving force for their research. Expectedly, the distribution of papers per label was uneven (Figure 1A): 1,137 studies (approximately 90%) did not mention the pandemic, whereas 101 studies (8%) used COVID-19 data and 37 studies mentioned COVID-19 as a driving force for their research. Interestingly, COVID-19-related publications significantly increased several months after the pandemic outbreak (Figure 1B, zoomed-in
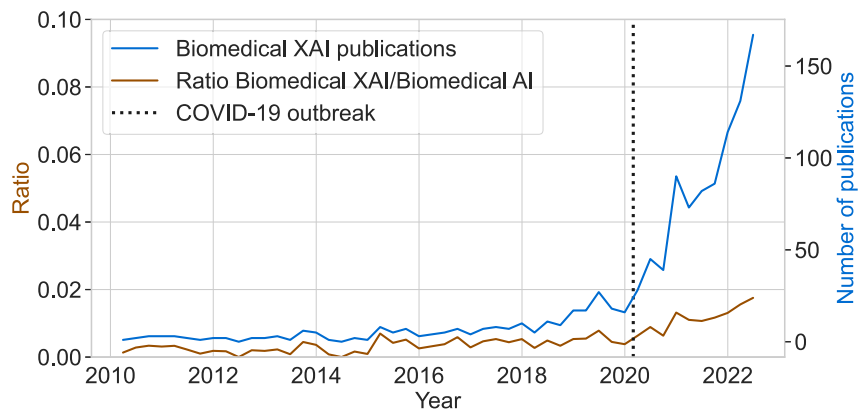
image), possibly representing the time period for researchers to collect enough data, design the studies, and start publishing them.

As an exploratory analysis, we investigated whether a change point is detected at the time of the pandemic outbreak or in the following months (Data S2). We found an inflection period in October 2020, when the quantity of XAI research in biomedical sciences surged upward significantly (Figure S4A). Pinpointing a specific cause for this acceleration is far from trivial. In the past 10 years, XAI has received increased attention in the form of funding, research programs, and legal and ethical requirements. Examples include the DARPA program[35] or the European Commission's White Paper on AI.[36] However, we speculate that the advent of the COVID-19 pandemic itself could have been an important factor in bringing the field of XAI to the forefront in many AI-related research fields. The well-documented difficulties of AI (lack of impact of AI) in COVID-19 analysis could have motivated the need to better understand how models work and the origin of predictions.[12,13] It is worth noting that across the most cited papers we reviewed, often the leading author (last co-author) was an AI researcher that expanded their horizon to the applied biomedical field.[37,38]

We then attempted to quantify the change observed and estimated that the field of biomedical XAI was pushed forward by about 25 months (Figure S4B). However, establishing the causal nature of this link would require a large-scale analysis of multiple research topics trends, which is beyond the scope of the current meta-analysis.

Finally, we compared the numbers of biomedical XAI publications to the ratio between biomedical XAI and biomedical AI publications a long time (Figure 2). We found an exponential growth of biomedical XAI papers, which is followed by an increase in the ratio of biomedical XAI/biomedical AI papers. This is a faster exponential growth than the one detected in publications of AI in biomedical sciences in previous studies,[39] as reflected by the constant increase in the ratio of biomedical XAI/biomedical AI papers. Importantly, the rapid increase in XAI publications is not only due to the increase in scientific literature using COVID-19 data (see, for instance, the exponential growth of the use of medical imaging and AI in the context of COVID-19[40]). On the contrary, our manual review showed that only 8% of the biomedical XAI papers were devoted to analysis of COVID-19 data, whereas 3% (37 papers) referred

to COVID-19 as a motivation to discuss or apply XAI methods. This further corroborates our hypothesis that the advent of the COVID-19 pandemic acted as a catalyst and contributed to pushing the level of interest in XAI methodologies to a critical point in an already evolving trend.

The papers list, the manual annotation, and the code used for all analysis can be found in Fghisoni and Visonà.[41]

## BIOMEDICAL XAI: STUDY TYPES

In order to get a more comprehensive view of the type of research performed in biomedical XAI, each abstract was annotated according to the type of study reported in the paper. We followed the approach discussed in Vilone and Longo[25] and labeled the abstracts with one of the following categories that best described the study: (1) review or meta-analysis, (2) discussion of XAI concepts, (3) introduction of novel XAI methods, or (4) evaluation or application of XAI methods and added a fifth category, (5) datasets or tools that support XAI.

In total, 642 papers (approximately 50% of the eligible studies) evaluated or applied existing XAI methods in their respective studies, followed by 214 papers (17%) that introduced a novel XAI technology (either by innovating the way they explain their AI algorithms' decisions or by adapting previous XAI ideas to a new research context; Figure 3A). Letters, comments, and narratives discussing XAI technologies comprised 13% of the cohort. Datasets or tools that support XAI technologies (e.g., software packages, AI frameworks and workflows, medical image datasets) and reviews on XAI methods composed together approximately 20% of the entire cohort.

We then tracked the Google Scholar citations of all eligible publications, considered as a proxy for public interest, and computed the number of citations per month since the date of publication (Figure 3B). As expected, review papers and novel methods attract most attention, as reflected by average citations per month, whereas application and evaluation of methods draw less attention. The relatively high average number of citations of papers with conceptual discussion is somewhat surprising and possibly reflects the heated debate on the role of XAI and its importance in the biomedical domain.

## BIOMEDICAL XAI: FUTURE PROSPECTS

The debate around XAI is an ongoing effort that requires input from experts in computer science, philosophy, ethics, and law. The largest challenges arise from the difficulty of defining and quantifying "explainability": how do we decide if the output of a model is interpretable? A very deep decision tree model, for
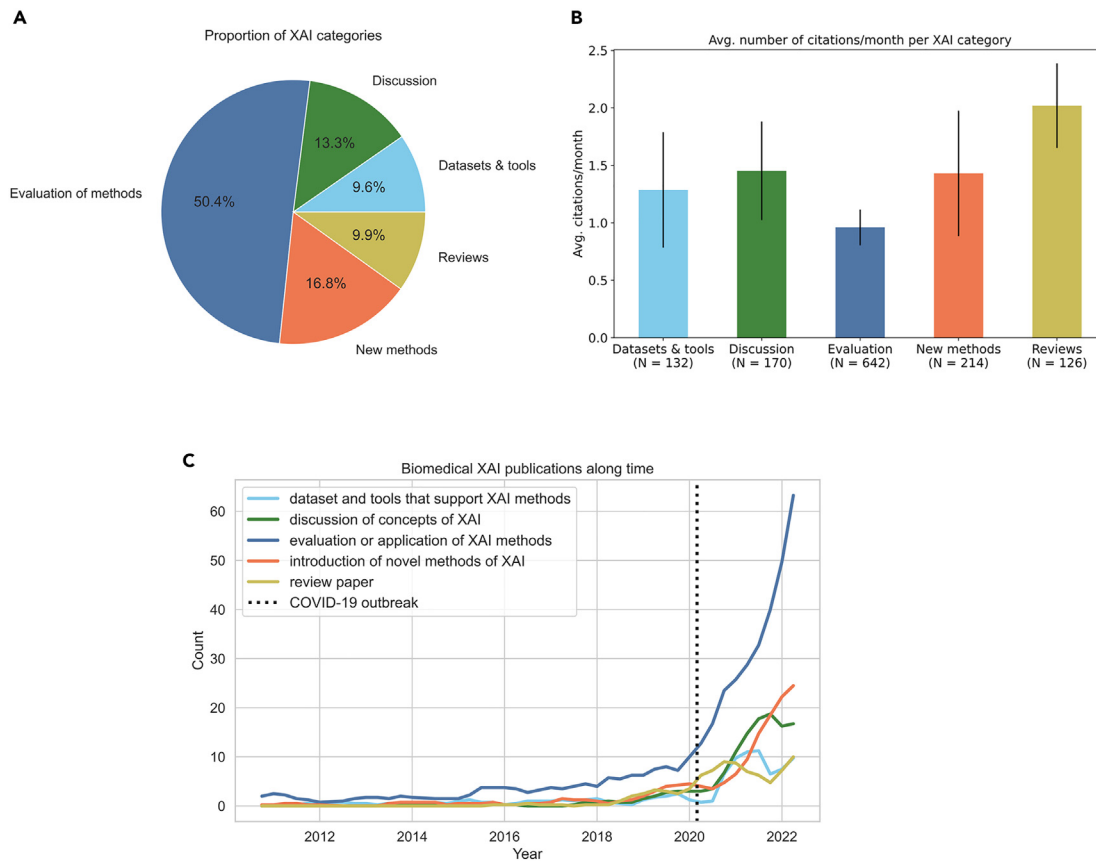
**Figure 3. Biomedical XAI statistics and trends of the cohort of eligible papers**
(A) Proportions of biomedical XAI categories.
(B) Average number of citations per month since published date. Error bars represent 95% confidence intervals.
(C) Counts of biomedical XAI papers published per quarter year. For each category (colored curve), we smoothed quarterly patterns by using the 1 year rolling average window.

example, would offer a direct explanation for the predicted outcome while being useless for human interpretation due to the sheer amount of split choices presented. And yet, despite these obstacles, ongoing discussion is necessary for the advancement of AI, especially in high-stakes settings, not only to improve robustness and reliability but also to satisfy the growing levels of legal requirements for the "right of explanation" that is being adopted to protect the stakeholders involved.[42]

Briefly, XAI algorithms can be divided into two main categories: transparent models that convey some intrinsic degree of interpretability and post hoc explainability techniques. One of the most prominent directions in the field is progressing toward leveraging the latter techniques, where explanation is provided to existing opaque models rather than the *de novo* design of methods rooted in causality. When analyzing the content of the ten most popular papers (based on their average citations), it was observed that Lundberg et al.'s study[43] was considerably the most cited one, averaging over 60 citations per month since its publication. Among these papers, four of them,[43–46] including Lundberg and colleagues' work, present post hoc explainability methods, while the remaining six papers[47–52] focus on developing inherently interpretable models. This tendency to explore post hoc solution could stem from the fact that several explana-

tion techniques can be applied to a variety of models with minimal added efforts. Lundberg's Shapley additive explanations (SHAP) analysis was frequently used in the biomedical literature that we reviewed. Although the majority of current post hoc methods only provide locally reliable explanations and can be misleading of model functionality,[53] Shapley values could be the sole approach that complies with legal requirements due to its foundation in a solid theory derived from axioms of a fair game.[54]

Above these variety of approaches, there is an ongoing debate regarding the tradeoff between performance and interpretability of machine-learning models. Widespread in the literature is the opinion that retaining or designing interpretability into an ML solution comes at a cost in accuracy of the predictions. It has been argued that such a tradeoff is but a myth,[55] and to an extent this holds true. Often, different models, including explainable ones, offer very similar performances.[56] Within these settings, it is therefore obvious to choose the solution that is also easier to interpret.

However, recent advances in large-scale deep-learning models have challenged this idea once again. The successes of foundation models,[57] which are those models trained on broad data and then adapted to specific tasks, have put in

question whether any explainable model could reach the same level of performance. The results produced by BERT,[58] GPT-3,[59] and DALL-E[60] are outclassing those of any transparent model by a wide margin. Recently, it was shown that large language models such as ChatGPT[61] may have the potential to assist with clinical reasoning.[62] Hence, it might become a tool to explain AI-based approaches and support AI-based decisions.

The paradigm shift produced by foundation models can offer completely different opportunities to make use of interpretations of their results. While a common approach with simpler models would be to diagnose the elements that lead to specific predictions (e.g., analyze the importance of features), with large-scale models, it is possible to obtain novel designs derived from the surprising emergent properties learned from the sheer quantity of data, which could expand our understanding of the task at hand in completely unexpected directions. Within the biomedical field, this could lead to, for example, the design of novel proteins that are not observed in nature[63] or to exploring new connections between diseases.[64]

Are the post hoc explainable technologies satisfying? Does the success of large-scale deep-learning models mean that it is futile to pursue non-post hoc explainability in AI? Perhaps it represents an unreachable goal. However, the increase in BXAI publications, highlighted in our survey, indicates that many research groups believe that the future of AI-based advances in medicine and health belongs to those who successfully address such questions.

## DATA AND CODE AVAILABILITY

All data analyzed and produced in this study, as well as the original code, have been deposited at https://github.com/gvisona/BiomedXAI and at the repository in Zenodo (https://doi.org/10.5281/zenodo.8207487). Data are publicly available as of the date of publication.

### REFERENCES

1. Tanne, J.H. (2021). Covid-19: FDA approves Pfizer-BioNTech vaccine in record time. BMJ 374, n2096.

2. (2022). A critical overview of current progress for COVID-19: development of vaccines, antiviral drugs, and therapeutic antibodies. J. Biomed. Sci. 29, 68. https://doi.org/10.1186/s12929-022-00852-9. https://link.springer.com/article/10.1186/s12929-022-00852-9.

3. Addressing Backlogs and Managing Waiting Lists during and beyond the COVID-19 Pandemic [Internet]. [cited 2023 Jan 11]. Available from: https://eurohealthobservatory.who.int/publications/i/addressing-backlogs-and-managing-waiting-lists-during-and-beyond-the-covid-19-pandemic.

4. Haldane, V., De Foo, C., Abdalla, S.M., Jung, A.S., Tan, M., Wu, S., Chua, A., Verma, M., Shrestha, P., Singh, S., et al. (2021 Jun). Health systems resilience in managing the COVID-19 pandemic: lessons from 28 countries. Nat. Med. 27, 964–980.

5. Martin, A., Nateqi, J., Gruarin, S., Munsch, N., Abdarahmane, I., Zobel, M., and Knapp, B. (2020). An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. Sci. Rep. 10, 19012.

6. Wang, M., Xia, C., Huang, L., Xu, S., Qin, C., Liu, J., Cao, Y., Yu, P., Zhu, T., Zhu, H., et al. (2020). Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. Lancet Digit. Health 2, e506–e515.

7. Jayatunga, M.K.P., Xie, W., Ruder, L., Schulze, U., and Meier, C. (2022). AI in small-molecule drug discovery: a coming wave? Nat. Rev. Drug Discov. 21, 175–176.

8. Palmisano, A., Vignale, D., Boccia, E., Nonis, A., Gnasso, C., Leone, R., Montagna, M., Nicoletti, V., Bianchi, A.G., Brusamolino, S., et al. (2022). AI-SCoRE (artificial intelligence-SARS CoV2 risk evaluation): a fast, objective and fully automated platform to predict the outcome in COVID-19 patients. Radiol. Med. 127, 960–972.

9. Ekman, I., Busse, R., van Ginneken, E., Van Hoof, C., van Ittersum, L., Klink, A., Kremer, J.A., Miraldo, M., Olauson, A., De Raedt, W., et al. (2016). Health-care improvements in a financially constrained environment. Lancet 387, 646–647.

10. Shen, J., Zhang, C.J.P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S.Y., Fang, P.H., and Ming, W.K. (2019). Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Med. Inform. 7, e10010.

11. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). AI in health and medicine. Nat. Med. 28, 31–38.

12. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat. Mach. Intell. 3, 199–217.

13. Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M.M.J., et al. (2020). Prediction

models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ *369*, m1328.

14. Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data *6*, 1–21.

15. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., and Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput. Biol. Med. *140*, 105111.

16. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. *267*, 1–38.

17. Adadi, A., and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access *6*, 52138–52160.

18. Doran, D., Schulz, S., and Besold, T.R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. Preprint at arXiv. https://doi.org/10.48550/arXiv.1710.00794.

19. Markus, A.F., Kors, J.A., and Rijnbeek, P.R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inf. *113*, 103655.

20. Cutillo, C.M., Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K.D., Beck, T., Collier, E., Colvis, C., Gersing, K., et al. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. npj Digit. Med. *3*, 47–55.

21. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inf. Decis. Making *20*, 310. https://doi.org/10.1186/s12911-020-01332-6. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01332-6.

22. Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X., and He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J. Am. Med. Inf. Assoc. *27*, 1173–1185.

23. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion *58*, 82–115.

24. Das, A., and Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.11371.

25. Vilone, G., and Longo, L. (2020). Explainable Artificial Intelligence: a Systematic Review. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.00093.

26. Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. *2*, 573–584.

27. Tjoa, E., and Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transact. Neural Networks Learn. Syst. *32*, 4793–4813.

28. Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., and Acharya, U.R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Comput. Methods Progr. Biomed. *226*, 107161.

29. Byrne, R.M.J. (2019). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (International Joint Conferences on Artificial Intelligence Organization) https://www.ijcai.org/proceedings/2019/876.

30. Kusner, M.J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual Fairness. In Advances in Neural Information Processing Systems (Curran Associates, Inc.) https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

31. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. WIREs Data Mining and Knowledge Discovery. https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1312.

32. (2022). Causal machine learning for healthcare and precision medicine. Preprint at arXiv. Royal Society Open Science. https://doi.org/10.48550/arXiv.2205.11402.

33. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. N. Engl. J. Med. *380*, 1347–1358.

34. Moher, D., Liberati, A., Tetzlaff, J., and Altman, D.G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ *6*, e1000097.

35. Gunning, D., and Aha, D.W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Mag. *40*, 44–58.

36. Paper on Artificial Intelligence: A European Approach to Excellence and Trust [Internet]. [cited 2023 Jun 13]. Available from: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

37. Wu, Y.H., Gao, S.H., Mei, J., Xu, J., Fan, D.P., Zhang, R.G., and Cheng, M.M. (2021). JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. IEEE Trans. Image Process. *30*, 3113–3126.

38. Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy *23*, 18.

39. Tran, B., Vu, G., Ha, G., Vuong, Q.H., Ho, M.T., Vuong, T.T., La, V.P., Ho, M.T., Nghiem, K.C., Nguyen, H., et al. (2019). Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. J. Clin. Med. *8*, 360.

40. Born, J., Beymer, D., Rajan, D., Coy, A., Mukherjee, V.V., Manica, M., Prasanna, P., Ballah, D., Guindy, M., Shaham, D., et al. (2021). On the role of artificial intelligence in medical imaging of COVID-19. Patterns (N. Y.) *2*, 100330.

41. Fghisoni, and Visonà, G. (2023). Gvisona/BiomedXAI: BiomedXAI - Patterns Submission v1.0 (v1.0) (Zenodo). https://doi.org/10.5281/ZENODO.8207487.

42. Aranovich, T.d.C., and Matulionyte, R. (2022). Ensuring AI explainability in healthcare: problems and possible policy solutions. Inf. Commun. Technol. Law *32*, 259–275.

43. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.I. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. *2*, 56–67.

44. Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S.H., Guerrier, C.E., Ebert, S.A., Pomerantz, S.R., Romero, J.M., Kamalian, S., et al. (2018). An explainable deep-learning algorithm for the detection of intracranial haemorrhage from small datasets. Nat. Biomed. Eng. *3*, 173–182.

45. Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., and Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. *11*, 3852.

46. Nagasubramanian, K., Jones, S., Singh, A.K., Sarkar, S., Singh, A., and Ganapathysubramanian, B. (2019). Plant disease identification using explainable 3D deep learning on hyperspectral images. Plant Methods *15*, 98.

47. Yang, J.H., Wright, S.N., Hamblin, M., McCloskey, D., Alcantar, M.A., Schrübbers, L., Lopatkin, A.J., Satish, S., Nili, A., Palsson, B.O., et al. (2019). A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell *177*, 1649–1661.e9.

48. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genom. *14*, S3.

49. Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E.B. (2022). Neural Granger Causality. IEEE Trans. Pattern Anal. Mach. Intell. *44*, 4267–4279.

50. Zhou, X., Zhou, J., Xin, F., Ma, J., Zhang, W., Wu, H., Jiang, M., and Dong, W. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *119*, 1171–1178.

### About the author

**Luca Maliverno** holds a PhD in physics. He is a data scientist in Porini SRL and a Microsoft Certified Trainer. He is part of the Porini BU Education, where he takes care of the Porini Innovation and Research Center for DataScience (PIRC). His interests lie in artificial intelligence for business application and digital transformation.

**Vesna Barros** is a research scientist at IBM Research. She graduated in electrical engineering at the University of Brasilia (BSc) and ETH Zurich (MSc). Before joining IBM, she worked as a data scientist at Novartis and Philips Healthcare. She carries out research at the intersection of machine learning and medicine with the goal of improving diagnosis and treatment outcomes to the benefit of the care and wellbeing of patients. She is currently pursuing a PhD at the Hebrew University of Jerusalem, Israel.

**Francesco Ghisoni** is a PhD candidate in quantum machine learning at the University of Pavia. He holds a master's degree in physics from King's College London, where his thesis focused on quantum Monte Carlo methods. Francesco has extensive international experience, having completed a 1 year fellowship at the University of California Berkeley and a research period in Canada at the Institute for Quantum Computing. He has won numerous awards, including recognition for academic achievement at KCL and first place at the ETH Quantum Hackathon in 2023. His professional experience includes working at Porini as a data science trainer and researcher.

**Giovanni Visonà** graduated in physics at the Universities of Turin (BSc) and Trent (MSc). After obtaining his MSc in artificial intelligence at the University of Edinburgh, he joined in December 2019 the Marie Curie Innovative Training Network "Machine Learning Frontiers in Precision Medicine." He is currently pursuing a PhD at the Max-Planck Institute for Intelligent Systems in Tübingen, Germany.

**Roman Kern** is chief scientific officer at Know-Center, a research center for trustworthy AI, and works at the Institute for Interactive Systems and Data Science at the Technical University of Graz. He was awarded his PhD in computer science by the Graz University of Technology. His research interests include natural language processing and machine learning, with a focus on causal data science. He applies these approaches in various application fields with the goal to achieve trustworthy AI methods.

**Philip J. Nickel** (PhD UCLA, 2002) is an associate professor in the philosophy and e thics group at Eindhoven University of Technology. He specializes in the philosophy of trust and the ethics of disruptive technologies. Some of his research is in the domain of biomedical ethics, focusing on issues of consent and the mediation of care through technology and data. He leads the project "Mobile Support Systems for Behavior Change" funded by the Dutch Research Council's Responsible Innovation program.

**Barbara Elvira Ventura** is a data analyst at Porini SRL. She holds a degree in bioengineering and biomedical engineering from University of Pavia.

**Ilija Šimić** is a researcher at Know-Center GmbH, Austria. He earned his bachelor's and master's degree in computer science at Graz University of Technology. In his research, he investigates explainable AI methods for debugging, verifying, and extracting knowledge from state-of-the-art deep-learning models for time series classification.

**Sarah Stryeck** has a PhD in biophysics and metabolism research from the Medical University of Graz, Austria. After her PhD, she worked for 3 years in data management/data governance at the Graz University of Technology and at Know-Center GmbH. Currently, Sarah is head of digitalization and data governance at the Research Center Pharmaceutical Engineering GmbH in Graz, Austria, and is consulting the Data Intelligence Initiative as data steward expert.

**Francesca Manni** is currently a clinical scientist at Philips. She has a background in biomedical engineering with a focus on AI for medical imaging. During her PhD at TUe Eindhoven, she focused on the development of novel imaging technologies for tumor detection and minimally invasive surgery. After that, she was an AI and data scientist at Philips Research. She has worked on enabling AI solutions in healthcare while leading the healthcare group at the BDVA. Francesca's research is reported in numerous international peer-reviewed scientific journals and top international conference proceedings in the field of image-guided interventions and AI.

**Cèsar Ferri** is a professor in the Department of Computer Systems and Computation at the èèUniversitat Politènica de València. He is also a member of the Valencian Graduate School and Research Network of Artificial Intelligence (ValgAI). Furthermore, he holds a long-standing research position in the DMIP team (a subgroup of VRAIN), which commenced in 1999. His research interest orients around machine learning and artificial intelligence. Cèsar has published papers in leading journals and at top conferences as a co-author on the aforementioned topics.

**Claire Jean-Quartier** has a PhD in the field of preclinical research regarding cell signaling with a master's in environmental system sciences and chemistry, as well as in molecular biomedical sciences. C.J. has been working in chemical, pharmaceutical, and environmental analytics and quality assurance and is engaged in bioinformatical cancer research and explainable AI at the Human-Centered AI Lab. As well as focusing on sustainability aspects in interdisciplinary research, she is currently working at the Graz University of Technology, fostering open science and research data management.

**Laura Genga** is an assistant professor in information systems at the Eindhoven University of Technology (TU/e). Her research revolves around

data-driven process analysis and enhancement. In particular, her work falls within the process mining discipline, the goal of which consists of turning event data collected by any sort of logging system, e.g., enterprise information systems, into insights on the corresponding processes. Her core topics involve automated discovery and analysis of flexible processes, process compliance analysis, and online and monitoring and prediction to support process managers in making decisions regarding current process executions, e.g., to react to potentially undesired situations.

**Dr. Gabriele Schweikert** previously completed a PhD at the Max Planck Institute Tuebingen (Schoelkopf, Weigel, Raetsch labs), developing machine-learning techniques for computational gene finding. She then moved on to the Wellcome Trust Center for Cell Biology in Edinburgh, joining the lab of Adrian Bird, one of the pioneers of epigenomic research. She later held a Marie Curie and an EMBO Fellowship at the School of Informatics, University of Edinburgh, prior to coming to the Division of Computational Biology at the School of Life Sciences, University of Dundee.

**Dr.sc. Mario Lovrić** leads the laboratory for chemical and biomedical informatics at the Institute for Anthropological Research in Zagreb and teaches as an assistant professor at the Faculty of Electrical Engineering in Osijek, both in Croatia. He is also scientific director of the Horizon EDIAQI project funded by the European Commission. Mario holds a PhD in computational chemistry with a focus on toxicology and has conducted his post-doc research in Know-Center, Austria, and RegionH, Denmark. Mario has published more than 30 peer-reviewed papers.

**Prof. Michal Rosen-Zvi** is the director of AI for drug discovery at IBM Research and an adjunct computational medicine professor at the Hebrew University at the Faculty of Medicine. She is also heading the AI for Accelerated HC&LS Discovery department at IBM Research, Israel. Michal holds a PhD in computational physics and completed her postdoctoral studies at UC Berkeley in machine learning. She joined IBM Research in 2005. Michal has published more than 50 peer-reviewed papers. She has served as a reviewer and guest editor in various leading conferences and journals and has served on various boards.