# The AIMe minimal information standard on artificial intelligence in biomedical research

Julian Matschinske[1], Olga Lazareva[1], Markus List[1], Jan Baumbach[1], Tim Kacprowski[1], Josh Pauling[1], Richard Röttger[2], You[3], and David B. Blumenthal[1]

[1]Chair of Experimental Bioinformatics, Technical University of Munich, Freising, Germany
[2]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[3]Your affiliation, your address

**Abstract**

Artificial intelligence (AI) and big data have found their way into basic and applied biological and biomedical research and demonstrate utility in manifold application scenarios. Numerous approaches for learning predictive, diagnostic, or exploratory models from all kinds of biomedical data emerge almost every day. However, varying evaluation schemes are employed and usually only selected aspects of the data analysis results are reported. This hinders comparability of different AI models and is detrimental for the development of novel, better AI methods. There is hence an urgent need to harmonize reporting of AI results in biomedical research. We present a concise reporting standard for AI methodology and results in biomedical research: the AIMe minimal information standard on **a**rtificial **i**ntelligence in bio**me**dical research. We intend AIMe to enhance accessibility, reproducibility, and usability of different AI models. Here, we give an overview of the questions contained in the standard, the rationales behind them, and provide clarifying details and best-practice examples. We also introduce a web-service available at https://aime-standard.org, which assists authors of new AI models in adhering to the AIMe standard and helps researchers to find existing AIs that are relevant for their use cases.

## 1 Introduction

The past two decades have seen massive advances and rapidly declining costs in high-throughput technologies that produce enormous amounts of biomedical data. This development has been accompanied by breakthroughs in the field of artificial intelligence (AI). With the help of AI, high-dimensional data can now be modeled in a mathematically robust and accurate way, which has led to numerous applications in biomedical research. For example, deep learning techniques allow to determine particles in cryogenic electron microscopy projection images [1], to infer proteins from mass spectrometry data [2], and to conduct exploratory analysis of single-cell data [3].

In spite of the obvious potential of AI in biomedical research, we observe trends that are detrimental for the development of new, improved AI methods and also constitute major hurdles for applying biomedical AIs in basic or translational biomedical research. Best practices of machine learning are not always adhered to [4], and often only selected aspects of the AI models and their evaluation are reported. Because of this, the decisions of biomedical AIs are often opaque, difficult to explain, and not fully reproducible [5, 6]. In particular in clinical research, it is crucial to instill trust in AI models and to report them in an explicit and transparent fashion that adheres to commonly used standards [4]. Or as put in [5]: "For widespread adoption to take place, AI systems must be approved by regulators [and] standardised to a sufficient degree [...]." However, such a reporting standard is missing for biomedical AIs.

To fill this gap, we present the AIMe minimal information standard for **a**rtificial **i**ntelligence in bio**me**dical research. It consists of a variety of questions, which require authors of new AIs to provide information about the purpose, data, method, and reproducibility of their AI system. To simplify adhering to the AIMe standard, we developed a web-service (`https://aime-standard.org`) guiding authors of new AIs through the questionnaire. The AIMe web-service provides detailed explanations of the AI concepts behind the questions, and gives pointers to relevant literature. Once the questionnaire has been filled in, a database entry and a report along with a unique AIMe identifier are created. The latter serves to keep the entry openly accessible and can be disseminated by the authors, e. g., by inclusion in a manuscript.

While, to the best of our knowledge, there is no formal reporting standard comparable to AIMe, several closely related efforts exist. [7] presents a retrospectively compiled collection of papers published in *PLOS Medicine*, *PLOS Computational Biology*, and *PLOS ONE* that meet minimal standards in code and data sharing. In [8], guidelines for reinforcement learning — a specific AI method — in healthcare are provided. Finally, in [9], guidelines for developing and reporting predictive machine learning models are presented. Of all existing approaches, this work has the most similarities with the AIMe standard. The main differences w. r. t. AIMe are, firstly, that the guidelines presented in [9] are designed to only cover predictive models, and, secondly, that they are normative rather than descriptive, i. e., they tell authors of new AIs what they should do instead of asking them what they did.

The remainder of this paper is organized as follows: In Sec. 2, we motivate why it is beneficial for the biomedical community to endorse the AIMe standard. In Sec. 3, we present the standard and provide rationales for the contained questions. In Sec. 4, we present a best-practice example. Sec. 5 concludes the paper.

## 2 Motivation

The AIMe standard serves a four-fold purpose. Firstly, it fosters scientific progress by enhancing reproducibility. Secondly, it helps practitioners and researchers who consider applying AI methods in real-world biomedical applications and instills trust on their side. Thirdly, it assists authors of new AIs. And fourthly, it supports journal editors, reviewers, and, more generally, the scientific audience. In the following, we will elaborate in detail on each of these four points. We also provide a motivating example that illustrates why a standard such as AIMe is called for.

**The AIMe standard fosters scientific progress by enhancing reproducibility.** Lack of reproducibility is a major issue in many scientific fields including biomedical sciences and AI research [10–13]. For the case of AI, in [12], three types of reproducibility are distinguished: method reproducibility, data reproducibility, and experiment reproducibility. An AI is *method reproducible* if it is sufficiently documented to allow re-implementation. It is *data reproducible* if it is method reproducible and the employed data is sufficiently documented to allow running an alternative re-implementation on it. Finally, an AI is *experiment reproducible* if it is data reproducible and its implementation is publicly available and sufficiently documented to allow re-running the same implementation on the same data. The authors of [12] randomly sampled 100 papers each from two editions of two top-tier AI conferences (IJCAI 2013 and 2016, AAAI 2014 and 2016) and then evaluated the papers w. r. t. three metrics that measure, respectively, degrees of method, data, and experiment reproducibility. The results are quite devastating; for all three metrics, the score is between 0.2 and 0.3, where 0 indicates no and 1 indicates full reproducibility. The AIMe standard is designed to enhance all three types of reproducibility: the "Data"-questions target data reproducibility (Sec. 3.3), the "Method"-questions target method reproducibility (Sec. 3.4), and the "Reproducibility"-questions target experiment reproducibility (Sec. 3.5).

**The AIMe standard helps practitioners and researchers who consider applying AI methods in real-world biomedical applications and instills trust on their side.** By issuing queries against AIMe's database, practitioners and researchers can find relevant AIs. Subsequently, they can consult the corresponding AIMe reports to gain a first, rough understanding of the AIs and to check whether or not a specific AI is suitable for their specific application scenario.

**The AIMe standard assists the authors of new biomedical AIs.** If AIMe's webservice is used to fill in the report *during* the design and evaluation phase, it raises awareness of potential pitfalls and provides pointers to possible solutions. Hence, it ultimately helps the authors to build better and more reliable AI systems.

**The AIMe standard supports journal editors, reviewers, and, more generally, the scientific audience.** When reading an AI paper — be it for review or out of scientific interest — it is often extremely tedious to search the main document and the supplements for basic information about the employed data, the selected method, and the availability of source code. For papers reporting the AIMe standard, this information is concisely available in one single document.

**A motivating example.** In a recent paper, it is shown that convolutional neural networks (CNNs) can be used to classify skin cancer types based on medical images [14]. While the obtained accuracy is impressive, the authors fail to explicitly mention that they trained and tested their CNN mainly on images of patients with Caucasian skin type. Instead, they only provide references to the employed publicly available datasets. This is unfortunate, because practitioners who are not familiar with the characteristics of these datasets might re-implement the CNN, train it on the referenced datasets, and then use it to predict skin cancer types for patients with non-Caucasian skin types. The consequences would be potentially catastrophic, because predictions of CNNs for data that is completely different from the training data are bound to be incorrect. Moreover, there is no empirical evidence that, even if fed with the right training data, the proposed CNN performs as well on non-Caucasian skin types as it does on Caucasian skin. In other words: it might be the case that the proposed CNN is actually of use only for a small fraction of the population and hence "exacerbate[s] health care disparities in dermatology" [15]. Note that we do not blame the authors of [14] for having trained and tested their CNN on biased data — after all, biased data is often all we have access to. However, we insist that they should have disclosed the data bias explicitly. If they had reported the AIMe standard, this would have happened in the answer to question (D.5).

# 3 The AIMe standard

In the following, we briefly describe each subsection of the AIMe standard and list the corresponding items from the questionnaire.

## 3.1 Metadata

The AIMe standard asks authors of biomedical AIs to report basic metadata of their methods. More specifically, the title (MD.1), a short description (MD.2), and information about the corresponding author (MD.3) should be provided. Moreover, a stable ID or URL to a paper describing the method should be provided if possible (MD.4). If available, authors should use the paper's DOI [16]. Otherwise, also other stable URLs such as arXiv identifiers [17] are admissible. Furthermore, the AIMe record must be marked as public or private (MD.5). While public records appear as results in queries against AIMe's database, private records can only be accessed via their AIMe identifiers. The authors can share these identifiers with editors or reviewers. Marking a record as private is hence especially useful when the reported AI system has not been published yet.

(MD.1) *Title of the AI.*

(MD.2) *Short description of the AI.*

(MD.3) *Corresponding author of the AI (name, institutional address, email, ORCID).*

(MD.4) *Stable ID or URL to a paper where the AI is described (optional).*

(MD.5) *Specification of whether the AIMe database entry should be public or private.*

## 3.2 Purpose

First of all, AIMe asks the authors to specify whether they are presenting a new AI method, new results (possibly achieved with existing techniques on new data), or both (P.1). Next, they should state what their AI is designed to learn or predict (P.2). This allows practitioners and researchers or authors of competing AI systems to easily verify if the reported AI is relevant for their problem setting. Note that the predicted outcome does not necessarily have to be a directly measurable response variable but can also be a surrogate marker. Whether or not this is the case is asked in question (P.3). If a surrogate marker is predicted, AIMe additionally asks the authors to clarify what is represented by the surrogate marker.

Furthermore, AIMe requests that the authors specify to which category their AI problem belongs (P.4). Typical categories are classification (assign discrete labels to all items), regression (predict a real-valued number for all items), clustering (partition a set of items into subsets of homogeneous groups), ranking (learn an ordering for a set of items), and dimensionality reduction (compress all items' initial high dimensional representations) [18].

(P.1) *Do you present a new AI method, new results, or both?*

(P.2) *What is your AI designed to learn or predict?*

(P.3) *Does your AI predict a surrogate marker? If so, what does it represent?*

(P.4) *To which category does your AI problem belong?*

## 3.3 Data

In biomedical research, it is common practice to use multiple datasets in the same pipeline to gain insights into complex biological processes. The AIMe standard therefore ask authors of new AIs to add each employed dataset separately and then characterize it in terms of data availability, possible biases, and applied transformations.

For each dataset, the authors should report the type of the data (D.1) — e. g., expression, methylation, or phenotype data. For instance, if an AI uses gene expression data to predict BMI, then the authors should add one dataset for the BMI data and a separate dataset for the expression data. Since there is often no "gold standard" data for biomedical AI problems [19], new AIs are often evaluated on simulated data. In view of this, AIMe asks the authors to specify whether their data is real or simulated (D.2). If the latter is the case, they should also provide a pointer to the employed simulator. Moreover, the authors should report whether the dataset is publicly available (D.3) and specify if it was used for training the AI method (D.4).

Biomedical data is often subject to various biases [20–22]. Even if these biases can be addressed appropriately, readers should be aware of them to avoid possible misinterpretations. Therefore, AIMe asks the authors to explicitly disclose whether their data is subject to *a priori* known biases (D.5) (e. g., class imbalance, batch effect, sample bias), and if so, to specify how these biases were addressed (D.6). AIMe also requests the authors to report the dimension of their data, i. e., to specify the number of samples and features (D.7). This is especially important because high dimensional data often exhibits multicollinearity and sparsity [23], which in turn tends to negatively affect the efficiency of AI systems [24] and often leads to overfitting (cf. (M.4)).

Since most AI methods are not scale-invariant, the data usually needs to be normalized during pre-processing. How exactly this is done is asked in question (D.8). For instance, some methods such as principal component analysis perform optimally with zero-centered data, while others such as typical neural network architectures require the data to be scaled between zero and one. There are various other pre-processing techniques such as feature engineering and pre-selection that can be used to make the data more suitable for AI-based algorithms. Whether or not such techniques have been employed should be reported in the answer to question (D.9).

**For each dataset:**

(D.1) *What is the type of the data?*

(D.2) *Is the data real or simulated? If it is simulated, is the simulator publicly available and, if so, where?*

(D.3) *Is the data publicly available? If so, where?*

(D.4) *Is this data used for training?*

(D.5) *Is the data subject to biases? If so, to which?*

(D.6) *Did you address your data biases? If so, how?*

(D.7) *How many samples and features do you have?*

(D.8) *Did you normalize your data? If so, how?*

(D.9) *Did you apply other pre-processing steps to your data? If so, which?*

## 3.4 Method

The next series of questions addresses the specific AI methods. The first question AIMe asks in this regard is which AI or optimization methods (e. g., logistic regression, random forest classification, deep neural networks, ant colony optimization, genetic programming) were used (M.1). Next, the authors should specify whether the employed methods have any hyper-parameters (e. g., number of trees and maximal depth of random forest models). If so, they should report how they picked the values for these hyper-parameters (M.2). This is important because they typically have a huge impact on the algorithm's performance [25, 26].

The AIMe standard also contains questions related to validation and verification of the employed AI method. The initial questions ask which test metrics (e. g., Gini coefficient, running time, mean squared error) were used to evaluate the method (M.3). Afterwards, the authors are asked to report how they prevented overfitting, i. e., how they ensured their AI model does not merely memorize the training data but can generalize to unseen, independent data (M.4). Overfitting can be prevented by using various techniques such as ensemble learning, cross-validation, and regularization. The authors should also mention if they validated their AI method on independent data (M.5).

Moreover, the AIMe standard asks the authors to clarify whether they checked their AI model has been affected by confounding factors (M.6). Confounding factors are variables that influence both the model input and output variables, and as a result, potentially distort the obtained results [27]. The authors are also required to report how they checked that their AI is robust, i. e., that the results are stable and that there are no random effects (M.7). Moreover, the authors should specify whether they made comparisons between their AI system and a simple baseline model, and if so, explain how they compared the two (M.8). Additionally, the authors should report if they compared their novel AI to state-of-the-art approaches tackling the same problem, and if so, justify their choice of the competitors (M.9).

(M.1) *Which AI or optimization methods did you use and how did you select them?*

(M.2) *Do your methods have hyper-parameters? If so, which ones and how did you select them?*

(M.3) *Which test metrics do you report?*

(M.4) *How do you prevent overfitting?*

(M.5) *Did you validate your AI on independent data, e.g., from a different cohort?*

(M.6) *Did you check if your AI model is affected by confounding factors?*

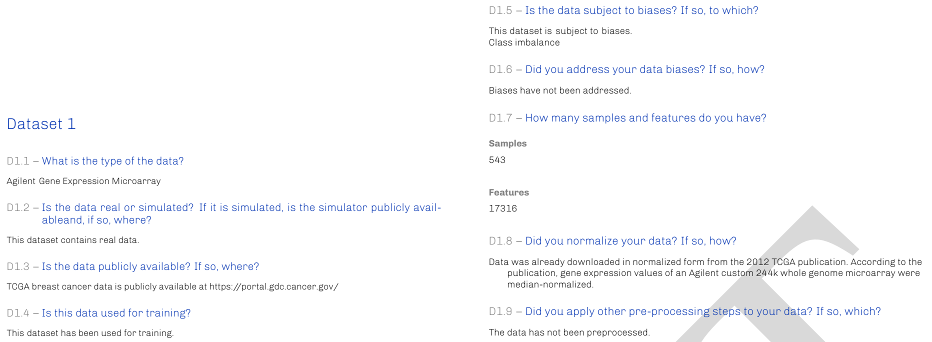
(M.7) *Did you check whether your AI is robust? If so, how?*

(M.8) *Did you compare against a simple baseline model? If so, how?*

(M.9) *Did you compare against state-of-the-art approaches? If so, against which, and how did you select the competitors?*

## 3.5 Reproducibility

The last five questions of the AIMe standard help increasing the reproducibility of the experiments that validate the proposed AI. First of all, the authors are requested to report which tools they employed for implementing their AI, specify whether these tools are publicly available, and, if so, provide links to them (R.1). In most cases, the authors will probably use existing tools (e. g., TensorFlow [28], PyTorch [29], LIBSVM [30], scikit-learn [31]), but it might also be the case that their AI system is built upon self-implemented software. Next, AIMe asks the authors to share the source code of their analysis, i. e., the code that has to be run to actually reproduce the experiments (R.2). To clearly see the difference between the questions (R.1) and (R.2), assume that an AI uses the $K$-means implementation of the widely used Python library scikit-learn for a biomedical clustering problem. While in this case a link to scikit-learn suffices as an answer to (R.1), for answering (R.2), the authors should provide a link to the Python script which they implemented on top of scikit-learn.



Dataset 1

D1.1 – What is the type of the data?

Agilent Gene Expression Microarray

D1.2 – Is the data real or simulated? If it is simulated, is the simulator publicly available and, if so, where?

This dataset contains real data.

D1.3 – Is the data publicly available? If so, where?

TCGA breast cancer data is publicly available at https://portal.gdc.cancer.gov/

D1.4 – Is this data used for training?

This dataset has been used for training.

D1.5 – Is the data subject to biases? If so, to which?

This dataset is subject to biases.
Class imbalance

D1.6 – Did you address your data biases? If so, how?

Biases have not been addressed.

D1.7 – How many samples and features do you have?

**Samples**

543

**Features**

17316

D1.8 – Did you normalize your data? If so, how?

Data was already downloaded in normalized form from the 2012 TCGA publication. According to the publication, gene expression values of an Agilent custom 244k whole genome microarray were median-normalized.

D1.9 – Did you apply other pre-processing steps to your data? If so, which?

The data has not been preprocessed.

Figure 1: Details for the first dataset in AIMe report u6bj7jpsqc.

Next, the authors are requested to state how they reported the dependencies of their experiments, e. g., via README or requirements files (R.3). They are also asked to specify the operating system and the hardware specifications of the machine on which they ran the experiments (R.4). Knowing on which hardware the AI was run is especially important if runtime is among the reported metrics (cf. question (M.3) above).

Finally, AIMe asks the authors to answer the question if, and if so why, their AI requires high-performance computing (R.5). Here, the intended reading is that an AI requires high-performance computing if it is infeasible to run it on a personal computer. This might be the case for various reasons such as huge memory consumption, massive parallelization, or simply an excessively long runtime that would block the personal computer for several days.

(R.1) *Which AI or optimization tools did you use for implementing your system? Are these tools publicly available and, if so, where?*

(R.2) *Is the source code of your analysis publicly available? If so, where?*

(R.3) *How did you report on your dependencies?*

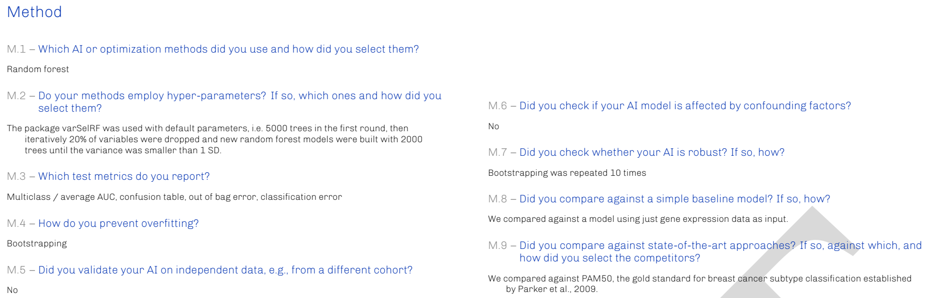
(R.4) *On which system (OS and hardware specifications) did you run your analysis?*

(R.5) *Does your analysis require high-performance computing? If so, why?*

# 4 A best-practice example

In the AIMe report u6bj7jpsqc, we showcase how the questions presented in the previous sections can be answered considering a previous study using AI to predict breast cancer subtypes [32].

No clinical data, e. g., measurements of blood pressure or blood glucose levels, but two types of omics data were used in this study. We thus add two data sets to the report. Figure 1 shows the details entered for the first data set. In this example, data was already normalized as part of the original TCGA publication [33]. While a class imbalance bias



Method

M.1 – Which AI or optimization methods did you use and how did you select them?

Random forest

M.2 – Do your methods employ hyper-parameters? If so, which ones and how did you select them?

The package varSelRF was used with default parameters, i.e. 5000 trees in the first round, then iteratively 20% of variables were dropped and new random forest models were built with 2000 trees until the variance was smaller than 1 SD.

M.3 – Which test metrics do you report?

Multiclass / average AUC, confusion table, out of bag error, classification error

M.4 – How do you prevent overfitting?

Bootstrapping

M.5 – Did you validate your AI on independent data, e.g., from a different cohort?

No

M.6 – Did you check if your AI model is affected by confounding factors?

No

M.7 – Did you check whether your AI is robust? If so, how?

Bootstrapping was repeated 10 times

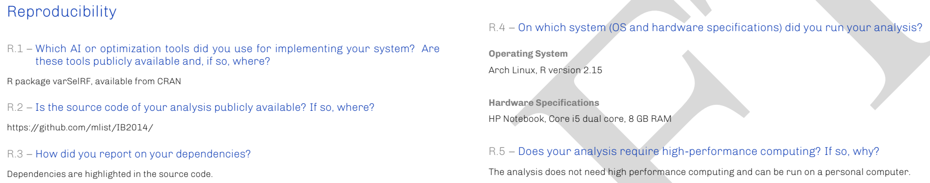
M.8 – Did you compare against a simple baseline model? If so, how?

We compared against a model using just gene expression data as input.

M.9 – Did you compare against state-of-the-art approaches? If so, against which, and how did you select the competitors?

We compared against PAM50, the gold standard for breast cancer subtype classification established by Parker et al., 2009.

Figure 2: Details for the employed AI method in AIMe report u6bj7jpsqc.



Reproducibility

R.1 – Which AI or optimization tools did you use for implementing your system? Are these tools publicly available and, if so, where?

R package varSelRF, available from CRAN

R.2 – Is the source code of your analysis publicly available? If so, where?

https://github.com/mlist/IB2014/

R.3 – How did you report on your dependencies?

Dependencies are highlighted in the source code.

R.4 – On which system (OS and hardware specifications) did you run your analysis?

**Operating System**

Arch Linux, R version 2.15

**Hardware Specifications**

HP Notebook, Core i5 dual core, 8 GB RAM

R.5 – Does your analysis require high-performance computing? If so, why?

The analysis does not need high performance computing and can be run on a personal computer.

Figure 3: Details for reproducibility in AIMe report u6bj7jpsqc.

was not addressed in the study, adding this information offers important insights into potential limitations and issues of the results.

Figure 2 shows the details entered for the method. In this example, we used a random forest classifier coupled with a recursive feature elimination strategy. Hyper-parameters were left at their defaults but nevertheless we explain the overall strategy briefly. Several test metrics were employed here to reflect the performance on this multi-class prediction task. General methods (e. g., multi-class AUC) can be named as well as method-specific metrics (e. g., out-of-bag error in random forest). While we highlight that we used bootstrapping to prevent overfitting and that we checked for robustness by repeating the entire procedure ten times, we also explicitly state that an independent data set from a different cohort was not available at the time of publication. This indicates that results and models from this study should not be used in clinical practice without additional validation.

Figure 3 shows the details entered for reproducibility. Here, we mention not only the R package varSelRF [34] that was used for this analysis, but also provide a link to the original code on a public repository. Dependencies are highlighted in the source code but version numbers are missing, possibly hampering exact reproducibility.

# 5 Conclusions

AI is on the rise in biology and medicine and demonstrates utility in numerous application scenarios. However, basic information about data, methods, and implementation is often incomplete in the respective publications. This makes it difficult to judge, comprehensively compare, and reproduce the results of biomedical AIs, which, in turn, constitutes a major hurdle for developing new AI methods and for applying AI in research and practice. The AIMe reporting standard presented in this paper is designed to address this problem and thereby improve the quality, reliability, and reproducibility of biomedical AIs. It comes with a web-service, which allows authors to easily register their AIs and assists researchers and practitioners in finding existing AI systems that are relevant for their application scenarios: https://aime-standard.org.

## References

1. Bepler, T. *et al.* Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods* **16,** 1153–1160. doi:10.1038/s41592-019-0575-8 (2019).
2. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17,** 41–44. doi:10.1038/s41592-019-0638-x (2020).
3. Amodio, M. *et al.* Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16,** 1139–1145. doi:10.1038/s41592-019-0576-7 (2019).
4. Liu, X. *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* **1,** e271–e297. doi:10.1016/s2589-7500(19)30123-2 (2019).
5. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthcare Journal* **6,** 94–98. doi:10.7861/futurehosp.6-2-94 (2019).
6. Stupple, A., Singerman, D. & Celi, L. A. The reproducibility crisis in the age of digital medicine. *NPJ Digit. Med.* **2,** 2. doi:10.1038/s41746-019-0079-z (2019).
7. Celi, L. A., Citi, L., Ghassemi, M. & Pollard, T. J. The PLOS ONE collection on machine learning in health and biomedicine: Towards open code and open data. *PLOS One* **14,** e0210232. doi:10.1371/journal.pone.0210232 (2019).
8. Gottesman, O. *et al.* Guidelines for reinforcement learning in healthcare. *Nat. Med.* **25,** 16–18. doi:10.1038/s41591-018-0310-5 (2019).
9. Luo, W. *et al.* Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J. Med. Internet Res.* **18,** e323. doi:10.2196/jmir.5870 (2016).
10. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454. doi:10.1038/533452a (2016).
11. Ioannidis, J. P. A. Why Most Clinical Research Is Not Useful. *PLOS Medicine* **13,** e1002049. doi:10.1371/journal.pmed.1002049 (2016).

12. Gundersen, O. E. & Kjensmo, S. *State of the art: reproducibility in artificial intelligence.* in *AAAI 2018* (eds McIlraith, S. A. & Weinberger, K. Q.) (AAAI Press, 2018), 1644–1651.

13. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359,** 725–726. doi:10.1126/science.359.6377.725 (2018).

14. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118. doi:10.1038/nature21056 (2017).

15. Adamson, A. S. & Smith, A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* **154,** 1247–1248. doi:10.1001/jamadermatol.2018.2348 (2018).

16. International DOI Foundation. *The DOI System* https://www.doi.org/.

17. Cornell University. *The arXiv archive* https://arxiv.org/.

18. Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning* 2nd ed. (The MIT Press, 2018).

19. Giannoulatou, E., Park, S.-H., Humphreys, D. T. & Ho, J. W. Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie. *BMC Bioinform.* **15 Suppl 16,** S15. doi:10.1186/1471-2105-15-S16-S15 (2014).

20. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35,** 498–507. doi:10.1016/j.tibtech.2017.02.012 (2017).

21. Schölz, C. *et al.* Avoiding abundance bias in the functional annotation of posttranslationally modified proteins. *Nat. Methods* **12,** 1003–1004. doi:10.1038/nmeth.3621 (2015).

22. Semmes, O. J. The "omics" Haystack: Defining Sources of Sample Bias in Expression Profiling. *Clin. Chem.* **51,** 1571–1572. doi:10.1373/clinchem.2005.053405 (2005).

23. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods* **15,** 399–400. doi:10.1038/s41592-018-0019-x (2018).

24. Indyk, P. & Motwani, R. *Approximate nearest neighbors: towards removing the curse of dimensionality* in *STOC 1998* (ACM, 1998), 604–613. doi:10.1145/276698.276876.

25. Van Rijn, J. N. & Hutter, F. *Hyperparameter importance across datasets* in *KDD 2018* (eds Guo, Y. & Farooq, F.) (ACM, 2018), 2367–2376. doi:10.1145/3219819.3220058.

26. Probst, P., Boulesteix, A.-L. & Bischl, B. Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.* **20,** 1–32 (2019).

27. Skelly, A. C., Dettori, J. R. & Brodt, E. D. Assessing bias: the importance of considering confounding. *Evid. Based Spine Care J.* **3,** 9–12. doi:10.1055/s-0031-1298595 (2012).

28. Abadi, M. *et al. TensorFlow: Large-scale machine learning on heterogeneous systems* Software available from tensorflow.org. 2015. https://www.tensorflow.org/.

29. Paszke, A. *et al. PyTorch: An imperative style, high-performance deep learning library* in *NIPS 2019* (eds Wallach, H. *et al.*) (Curran Associates, Inc., 2019), 8024–8035. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

30. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2,** 27:1–27:27. doi:10.1145/1961189.1961199 (2011).

31. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

32. List, M. *et al.* Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J. Integr. Bioinform.* **11,** 236:1–236:14. doi:10.2390/biecoll-jib-2014-236 (2014).

33. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70. doi:10.1038/nature11412 (2012).

34. Diaz-Uriarte, R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinform.* **8,** 328:1–328:7. doi:10.1186/1471-2105-8-328 (2007).

## 6 Acknowledgements

TK is grateful for financial support from H2020 project RepoTrial (no. 777111). JM and RR received funding from the European Union's Horizon2020 research and innovation programme under grant agreement No 826078. JB, TK, and ML are grateful for financial support from BMBF grant Sys_CARE (no. 01ZX1908A) of the Federal German Ministry of Research and Education. JB's BMBF grant SyMBoD (no. 01ZX1910D) also financed parts of this project. JB was partially funded by his VILLUM Young Investigator Grant no. 13154. Contributions by JKP and OL are funded by the Bavarian State Ministry of Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B, grant LipiTUM).