PLOS Computational Biology

EDUCATION

Nine quick tips for trustworthy machine learning in the biomedical sciences

Luca Oneto[1], **Davide Chicco**[2,3]*

1 Dipartimento di Informatica Bioingegneria Robotica e Ingegneria dei Sistemi, Università di Genova, Genoa, Italy, **2** Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy, **3** Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

* davidechicco@davidechicco.it

OPEN ACCESS

Citation: Oneto L, Chicco D (2025) Nine quick tips for trustworthy machine learning in the biomedical sciences. PLoS Comput Biol 21(10): e1013624. https://doi.org/10.1371/journal.pcbi.1013624

Editor: Patricia M Palagi, SIB Swiss Institute of Bioinformatics, SWITZERLAND

Published: October 30, 2025

Funding: The work of L.O. is partially supported by (i) project ELSA—European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617, (ii) EU—NGEU National Sustainable Mobility Center (CN00000023) Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10), (iii) project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NGEU, (iv) project FAIR (PE00000013) under the NRRP

Abstract

As machine learning (ML) becomes increasingly central to biomedical research, the need for trustworthy models is more pressing than ever. In this paper, we present nine concise and actionable tips to help researchers build ML systems that are technically sound but ethically responsible, and contextually appropriate for biomedical applications. These tips address the multifaceted nature of trustworthiness, emphasizing the importance of considering all potential consequences, recognizing the limitations of current methods, taking into account the needs of all involved stakeholders, and following open science practices. We discuss technical, ethical, and domain-specific challenges, offering guidance on how to define trustworthiness and how to mitigate sources of untrustworthiness. By embedding trustworthiness into every stage of the ML pipeline – from research design to deployment – these recommendations aim to support both novice and experienced practitioners in creating ML systems that can be relied upon in biomedical science.

Introduction

The current generation of Artificial Intelligence (AI) tools in the biomedical sciences is predominantly based on inductive AI, specifically machine learning (ML) [1], which leverages data to generate new knowledge [2]. Applications range from the use of large language models for clinical diagnosis [3] to addressing the protein folding problem with advanced deep learning architectures [4]. In contrast, deductive AI [5] operates by reasoning over well-established knowledge, for example, to plan hospital procedures [6] or define clinical pathways [7]. As ML discovers patterns by analyzing data and identifying correlations [8], its trustworthiness can be limited [9], and its performance is highly dependent on the quality and quantity of the input data [10]. The widespread adoption of AI, along with the resulting surge in data usage and rising societal concerns about its untrustworthiness, are among the main reasons behind the

MUR program funded by the EU—NGEU, and (v) project FISA-2023-00128 funded by the MUR program "Fondo italiano per le scienze applicate." The work of D.C. is partially funded by the Italian Ministero Italiano delle Imprese e del Made in Italy under the Digital Intervention in Psychiatric and Psychologist Services (DIPPS) programme and is partially supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

development of regulations such as the Data-Act [11] and the AI-Act [12] in the European Union. Beyond the obvious risks of intentional misuse, the current generation of AI-based systems, being highly untrustworthy, may lead to unintentional misuse [13]. Nevertheless, striking an appropriate balance between the need for regulation and the advancement of biomedical sciences remains a significant challenge [14].

Trustworthiness is the property of an ML-based system that emerges from the integration of technical robustness [15], ethical responsibility [16], and domain awareness [17], ensuring that its behavior is reliable, transparent, and contextually appropriate for biomedical applications. From a technical standpoint [15], ML-based system can accumulate technical debt, such as fragile data pipelines, poorly versioned models, and lack of robust monitoring, that makes them difficult to maintain and prone to failure, particularly in operational environments. Ethically [16], ML-based system may violate principles of privacy through excessive data collection or poor anonymization, exhibit bias and lack fairness due to skewed training data, and lack robustness and explainability, making them unreliable in unpredictable settings and opaque to end-users and stakeholders. Lacks of domain knowledge [17] result in ML-based system that might capture statistical correlations but miss clinically meaningful insights, potentially leading to unsafe or ineffective recommendations. In fact, if contextual information is not incorporated to appropriately support the collected data, we might identify non-causal relationships that could lead to the model failing to generalize or, even more concerning, producing unsafe or ineffective recommendations.

To address the technical, ethical, and domain-specific challenges of deploying ML-based systems, we propose a series of quick tips for fostering trustworthy ML in the biomedical sciences. These concise and practical recommendations are not intended to be exhaustive; rather, they aim to raise awareness among novices while remaining equally valuable for experienced practitioners.

Tip 1: Think about the consequences of your actions

Our advice is straightforward: before applying an ML method to biomedical data, carefully evaluate all potential consequences, with particular attention to possible negative outcomes. This advice is intended for all machine learning researchers and to anyone reading this article and planning to perform a machine learning analysis of biomedical data. Scientific projects, despite their intentions, can lead to unintended consequences, some of which may cause significant harm.

Examples of such unintended negative outcomes include adverse drug reactions and misinterpretations of data. For instance, drugs discovered through computational methods may present unforeseen side effects not evident during clinical trials, potentially leading to health problems for patients [18]. Likewise, findings from medical informatics studies may be misinterpreted or overgeneralized, resulting in inappropriate clinical practices or misguided public health policies [19].

Some biomedical informatics studies may be problematic not because of their findings, but due to their origins. A well-known example is the thousands of studies conducted using HeLa cells, derived from the tissue of Henrietta Lacks without her consent [20].

To mitigate such risks, we recommend considering all stakeholders involved in a study (such as ML researchers, biomedical scientists, clinicians, study designers, patients, grant holders, and legal or administrative personnel) and carefully reflecting on the potential consequences—positive and negative, intended and unintended—of the research outcomes. The boundary between good and bad is often ambiguous. Even outcomes that violate an individual's rights may be tolerated when framed as low-risk AI use or as serving the greater good.

The AMIA 2009 Health Policy Meeting [21] proposed a useful framework for categorizing consequences: anticipated (direct or indirect) and unanticipated (direct or indirect). Each of these categories can be further classified as desirable or undesirable. Additionally, consequences may be temporally segmented into short-term (up to six months), medium-term (six months to three years), and long-term (beyond three years). Again, however, the distinction between short-term and long-term is fuzzy. Some consequences might come with undesirable short-term change problems which can also affect patients negatively soon, but ultimately in long run benefits more people. Further insights into the unintended consequences of biomedical information technologies can be found in the specialized literature [22,23].

Tip 2: Define the concept of trustworthiness for your specific biomedical applications

The meaning of trustworthiness can vary depending on the domain, data, and other contextual factors. Importantly, trustworthiness is also dynamic: what is deemed trustworthy may shift over time as technologies evolve, regulatory frameworks change, or societal expectations develop.

For example, Cutillo and colleagues [24] define trustworthiness as "the ability to assess the validity and reliability of a machine intelligence-derived output across varying inputs and environments." Park and colleagues [25] associate trustworthiness with providing users interpretable information that enhances their confidence in diagnostic results and potentially influences their decisions. Similarly, Huang and colleagues [26] describe trustworthiness as the provision of reliable and understandable information extracted from magnetic resonance images, thereby supporting diagnostic differentiation for medical doctors.

Given these differing perspectives, we caution against adopting a single, rigid definition of trustworthiness. Instead, researchers should explore and adapt context-appropriate notions of trustworthiness to their specific scientific setting. Two main strategies can be pursued. First, biomedical informatics may be divided into three principal subfields (bioinformatics, health informatics, and chemoinformatics [27]) and the role of trustworthiness within each. Alternatively, one could classify the field according to data types—such as genomics, proteomics, transcriptomics, medical images, electronic health records, drug compounds, and biosignals—and analyze what "trustworthy" means in the context of each category.

In either case, the analysis should begin by identifying the actors involved with the ML-based tool (Tip 4). Subsequently, the three primary dimensions of trustworthiness should be considered: technical (Tip 5), ethical (Tip 6), and domain-specific (Tip 7). Clarifying trustworthiness in this way directly shapes system design, evaluation criteria, and ultimately the likelihood of stakeholder adoption [28–31].

It is important to notice that the meaning and definition of trustworthiness also differ from the perspective of each actor (Tip 4).

Tip 3: Be aware of what can and what cannot be done to achieve trustworthiness

This tip defines theoretical boundaries of trustworthiness metrics. Trustworthiness in biomedical ML can be pursued through two conceptually distinct routes, each with its own strengths and limitations [32].

The first route involves embedding trustworthiness directly into the learning objective, instructing the model itself to discover and exhibit trustworthy behavior [33], for example, through approaches such as neuro-symbolic ML [34]. Although conceptually appealing, this strategy remains largely aspirational: it faces unresolved philosophical questions and the practical challenge of translating abstract notions of trustworthiness into precise mathematical formalisms.

The second route takes a pragmatic approach: trustworthiness is operationalized through a portfolio of carefully human-defined quantitative metrics [16]. These include ethical metrics: fairness [35], including demographic parity or counterfactual fairness; explainability [36], encompassing intrinsic or post-hoc approaches, model-specific or model-agnostic methods, and global or local perspectives, whether feature- or example-based; robustness [37], with respect to both natural and adversarial perturbations; and privacy guarantees [38], ranging from mitigation strategies based on differential privacy to stronger protections relying on cryptographic protocols. Additional technical dimensions [39] involve compliance with regulations, validity, accountability, replicability, and computational or energy requirements. Finally, domain-specific considerations [40] address clinical validity and utility, alignment with biomedical knowledge, conformity with medical standards and regulations, and the integration of expert feedback.

This second strategy has achieved remarkable results, yet it is not without limitations. First, most metrics are aggregated, capturing only average behavior. Pointwise guarantees, which would ensure safe performance for every individual patient or rare yet clinically critical sub-populations, generally require symbolic reasoning and remain elusive [41]. Second, many desiderata are mutually incompatible, as illustrated by fairness impossibility results [42] and tensions between competing trustworthiness metrics [16]. Certain dimensions of trustworthiness are actually binary. For instance, when strict confidentiality is required [38], differential privacy or anonymization can only mitigate risk. Stronger guarantees demand cryptographic protocols such as secure multi-party computation or homomorphic encryption, which severely restrict the set of feasible algorithms and inflate computational costs. Some metrics are also ill-posed. Explainability exemplifies this challenge: post-hoc techniques provide useful probes of decision surfaces but only approximate, rather than faithfully reveal, the internal causal logic of opaque models [36]. Unless interpretability is built in by design—through sparse rule sets or symbolic programs—explanations should be presented as hypotheses, not ground truths, and validated through user studies or counterfactual tests. Finally, optimizing any single dimension of trustworthiness can expose hidden vulnerabilities elsewhere, triggering cascades of additional diagnostics [43]. Iterative, multi-faceted evaluation, therefore, becomes a continuous obligation rather than a one-time certification step [44].

Nevertheless, while human oversight should be embedded in ML-based systems, human decisions still rely on the quality of the underlying technical components; relying solely on qualitative, human-based judgments can be equally risky. Striking the right balance is crucial—human-defined criteria may take the lead, but technical rigor remains essential.

Tip 4: Look for a compromise between all involved parties

Biomedical informatics projects often involve multiple stakeholders, including ML researchers, biomedical scientists, clinicians, study designers, patients, grant holders, and legal or administrative personnel. These roles bring distinct priorities and interpretations of trustworthiness. For example, patients may prioritize data privacy [45], mental health professionals may emphasize transparency and explainability [46], while researchers and editors may focus on methodological rigor and adherence to publication standards [47].

We recommend that these perspectives be carefully balanced, with patients placed at the top of the priority list [48, Rule 4]. Not all constraints are negotiable: legal and ethical frameworks—such as Data- and AI-act—may impose strict limitations, while budgetary or methodological restrictions can preclude otherwise desirable analyses. Compromise must therefore extend beyond scientific objectives to encompass practical implementation, communication strategies, and publication venues. To make this process actionable, concrete engagement techniques such as stakeholder interviews, co-design workshops, consensus-building approaches, or Delphi panels can be employed to elicit and integrate diverse perspectives [49]. Open and inclusive discussion remains essential, with the recognition that perfect consensus is rare and some dissatisfaction is inevitable [50]. To avoid analysis paralysis, a transparent decision-making process should be established, ensuring that trade-offs are openly documented and justified.

Tip 5: Take care of the technical aspects

This tip provides technical tools to address Tip 3's limits. Biomedical datasets are outpacing Moore's law, fueled by high-resolution imaging, next-generation sequencing, and continuous remote monitoring [51]. Harnessing hardware accelerators and distributed frameworks enables laboratories to iterate on large models while keeping training times practical [52,53]. Yet raw speed is only one side of the coin: computation, energy, and memory footprints also matter [54]. Techniques such as mixed-precision training, quantization, pruning, and knowledge distillation can shrink model size without sacrificing diagnostic accuracy [55].

Prototype notebooks often graduate directly into production dashboards, accumulating hidden technical debt that hampers future innovation [39]. Adopting modular, version-controlled pipelines, augmented with automated unit and integration tests and infrastructure-as-code templates, helps systems remain malleable while satisfying good machine-learning-practice guidelines [56].

Biomedical ML must satisfy both domain-agnostic and domain-specific regulation [11,12,57]. Core practices include strict data-provenance logging, privacy-preserving pipelines, and routine model-governance audits that trace every prediction back to a versioned model artifact and dataset snapshot [56,58,59]. This is tightly connected with the concept of accountability [60,61] requiring continuous monitoring, standard governance artifact, sign-off from domain experts, data stewards, and compliance officer.

Classic benchmarks seldom capture the nuances of real-world clinical deployment [62]. In addition to accuracy, teams should track diverse metrics and log every experimental choice to facilitate audits and exact reproduction during publication or regulatory review [63,64]. The rise of billion-parameter foundation models introduces further operational burdens: teams must track the provenance of web-scale training corpora, expunge sensitive content, and apply safe, domain-specific fine-tuning [56]. Table 1 presents typical scopes along with the associated tasks and libraries that handle the technical aspects.

Tip 6: Take care of the ethical aspects

This tip offers ethical tools to complement Tip 3's boundaries. Systematic biases in electronic health records, biobanks, or imaging repositories can propagate directly into clinical-decision tools, widening existing health disparities [35,65]. Before model development, audit cohort composition with respect to sensitive attributes and simulate counterfactual performance gaps using a-priori fairness metrics. If substantial imbalance is detected, apply pre-, in-, or post-processing mitigation strategies.

Regulators increasingly expect human-interpretable rationales for ML-driven recommendations in patient care [36,66]. Use explainable models by design where feasible, or apply post-hoc methods when interpretability must be added after training. Combine model-specific tools (for example, gradient-based) with model-agnostic ones (for example, perturbation-based) to strengthen trust. Global explanations should confirm alignment with biomedical mechanisms, while local explanations—whether feature-based or example-based—should provide per-patient justifications that clinicians can assess. Natural and adversarial perturbations to medical images can flip a cancer diagnosis, and data-poisoning attacks can smuggle malicious correlations into federated genomic models [37]. Before clinical validation, subject the entire inference stack to testing, formal verification for worst-case guarantees on high-risk thresholds, and scenario-based simulations that probe rare but plausible edge cases [67].

Strict data-protection regimes restrict the handling of personal biomedical data [38]. Differential privacy and secure multiparty computation have matured to the point where they can be embedded into everyday ML pipelines. If data sovereignty laws prevent centralization, consider federated or split-learning protocols backed by secure multiparty computation or homomorphic encryption. Table 2 presents typical scopes along with the associated tasks and libraries that handle the ethical aspects.

Table 1. Typical scopes along with the associated tasks and libraries that handle the technical aspects.

Scope	Tasks	Libraries
Hardware accelerators and distributed training	GPU/TPU utilization, multi-node scaling, memory sharding	CUDA/cuDNN, ROCm, oneAPI, PyTorch DDP/FSDP, DeepSpeed (ZeRO), Horovod, Ray Train, Lightning Fabric, Hugging Face Accelerate, Megatron-LM, JAX + XLA
Efficient training (mixed precision, quantization)	Automatic mixed precision, post-training, or quantization-aware training	PyTorch AMP, NVIDIA Apex, JAX bf16, TF mixed precision, ONNX Runtime Quantization, TensorRT INT8, OpenVINO, Intel Neural Compressor, bitsandbytes (8/4-bit)
Efficient inference and serving	Low-latency serving, graph optimization	TensorRT, Triton Inference Server, TorchTensorRT, TF-TRT, ONNX Runtime, vLLM, Text Generation Inference (TGI), TorchServe, TF Serving, BentoML, KServe
Pruning and sparsity	Structured/unstructured pruning, sparse kernels	PyTorch pruning, TensorFlow Model Optimization Toolkit (TF-MOT), SparseML/DeepSparse, NNI pruning
Parameter-efficient fine-tuning	LoRA/QLoRA, adapters	PEFT (HF), LoRA, QLoRA, Adapters (adapter-transformers)
Knowledge distillation	Teacher-student training	torchdistill, Hugging Face Transformers distillation scripts, TextBrewer (NLP), TF-MOT distillation
Notebook to maintainable pipeline	Pipelines, orchestration, modularity	Dagster, Prefect, Airflow, Kubeflow Pipelines, Metaflow, Luigi
Version control (code and data)	Git, data artifacts, datasets	Git/Git LFS, DVC, lakeFS, Pachyderm, Quilt, Delta Lake/Iceberg/Hudi
Testing (unit/integration/data)	CI, data quality, ML-specific tests	pytest, hypothesis, Great Expectations, Soda, DeeChecks, Giskard
Infrastructure as code and environments	Reproducible infrastructure, kubernetes, packaging	Terraform, Pulumi, CloudFormation, Helm/Kustomize, Docker/Podman, Poetry/Conda
Experiment tracking and reproducibility	Runs, params, artifacts, registries	MLflow (Tracking + Model Registry), Weights and Biases, Neptune, Comet, Sacred
Data provenance and lineage	Track sources, transformations	OpenLineage/Marquez, DataHub, OpenMetadata, Amundsen, MLMD
Governance artifacts	Model/dataset cards, documentation	Model Card Toolkit, Responsible AI Dashboard (Azure), Fairlearn, AIF360
Security and confidential compute	Encrypted/isolated training and inferencing	OpenFHE/SEAL/PALISADE (HE), SGX/TDX runtimes (e.g., Gramine), HashiCorp Vault, OPA
Compliance workflow and audits	Evidence collection, approvals, trials	Jira/ServiceNow integrations, Confluence, Audit log exporters, Schellman Evidence, Vanta (general GRC)
Monitoring and drift	Data/label drift, performance, alerts	Evidently, whylogs/WhyLabs, Arize AI, Fiddler, Seldon Alibi Detect, SageMaker Model Monitor, Vertex Model Monitoring
Benchmarking and eval (beyond accuracy)	Robustness, fairness, clinical realism	MedPerf (medical eval), RobustBench, HoloClean (data), Fairlearn/AIF360, HELICON/evals (LLM evals), Promptfoo
Foundation-model data curation	Corpus filtering, dedup, licensing	ccnet, datatrove (HF), Dolma (AI2), Cleanlab, datasketch (MinHash), scancode-toolkit, reuse-tool
Safety and content filtering	PII/toxicity/medical-risk filters	Presidio, Detoxify, Perspective API (service), OpenAI Safety Spec tooling analogs, LAION NSFW detectors
Deployment in clinical settings	API/serving, rollouts, A/B testing, canary releases	BentoML, KServe, Seldon Core, Istio (canary), FastAPI

https://doi.org/10.1371/journal.pcbi.1013624.t001

Tip 7: Leverage the domain knowledge

In biomedical ML, integrating domain knowledge is essential for trustworthiness [17]. Biomedical expertise shapes datasets, model design, and interpretation, provides insights that align ML with biomedical logic, and grounds models in scientific reasoning.

Domain knowledge can be integrated at three key stages: pre-processing, in-processing, and post-processing [17].

Table 2. Typical scopes along with the associated tasks and libraries that handle the ethical aspects.

Scope	Tasks	Libraries
Bias and fairness auditing	Audit cohort composition, simulate counterfactuals, compute fairness metrics	AIF360 (IBM), Fairlearn, Themis-ML, EthicalML/ Fairness Indicators, aequitas, Giskard
Bias mitigation (pre/ in/post)	Rebalancing, adversarial debiasing, reweighting, equalized odds	AIF360, Fairlearn, TensorFlow Constrained Optimization, FairTorch
Explainability (global and local)	Post-hoc explanations, interpretable models	SHAP, LIME, Captum (PyTorch), ELI5, InterpretML, inherently interpretable models: Explainable Boosting Machine, GLRM
Biomedical-specific explainability	Align explanations with biomedical knowledge, feature attributions	BioSHAP (biomedical SHAP adaptations), Klarity AI (radiology), Integrated Gradients (Captum/ JAX), Grad-CAM/Score-CAM for imaging
Robustness and adversarial testing	Image perturbations, poisoning simulations, adversarial robustness	CleverHans, Adversarial Robustness Toolbox (ART), Foolbox, TextAttack (for NLP), DeepRobust
Formal verification and safety checks	Provable guarantees, threshold certification	ERAN (ETH Robustness Analyzer), Marabou, AI2's Reluplex, DeepCert, DeepPoly
Scenario-based simulations	Rare/edge-case probing, stress testing	Carla (causal simulations), scikit-multiflow (stream scenarios), custom Monte Carlo sims, PyRIT (Red Teaming AI)
Differential privacy	Differentially private stochastic gradient descent, noise addition, private queries	Opacus (PyTorch), TensorFlow Privacy, SmartNoise (OpenDP)
Secure multiparty computation	Training without data centralization	CrypTen (Meta), PySyft (OpenMined), TF Encrypted, MOTION2NX
Federated and split learning	Distributed learning across sites	Flower, TensorFlow Federated, FedML, OpenFL (Intel)
Homomorphic encryption	Encrypted computation	HElib, PALISADE, SEAL (Microsoft), Concrete (Zama)
Data sovereignty/ governance	Compliance with data residency laws	Gaia-X connectors, lakeFS with region-based storage policies, DataHub lineage + policy enforcement

https://doi.org/10.1371/journal.pcbi.1013624.t002

In pre-processing, biological priors guide feature selection, engineering, and cleaning. For example, biomarkers or gene pathways help prioritize variables in genomics, while domain expertise informs experiments that capture disease heterogeneity and patient subpopulations [68].

In-processing embeds constraints directly into the learning objective [17]. This includes regularization enforcing monotonicity or conservation laws, and hybrid models that combine empirical data with physics-based simulators. Such strategies improve generalization and reduce biologically implausible results under distribution shifts [69].

Post-processing ensures outputs comply with clinical rules or logical constraints, such as disease taxonomies [70]. This step makes predictions more actionable and reliable. For example, one could exploit KEGG pathways in genomics [71] to validate prediction of biomedical annotations made on the Gene Ontology database [72]. Table 3 presents typical scopes along with the associated tasks and libraries to leverage domain knowledge.

Tip 8: Follow open science best practices, and document everything

Trustworthy ML requires open technologies and data, making adherence to open science best practices essential. We recommend using open-source programming languages, openly sharing data (with proper authorization), releasing all code publicly, and publishing in open-access journals. Open-source tools such as Python or R enable free and reproducible

Table 3. Typical scopes along with the associated tasks and libraries to leverage domain knowledge.

Scope	Tasks	Libraries
Pre-processing (domain-guided)	Feature selection, biological priors, biomarkers, cleaning, stratification	BioPython, Scanpy (single-cell), Bioconductor (R), scikit-learn feature selection, SHAP (feature importance), Omics-specific DBs (KEGG, Reactome, Gene Ontology), pandas-profiling/ ydata-profiling, Great Expectations
Experiment design/ cohort definition	Stratifying by disease subpopulations, handling heterogeneity	EHR data frameworks: FHIR APIs, OHDSI/OMOP CDM, pyomop, i2b2, DataSHIELD (privacy-preserving cohorting)
In-processing (knowledge-embedded)	Embedding priors/constraints in the model, hybrid physics-ML models	PyTorch constraints/ monotonic networks, TensorFlow Lattice, scikit-monotonic, DeepXDE (physics-informed), SimulAI/ SciML (Julia), PINNs frameworks, GPyTorch with structured kernels
Hybrid biomedical + ML models	Coupling simulators with ML	COPASI (biochem simulation), PySB (systems biology), NEURON (neuroscience), CellML/OpenCOR
Regularization with priors	Encourage plausible behavior under shifts	PyTorch custom regularizers, Keras custom loss layers, GPflow (structured priors)
Post-processing (knowledge-based validation)	Constraining outputs, aligning with biomedical logic, ontologies	OntoBio, PyOBO, OWLready2 (ontologies), SNOMED CT/ UMLS APIs, BioPortal, GOATOOLS
Rule-based combination with ML outputs	Logical/clinical constraints after model prediction	Drools (rule engine), PyKnow, Experta, Logica (Google)
Explainability aligned with biomedical knowledge	Making sure explanations match biological mechanisms	BioSHAP, Captum (with pathway grouping), Pathway2Vec, Knowledge Graph Embeddings (DGL-KE, PyKEEN)

https://doi.org/10.1371/journal.pcbi.1013624.t003

experimentation, while proprietary software restricts access to those with licenses [73]. Where informed consent and ethics approval allow, patient data should be released through repositories like Zenodo, Figshare, Gene Expression Omnibus, or ArrayExpress. Both raw and processed data should be shared to maximize reuse and accelerate discovery [74]. Reproducibility further requires making all analysis code freely available on platforms such as GitHub or GitLab [75]. Open access publishing ensures that results are freely available worldwide, increases visibility in under-resourced regions, and is associated with higher citation rates [76].

Finally, we stress the importance of clear documentation, including well-annotated code [77] and a scientific diary tracking study progress and key methodological decisions [78]. Documenting choices such as evaluation metrics enhances transparency and strengthens trustworthiness. Of course, we know that sometimes it is impossible to share data in hospitals and biomedical research centers for privacy reasons, and we agree that patients' rights should always be protected [48].

Tip 9: Make your whole machine learning pipeline trustworthy

In this final tip, we argue that trustworthy biomedical ML requires rethinking the entire pipeline [79], from research question formulation to data collection, feature engineering, model training, evaluation, deployment, monitoring, and maintenance [16,41,44,80,81] (Fig 1).

While all stages benefit from the proposed tips, some have particular relevance in specific contexts. Tip 1 stresses the need to anticipate long-term consequences of design choices. Tip 3 highlights data and model limitations when defining system capabilities. Tip 4 calls for stakeholder involvement—including patients, clinicians, data scientists, and regulators—to ensure diverse perspectives. Tip 8 promotes open science practices, emphasizing transparency, documentation,

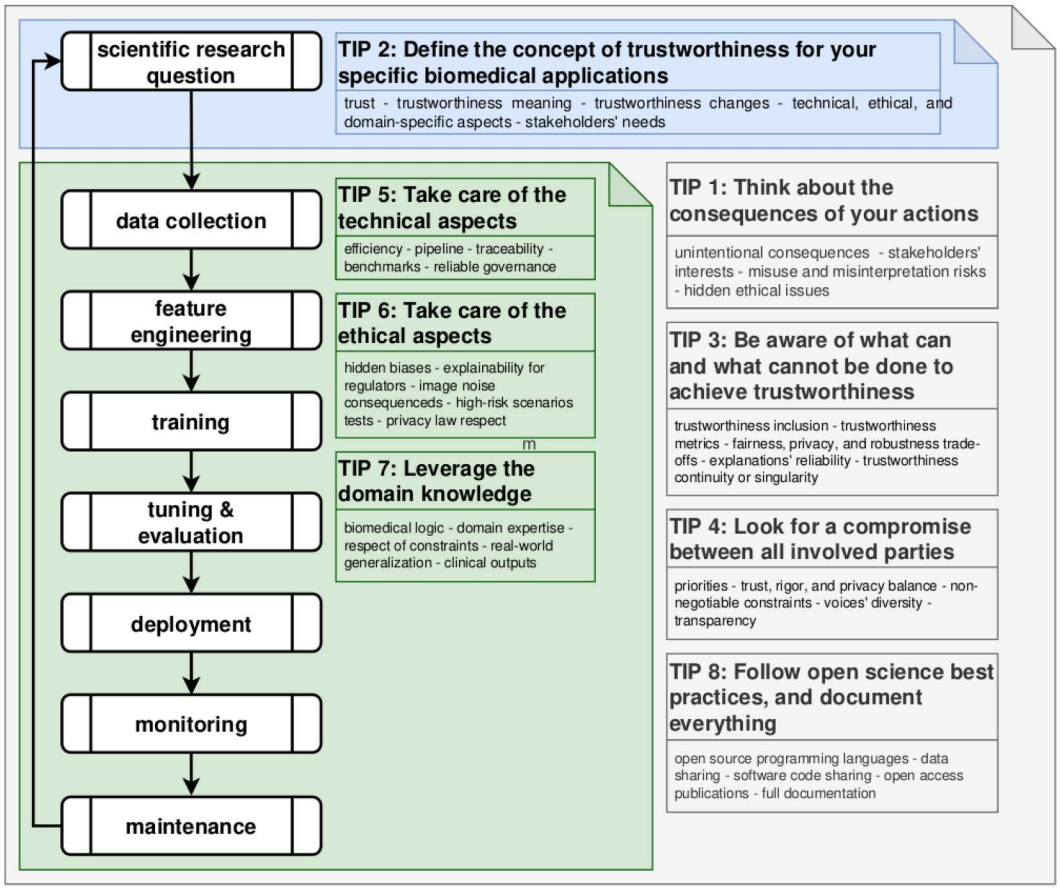

Fig 1. Representation of our tips for a trustworthy machine learning pipeline cycle in the biomedical sciences.

https://doi.org/10.1371/journal.pcbi.1013624.g001

and reproducibility. At the problem-definition stage, Tip 2 is crucial: it urges practitioners to define what trustworthiness means in context, aligning goals with ethical, clinical, and societal expectations.

During data collection and feature engineering, Tip 7 underscores the integration of domain expertise to embed relevant prior knowledge and ensure meaningful features. In model training, deployment, and maintenance, Tips 5, 6, and 7 collectively address technical, ethical, and domain-specific requirements, guiding the design of systems that are not only performant but also trustworthy. Trustworthiness is therefore not an afterthought, but a principle to be embedded at every stage of the pipeline [16,41,44,80,81].

A detailed workflow explaining the differences between a trustworthy machine learning pipeline and a traditional pipeline can be found in the study by Rasheed and coauthors [82]. An illustrative example of a wise usage of our eight tips is the study by Gardiner and colleagues [83]: they applied a transparent, trustworthy machine learning approach to genomics data and chemical structure information to predict kidney dysfunction, as a proxy for drug-induced renal toxicity, in rats.

Conclusions

With the rapid growth of ML in biomedical sciences, there is increasing recognition of the need for trustworthy approaches among researchers, professionals, and regulators. Black-box effectiveness is no longer sufficient; evaluation must also consider technical, ethical, and domain-specific criteria. Yet, despite widespread efforts, few studies fully achieve trustworthiness.

In response, we present nine practical recommendations to strengthen ML in biomedical sciences. These tips, if adopted, can enhance the trustworthiness of any study involving ML, and are equally relevant to other scientific fields where ML plays a central role.

Author contributions

Conceptualization: Luca Oneto.

Formal analysis: Luca Oneto, Davide Chicco.

Investigation: Luca Oneto.

Methodology: Luca Oneto, Davide Chicco.

Resources: Davide Chicco.

Validation: Davide Chicco.

Writing – original draft: Luca Oneto, Davide Chicco.

Writing – review & editing: Luca Oneto, Davide Chicco.

References

1. Shalev-Shwartz S, Ben-David S. Understanding machine learning: from theory to algorithms. Cambridge, England, United Kingdom: Cambridge University Press; 2014.
2. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773–84. https://doi.org/10.1038/s41591-022-01981-2 PMID: 36109635
3. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969 PMID: 39466245
4. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 2024;52(D1):D368–75. https://doi.org/10.1093/nar/gkad1011 PMID: 37933859
5. Norvig P. Artificial intelligence: a modern approach. 4th ed. Pearson; 2020.
6. Spyropoulos CD. AI planning and scheduling in the medical hospital environment. Artif Intell Med. 2000;20(2):101–11. https://doi.org/10.1016/s0933-3657(00)00059-2 PMID: 10936748
7. Wang H, Zhou T, Tian L, Qian Y, Li J. Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data. J Biomed Inform. 2014;52:354–63. https://doi.org/10.1016/j.jbi.2014.07.017 PMID: 25109270
8. Marwala T. Causality, correlation and artificial intelligence for rational decision making. World Scientific; 2015.
9. Eshete B. Making machine learning trustworthy. Science. 2021;373(6556):743–4. https://doi.org/10.1126/science.abi5052 PMID: 34385384
10. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst. 2009;24(2):8–12. https://doi.org/10.1109/mis.2009.36
11. European Commission. Data Act. [cited 2025 June 1]. Available from: https://digital-strategy.ec.europa.eu/en/policies/data-act
12. European Commission. AI Act. [cited 2025 June 1]. Available from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
13. McGregor S. Preventing repeated real world AI failures by cataloging incidents: the AI incident database. In: AAAI Conference on Artificial Intelligence; 2021.
14. Gilbert S. The EU passes the AI Act and its implications for digital medicine are unclear. NPJ Digit Med. 2024;7(1):135. https://doi.org/10.1038/s41746-024-01116-6 PMID: 38778162
15. Tantithamthavorn CK, Palomba F, Khomh F, Chua JJ. MLOps, LLMOps, FMOps, and beyond. IEEE Software. 2025;42(01):26–32.
16. Oneto L, Ridella S, Anguita D. Towards algorithms and models that we can trust: a theoretical perspective. Neurocomputing. 2024;592:127798. https://doi.org/10.1016/j.neucom.2024.127798
17. Oneto L, Chicco D. Eight quick tips for biologically and medically informed machine learning. PLoS Comput Biol. 2025;21(1):e1012711. https://doi.org/10.1371/journal.pcbi.1012711 PMID: 39787089
18. Huang L-C, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC Genomics. 2011;12 Suppl 5(Suppl 5):S11. https://doi.org/10.1186/1471-2164-12-S5-S11 PMID: 22369493
19. Duckett SJ, Breadon P, Romanes D. Identifying and acting on potentially inappropriate care. Med J Aust. 2015;203(4):183e.1-6. https://doi.org/10.5694/mja15.00025 PMID: 26268287

20. Baptiste D-L, Caviness-Ashe N, Josiah N, Commodore-Mensah Y, Arscott J, Wilson PR, et al. Henrietta Lacks and America's dark history of research involving African Americans. Nurs Open. 2022;9(5):2236–8. https://doi.org/10.1002/nop2.1257 PMID: 35700235
21. Bloomrosen M, Starren J, Lorenzi NM, Ash JS, Patel VL, Shortliffe EH. Anticipating and addressing the unintended consequences of health IT and policy: a report from the AMIA 2009 Health Policy Meeting. J Am Med Inform Assoc. 2011;18(1):82–90. https://doi.org/10.1136/jamia.2010.007567 PMID: 21169620
22. Ní Shé É, Harrison R. Mitigating unintended consequences of co-design in health care. Health Expect. 2021;24(5):1551–6. https://doi.org/10.1111/hex.13308 PMID: 34339528
23. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A survey of the literature on unintended consequences associated with health information technology: 2014-2015. Yearb Med Inform. 2016;(1):13–29. https://doi.org/10.15265/IY-2016-036 PMID: 27830227
24. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med. 2020;3:47. https://doi.org/10.1038/s41746-020-0254-2 PMID: 32258429
25. Park C, Awadalla A, Kohno T, Patel S. Reliable and trustworthy machine learning for health using dataset shift detection. In: Neural information processing systems; 2021.
26. Huang J, Xin B, Wang X, Qi Z, Dong H, Li K, et al. Multi-parametric MRI phenotype with trustworthy machine learning for differentiating CNS demyelinating diseases. J Transl Med. 2021;19(1):377. https://doi.org/10.1186/s12967-021-03015-w PMID: 34488799
27. López-López E, Bajorath J, Medina-Franco JL. Informatics for chemistry, biology, and biomedical sciences. J Chem Inform Model. 2020;61(1):26–35.
28. Kabir MN, Wang LR, Goh WWB. Exploiting the similarity of dissimilarities for biomedical applications and enhanced machine learning. PLoS Comput Biol. 2025;21(1):e1012716. https://doi.org/10.1371/journal.pcbi.1012716 PMID: 39854337
29. Lee BD, Gitter A, Greene CS, Raschka S, Maguire F, Titus AJ, et al. Ten quick tips for deep learning in biology. PLoS Comput Biol. 2022;18(3):e1009803. https://doi.org/10.1371/journal.pcbi.1009803 PMID: 35324884
30. Kherroubi Garcia I, Erdmann C, Gesing S, Barton M, Cadwallader L, Hengeveld G, et al. Ten simple rules for good model-sharing practices. PLoS Comput Biol. 2025;21(1):e1012702. https://doi.org/10.1371/journal.pcbi.1012702 PMID: 39792790
31. Goh WWB, Kabir MN, Yoo S, Wong L. Ten quick tips for ensuring machine learning model validity. PLoS Comput Biol. 2024;20(9):e1012402. https://doi.org/10.1371/journal.pcbi.1012402 PMID: 39298376
32. Winfield AF, Michael K, Pitt J, Evers V. Machine ethics: the design and governance of ethical AI and autonomous systems [scanning the issue]. Proc IEEE. 2019;107(3):509–17. https://doi.org/10.1109/jproc.2019.2900622
33. Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. J Exp Theor Artif Intell. 2000;12(3):251–61. https://doi.org/10.1080/09528130050111428
34. Dingli A, Farrugia D. Neuro-symbolic AI: design transparent and trustworthy systems that understand the world as you do. Packt Publishing; 2023.
35. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1–35. https://doi.org/10.1145/3457607
36. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
37. Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recognit. 2018;84:317–31. https://doi.org/10.1016/j.patcog.2018.07.023
38. Al-Rubaie M, Chang JM. Privacy-preserving machine learning: threats and solutions. IEEE Secur Privacy. 2019;17(2):49–58. https://doi.org/10.1109/msec.2018.2888775
39. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D. Hidden technical debt in machine learning systems. In: Neural Information Processing Systems; 2015.
40. Chicco D, Oneto L, Tavazzi E. Eleven quick tips for data cleaning and feature engineering. PLoS Comput Biol. 2022;18(12):e1010718. https://doi.org/10.1371/journal.pcbi.1010718 PMID: 36520712
41. Kemmerzell N, Schreiner A, Khalid H, Schalk M, Bordoli L. Towards a better understanding of evaluating trustworthiness in AI systems. ACM Comput Surv. 2025;57(9):1–38. https://doi.org/10.1145/3721976
42. Green B. Escaping the impossibility of fairness: from formal to substantive algorithmic fairness. Philos Technol. 2022;35(4). https://doi.org/10.1007/s13347-022-00584-6
43. Xu H, Liu X, Li Y, Jain A, Tang J. To be robust or to be fair: towards fairness in adversarial training. In: International conference on machine learning; 2021.
44. Vetter D, Amann J, Bruneault F, Coffee M, Düdder B, Gallucci A, et al. Lessons learned from assessing trustworthy AI in practice. DISO. 2023;2(3). https://doi.org/10.1007/s44206-023-00063-1
45. Milne R, Morley KI, Almarri MA, Anwer S, Atutornu J, Baranova EE, et al. Demonstrating trustworthiness when collecting and sharing genomic data: public views across 22 countries. Genome Med. 2021;13(1):92. https://doi.org/10.1186/s13073-021-00903-0 PMID: 34034801

46. Chandler C, Foltz PW, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. Schizophr Bull. 2020;46(1):11–4. https://doi.org/10.1093/schbul/sbz105 PMID: 31901100
47. Yarborough M. Taking steps to increase the trustworthiness of scientific research. FASEB J. 2014;28(9):3841–6. https://doi.org/10.1096/fj.13-246603 PMID: 24928193
48. Chicco D, Jurman G. Ten simple rules for providing bioinformatics support within a hospital. BioData Min. 2023;16(1):6. https://doi.org/10.1186/s13040-023-00326-0 PMID: 36823520
49. Lloyd-Williams F, Hyseni L, Guzman-Castillo M, Kypridemos C, Collins B, Capewell S, et al. Evaluating stakeholder involvement in building a decision support tool for NHS health checks: co-producing the WorkHORSE study. BMC Med Inform Decis Mak. 2020;20(1):182. https://doi.org/10.1186/s12911-020-01205-y PMID: 32778087
50. Ehrmann DE, Joshi S, Goodfellow SD, Mazwi ML, Eytan D. Making machine learning matter to clinicians: model actionability in medical decision-making. NPJ Digit Med. 2023;6(1):7. https://doi.org/10.1038/s41746-023-00753-7 PMID: 36690689
51. Kocheturov A, Pardalos PM, Karakitsiou A. Massive datasets and machine learning for computational biomedicine: trends and challenges. Ann Oper Res. 2018;276(1–2):5–34. https://doi.org/10.1007/s10479-018-2891-2
52. Martínez-Fernández S, Bogner J, Franch X, Oriol M, Siebert J, Trendowicz A, et al. Software engineering for AI-based systems: a survey. ACM Trans Softw Eng Methodol. 2022;31(2):1–59. https://doi.org/10.1145/3487043
53. Talib MA, Majzoub S, Nasir Q, Jamal D. A systematic literature review on hardware implementation of artificial intelligence algorithms. J Supercomput. 2020;77(2):1897–938. https://doi.org/10.1007/s11227-020-03325-8
54. van Wynsberghe A. Sustainable AI: AI for sustainability and the sustainability of AI. AI Ethics. 2021;1(3):213–8. https://doi.org/10.1007/s43681-021-00043-6
55. Zhang B, Wang T, Xu S, Doermann D. Neural networks with model compression. Springer; 2024.
56. Gift N, Deza A. Practical MLOps. O'Reilly Media, Inc. 2021.
57. US Centers for Disease Control and Prevention (CDC). Health Insurance Portability and Accountability Act of 1996 (HIPAA). [cited 2025 June 1]. Available from: https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html
58. Masum H, Bourne PE. Ten simple rules for humane data science. PLoS Comput Biol. 2023;19(12):e1011698. https://doi.org/10.1371/journal.pcbi.1011698 PMID: 38127691
59. Malik A, Patel P, Ehsan L, Guleria S, Hartka T, Adewole S, et al. Ten simple rules for engaging with artificial intelligence in biomedicine. PLoS Comput Biol. 2021;17(2):e1008531. https://doi.org/10.1371/journal.pcbi.1008531 PMID: 33571194
60. Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: what it is and how it works. AI & Soc. 2023;39(4):1871–82. https://doi.org/10.1007/s00146-023-01635-y
61. Busuioc M. Accountable artificial intelligence: holding algorithms to account. Public Adm Rev. 2021;81(5):825–36. https://doi.org/10.1111/puar.13293 PMID: 34690372
62. Mahmood F. A benchmarking crisis in biomedical machine learning. Nat Med. 2025;31(4):1060. https://doi.org/10.1038/s41591-025-03637-3 PMID: 40200055
63. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics. 2020;26(4):2141–68. https://doi.org/10.1007/s11948-019-00165-5 PMID: 31828533
64. Sohn E. The reproducibility issues that haunt health-care AI. Nature. 2023;613(7943):402–3. https://doi.org/10.1038/d41586-023-00023-2 PMID: 36624237
65. Chen RJ, Chen TY, Lipkova J, Wang JJ, Williamson DFK, Lu MY. Algorithm fairness in AI for medicine and healthcare. arXiv preprint. 2021. https://doi.org/10.48550/arXiv.2110.00603
66. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):1–33.
67. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Man´e D. Concrete problems in AI safety. 2016. https://doi.org/arXiv:160606565
68. Chicco D, Sanavia T, Jurman G. Signature literature review reveals AHCY, DPYSL3, and NME1 as the most recurrent prognostic genes for neuroblastoma. BioData Min. 2023;16(1):7. https://doi.org/10.1186/s13040-023-00325-1 PMID: 36870971
69. Haig BD. What is a spurious correlation?. Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences. 2003;2(2):125–32.
70. Giunchiglia E, Imrie F, Van Der Schaar M, Lukasiewicz T. Machine learning with requirements: a manifesto. Neurosymbolic Artif Intell. 2025;1:240767.
71. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662
72. Chicco D, Masseroli M. Ontology-based prediction and prioritization of gene functional annotations. IEEE/ACM Trans Comput Biol Bioinform. 2016;13(2):248–60. https://doi.org/10.1109/TCBB.2015.2459694 PMID: 27045825

73. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285 PMID: 24204232
74. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
75. Barnes N. Publish your computer code: it is good enough. Nature. 2010;467(7317):753. https://doi.org/10.1038/467753a PMID: 20944687
76. Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, et al. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ. 2018;6:e4375. https://doi.org/10.7717/peerj.4375 PMID: 29456894
77. Karimzadeh M, Hoffman MM. Top considerations for creating bioinformatics software documentation. Brief Bioinform. 2018;19(4):693–9. https://doi.org/10.1093/bib/bbw134 PMID: 28088754
78. Noble WS. A quick guide to organizing computational biology projects. PLoS Comput Biol. 2009;5(7):e1000424. https://doi.org/10.1371/journal.pcbi.1000424 PMID: 19649301
79. Hapke H, Nelson C. Building machine learning pipelines. O'Reilly Media; 2020.
80. Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: from principles to practices. ACM Comput Surv. 2023;55(9):1–46. https://doi.org/10.1145/3555803
81. Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. ACM Comput Surv. 2022;55(2):1–38. https://doi.org/10.1145/3491209
82. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Comput Biol Med. 2022;149:106043. https://doi.org/10.1016/j.compbiomed.2022.106043 PMID: 36115302
83. Gardiner L-J, Carrieri AP, Wilshaw J, Checkley S, Pyzer-Knapp EO, Krishna R. Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. Sci Rep. 2020;10(1):9522. https://doi.org/10.1038/s41598-020-66481-0 PMID: 32533004