Curating Benchmark Data for Healthcare AI: Four Axioms for Success and New Datasets

Engy Ziedan* Richard Ho†

October 23, 2025

Abstract

Healthcare is incredibly complex. Many scenarios and interactions between patients and systems or patients and physicians lack closed form solutions. As such, evaluations of AI on real world healthcare capabilities must mimic such complexity. In this brief, we first analyze the landscape of healthcare benchmarks on Hugging Face and related repositories, we also review FDA 510(k) submissions for AI-enabled medical devices. We found that while multimodal benchmarks that offer realistic, non-synthetic healthcare scenarios remain limited (approximately <10%), there is a rising share of such benchmarks in recent years. We then propose four axioms that define what makes a benchmark dataset optimal for AI evaluations: *internally valid*, *externally valid*, *uncontaminated* (independent of training), and *sufficient in sample size* for heterogeneous effects testing. Finally, we provide examples of benchmark datasets we have constructed—spanning dermatology portal Q&A, nurse narratives with triage, multi-modal oncology staging and tumor board cases, and early diagnostic/progression datasets in cardiology and oncology.

1 Introduction

Healthcare AI evaluations today lack realism—defined here as multi modality and external validity to real-world settings. We summarize how the number and type of public benchmarks have evolved over time. From over 800 Hugging Face–posted relevant healthcare benchmarks, we found only a smaller subset to have publications (peer-reviewed or not) that describe data construction and sample sizes, yielding a working set of approximately 270 benchmarks. We describe this list in Appendix A.

Most early benchmarks were synthetic or text-only, while multi-modal datasets—those combining notes, imaging, and structured data—remain rare. That share is increasing but still limited. A handful of international datasets exist, yet many reuse overlapping internet data or translated variants (e.g., multiple language versions of MedQA), risking overlap between training and test sets. Recently, benchmarks such as [10] represent effort to build a diverse, non-contaminated benchmark set for healthcare AI covering over 5000 questions by over 250 physicians—although it remains primarily text-only rather than multi-modal and not multi turn.

This limited external validity in evaluations is a barrier to approving more generative, multitask AI systems in healthcare. Most regulatory clearances (such as FDA 510(k)) of AI tools today focus on narrowly defined diagnostic applications - often single-modality models evaluated in isolated tasks. Even there, validation datasets rarely reflect real-world multi-modal patient contexts that combine imaging, notes, and longitudinal data. This has likely slowed some approvals and overall technology penetration.

*Assistant Professor of Economics, Tulane University; Affiliate Researcher, O'Neill School of Environmental and Public Affairs, Indiana University; and Chief Scientific Officer, Protege. engy@withprotege.ai

†Chief Technology Officer, Protege. richard@withprotege.ai

Figure 1 shows the count of public healthcare benchmarks by year and modality (multi-modal, text-only, or other).

- **Multi-modal:** Uses electronic medical records (EMR) together with imaging, pathology, or other biometric data—for example, SlideChat-2024 and TCGA-PRAD.
- **Text-only:** EMR or medical exam datasets without other modalities—for example, HealthBench-2025, MedQA-USMLE, and MEDS-BENCH.
- **Other:** Audio transcripts, administrative, or revenue-cycle datasets that can be text-based but are not clinical narratives.

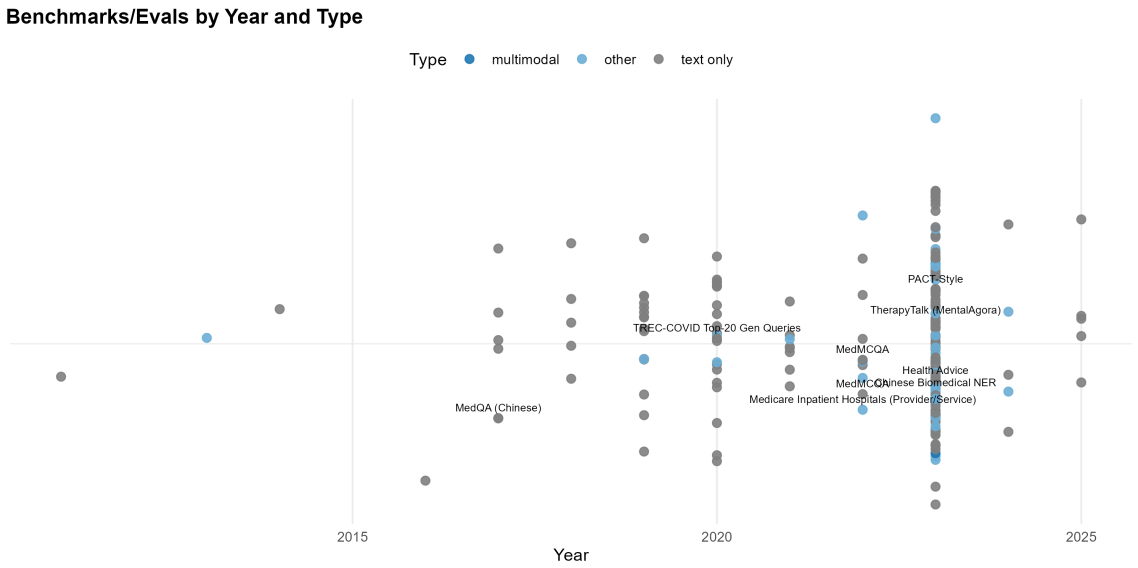

Figure 1: Benchmarks and evaluations by year and type (one dot per dataset).

Figure 2 shows the cumulative growth of benchmarks. Most early benchmarks were synthetic or text-only, while multimodal datasets—those combining notes, imaging, and structured data—remain rare (under 10% of all healthcare benchmarks as of 2025). That share is increasing but still limited.

A handful of international datasets exist, yet many reuse overlapping internet data or translated variants (e.g., seven language versions of MedQA-USMLE [12]). This overlap means the test sets may not be fully independent."

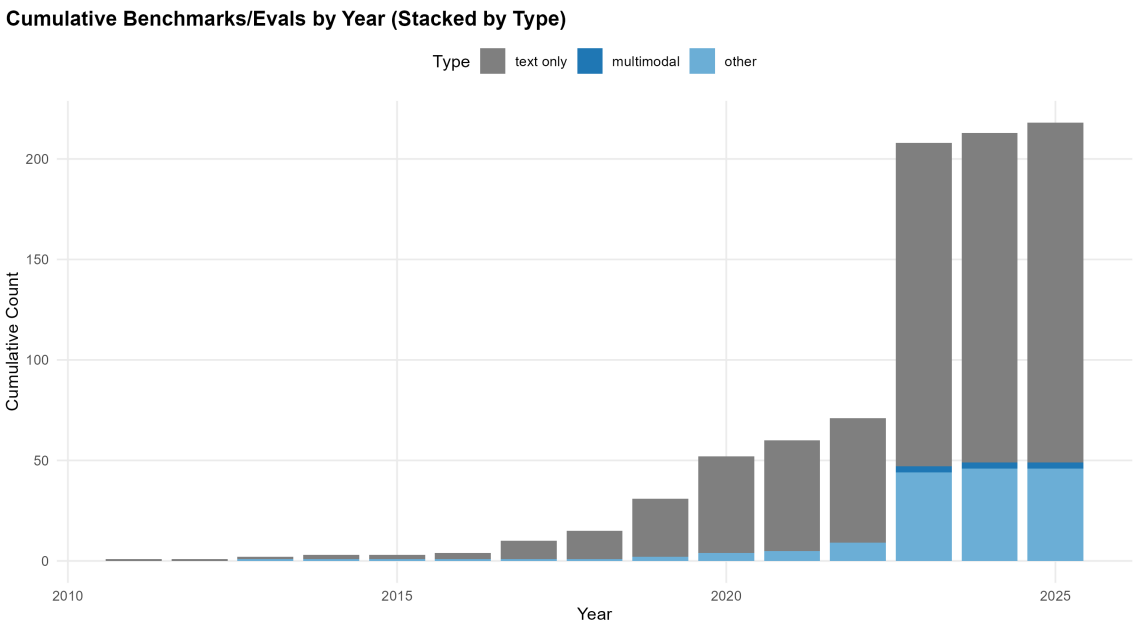

Figure 2: Cumulative benchmarks and evaluations by year (stacked by type).

Figure 3 shows annual FDA approvals of AI-enabled devices based on 510(k) public summaries. As of October 14, 2025, 956 out of 1,167 approved AI devices are in radiology, followed by 116 in cardiology. The 2025 data is incomplete due to lag in publication by the FDA. The pattern in Figure 3 underscores how evaluation strategies have favored controlled, domain-specific use cases and especially radiology. Many other clinical domains remain without FDA approved AI tools. Even among approved technologies the task domain is very narrow. A recent review [29] found that out of nearly 900 FDA-approved AI devices, many lacked broad clinical generalization—most were tested in narrow or highly controlled environments.

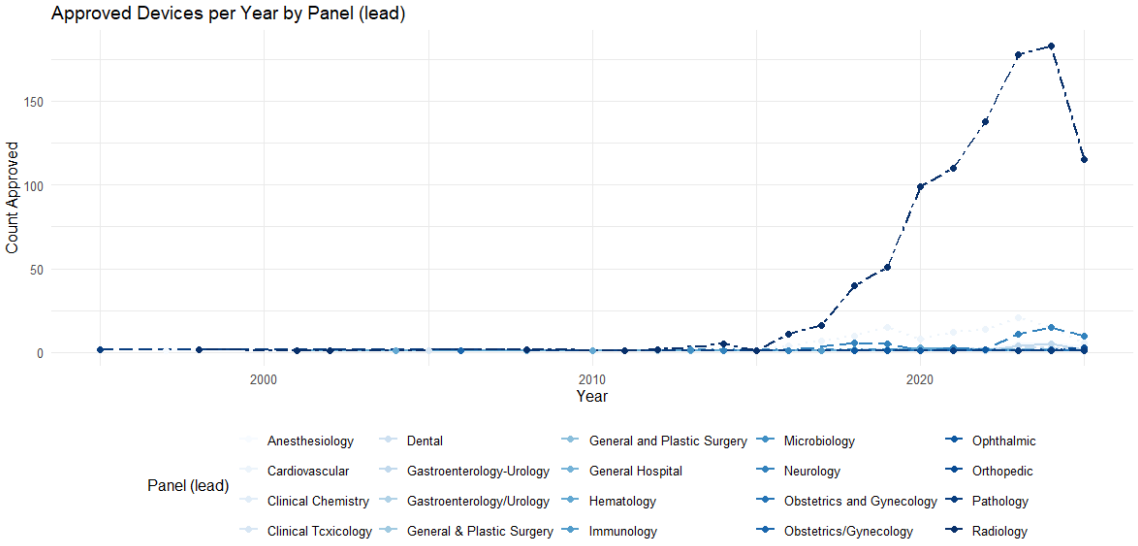

Figure 3: FDA-approved AI medical devices by year and clinical panel (510(k) summaries, 2000–2024). Radiology dominates approvals, followed by cardiology and orthopedics.

A particular failure of vignette-style evaluations is also the lack of patient physician interactions in the real world. For example, healthcare provision that is neither driven solely by provider knowledge nor just by demand (patient trust in the provider or financial means), but

is bilateral and mediated by the patient's belief in the quality of the provider and the physicians belief in what is holistically best for a patient. Recent work building tumor-board style benchmarks has leaned on synthetic vignettes [3], and large-scale Q&A sets have even transformed detailed *NEJM* cases into automatic evaluation items [9]. Unfortunately, patients in the real world do not present as neatly curated vignettes or fully elaborated *NEJM* cases. In a randomized experiment on physician training in India, Banerjee et al. [4] document that vignette-style assessments failed to update knowledge: *"Among those who 'know' the correct treatment, wide variation in 'do.'"* Real-world care is not only about textbook recommendations; it is intertwined with patient preferences, incentives, and constraints. Ethnographic accounts [6, 20, 21] describe low-trust environments where clinical encounters resemble negotiation: patients may resist costly tests, and clinicians anticipate those reactions when forming recommendations.[1] Thus, beyond "what do medical textbooks recommend," credible healthcare evaluations must consider how care is practiced in the real world.

2 Benchmark Typology and Internal vs External Validity

We now summarize categories of evaluations from least advanced to most advanced. Figure 4 illustrates a typology of benchmark designs arranged by their source realism and evaluation setting. The top row (A–C) shows the main spectrum of benchmarks—from controlled offline environments to fully deployed real-world evaluations—while the second row (D–E) further breaks down the offline group into expert-constructed and user-generated variants. This is similar to the four levels of healthcare AI evaluations proposed by Singhal [22]. Much like other medical care technologies (e.g., pharmaceuticals), AI technology would benefit from randomized controlled trial (RCT)-style assessments that can capture the true effectiveness of these tools in the real world. However, even traditional RCTs face important limitations in external validity [7]. Despite interest in more randomized clinical trial–style evidence with clean treatment and control arms, testing new or risky tools on the broader population is often not feasible. In addition, trial recruitment is time-consuming and integration within health systems may suffer from *complier or sign up bias*. As such, there is clear interest in developing quasi-experiments or evaluations from existing datasets.

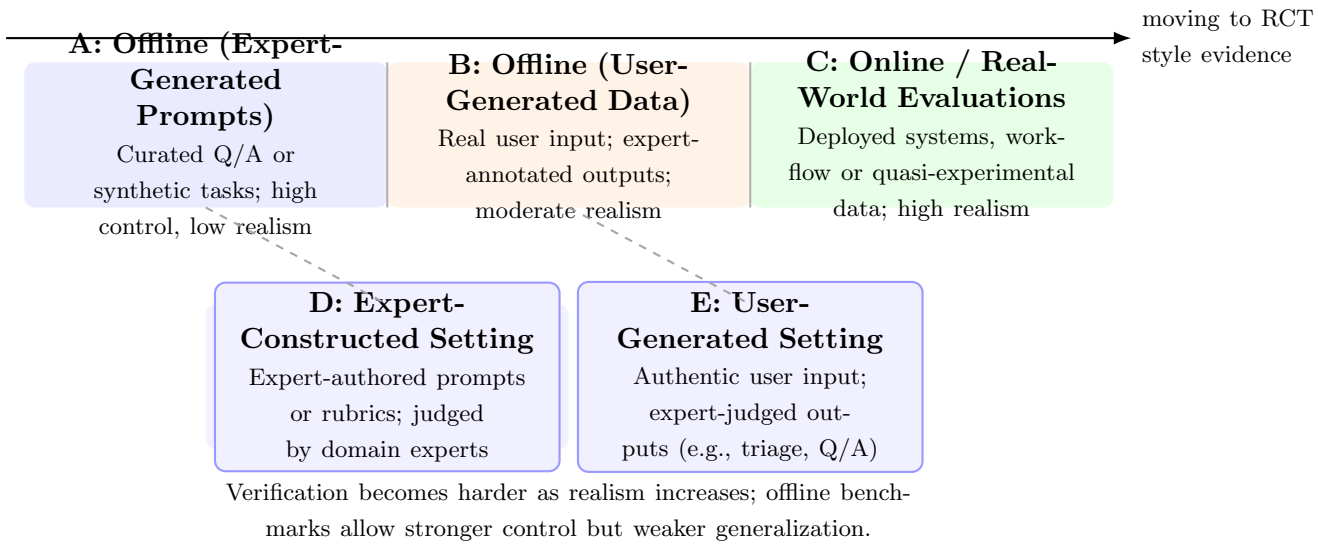

Figure 4: Benchmark typology (A–E) organized by evaluation setting and realism.

[1] Example from (*The Fallen Idol: Mistrust and Medicine in India*, 2020): "If I tell the patient straightaway that they need tests costing 10,000 rupees, they will run away and not return; so I start with medicines and only later ask for tests."

3 Axioms of Credible Benchmark Data

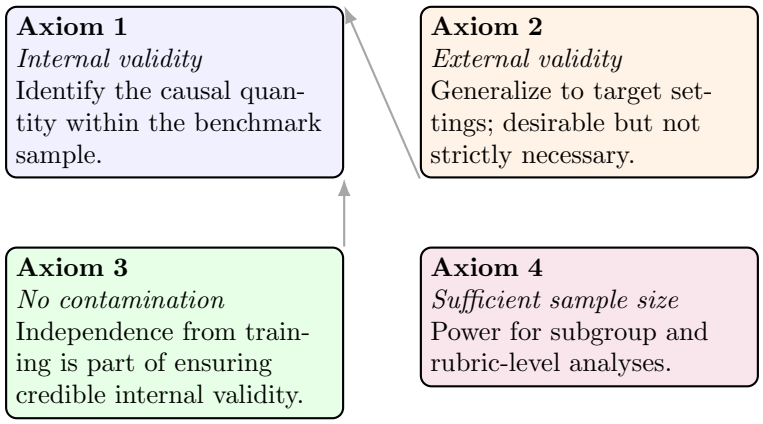

Figure 5: Axioms at a glance: Internal validity is necessary; external validity is desirable; non-contamination supports internal validity.

4 Description of Axioms of Credible Benchmark Data

We define four axioms to guide the curation of benchmark datasets for AI evaluation. We discuss each axiom and give examples of when its breached.

Axiom 1: Benchmark Population is not Selected on Outcomes. The benchmark population must not be systematically more or less likely to exhibit higher or lower outcomes on the AI evaluation than the non-benchmark population. In other words, individuals chosen for benchmarking should be randomly drawn with respect to evaluation difficulty or outcome likelihood. This Axiom may not be required if for example particularly difficult cases are desirable. An insidious form is when the benchmark population is more likely to perform better in an evaluation than the wider population. An example of this is minorities (e.g., race Black) tend to have lower-fidelity medical records, which make future predictions of their outcomes harder [27]. High-share Black is also less attainable in rare disease therapies, skewing curated benchmark datasets toward easier modal race groupings.

Formally, let $S_i \in \{0, 1\}$ indicate whether case i is included in the benchmark ($S_i = 1$) or not ($S_i = 0$). Let $Y_i^\star$ denote the *intrinsic* per-case evaluation outcome (e.g., correctness indicator or continuous score that is a weighted average of several factors) that the model would achieve on case i, regardless of whether i is selected.

The benchmark reports:

$$\mu_{\text{bench}} \equiv \mathbb{E}[Y_i^\star \mid S_i = 1],$$

while the target population performance is:

$$\mu_{\text{pop}} \equiv \mathbb{E}[Y_i^\star].$$

The benchmark is *unbiased* for population performance when selection is independent of per-case difficulty:

$$Y_i^\star \perp S_i \quad \text{(or, more generally, } Y_i^\star \perp S_i \mid X_i),$$

so that $\mu_{\text{bench}} = \mu_{\text{pop}}$

Concrete example (classification accuracy). Consider an AI system that identifies lung cancer on CT scans. Define $Y_i^\star = \{\text{model correctly classifies case } i\}$, a 0/1 per-case correctness indicator. Inclusion in the benchmark ($S_i = 1$) does not change whether the model would be correct on case i; it only determines whether $Y_i^\star$ is *observed*. If the benchmark over-represents "easy" scans (e.g., high-quality images with large, well-demarcated lesions) and under-represents "hard" scans (e.g., motion artifacts, subtle nodules, complex comorbidities), then:

$$\mathbb{E}[Y_i^\star \mid S_i = 1] \;>\; \mathbb{E}[Y_i^\star],$$

inflating reported performance. Axiom 1 requires that selected cases be no easier or harder, on average, than non-selected cases with respect to $Y_i^\star$.

Internal validity asks whether we are correctly identifying the causal impact of the model or intervention *within* the context of the benchmark sample. Factors that impact internal validity include omitted-variable bias (missing data), measurement error in labels, and reverse causality. A form of reverse causality is *look-ahead bias*, in which information from the future or from the model's target output leaks into predictors. **In general, internal validity is more important than external validity. If an evaluation is internally invalid, we have learned nothing; if it is internally valid but externally invalid, we have at least learned something about our specific context.**

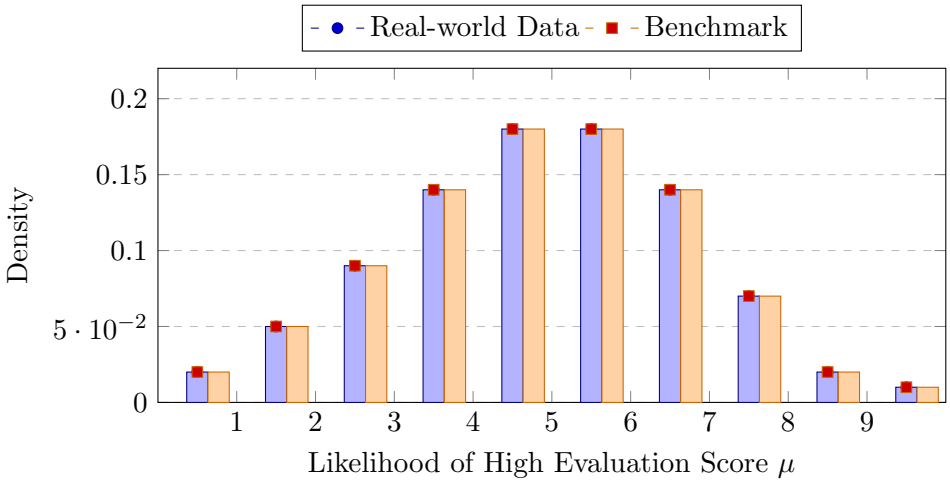

Figure 6: Distribution of cases by likelihood of high evaluation score (μ). Benchmark and real-world data are similarly distributed, satisfying Axiom 1.

Axiom 2: External Validity and Realistic Context. Benchmarks should reflect real-world clinical workflows and multi-turn events rather than scripted or artificially generated interactions. Data realism—grounded in real-world records—enhances both external validity and clinical relevance.

Assuming internal validity holds (see Axiom 1), **external validity** concerns generalizability: whether causal relationships or evaluation metrics estimated on the benchmark sample can be transported to other populations or settings. Even when internal validity holds, external validity may fail if the benchmark cohort is not representative of the broader population of intended AI tool users, or if the benchmark context is atypical (e.g., single-institution data, restricted demographic strata, or curated subsets of easy cases).

A benchmark attains external validity when, across both observable and unobservable covariates, the joint distribution of characteristics in the benchmark cohort closely matches that of the real-world population. One particular covariate is the likelihood of exposure to the AI tool

in the real world if released. A very similar context is the gap between efficacy and effectiveness that is evident in pharmaceutical testing: about 58% of Phase II trials fail in Phase IIIas populations broaden (see [25]).

Graphically, this can be shown as:

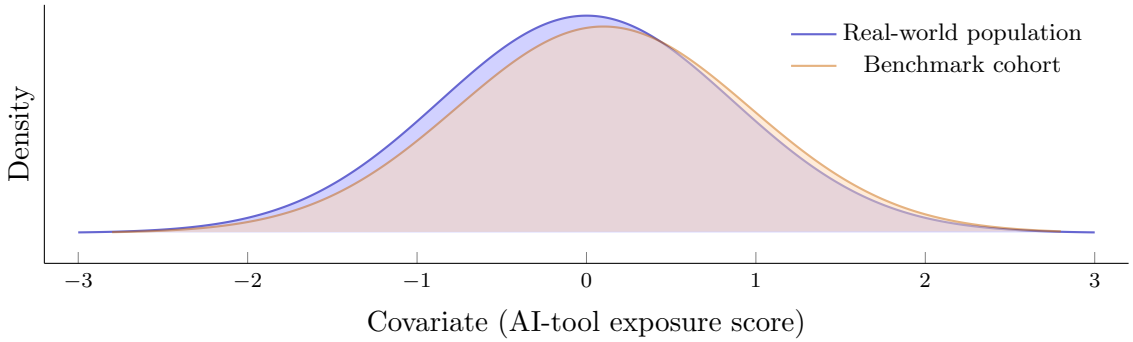

Figure 7: Benchmark representativeness: the distribution of covariates in the benchmark cohort should mirror the real-world population to ensure external validity.

While Axiom 2 establishes the representational foundation for *external validity*—ensuring the benchmark captures the same distribution of contexts as the real world—Axiom 1 complements it by ensuring *performance independence*, i.e., that benchmark inclusion is uncorrelated with evaluation difficulty or outcome likelihood.

Axiom 3: Independence of Training and Benchmark Data i.e No Contamination. Contamination between training and benchmark data parallels the problem of *serial correlation* in panel data, where repeated observations on the same entities introduce dependence.

In a basic linear model, this is seen when estimating:

$$\text{health}_{it} = \beta\, \text{doctor visits}_{it} + \varepsilon_{it},$$

with i indexing patient and t time. Increasing the number of time periods per patient does not proportionally increase the true sample size because the errors ε_{it} are correlated within i. Analogously, in medical data, repeated records for the same patient, provider, or hospital system can produce artificially high apparent performance if overlapping information leaks between training and benchmark sets.

Degrees of Contamination. Following Xu et al. [26], there are degrees of benchmark contamination that can be defined. From least to most extreme. We provide clinical examples of how this would occur:

1. **Semantic level Contamination.** Contamination due to content on the same topic or by the same upstream source. This form of contamination is common and highly likely in healthcare. Clinical practice and note-taking styles are often serially correlated across physicians and there is a small number of doctors within specialties. *Example:* with only about 10,000 nephrologists in the U.S., overlap in practice or diagnostic patterns can inflate measured performance across training and benchmark sets. This is a similar concept to *spatial correlation*, where a group of records are more correlated within geography due to the network of physicians servicing that geography.

2. **Data level Contamination.** Partial exposure of benchmark data without labels—such as unlabeled images, free-text notes, or patient sequences later used in evaluation. *Example:* Data from The Cancer Genome Atlas (TCGA) used in model pretraining without labels can later reappear in labeled form within downstream benchmarks (e.g., [28]). The

evaluation accuracy then is inflated by the co-linearity between labeled and unlabeled exact scans.

3. **Label level Contamination.** The most severe form of contamination occurs when benchmark data, including labels, are fully exposed during training. *Example:* patients and their exact medical records appear in both the training and evaluation sets, often due to the absence of a universal patient identifier across data vendors providing non-deterministic IDs. This issue has been partially mitigated through the use of hashed patient identifiers from high fidelity probabilistic attributes (e.g., last name, first name, date of birth, and gender) or deterministic attributes (eg: social security numbers). If these identifiers are tracked across datasets in training vs benchmarking patient separation can avoid this type of contamination.

   More broadly, contamination has grown increasingly complex as models acquire reasoning and retrieval capabilities. Recent research by Kapoor et al. [13] shows that world models shows that world models can recognize benchmark questions and directly retrieve answers from the web or public repositories such as Hugging Face—effectively short-circuiting genuine evaluation. The figure below provides an illustrative example of how medical records under each level of contamination would appear.

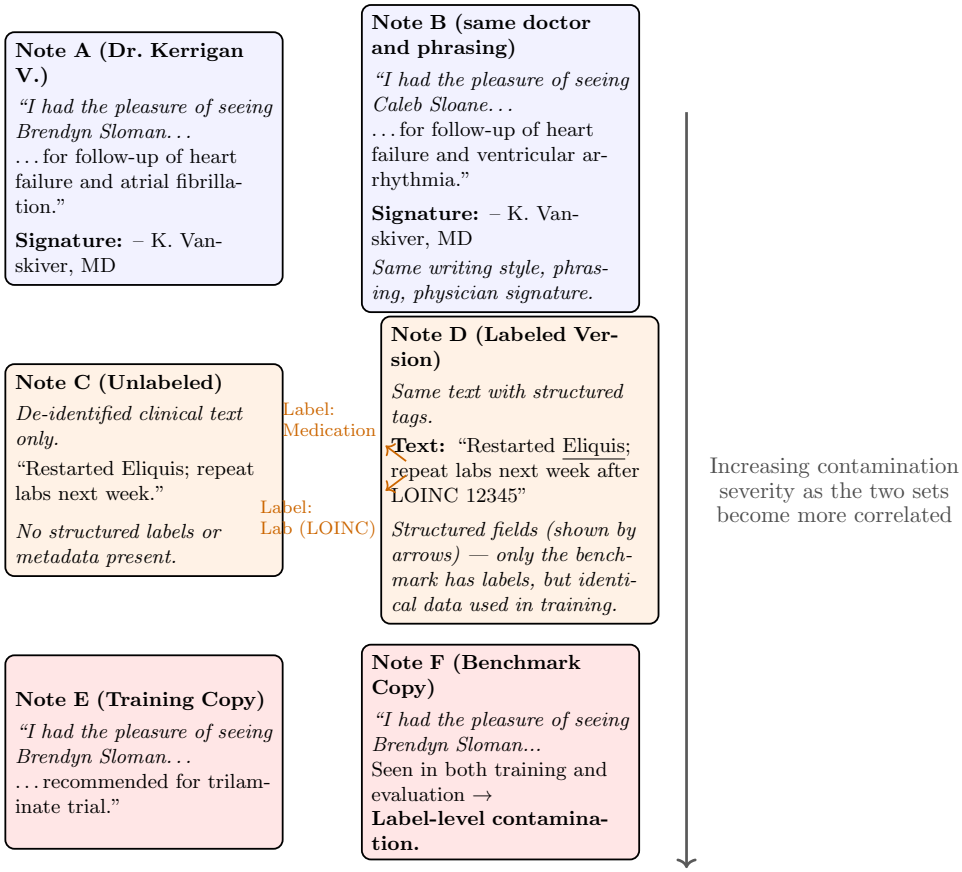

Figure 8: Illustration of contamination pathways across semantic, data, and label levels. Top row: stylistic (semantic) similarity; middle row: data-level exposure with and without structured labels; bottom row: label-level duplication of identical notes.

Axiom 4: Sufficient Sample Size for Detecting Heterogeneous Effects. Small benchmark datasets do not detect meaningful differences across groups or subpopulations. A remedy often proposed is to treat each graded criterion within a case as a data point. Because these crite-

ria are correlated, they should be treated as *multiple outcomes*. Testing many outcomes inflates the familywise error rate (FWER). A standard step-down correction is the *Holm–Bonferroni* procedure [11]:

Often described as if m is the number of hypothesis tests with ordered p-values $p_{(1)} \leq \cdots \leq p_{(m)}$. we test sequentially for $i = 1, \ldots, m$ using thresholds

$$p_{(i)} \leq \alpha_{(i)} \quad \text{where} \quad \alpha_{(i)} = \frac{\alpha}{m - i + 1}.$$

and we reject at step i only if all earlier hypotheses have been rejected. As m grows (the number of criteria tested in the benchmark), each effective significance level $\alpha_{(i)}$ becomes smaller, implying that reliable detection of heterogeneous effects requires *larger sample sizes* when many rubric criteria (or subgroups) are tested.

This in particular is a silent mode of failure in most benchmarks and despite the well rooted ideas of statistics in machine learning, adding error bars to AI evaluations—something that should have been standard practice—had to be explicitly emphasized in 2024 [17] after many evaluations were published without any form of statistical significance metrics.

A secondary component of axiom 4 is standard error clustering, when benchmarks contain non independent data for the same setting or patient, the standard errors are then correlated. eg: MedQA in english and MedQA in french are the exact questions in different languages. Similarly, often from the same EMR corpora researchers will generate tasks that assess factual recall from structured text (e.g., "What drug was prescribed?") and tasks that assess reasoning across values (e.g., "why was this drug prescribed?") . The correlation across data points if not accounted for systematically shrinks the standard errors and leads to statistically significant but erroneous results [1].

The discussion above sheds light on some of the common limitations of benchmark datasets today. We now describe a suite of new benchmark datasets offered by Protege in collaboration with health systems and domain experts.

5 Curated Datasets for Benchmarks

We curated **real-world, multimodal, diverse, and longitudinal cohorts** that are fully held out from any current or future training datasets. Each benchmark dataset is separated using a deterministic, patient-level hash derived from last name, first name, date of birth, and gender—to minimize contamination across training and evaluation sets. All benchmark datasets include a sufficient number of samples to enable subgroup-level statistical power. Except for the committee-style oncology diagnostic benchmark, each cohort includes no fewer than 2,000 patients and can scale to tens of thousands. This emphasis on sufficient sample sizes is to support a. multiple hypothesis testing across various rubric attributes (since rejection P value thresholds must be lower with multiple hypothesis testing) and b. analysis of heterogeneous effects within smaller patient subgroups.

4.1. Multimodal Cancer Journey- Varied

Description: Dataset contains oncology and non-oncology notes, radiology reports, pathology reports, NGS test results (PDF) when available, whole-slide pathology images, and DICOM imaging pre-, during-, and post-treatment. Mortality outcomes and registry tables (approximately 18 variables) are included.
Sample size: 2,500 patients across lung and breast cancer. **Held-out identifier:** deterministic hash of (last name, first name, DOB, gender).
Purpose: Oncology patients generate over 250,000 text tokens (approximately 1M characters) per care journey. High-accuracy AI tools are needed to summarize longitudinal records, extract

structured staging data, and integrate with EMR text information from PDF reports, radiology imaging and pathology results. Cancer diagnosis is also a task that is slowed down by information inertia (the volume of background on the patient).

Problem AI would Solve:

- Diagnostic intervals from initial suspicion to definitive staging can extend up to 180 days in real-world settings due to workflow and scheduling delays (Patel et al., 2023).
- Staging quality varies across community sites, and clinicians face competing workload pressures that affect accuracy and timeliness.

Input: Pathology (both reports and images) and radiology (both reports and images), genetic testing report, biomarker status, oncology and non oncology notes.
Output: Structured registry fields (primary site, histology, stage, grade), diagnosis reached, treatments offered and executed, long term outcomes (hospice, mortality, remission, metastasis etc

Illustrative Example of the Multimodal Cancer Journey Benchmark Dataset:

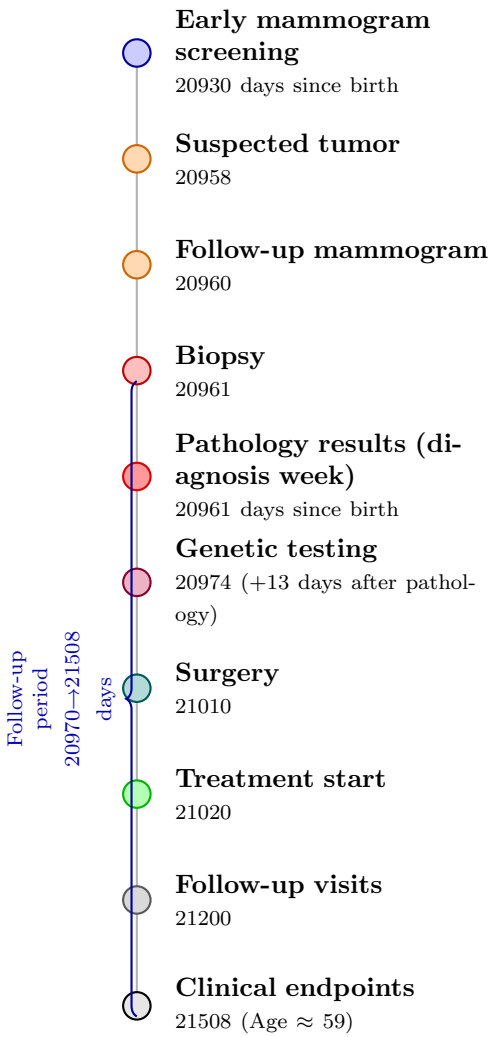

Figure 9: Timeline for a female breast cancer case showing screening, suspicion, imaging, biopsy, pathology (20961), genetics (+13 days), surgery, treatment, follow-up, and clinical endpoints. The patient is followed from 20970 (age ≈ 57) to 21508 (age ≈ 59) days since birth.

4.2 Multimodal Cancer Journey plus Physician Conference - Complex Cases

Description: Case summaries integrating notes, pathology, genetics and imaging to evaluate complex oncology cases. These patients are typically metastatic liver, lung and brain patients eligible for clinical trials.
Sample size: 300 cases. **Held-out identifier:** deterministic hash of (last name, first name, DOB, gender).

Problem AI would Solve:

- oncologists tumor case reviews see 15–50 cases per week, often allocating only 2–8 minutes per case (Lai et al., 2023).
- Digital workflows can reduce preparation burden (Dong et al., 2024). For example in the above dataset we found—up to 6,800 PDFs for 168 patients. The volume of PDFs and documents that have to be reviewed slows patient discussion in the conference and treatment progression.

Input: Longitudinal EMR notes, pathology notes, radiology notes, radiology and pathology images, genetic testing, case summary ahead of meeting, doctor level recommendation (>2 physicians in discussion and up to 6).
Output: Suggested management plan; final summary, clinical trials suggested.

Illustrative Example of the Multimodal Cancer Journey plus Physician Conference-Complex Cases Benchmark Dataset:

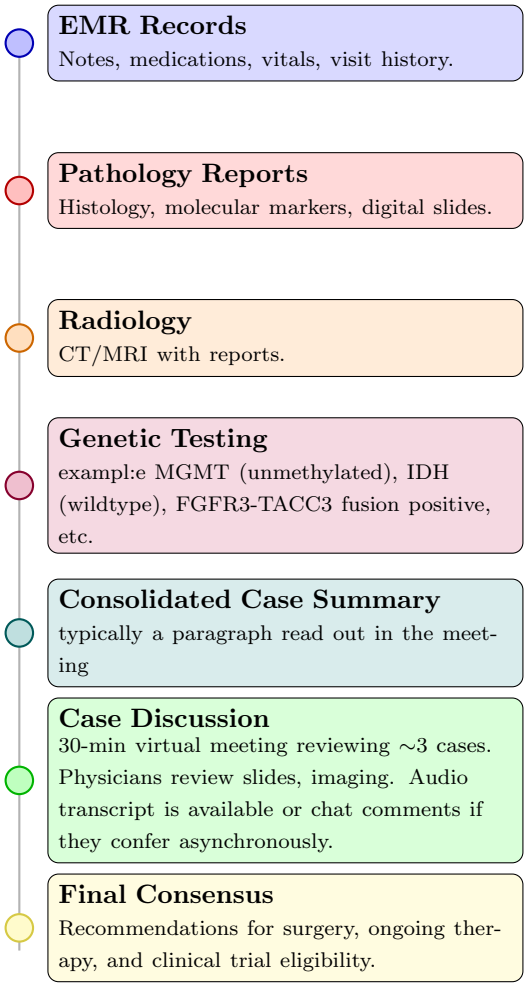

Figure 10: Dataset flow: multimodal inputs (EMR, pathology, radiology, genetics) are inputs, discussions by physicians and decision are outputs

4.3 Dermatology Portal Q&A Benchmark Dataset

Description: Photo + text triage from dermatology patient portal messages paired with EMR. Evaluates AI triage accuracy in dermatology settings.
Sample size: 10,000 bundles of Question, Answer and EMR. **Held-out identifier:** deterministic patient hash.

Problem:

- Dermatology portal messages are among the most frequent across all specialties (AMA, 2024).
- Median dermatology appointment wait times range from 13 to 30 days depending on insurance status, increasing patient reliance on asynchronous portals (Ramaswamy et al., 2024).
- Providers often depend on EHR context beyond images alone to respond effectively.

Input: Patient portal messages with photos and structured metadata.
Output: Triage with rationale. Physician responds with an answer or an invite to visit the clinic. When a visit occurs the medical record of the visit (EMR notes) are captured.

Illustrative Example of the Dermatology Portal Q&A Benchmark Dataset:

EMR Context (available from prior visit notes)
Prior clinic summary, meds/allergies, problem list, derm history, recent labs.

Patient Message with Image Attachment
Sent: 1/28/2025 7:47:40 PM MST *Subject:* RE: Skin update
Body: "rash is spreading to more of my body (arms and legs). I wasn't able to get Benadryl until tonight but just took it and also bought Benadryl cream that I put on my chest. We'll see if that helps."

Physician Triage
Paths:

- Ask clarifying questions (onset, itch/pain, fever, new meds).
- **Direct to ER** if red flags.
- **Schedule follow-up visit** (in-person or telederm).

Follow-up Visit (if booked)
Prior notes auto-retrieved; new image captured if available; brief ROS and skin exam documented.

Figure 11: Dermatology QA flow: EMR context available → portal message → physician triage → follow-up visit if booked.

4.4 Nurse Narratives and Triage Benchmark

Description: Telephone narratives between nurse and patient (recorded by the nurse in the EMR), with relay to MD and decision from MD. EMR notes pre and post patient questions are bundled in.
Sample size: $>$ 1 million conversation narratives between nurse and patient, with physician responding in the EMR and followup recorded. **Held-out identifier:** deterministic patient hash.

Problem:

- Major healthcare systems report millions of portal messages annually, rising sharply after the pandemic (Crotty et al., 2020); (AMA, 2024).
- Response quality varies, and AI-based triage can reduce burden.

Input: Patient messages and relevant pre message EHR data.
Output: response to patient, next steps taken and EMR notes if visit occurs as a result of the question.

Illustration of Nurse Narrative (from call with patient) and MD Triage Benchmark

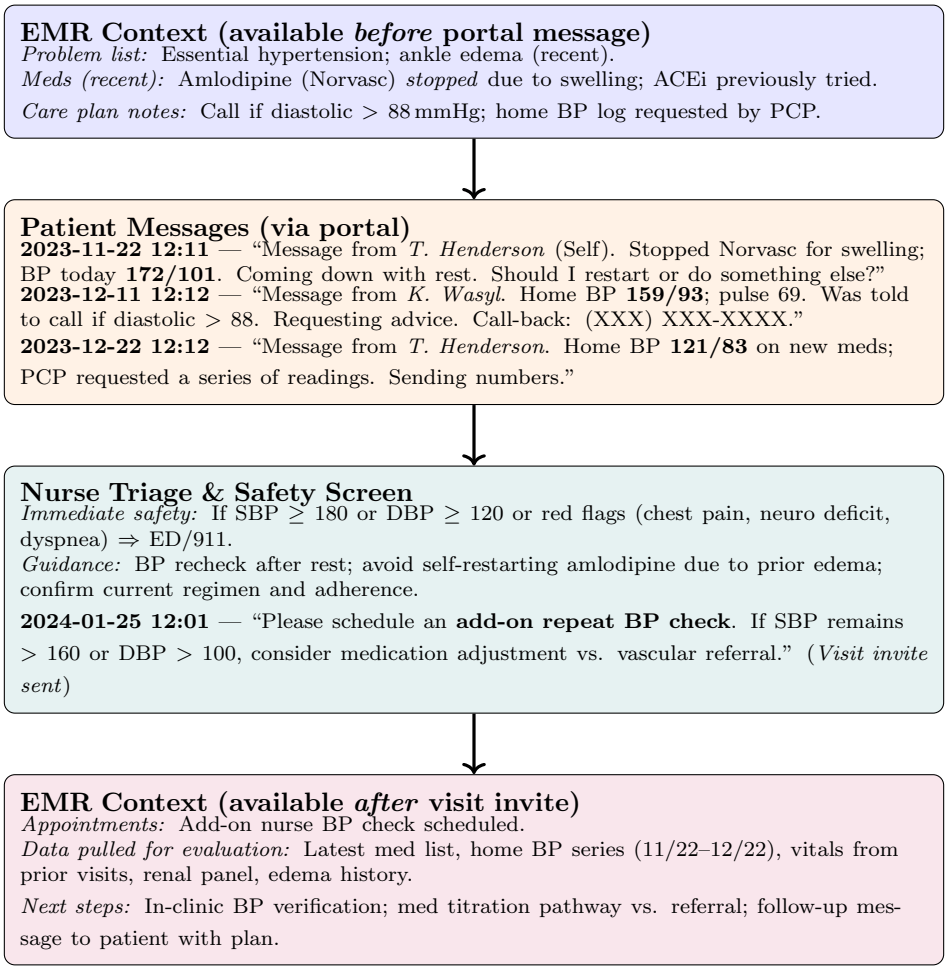

Figure 12: Nurse Narratives benchmark flow: EMR context is available before the message, a series of patient portal messages arrive about home blood pressure, nurse triage to MD who invites patient for a visit, and updated EMR context becomes available after the visit with the new medication decision reached. These are narratives because some of the communication occurs on the phone and is described by the nurse to an MD in a note. The first person communication is occurring between the nurse and the MD about the patient.

Additional Early Diagnostic and Progression Benchmarks

In addition to the oncology and triage benchmarks described above, we have curated three multimodal benchmark datasets focused on early diagnosis and disease progression. Each dataset

is held out using deterministic patient-level hashes and supports subgroup-level evaluation for robustness and generalization.

- **Acute Myocardial Infarction (AMI) Prediction Benchmark (Cardiology)** — *2500 patients.* Predicts the likelihood of major adverse cardiac events such as acute myocardial infarction using multimodal cardiology inputs including EKG PDFs or CTA scans, and longitudinal EMR and lab results.
- **Breast Cancer Progression Prediction Benchmark (Oncology)** — *2000 patients.* Uses pathology slides, genetic biomarkers, and EMR data to predict disease progression and therapy response.
- **Early Lung Cancer Detection Benchmark (Radiology)** — *10,000 longitudinal imaging series.* Links multiple low-dose chest CT scans per patient across years to predict lung cancer. Multiple Lung Rads paired with EMR over the years and final ICD C34 status.

Conclusion

In this overview, we documented four main axioms for benchmarking datasets that we believe accelerate AI evaluations in healthcare. We caution that not all four axioms must always hold, and emphasize that internal validity is far more important than external validity. In other words, uncontaminated benchmarks that are not selected to skew evaluation outcomes take much higher priority than benchmarks designed to capture the global patterns of all patients receiving care.

We also note that the very nature of accessible data—whether used for training or benchmarking—heightens the risk of contamination. As researchers, we rely on the data available to us, which often results in limited datasets or a niche cohort of sites across available data. This *selection bias* in data participation threatens internal validity and increases the risks of contamination. Addressing these limitations should be a key focus for data entities aiming to advance the AI frontier in benchmarks and evaluations.

Finally, we provided a list of several new datasets available for evaluation that are held out from training data. These datasets span oncology, cardiology, dermatology, and general patient triage. All curated datasets are multimodal and multi-turn, featuring realistic, non-synthetic clinical encounters. Our future focus is to expand evaluations beyond traditional question–answer formats, applying the four axioms discussed to specialized clinical domains.

References

[1] Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. When should you adjust standard errors for clustering? Technical Report Working Paper No. 24003, National Bureau of Economic Research, 2017. URL https://www.nber.org/papers/w24003.

[2] American Medical Association. Phone calls stable, patient portal messages keep piling up. https://www.ama-assn.org/practice-management/digital-health/phone-calls-stable-patient-portal-messages-keep-piling, 2024.

[3] A. Author and B. Author. Building tumor board benchmarks with synthetic vignettes, 2024. URL https://arxiv.org/abs/2411.03395. arXiv:2411.03395.

[4] Abhijit Banerjee et al. Training doctors, improving practice? evidence from india. https://www.hbs.edu/ris/Publication%20Files/RHCP_Paper_July2023_e15e8fe9-dd5e-4188-9b04-eb4e04e7f010.pdf, 2023.

[5] B. H. Crotty et al. Patient portal messaging trends before and after the pandemic. *JMIR Medical Informatics*, 8(7):e16521, 2020. URL https://medinform.jmir.org/2020/7/e16521/.

[6] Jishnu Das, Abhijit Chowdhury, Reshmaan Hussam, and Abhijit V. Banerjee. The impact of training informal health care providers in india: A randomized controlled trial. *Science*, 354(6308), 2016. URL https://science.sciencemag.org/content/354/6308/aaf7384.

[7] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, 2018. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10565197/.

[8] S. Dong et al. Digital preparation workflows for oncology case review. *JCO Clinical Cancer Informatics*, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11815928/.

[9] NEJM AI Editors. Nejm medical cases as q&a benchmarks for ai evaluation. *NEJM AI*, 2025. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500120.

[10] HealthBench-2025 Dataset. Healthbench. https://huggingface.co/datasets/openai/healthbench, 2025.

[11] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[12] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020. URL https://arxiv.org/abs/2009.13081.

[13] S. Kapoor et al. Agents that game benchmarks via web retrieval, 2025. URL https://arxiv.org/abs/2510.11977.

[14] A. G. Lai et al. Workflow analysis of tumor board meetings. *JCO Oncology Practice*, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10336866/.

[15] MedQA-USMLE Dataset. Medqa-usmle benchmark. https://huggingface.co/datasets/shuyuej/MedQA-USMLE-Benchmark, 2020.

[16] MEDS-BENCH Dataset. Meds-bench. https://huggingface.co/datasets/Henrychur/MedS-Bench, 2024.

[17] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024. URL https://arxiv.org/abs/2411.00640.

[18] M. Patel et al. Timeliness of cancer staging in real-world oncology practice. *JCO Oncology Practice*, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10388018/. PMCID: PMC10388018.

[19] R. Ramaswamy et al. Wait times for dermatology care by insurance type in new york. *Archives of Dermatological Research*, 2024. URL https://link.springer.com/article/10.1007/s00403-024-03249-w.

[20] Vaibhav Saria. *The Fallen Idol: Mistrust and Medicine in India*. Stanford University Press, 2020.

[21] Vaibhav Saria. New machine, old cough: Technology and tuberculosis in patna. *Frontiers in Sociology*, 5, 2020. URL https://www.frontiersin.org/articles/10.3389/fsoc.2020.00018/full.

[22] Karan Singhal. Levels of clinical evaluation for llms. https://www.karansinghal.com/notes/levels-of-clinical-evaluation/, 2025.

[23] SlideChat-2024 Dataset. Slidechat-2024. https://huggingface.co/datasets/slidechat/slidechat-2024, 2024.

[24] The Cancer Genome Atlas (TCGA) PRAD. Tcga-prad project. https://portal.gdc.cancer.gov/projects/TCGA-PRAD.

[25] Gail Van Norman. Limitations of clinical trials: Is efficacy really different from effectiveness? *JACC: Basic to Translational Science*, 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6609997/.

[26] H. Xu et al. A taxonomy of benchmark contamination for llms, 2024. URL https://arxiv.org/abs/2406.04244.

[27] D. Yan et al. Association of race and ethnicity with data quality in ehrs. *JAMA Network Open*, 2023. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2810366.

[28] J. Yang et al. Cmob: A comprehensive benchmark for multimodal oncology, 2024. URL https://arxiv.org/abs/2409.02143.

[29] Y. Zhang et al. Landscape of fda-approved ai/ml-enabled medical devices: Generalizability and validation. *npj Digital Medicine*, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12044510/.

Appendix: Filtered Benchmarks Table

Table 1: Hugging Face–filtered benchmarks with samples and papers by modality and year.

Name	Year	Modality Label
MedQA USMLE	2020	text
MedMCQA	2022	text
Asclepius Synthetic Clinical Notes	2023	text
Augmented Clinical Notes	2022	text
Clinical Synthetic Text LLM	2023	text
Multilingual Medical Corpus	2024	text
Medical Multimodal Evaluation Data	2024	text plus images rad
ClusTREC-Covid	2024	text
SlideChat	2024	text plus images rad
Medical O1 Reasoning SFT	2024	text
Medical QA (MTEB)	2022	text
MediQ AskDocs Preference	2025	text
II Medical Reasoning SFT	2025	text
DiagnosisArena	2025	text
ReasonMed	2025	text
Drug Approval Prediction	2025	text
n2c2 2014 De-identification	2014	text
Hallmarks of Cancer	2016	text
Long-COVID Classification Data	2022	text
i2b2 2010 Relation Extraction	2011	text
BIOSSES	2017	text
BioASQ 7b – Factoid	2019	text
BioASQ 7b – Yes/No	2019	text
BioASQ 7b – List	2019	text
CORD-19 (Kaggle)	2020	text
CovidQA-MediSYS	2020	text
CovidQA-FAKTA	2020	text
Covid19-QA	2020	text
EmoryNLP Covid-19 Discourse	2020	text
MedDialog (Chinese)	2020	text
CovidDialog (Chinese)	2020	text
CovidDialog (English)	2020	text
COVID-19 Twitter (MediRet)	2020	text
CARES (Catalysis AI COVID-19)	2020	text
COMETA	2020	text
PACT-1	2019	text
PACT-2	2020	text
PACT-3	2021	text
PACT-4	2022	text
PACT-5	2023	text
PACT-MS	2023	text
PACT-XS	2023	text
PACT-XSD	2023	text
PACT-D	2023	text
PACT-Misc	2023	text
PACT-Domain	2023	text
PACT-Language	2023	text
PACT-Clinical	2023	text
PACT-Style	2023	text
PACT-Task	2023	text
MedDialog (English)	2020	text
MedNLI	2018	text
MedicationQA	2021	text
LiveQA-Med	2017	text
MEDIQA-AnS	2019	text
MEDIQA-RQE	2019	text
MEDIQA-QA	2019	text
PubMedQA	2019	text
BioClinicalQA	2021	text
HEAD-QA (En)	2019	text
HEAD-QA (Es)	2019	text
MEDQA	2018	text
MedMCQAI	2023	text
MedEval	2023	text

Continued on next page

Name	Year	Modality Label
MedQA (Chinese)	2017	text
WebMedQA	2018	text
MedQA (Simplified Chinese)	2017	text
MedQA (Traditional Chinese)	2017	text
CMB	2020	text
MedQA (Vietnamese)	2023	text
MedQA (French)	2023	text
MedQA (Spanish)	2023	text
Fictive Medical Reports & Summaries (French)	2023	text
Chinese Biomedical NER	2023	text
MRI-sym2	2023	text plus images rad
MedSynth	2023	text
Asclepius Synthetic Clinical Notes (v2)	2023	text
X-Ray Chest Images	2023	text plus images rad
TREC Clinical-Trials (TREC-PM 2019)	2019	text
Breast Cancer Ultrasound Classification	2023	text plus images rad
MedSum	2023	text
MEDIQA-QA (BigBio)	2019	text
CodiEsp Corpus	2020	text
SPACCC Sentence Splitter	2018	text
Italian Parkinson's Voice & Speech	2019	other
MedExQA	2023	text
Biomedical CPG-QA	2023	text
PubMed QA (chungimungi)	2023	text
XLingHealth	2023	text
Refined TCGA PRAD Pathology Dataset	2023	text plus pathology
Enhanced MedMNIST	2023	text plus images rad
PubMed (cyrilzakka)	2023	text
VERI-Emergency	2023	text
Region Hovedstaden Clinical Text	2023	text
CleanPatrick Dermatology	2023	text plus images rad
MMLU-Medical (MedGENIE)	2023	text
Clinical Trial Texts (Rosati)	2023	text
HealthVer Entailment	2023	text
Eka Medical ASR Evaluation	2023	other
MedRescue	2023	text
Pulmonary Disease Airway & Lung Function	2023	other
PharmaER.IT	2023	text
Thyroid Ultrasound Images	2023	text plus images rad
MedicalQuestions (fhirfly)	2023	text
ICD10GM-Alpha	2023	text
FOMO MRI	2023	text plus images rad
Abdomen MRI	2023	text plus images rad
ACL X-ray	2023	text plus images rad
Axial MRI	2023	text plus images rad
Gynecology MRI	2023	text plus images rad
X-ray Rheumatology	2023	text plus images rad
COVID-19 HEALTH Wikipedia (FrancophonIA)	2021	text
PubMedVision	2023	text plus images rad
IMed 361M	2023	text
MedMentions NER	2019	text
Psychology-Therapy	2023	text
Healthcare Disease Knowledge	2023	text
BASED-FDA	2023	text
Mental Health Chatbot	2023	text
MedS-Bench	2023	text
Medicare COVID-19 Hospitalization Trends	2020	other
Medicare Inpatient Hospitals (Provider/Service)	2022	other
Medicare Outpatient Hospitals (Geography/Service)	2022	other
Medicare Physician & Practitioner (Provider)	2022	other
Weekly Lab-Confirmed RSV Hospitalization	2022	other
SNOMED CT Hierarchy-Transformers	2023	other
Vietnamese Medical QA	2023	text
TREC-COVID Top-20 Gen Queries	2020	text
Genomics Long-Range Benchmark	2023	text
IRDS ClinicalTrials (TREC-PM 2017)	2017	text
IRDS ClinicalTrials (TREC-PM 2018)	2018	text
IRDS ClinicalTrials (TREC-PM 2019)	2019	text
Evidence Inference – Treatment	2020	text

Continued on next page

Name	Year	Modality Label
Belgian Entrance Exam (Physician)	2023	text
Race-Based Medicine Questions	2023	text
Chest Xray Classification	2023	text plus images rad
Dermatology-QA	2023	text
ICD Dx Description Map	2023	text
ICD-11 QA	2023	text
Taiwan Epilepsy Guidelines QA	2023	text
Clinical Trials (louisbrulenaudet)	2023	text
RCL Breast Cancer Biopsy 7500	2023	text plus pathology
RCL Lymph Node Biopsy 100K	2023	text plus pathology
MEDAL (McGill-NLP)	2023	text
CORD-19	2020	text
Health Advice	2023	text
Medical Meadow Flashcards	2023	text
MEDIQA (2019 Challenge)	2019	text
WikiDoc (Medical Meadow)	2023	text
MedGUIDE-MCQA-8K	2023	text
MedRAG PubMed	2023	text
MedRAG Textbooks	2023	text
TherapyTalk (MentalAgora)	2023	text
Pneumonia X-ray	2023	text plus images rad
CDS-BART Vaccine Degradation	2021	text
CoT Reasoning – Clinical Diagnosis	2023	text
MedWiki (mvarma)	2023	text
COVID Fake News (nanyy1025)	2021	text
Open Patients	2023	text
MedNurse-QA	2023	text
Hospital Financial Reasoning	2023	text
MedBench Resident	2023	text
Medical-Gen-VQA	2023	text plus images rad
ICD10 e5 Small v2 Embeddings	2023	other
SA-Med2D-20M (OpenGVLab)	2023	text plus images rad
m1-MedBench (OpenMedical)	2023	text
Clinical Persian QA I	2023	text
Clinical Persian QA II	2023	text
CT-ScanGaze	2023	text plus images rad
MedSSS-data	2023	text
PMC-Treatment	2023	text
MedMNIST (mirror)	2021	text plus images rad
Recurv Clinical Dataset	2023	text
Recurv Medical Dataset	2023	text
BioLeaflets Biomedical NER	2023	text
SFT Dataset (medical)	2023	text
DisEmbed Symptom–Disease v1	2023	text
Thai Depression (SEACrowd)	2023	text
Medical Speech Transcription & Intent	2023	other
MedKGent-KG	2023	other
HealthBench	2025	text
MedKGent-KG	2023	other
Brain Tumour MRI Scan	2023	text plus images rad
MACCROBAT Biomedical NER	2023	text
Synthetic Chest X-ray	2023	text plus images rad
Synthetic Mammography	2023	text plus images rad
Biomedical EN–FR Corpus	2022	text
ClinicalQA (SNUH)	2023	text
CMS Federal Medicare Data (stigsfoot)	2023	other
Greengenes	2013	other
Real Clinical Cases (TCM Doctors)	2023	text
Synthetic Clinical Notes (Embedded)	2023	text
PMC-Patients-ReCDS	2023	text
MedCT Clinical Notes	2023	text
Mental Health FAQ	2023	text
Internal Medicine Binary Questions	2023	text
Pediatrics Questions	2023	text
MedXpertQA	2023	text
MedReason (UCSC-VLAA)	2023	text
MedTrinity-25M	2023	text
Medical Dialog (UCSD)	2021	text
Chest CT Images	2023	text plus images rad

Continued on next page

Name	Year	Modality Label
Turkish Hospital Medical Articles	2023	text
Turkish Medical Articles	2023	text
Medical Masks Image Dataset	2020	other
Vietnamese Healthcare	2023	text
HNSCC Multi-Omics (Gene Networks)	2023	other
Biomedical Lectures (ENG) v2	2023	text
BTCV-CT-as-video (MedSAM2)	2023	text plus images rad
LLD-MMRI (MedSAM2)	2023	text plus images rad
CT_DeepLesion (MedSAM2)	2023	text plus images rad
RVENet (MedSAM2)	2023	text plus images rad
Dental-2.5k-Instruct	2023	text plus images rad
DrugMap Ligandability	2023	text
Pharmacology LLM Test Set	2023	text
PMC Patients	2023	text
Alpaca PubMed Summarization	2023	text
Medical-O1-Reasoning SFT (Thai)	2023	text
Patient–Doctor QA (ZoneTwelve)	2023	text
MedMCQA	2022	text

Note. Non-exhaustive list of filtered benchmarks from Hugging Face with clear samples and papers by modality and year.