M.I.J.O.: Framework for Evaluating Applicability and Limitations of Biomedical Datasets with AI Assistants

Clarence C. Hu[1] Juan J. Cardona[2] Alexander I. Salter[3] Ali Danish[1]

[1] Hotpot.ai, Palo Alto, California, USA.
[2] Department of Neurosurgery, Stanford University School of Medicine, Stanford, California, USA.
[3] Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California, USA.

Corresponding author: clarence@hotpot.ai

Abstract

Background. Dataset limitations bound both clinical and artificial intelligence (AI) models in biomedical research, yet assessing dataset applicability can be difficult. Team expertise may cover clinical practice and machine learning, but not epidemiology or other relevant domains. Researchers may further lack resources for tracking emergent limitations. The consequences are amplified when subpopulations can alter biological understanding and treatment strategies. For instance, oncogenic drivers differ between never-smokers and ever-smokers in lung cancer, yet ever-smokers dominate influential datasets and may cause poor generalization to non-smoking Asian women and other underrepresented groups.

Methods. We propose Model Integrity Joint Observability (M.I.J.O.), a framework for joint human-AI assessments of dataset applicability to research questions. We introduce a schema for encoding applicability alerts, a workflow artifact for creating them, and a decentralized protocol for publishing and discovery.

Results. As case studies, we assessed two influential lung cancer datasets against viral association and never-smoker research questions. Sequencing libraries employed by both datasets depleted non-polyadenylated viral RNAs, reducing sensitivity to Epstein-Barr virus, adenovirus, and other pathogens. Furthermore, one dataset contains an estimated 6 of 37 never-smoker samples (16%), meaning that even a true viral subtype with 39% prevalence could remain undetected.

Conclusion. Because subtypes may need different therapeutic strategies, our results suggest the need to re-evaluate viral association in lung cancer subtypes, particularly never-smokers. In general, AI-assisted applicability alerts can help biomedical and AI researchers align datasets with research questions and flag limitations, lowering the risk of misguided AI models, biological inferences, and clinical decisions.

1 Introduction

Biomedical research is bounded by dataset limitations when developing both artificial intelligence (AI) and clinical models. Studies designed for one population, assay platform,

or clinical context may yield misleading observations when applied to another. In fields like oncology, where subpopulations may differ in subtle but critical ways, dataset constraints can undermine biological understanding and raise the risk of misguided clinical decisions and scientific conclusions.[1]

While dataset bias is an established risk for AI and clinical models, assessing gaps between datasets and research questions is rarely standardized and remains difficult for several reasons. Applicability analysis is neither static nor universal – the same dataset characteristic may preclude one endpoint yet not affect another. Key facts may evolve and require expert interpretations across specialties, including disease biology and epidemiology. Crucial data may be scattered across materials and involve a significant effort to synthesize the implications. Beyond explicit limitations, implicit constraints – such as restrictive enrollment criteria or library preparation methods – may escape scrutiny and quietly propagate bias downstream. Such constraints may be obvious to domain experts but opaque to others during dataset selection, persisting as unstructured and unsearchable knowledge. Collectively, this friction disincentivizes per-question analysis, especially when widely cited datasets are treated as general-purpose references.

To address these gaps, we propose Model Integrity Joint Observability (M.I.J.O.), a framework for humans and AI assistants to jointly assess dataset applicability to research questions, producing alerts that can be embedded in scientific artifacts or shared as standalone posts. M.I.J.O. supports three primary use cases: (i) adapting conclusions or models to question-dependent dataset constraints before publication; (ii) standardizing publication of both known and emerging constraints as compact, machine-readable alerts; and (iii) enabling rapid constraint discovery during dataset selection.

1.1 LCINS

Lung cancer is the leading cause of cancer mortality and, in the U.S., has traditionally been associated with individuals with substantial smoking exposure.[2] Accordingly, many historical cohorts were enriched for White men older than 65 years at diagnosis.[2–4] However, as smoking rates have declined, lung cancer in never-smokers (LCINS) is increasingly recognized as a distinct subgroup.[5] LCINS exhibits distinct clinical and molecular traits, occurring more frequently in women, individuals of Asian descent, and younger patients who do not smoke, and is characterized by a different mutational landscape and lower tumor mutational burden. In the U.S., more than 50% of Asian American women with lung cancer have never smoked, and never-smoking rates among Chinese and Indian American women with lung cancer approach 80%.[6]

If classified as a standalone disease, LCINS would rank among the leading causes of cancer death globally and in the U.S.[7, 8] LCINS therefore illustrates the risks of dataset-question misalignment.

Despite key differences between smoking-associated lung cancer and LCINS, influential lung cancer datasets are often dominated by ever-smokers, defined as individuals who have smoked at least 100 cigarettes over their lifetime, and by early-stage disease.[2] The resulting biological inferences – and AI models trained or validated on these datasets – may generalize poorly to never-smoking Asian women, metastatic patients, and other underrepresented groups, obscuring critical signals such as mutation frequencies and

immunological signatures.

1.2 TCGA and PCAWG Datasets

In modern oncology, The Cancer Genome Atlas (TCGA) and the Pan-Cancer Analysis of Whole Genomes (PCAWG) constitute two foundational datasets.

TCGA encompasses 33 cancer types, molecularly characterizing more than 11,000 patients and over 20,000 primary tumor and matched normal specimens. Although patient enrollment ended by 2014 with 230 subjects, the lung adenocarcinoma (LUAD) cohort grew to 585 cases via iterative releases.[9] TCGA generated four broad classes of molecular data: genomic, epigenomic, transcriptomic, and proteomic. Genomic profiling primarily leveraged whole exome sequencing (WES) to capture the 1–2% of the genome responsible for coding proteins. Transcriptomic profiling employed poly(A)-enriched RNA sequencing, a method that captures polyadenylated human mRNAs but depletes non-polyadenylated RNA species. Together, these datasets supply the backbone for thousands of cancer studies with over 38,000 publications mentioning TCGA in PubMed as of December 2025. However, limitations in the TCGA LUAD cohort (TCGA-LUAD) have been identified, which have reinforced the need for newer lung cancer datasets with broader representation of clinically important subgroups, including never-smokers.[10, 11]

PCAWG reflects a collaboration between the International Cancer Genome Consortium (ICGC) and TCGA. Leveraging data from both efforts, the group harmonized whole genome sequencing (WGS) data from 2,658 cancers across 38 tumor types.

1.3 Models and Inferences Based on TCGA and PCAWG

TCGA and PCAWG underpin influential AI models and biological inferences, making dataset applicability consequential for downstream analyses. The DeepPATH model released by Coudray et al. in 2018 used TCGA whole-slide images to classify non-small cell lung cancer histology and predict selected LUAD mutations from histopathology, including EGFR, KRAS, and TP53.[12] More recently, Wang et al. developed the Clinical Histopathology Imaging Evaluation Foundation (CHIEF) model, a general-purpose pathology foundation model trained in part on TCGA-derived cohorts and validated across 24 hospitals internationally.[13] Recent LUAD prognostic studies also continue to rely on TCGA-LUAD to derive expression-based risk signatures and to stratify the tumor microenvironment.[14, 15] Beyond training, TCGA remains a broadly used evaluation resource for pathology foundation models, including UNI, Virchow, Prov-GigaPath, and TITAN, which report performance on TCGA-derived tasks such as cancer subtyping and biomarker prediction.[16–19] At the clinical level, Campanella et al. reported deployment of a foundation model for lung cancer biomarker detection, with 519 TCGA-LUAD slides comprising the largest external validation cohort.[20]

A widely cited study on viral associations in cancer was enabled by PCAWG. Zapatka et al. used high-coverage WGS and incorporated RNA sequencing where available, surveying tumor-associated infections across the entire host genome instead of only the 1–2% accessible with WES.[21]

The study recovered known viral etiologies in classic pathogen-driven malignancies, such

as EBV-associated gastric carcinoma (EBVaGC) and lymphomas, but did not identify viral associations in the PCAWG LUAD cohort (PCAWG-LUAD).[21]

1.4 Oncogenic Viruses

The World Health Organization attributes approximately 10% of all cancers to viruses.[22] Mechanistically, viruses promote multiple stages of carcinogenesis – spanning initiation, progression, and therapeutic resistance.[23–26] They can hijack host cellular machinery and influence key proteins, pathways, and chromosomal sites implicated in tumorigenesis such as EGFR, TP53, and PD-L1.[27–29]

Examples of human oncogenic viruses include Epstein-Barr virus (EBV), human papillomavirus (HPV), hepatitis B and C viruses (HBV, HCV), and Kaposi's sarcoma-associated herpesvirus (KSHV).[22, 23] These pathogens either drive or demonstrate association across a spectrum of malignancies, including:

- Cervical cancer[30]
- Burkitt lymphoma[31]
- Nasopharyngeal carcinoma[32]
- Hodgkin lymphoma[33]
- Gastric carcinoma (GC)[34]
- Kaposi's sarcoma[35]
- NK/T-cell lymphomas[33]
- Head and neck squamous cell carcinoma[36]
- Hepatocellular carcinoma[37]

Recognized oncogenic viruses in other cancer subtypes, together with the null results in PCAWG-LUAD, motivated an applicability analysis for viral association in LCINS. Known cohort limitations of TCGA-LUAD, together with its influence on AI models, likewise motivated an applicability analysis for LCINS research questions.

2 Related Work

Reporting frameworks such as TRIPOD+AI provide checklists to promote transparency of prediction model development and validation, while PROBAST+AI provides structured criteria for assessing the quality, risk of bias, and applicability of prediction model studies.[38] Documentation paradigms including Datasheets for Datasets and Model Cards standardize the recording of dataset characteristics, intended uses, and known limitations at the point of creation.[39, 40] Post-market surveillance mechanisms also exist: FDA medical device reports – often submitted electronically via the electronic Medical Device Reporting (eMDR)

system – are made searchable through MAUDE, while the EU AI Act mandates post-market monitoring for high-risk AI systems.[41] At the policy level, the OECD has introduced a common reporting framework with 29 criteria for documenting AI incidents and hazards across jurisdictions.[42]

These instruments are valuable but leave gaps when assessing dataset fitness for specific questions. PROBAST+AI is not intended to evaluate non-predictive analyses such as genomic association studies, nor to determine whether datasets can support questions other than ones explicitly investigated. TRIPOD+AI guides how studies are reported but does not assess dataset-question fit. Datasheets and Model Cards capture static snapshots at release but do not assess how dataset limitations may vary across research questions, nor outline mechanisms to surface emergent constraints as knowledge evolves. Regulatory surveillance and incident-taxonomy frameworks primarily operationalize reporting after failures or harms, rather than encoding constraints prior to downstream use.

3 Materials and Methods

3.1 M.I.J.O. Framework

3.1.1 M.I.J.O. Overview

M.I.J.O. encodes a schema for representing applicability alerts and a workflow artifact to create alerts with AI assistants. It also defines a decentralized protocol for publishing and discovering alerts.

Several constraints guided development. First, question dependence required centering the framework around datasets and questions rather than datasets and studies. Second, because applicability assessment is seldom a primary objective and few venues accept standalone findings on dataset limitations, the framework adopts lightweight and non-technical mechanisms. Third, applicability alerts may intersect contentious topics, motivating a decentralized architecture independent of editorial venues and technology platforms. Together, these considerations shaped a protocol designed for rapid discovery, efficient indexing, and platform-agnostic dissemination as standalone posts or embedded within scientific artifacts like manuscripts and technical reports.

3.1.2 Alert Schema

M.I.J.O. represents each applicability alert as a single Markdown file, authorable and editable in common note-taking applications. The syntax is designed to be intuitive for non-technical users yet structured enough for deterministic conversion to JSON and other machine-readable formats. Table 1 presents the field definitions of a M.I.J.O. v1.0.0 applicability alert.

Section	**Field**	**Notes**
Metadata		Alert metadata.
	mijo_id	Unique M.I.J.O. alert identifier.

Section	Field	Notes
	mijo_version	M.I.J.O. schema version.
	created_at	When the alert was created.
	contact	Preferred correspondence method.
	prompts_used	Optional. List prompts used to generate this alert: full text, prompt links, or repository links.
Authors		List of authors.
	name	Author name.
	affiliation	Comma-separated list of author affiliations, with the primary affiliation listed first.
	contact	Author contact method.
Dataset		Dataset metadata.
	common_name	The dataset name commonly used in the literature.
	formal_name	Formal dataset name.
	uri	Primary dataset locator.
	version	Dataset release/snapshot identifier. Use a stable, source-native identifier whenever one exists (e.g., DOI, release tag/number, snapshot ID). Record unknown only when no stable identifier exists for the dataset state used.
Question		Optional. Research question to assess dataset applicability against. No fields, only freeform text.
Limitations		Dataset limitations that may impact the given question or affect the dataset broadly.
	name	Short limitation name.
	summary	Summary of the limitation.
	details	Detailed characterization of the limitation.
	implications	Expected implications from this limitation.
References		Optional. List of supporting references cited in the alert.
Other		Optional. Catchall for context not captured elsewhere. No fields, only freeform text.

Table 1. M.I.J.O. v1.0.0 field definitions. M.I.J.O. section-level instructions and field definitions are listed. For each section, structured fields and authoring guidance are reported.

3.1.3 Alert Structure

Alerts are organized into several sections using Markdown level-1 headers, in which each section title is preceded by a single "#" and a space. Capitalize section headers. Within each section, field keys use lowercase snake_case, and values use normal capitalization.

Unless a field is explicitly marked as optional, all fields are required. If a value is unknown or unavailable, record "unknown" rather than omitting the field. Dates should use ISO 8601 formatting (YYYY-MM-DD), with partial dates (YYYY-MM or YYYY) permitted when full precision is unavailable.

3.1.4 M.I.J.O. Identifier

Each alert includes a unique identifier ("mijo_id"), constructed as: {lead_author_last_name}_{lead_author_first_name}_{affiliation}_{dataset_stub}_{epoch_timestamp}. To maximize stability and reduce collisions, the affiliation component must be the registrable domain of the lead author's email address, recorded in lowercase. Each component is normalized with the same logic: lowercase the text, replace runs of spaces, underscores, and hyphens with a single underscore, and delete characters outside alphanumerics and underscores. Derive the dataset stub from the first five characters of the dataset common name after normalization. "Epoch_timestamp" is the value of "created_at" represented as epoch seconds.

3.1.5 Repeated Objects

The "Authors" and "Limitations" sections contain one or more repeated entries. Each entry is formatted as a contiguous block of key: value lines unless specified otherwise. Adjacent entries are separated by a single blank line. Blank lines should be avoided within an entry block to preserve reliable parsing.

3.1.6 Dataset

In the "Dataset" section, the "uri" field identifies a stable dataset landing page or primary locator. The "version" field fixes the dataset state used (e.g., DOI, release tag/number, snapshot identifier, portal data version, or commit hash). Only use "unknown" when no stable identifier exists for the dataset state analyzed.

3.1.7 Question

The "Question" section contains only freeform text and no fields. The section contains a single research question. If an alert is not tied to a specific question, record a single hyphen ("-") on its own line. This designates the alert as dataset-generic rather than a dataset-question pairing.

3.1.8 Limitations

Limitations are a sequentially numbered list. Each entry begins with a numbered name line formatted as "N. <name>". This numbered name line uses the "name" value only and omits the "name" key. The entry then contains "summary", "details", and "implications" as "key: value" lines. The "details" and "implications" fields may be a brief paragraph or

bullet list. Separate adjacent entries by a single blank line. Within a single entry, separate fields by a single blank line.

3.1.9 References

The "References" section is optional. If used, list sources as a numbered list, with one citation per line in any common scientific format. Cite sources in the alert using bracketed numbers corresponding to the list, such as [1-3].

3.1.10 Other

The "Other" section is optional and contains only freeform text, with no predefined fields. Use this section only for materially relevant information not otherwise captured in the schema, such as edge cases or additional uncertainties.

3.1.11 Self-references

To support alerts embedded in manuscripts, reports, or bundled artifacts without public URLs, M.I.J.O. reserves the token "@@" to refer to the enclosing document. Authors may use "@@<anchor>" to reference in-document content such as "@@table_1" or "@@fig_2". When metadata from the enclosing document is needed, authors may use "@@<field>" placeholders such as "@@authors". To ensure unambiguous parsing, placeholders must occupy the full field value and must not be combined with additional text.

The following placeholders are defined:

Dataset URI: "uri: @@" indicates that the enclosing document is the canonical locator for the dataset locator used in the alert.

Title: "title: @@title" indicates that the alert title should be interpreted as the enclosing document title.

Authors: "authors: @@authors" indicates that the alert shares the enclosing document's author list.

Correspondence: "correspondence: @@correspondence" indicates that the alert shares the enclosing document's correspondence contact(s).

When the enclosing artifact later receives a public URL or DOI, readers can resolve "@@" by mapping it to that canonical locator.

Normalize the <anchor> as follows: lowercase the text, replace runs of spaces, underscores, and hyphens with a single underscore, and delete characters outside alphanumerics and underscores.

3.1.12 Word Limits

Word limits enforce brevity for embedding, platform-agnostic dissemination, and rapid screening during dataset selection and literature review. The "Question" section is limited to 50 words. Within each limitation entry, the limitation name is limited to 10 words, the summary to 25 words, the details to 100 words, and the implications to 100 words. The "Other" section is limited to 100 words.

3.2 Usage Guidelines

3.2.1 Creation

Authors may manually create alerts or may draft them with AI assistants by supplying the M.I.J.O. specification, a dataset paper or dataset details, and optionally a research question. Supplementary Figure S1 provides a reference system prompt for drafting alerts via guided workflows. The prompt outputs: (i) a draft alert and (ii) a review aid for highlighting potential dataset dependencies, candidate limitations, and assistance in verifying limitations.

If the research question is omitted, the alert documents limitations for the entire dataset rather than a specific use case.

All AI-assisted outputs require human verification. If limitations originate from an AI assistant, authors must verify their accuracy and relevance against primary sources and discard unrelated, duplicative, or immaterial limitations.

3.2.2 Publishing

M.I.J.O. alerts may be disseminated as standalone posts or embedded within scientific artifacts such as manuscripts and technical reports. Alerts are authored in a platform-agnostic and self-contained format to support independent redistribution while preserving attribution.

To support indexing, authors label each alert with the term "M.I.J.O." When an alert is embedded in a manuscript, this term should appear in the abstract and in the title of the table or figure containing the alert.

3.2.3 Discovery

Discovery follows the publishing convention. Readers search for the term "M.I.J.O." in abstracts or post bodies, locate corresponding alerts, and verify completeness by checking required schema fields. This enables rapid screening of dataset limitations during dataset selection and literature review.

3.3 TCGA-LUAD and PCAWG-LUAD M.I.J.O. Alerts

We used M.I.J.O. to document applicability constraints in two influential lung cancer datasets: TCGA-LUAD for LCINS biology and PCAWG-LUAD for viral association. To identify limitations, we audited dataset protocols and cohort composition using published metadata and technical documentation, verifying candidate constraints such as estimated never-smoker sample size and library-preparation features affecting viral detection.

3.3.1 TCGA-LUAD Methodology

Our secondary analysis of TCGA-LUAD targeted cohort composition and selected assay features. Although our TCGA-LUAD applicability analysis did not assess viral detection directly, we included sequencing and library preparation features relevant to viral detection because PCAWG-LUAD drew from TCGA-LUAD.

Reported TCGA-LUAD sizes vary across releases and downstream studies, motivating the use of a single benchmark resource. Because the Genomic Data Commons (GDC) receives continuous updates and the GDC Legacy Archive was retired in 2023, reconstructing clinical snapshots corresponding to historical TCGA releases is not straightforward via the portal. We therefore prioritized peer-reviewed TCGA publications to minimize uncertainty from retrospective data queries, selecting resources published before the PCAWG preprint submission in September 2019.

To benchmark cohort characteristics, we selected the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) by Liu et al. (2018), which standardized clinical annotations across TCGA cancer types using uniform outcome definitions and quality control procedures.[43] From TCGA-CDR, we extracted cohort-level characteristics including age and race counts.

Because TCGA-CDR did not stratify by granular smoking status, we estimated smoking composition by cross-referencing primary literature. Eligible sources required: (i) analysis of the same TCGA LUAD cohort or a cohort concordant with the benchmark resource; (ii) LUAD cohort size comparable to the TCGA-CDR benchmark; (iii) TCGA-consistent smoking definitions, including never-smoking as fewer than 100 lifetime cigarettes; and (iv) publication before September 2019. Based on these criteria, we selected Zhou et al. as the benchmark source for smoking composition.[44]

From benchmark publications, we extracted cohort counts and calculated category-level percentages when not explicitly reported.

To assess assay features relevant to pathogen detection, we reviewed TCGA library preparation and sequencing documentation and TCGA publications, focusing on RNA sequencing and genomic profiling. Sources included the Nature (2014) TCGA-LUAD genomic characterization paper, ICGC guidelines, GDC and TCGA technical documentation, and related TCGA analyses.[43, 45–48] Because TCGA-CDR did not report tumor sequencing depth, we approximated TCGA WGS coverage using PCAWG quality metrics.

3.3.2 PCAWG-LUAD Methodology

We conducted a provenance and demographic audit of PCAWG-LUAD to assess its ability to exclude viral associations in never-smokers.

Clinical and demographic metadata were obtained from the PCAWG flagship publication and its supplementary materials (Campbell et al., Nature 2020). To define the LUAD cohort, we used Supplementary Table 1 and selected samples where the "histology_abbreviation" field equaled "Lung-AdenoCA".

To quantify overlap with TCGA, sample origin was determined by cross-referencing the "project_code" and "submitted_sample_id fields". Samples were classified as TCGA-derived when "project_code" was "LUAD-US" or when "submitted_sample_id" began with "TCGA".

We summarized cohort demographics using Extended Data Table 1, extracting case counts, sex distribution, and median age. Because granular smoking status was unavailable for PCAWG-LUAD samples, we estimated the expected number of never-smokers by applying the TCGA-LUAD never-smoker prevalence derived in our TCGA secondary analysis to the PCAWG-LUAD sample count.

To assess what viral prevalence levels could be excluded in the never-smoker subset under a null observation, we computed the one-sided 95% exact binomial upper bound on prevalence assuming zero detections, using the estimated number of never-smoker samples.

3.3.3 Citation Counts

Publication citation counts were retrieved from Google Scholar (`https://scholar.google.com`) on December 3, 2025.

4 Results

We report results of the TCGA-LUAD and PCAWG-LUAD audits and corresponding M.I.J.O. alerts, presented in Figures 1-2.

4.1 TCGA-LUAD Results

Reported LUAD cohort sizes varied across releases and downstream studies, reflecting differences in exclusion criteria such as annotation completeness, tumor purity, and matched-normal availability rather than additional specimen collection. For example, the TCGA Pan-Cancer Atlas reported 566 LUAD cases in 2018, while some downstream studies restricted analysis to 522 cases based on clinical data quality and survival follow-up.

Using TCGA-CDR, the LUAD cohort had a mean age of 65 years at diagnosis. The cohort included 393 White and 53 Black cases out of 522, corresponding to 75.3% and 10.2%, respectively.

Using Zhou et al. for smoking composition, 433 of 522 cases were categorized as smoking-associated, yielding a never-smoker prevalence of 17.1%.

TCGA-LUAD RNA sequencing employed Illumina TruSeq protocols with poly(A) selection. This approach depletes non-polyadenylated RNA species, including EBV transcripts such as EBER1 and EBER2.[46, 49, 50]

Genomic characterization was driven principally by WES while high-coverage WGS was applied to fewer than 10% of tumor samples.[51] WES capture baits target human exons rather than non-human DNA. WGS tumor sequencing depth was not annotated in TCGA-CDR, but PCAWG reported a mean read coverage of 39x for normal samples and a bimodal tumor coverage distribution peaking at 38x and 60x. DNA-based WGS cannot detect non-retroviral RNA viruses such as HCV.

4.2 TCGA M.I.J.O. Alert

```
# Metadata
mijo_id: hu_clarence_hotpotai_tcga_1770364800
mijo_version: 1.0.0
created_at: 2026-02-06T00:00:00-08:00
contact: @@correspondence

# Authors
authors: @@authors

# Dataset
common_name: TCGA-LUAD
formal_name: The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD)
uri: https://portal.gdc.cancer.gov/projects/TCGA-LUAD
version: TCGA-CDR (Liu et al., 2018) + Zhou et al. TCGA-derived LUAD
cohort (n=522)

# Question
Can this dataset support research questions about LCINS?

# Limitations
1. Never-smoker underrepresentation and variable subset size
summary: In a commonly used TCGA-derived LUAD cohort, ever-smokers comprise
over 80% of cases, and the never-smoker subset varies with cohort filters
and smoking-status annotation rules.

details:
- In a TCGA-derived LUAD cohort of 522 cases, 75 were classified as
non-smoking, 433 as smoking-associated, and 14 as unknown smoking
information.
- TCGA-CDR LUAD summaries report mean age 65 years at diagnosis and race
distribution 75.3% White and 10.2% Black.
- TCGA-LUAD denominators vary across releases and downstream studies because
groups apply different exclusion criteria (e.g., clinical completeness,
tumor purity thresholds, matched-normal availability).
- Never-smoker counts depend on smoking-status definitions and how missing or
ambiguous smoking data are handled.

implications: Models and studies based on TCGA-LUAD may generalize poorly
to never-smoker populations due to ever-smoker enrichment and inconsistent
never-smoker cohorts.
```

```
2.  Pathogen detection gaps
summary:  TCGA-LUAD sequencing is designed for cancer genomic profiling, not
low-copy pathogen detection or comprehensive viral discovery.

details:
- TCGA-LUAD molecular profiling emphasizes tumor genomics and transcriptomics
intended to characterize human cancer biology rather than pathogen
discovery.
- Exome-focused DNA sequencing used in TCGA-LUAD analyses targets human exons
and does not enrich pathogen genomes.  Viral sequences are expected
primarily among off-target reads and can be sparse for low-copy,
non-integrated, or subclonal infections.
- TCGA-LUAD RNA sequencing uses poly(A)-enriched libraries, which can
underdetect non-polyadenylated viral RNAs.

implications:  Null viral findings may reflect assay limitations and should
not, by themselves, be interpreted as evidence of pathogen absence in LUAD.

3.  Original 2014 TCGA LUAD cohort is resected and previously untreated
summary:  The 2014 TCGA LUAD cohort characterization analyzed resected,
previously untreated primary tumors.  This sampling may differ from small
biopsies, cytology specimens, metastatic specimens, or post-treatment
material.

details:
- The 2014 TCGA LUAD publication reports analysis of 230 previously untreated
LUAD cases and describes the tumors as resected with multi-assay profiling
[1].

implications:  Models and studies derived from the 2014 cohort may
generalize poorly to small-biopsy, metastatic, or post-treatment contexts.

# References
1.  Comprehensive molecular profiling of lung adenocarcinoma.  Nature.
2014.
```

Figure 1. M.I.J.O. alert for TCGA-LUAD. M.I.J.O. alert documenting TCGA dataset limitations for LCINS research questions.

4.3 PCAWG-LUAD Results

The PCAWG consortium reported 38 LUAD samples in Extended Data Table 1 of the core publication (Campbell et al., Nature 2020).[47] However, Supplementary Table 1 contained clinical metadata for only 37 LUAD samples. Provenance tracing showed that 37/37 samples (100%) originated from TCGA.

The cohort comprised 20 females (54%) and 17 males (46%), with a median age of 66 years. Granular smoking status was not available in PCAWG clinical metadata for LUAD samples. Applying the TCGA-derived never-smoker prevalence to this cohort yielded an estimate of 6 never-smoker samples.

4.4 PCAWG M.I.J.O. Alert

```
# Metadata
mijo_id:  hu_clarence_hotpotai_pcawg_1770336000
mijo_version:  1.0.0
created_at:  2026-02-06T00:00:00-08:00
contact:  @@correspondence

# Authors
authors:  @@authors

# Dataset
common_name:  PCAWG-LUAD
formal_name:  Pan-Cancer Analysis of Whole Genomes - Lung adenocarcinoma
subset
uri:  https://www.nature.com/articles/s41586-020-1969-6
version:  Campbell et al., Nature 2020 (PCAWG flagship)

# Question
Is this dataset applicable to viral association research?

# Limitations
1.  Limited exclusion of low-prevalence viral subtypes
summary:  The never-smoker subset is estimated at six cases, limiting
statistical power for viral association questions in never-smokers.

details:
- PCAWG reports 38 LUAD cases in the core manuscript, but Supplementary Table
1 contains clinical metadata for 37 LUAD cases.
- Provenance tracing indicates the LUAD cases with clinical metadata derive
from TCGA.
- Granular smoking status is not reported for PCAWG-LUAD in the available
clinical metadata, so never-smoker counts must be estimated rather than
directly measured.
- Applying the TCGA-derived never-smoker prevalence yields an estimated six
never-smoker cases in PCAWG-LUAD.

implications:  With six never-smokers and zero detections, the one-sided 95%
exact binomial upper bound on prevalence is 39%, meaning lower-prevalence
subtypes cannot be excluded.

2.  Poly(A)-biased transcriptome limits non-polyadenylated virus detection
summary:  RNA sequencing captured with poly(A)-enriched library preparation
can underdetect viruses with non-polyadenylated RNAs.

details:
- Poly(A)-enriched RNA-seq is designed to capture polyadenylated transcripts
and can deplete non-polyadenylated viral RNAs.
- This increases the false-negative risk for RNA-level viral association
questions, particularly when the hypothesized signal is transcriptional
activity or low-level expression.

implications:  Null RNA-level viral findings in LUAD may reflect
library-capture bias and should not, on their own, be interpreted as
evidence of viral absence.
```

Figure 2. M.I.J.O. alert for PCAWG-LUAD. M.I.J.O. alert documenting PCAWG dataset limitations for viral association research questions in LCINS.

5 Discussion

The M.I.J.O. framework may help researchers assess dataset applicability, publish limitations as machine-readable alerts, and discover prior alerts, facilitating knowledge reuse and reducing the risk that dataset bias propagates into AI models or biomedical research.

Our case studies illustrate such risks in LCINS, which would rank among the deadliest cancers if classified separately and in which a theoretical oncovirus-driven subtype could be clinically meaningful even at 5% prevalence. Our results suggest that null pathogen findings in landmark LUAD cohorts reflect dataset and assay constraints rather than robust evidence of biological absence. Both TCGA-LUAD and PCAWG-LUAD rely on measurements optimized for human cancer profiling rather than pathogen discovery. Poly(A)-enriched RNA sequencing depletes non-polyadenylated viral RNAs, increasing false-negative risk for viruses with non-polyadenylated transcripts such as EBV. Exome-focused DNA sequencing enriches human coding regions rather than pathogen genomes, leaving viral sequences primarily in sparse off-target reads. In PCAWG, tumor WGS coverage targets of 38× and 60× were optimized for somatic variant discovery rather than pathogen discovery, limiting sensitivity for low-copy, non-integrated, or subclonal infections. Low-purity tumors, where many non-tumor cells dilute the sample, can further amplify this limitation and are frequently observed in LUAD.

Small cohorts further weaken interpretation of null results. With 37 LUAD cases in PCAWG-LUAD, a true viral subtype with prevalence up to 7.8% could remain undetected. More broadly, the minimum sample sizes to exclude a viral subtype at low prevalence under null observations – 59 for 5%, 99 for 3%, 299 for 1% – exceed what the PCAWG-LUAD cohort can support. This limitation is amplified for never-smokers: our analysis estimated that only 6 of 37 PCAWG-LUAD cases (16%) were never-smokers, so even a true viral subtype with 39% prevalence could remain undetected.

Our results also emphasize the risk of mismatch between dataset composition and target populations. Several influential AI models and biological inferences are grounded in TCGA and PCAWG, yet may not generalize reliably to never-smokers and other underrepresented patient groups. For instance, the smoker bias in TCGA-LUAD means biological inferences and ML models trained or benchmarked on TCGA-LUAD, including widely used DeepPATH, may overlook critical signals in never-smokers such as oncogenic drivers and immunological profiles.

Finally, M.I.J.O. presents a complementary approach to dataset documentation. Constraints known to specialists are often opaque to others when selecting datasets. M.I.J.O. turns question-specific applicability limits into structured, queryable artifacts that can be embedded in manuscripts or shared independently, promoting knowledge reuse and simplifying dataset assessment before conclusions harden into precedent.

Although originally conceived to help AI researchers select datasets in biomedicine, M.I.J.O. may also support clinicians and other domain experts in oncology and beyond, as well as researchers in fields such as natural language processing, where dataset applicability depends on the research question.

Future work could extend M.I.J.O. in three directions. First, impact tracing could link dataset alerts to derived models and conclusions, enabling investigators to assess whether limitations have downstream consequences. Second, a reporting checklist could help manuscripts demonstrate verification of dataset-question applicability, improving transparency and reproducibility. Third, a community-maintained repository of searchable alerts, along with affected studies and models, could facilitate discovery and knowledge sharing.

6 Limitations

Several limitations influence the M.I.J.O. framework and our findings.

We did not measure whether M.I.J.O. changes dataset selection decisions, model performance, or interpretation quality relative to standard review practices. Generalizability beyond these case studies needs investigation. Because the GDC Legacy Archive was retired, we could not reconstruct historical TCGA-LUAD clinical snapshots, relying instead on peer-reviewed benchmarks to approximate cohort composition. Consequently, smoking-status prevalence and denominator sizes may vary across downstream studies.

Similarly, PCAWG-LUAD samples lacked granular smoking data, forcing us to infer smoking status composition from provenance and the parent TCGA cohort rather than from direct annotation. Furthermore, our viral detectability assessment is based on reasoning from library protocols rather than empirical re-calling. This highlights assay-specific blind spots, such as poly(A) depletion of canonical EBV transcripts, but does not quantify absolute sensitivity or specificity.

Additionally, the decentralized nature of the M.I.J.O. framework carries risks. Allowing standalone publishing without centralized peer review may introduce noise or misleading claims if authors do not verify limitations and readers do not assess author credibility. While AI assistants can accelerate the identification of implicit constraints, they require expert oversight to prevent the propagation of errors and spurious claims. Structured alerts may create an illusion of exhaustiveness, leading researchers to treat an incomplete set as a complete inventory of technical and biological constraints. Finally, our power calculations assume simple binomial sampling and may not fully account for biological heterogeneity in tumor purity, subclonal infection, or demographic factors.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Funding

This work was supported by Hotpot.ai.

References

[1] A. K. Amatya, M. H. Fiero, E. W. Bloomquist, A. K. Sinha, S. J. Lemery, H. Singh, et al. Subgroup analyses in oncology trials: regulatory considerations and case examples. *Clinical Cancer Research*, 27(21):5753–5756, November 2021. doi: 10.1158/1078-0432.CCR-20-4912.

[2] A. J. Alberg and J. M. Samet. Epidemiology of lung cancer. *Chest*, 123(1 Suppl):21S–49S, January 2003. doi: 10.1378/chest.123.1_suppl.21s.

[3] X. Wang, J. T. Steensma, M. H. Bailey, Q. Feng, H. Padda, and K. J. Johnson. Characteristics of The Cancer Genome Atlas cases relative to U.S. general population cancer cases. *British Journal of Cancer*, 119(7):885–892, October 2018. doi: 10.1038/s41416-018-0140-8.

[4] SEER. Cancer of the lung and bronchus - cancer stat facts. `https://seer.cancer.gov/statfacts/html/lungb.html`, 2026. Accessed 2026-05-03.

[5] C. Murphy, T. Pandya, C. Swanton, and B. J. Solomon. Lung cancer in nonsmoking individuals: A review. *JAMA*, 334(20):1836, November 2025. doi: 10.1001/jama.2025.17695.

[6] S. L. Gomez, M. DeRouen, M. S. Chen, H. Wakelee, J. B. Velotta, L. C. Sakoda, et al. Elevated risk of lung cancer among Asian American women who have never smoked: an emerging cancer disparity. *Journal of the National Cancer Institute*, 117(6):1104–1109, June 2025. doi: 10.1093/jnci/djae299.

[7] W. J. McCarthy, R. Meza, J. Jeon, and S. Moolgavkar. Lung cancer in never smokers: epidemiology and risk prediction models. *Risk Analysis*, 32(Suppl 1):S69–S84, July 2012. doi: 10.1111/j.1539-6924.2012.01768.x.

[8] SEER. Common cancer sites - cancer stat facts. `https://seer.cancer.gov/statfacts/html/common.html`, 2026. Accessed 2026-05-03.

[9] National Cancer Institute. TCGA timeline and milestones. `https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history/timeline-milestones`, 2018. Accessed 2026-05-03.

[10] M. T. Landi, N. C. Synnott, J. Rosenbaum, T. Zhang, B. Zhu, J. Shi, et al. Tracing lung cancer risk factors through mutational signatures in never-smokers. *American Journal of Epidemiology*, 190(6):962–976, October 2020. doi: 10.1093/aje/kwaa234.

[11] M. Díaz-Gay, T. Zhang, P. H. Hoang, C. Leduc, M. K. Baine, W. D. Travis, et al. The mutagenic forces shaping the genomes of lung cancer in never smokers. *Nature*, 644(8075):133–144, August 2025. doi: 10.1038/s41586-025-09219-0.

[12] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, October 2018. doi: 10.1038/s41591-018-0177-5.

[13] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, October 2024. doi: 10.1038/s41586-024-07894-z.

[14] A. Li, P. Cui, L. Liu, J. Liu, X. Zhou, W. Wu, et al. Multi-omics integration and machine learning identify NPC2 as a prognostic and treatment-responsive regulator in lung adenocarcinoma. *Frontiers in Immunology*, 16:1697560, 2025. doi: 10.3389/fimmu.2025.1697560.

[15] Q. Cao, C. Li, Y. Li, X. Kong, S. Wang, and J. Ma. Tumor microenvironment and drug resistance in lung adenocarcinoma: molecular mechanisms, prognostic implications, and therapeutic strategies. *Discover Oncology*, 16:238, February 2025. doi: 10.1007/s12672-025-01981-x.

[16] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, October 2024. doi: 10.1038/s41591-024-03141-0.

[17] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. doi: 10.1038/s41591-024-02857-3.

[18] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. doi: 10.1038/s41586-024-07441-w.

[19] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, November 2025. doi: 10.1038/s41591-025-03982-3.

[20] G. Campanella, N. Kumar, S. Nanda, S. Singi, E. Fluder, R. Kwan, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine*, 31(9):3002–3010, September 2025. doi: 10.1038/s41591-025-03780-x.

[21] M. Zapatka, I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, M. Alawi, et al. The landscape of viral associations in human cancers. *Nature Genetics*, 52(3):320–330, March 2020. doi: 10.1038/s41588-019-0558-9.

[22] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi. Global burden of cancers attributable to infections in 2012: a synthetic analysis. *The Lancet Global Health*, 4(9):e609–e616, September 2016. doi: 10.1016/S2214-109X(16)30143-7.

[23] P. J. Farrell. Epstein-barr virus and cancer. *Annual Review of Pathology*, 14:29–53, January 2019. doi: 10.1146/annurev-pathmechdis-012418-013023.

[24] L. S. Young, L. F. Yap, and P. G. Murray. Epstein-barr virus: more than 50 years old and still providing surprises. *Nature Reviews Cancer*, 16(12):789–802, December 2016. doi: 10.1038/nrc.2016.92.

[25] J. Sadri Nahand, N. Rabiei, R. Fathazam, M. Taghizadieh, M. S. Ebrahimi, M. Mahjoubin-Tehran, et al. Oncogenic viruses and chemoresistance: What do we know? *Pharmacological Research*, 170:105730, August 2021. doi: 10.1016/j.phrs.2021.105730.

[26] J. Chen, S. Kendrick, and Z. Qin. Mechanistic insights into chemoresistance mediated by oncogenic viruses in lymphomas. *Viruses*, 11(12):1161, December 2019. doi: 10.3390/v11121161.

[27] K. M. Lai and W. L. Lee. The roles of epidermal growth factor receptor in viral infections. *Growth Factors*, 40(1–2):46–72, June 2022. doi: 10.1080/08977194.2022.2063123.

[28] R. Aloni-Grinstein, M. Charni-Natan, H. Solomon, and V. Rotter. p53 and the viral connection: Back into the future. *Cancers*, 10(6):178, June 2018. doi: 10.3390/cancers10060178.

[29] G. Schönrich and M. J. Raftery. The PD-1/PD-L1 axis and virus infections: A delicate balance. *Frontiers in Cellular and Infection Microbiology*, 9:207, 2019. doi: 10.3389/fcimb.2019.00207.

[30] M. Schiffman, P. E. Castle, J. Jeronimo, A. C. Rodriguez, and S. Wacholder. Human papillomavirus and cervical cancer. *The Lancet*, 370(9590):890–907, September 2007. doi: 10.1016/S0140-6736(07)61416-0.

[31] G. Kennedy, J. Komano, and B. Sugden. Epstein-barr virus provides a survival factor to burkitt's lymphomas. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):14269–14274, November 2003. doi: 10.1073/pnas.2336099100.

[32] T. Yang, C. You, S. Meng, Z. Lai, W. Ai, and J. Zhang. EBV infection and its regulated metabolic reprogramming in nasopharyngeal tumorigenesis. *Frontiers in Cellular and Infection Microbiology*, 12:935205, 2022. doi: 10.3389/fcimb.2022.935205.

[33] D. G. Sausen, A. Basith, and S. Muqeemuddin. EBV and lymphomagenesis. *Cancers*, 15(7):2133, April 2023. doi: 10.3390/cancers15072133.

[34] K. Sun, K. Jia, H. Lv, S. Q. Wang, Y. Wu, H. Lei, et al. EBV-positive gastric cancer: Current knowledge and future perspectives. *Frontiers in Oncology*, 10:583463, 2020. doi: 10.3389/fonc.2020.583463.

[35] A. Jary, M. Veyri, A. Gothland, V. Leducq, V. Calvez, and A. G. Marcelin. Kaposi's sarcoma-associated herpesvirus, the etiological agent of all epidemiological forms of kaposi's sarcoma. *Cancers*, 13(24):6208, December 2021. doi: 10.3390/cancers13246208.

[36] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis. Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*, 6(1):92, November 2020. doi: 10.1038/s41572-020-00224-3.

[37] J. M. Llovet, R. K. Kelley, A. Villanueva, A. G. Singal, E. Pikarsky, S. Roayaie, et al. Hepatocellular carcinoma. *Nature Reviews Disease Primers*, 7(1):6, January 2021. doi: 10.1038/s41572-020-00240-3.

[38] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, April 2024. doi: 10.1136/bmj-2023-078378.

[39] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, et al. Datasheets for datasets. arXiv, 2021. URL `http://arxiv.org/abs/1803.09010`. Accessed 2026-05-05.

[40] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, et al. Model cards for model reporting. arXiv, 2019. URL http://arxiv.org/abs/1810.03993. Accessed 2026-05-05.

[41] Health C for D and R. FDA. eMDR – electronic medical device reporting, 2026. URL https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/emdr-electronic-medical-device-reporting. Accessed 2026-05-05.

[42] OECD. Towards a common reporting framework for AI incidents. OECD Artificial Intelligence Papers, February 2025. URL https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html. Accessed 2026-05-05.

[43] J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, April 2018. doi: 10.1016/j.cell.2018.02.052.

[44] D. Zhou, Y. Sun, Y. Jia, D. Liu, J. Wang, X. Chen, et al. Bioinformatics and functional analyses of key genes in smoking-associated lung adenocarcinoma. *Oncology Letters*, 18(4):3613–3622, October 2019. doi: 10.3892/ol.2019.10733.

[45] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, July 2014. doi: 10.1038/nature13385.

[46] G. F. Gao, J. S. Parker, S. M. Reynolds, T. C. Silva, L. B. Wang, W. Zhou, et al. Before and after: Comparison of legacy and harmonized TCGA Genomic Data Commons data. *Cell Systems*, 9(1):24–34.e10, July 2019. doi: 10.1016/j.cels.2019.06.006.

[47] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, L. A. Aaltonen, F. Abascal, A. Abeshouse, H. Aburatani, D. J. Adams, et al. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, February 2020. doi: 10.1038/s41586-020-1969-6.

[48] J. P. Whalley, I. Buchhalter, E. Rheinbay, K. M. Raine, M. D. Stobbe, K. Kleinheinz, et al. Framework for quality assessment of whole genome cancer sequences. *Nature Communications*, 11(1):5040, October 2020. doi: 10.1038/s41467-020-18688-y.

[49] A. Solovyov, N. Vabret, K. S. Arora, A. Snyder, S. A. Funt, D. F. Bajorin, et al. Global cancer transcriptome quantifies repeat element polarization between immunotherapy responsive and T cell suppressive classes. *Cell Reports*, 23(2):512–521, April 2018. doi: 10.1016/j.celrep.2018.03.042.

[50] S. R. Selitsky, D. Marron, D. Hollern, L. E. Mose, K. A. Hoadley, C. Jones, et al. Virus expression detection reveals RNA-sequencing contamination in TCGA. *BMC Genomics*, 21(1):79, January 2020. doi: 10.1186/s12864-020-6483-6.

[51] C. Ganini, I. Amelio, R. Bertolo, P. Bove, O. C. Buonomo, E. Candi, et al. Global mapping of cancers: The Cancer Genome Atlas and beyond. *Molecular Oncology*, 15 (11):2823–2840, November 2021. doi: 10.1002/1878-0261.13056.

Supplementary Materials

```
# System Role
You are an elite machine learning and biomedical researcher who specializes
in dataset evaluation and provides deep, insightful analysis without ever
hallucinating or inventing facts.

# Goal
Draft M.I.J.O. applicability alerts that help researchers assess dataset
limitations broadly or against a given research question.

# Outputs
The user may request any subset of the following outputs.  Produce only what
the user requests.
- Alert:  A M.I.J.O. applicability alert in Markdown that strictly follows
the full M.I.J.O. specification provided in the conversation.
- Review:  A compact review aid for human verification.
- Audit:  A compliance audit of candidate alerts, with pass/fail checks and
minimal fixes.

# Required Inputs
- Full M.I.J.O. specification text.
- Dataset paper or dataset details sufficient to draft an alert.
- Research question is optional and may be omitted for dataset-generic alerts
per spec.

# Requirements
- Never invent facts or details.
- If a required value is missing or not verifiable from user-supplied
materials, represent it using the spec-defined missing-value convention or
the user-provided placeholder token where permitted.
- Label uncertain statements as candidate limitations that need verification
and avoid phrasing them as established facts.
- When interacting with users, be concise to reduce their cognitive load.
- Alerts must follow the M.I.J.O. specification, but if the user requests
deviations from the spec, flag the result as a non-compliant Alert draft.

# Workflow

## Step 1 Requirements Check
Confirm the minimum requirements needed to proceed.
Check that the full M.I.J.O. specification is present in the conversation
and that the user has provided a dataset paper or sufficient dataset
details.
If no research question is provided, confirm that the user wants to analyze
the dataset broadly and not against a research question.
If required inputs are missing, request them and stop.
End with:
- Next step:  Provide missing inputs, start next step, or stop.

## Step 2 Limitations Decision
Ask if the user wants:
- Help identifying limitations from dataset materials and first principles,
or
```

```
- Help documenting limitations the user will provide.
If the user provided limitations, assume documentation mode and request
supporting evidence if missing.
If the user wants documentation help, request the limitations and supporting
evidence, then create alerts.
If the user wants discovery help, proceed to Step 3.
End with:
- Next step:  Start limitation discovery, create alerts, or stop.

## Step 3 Limitations Discovery
Execute only if the user requests limitation discovery.
Begin by identifying datasets upstream of the chosen dataset, but only list
upstream datasets explicitly referenced in the dataset paper or datasheet.
Ask if limitation analysis should include these upstream datasets or only
the chosen dataset.
Then identify limitations that could affect applicability to the question or
the dataset broadly, drawing from both explicit statements in dataset
materials and implicit constraints suggested by material factors such as
cohort composition, label definitions, assay design, preprocessing,
missingness, provenance, temporal shift, and access constraints.
- Use a single limitation per entry, de-duplicate overlaps, and prioritize by
expected impact on the question and by the likelihood of misleading
inference if ignored.
- Note maintenance and version risks when limitations may change across
releases, pipelines, or portal snapshots.
- End the response with a concise reminder to verify candidate limitations.
End with:
- Next step:  Request verification guidance, create alerts, or stop.

## Step 4 Verification Guidance
Execute only if requested by the user.
Provide verification guidance describing what to check, where to check, and
how to document evidence.
Create a short checklist for each limitation, focused on primary sources and
the most efficient way to sufficiently verify limitations since researchers
are resource-constrained.
End with:
- Next step:  Create alerts, or stop.

## Step 5 Outputs
Before generating alerts, ask for any missing information required to
generate the alert and ask whether the alert is meant to be embedded or
published alone.
If the alert is embedded, use the @@ placeholder mechanism for relevant
fields.
Then produce only the outputs the user requests.
If Alert is requested, draft the alert according to the provided M.I.J.O.
specification.
If Review is requested, provide a compact review aid to support human
verification, focused on what inputs were used, what is missing, the
prioritized limitations, any upstream dependency notes if assessed, and any
verification guidance if requested.
End with:
```

````
- Next step:  Request edits, request audit, recreate alerts, or stop.

## Step 6 Audit
Execute only if requested by the user.
Audit alerts against the supplied M.I.J.O. specification.
By default, audit the most recent Alert generated during the session unless
the user provides specific alerts to audit.
Output a concise checklist with pass or fail per requirement and a short
justification for each failure.
For failed items, propose the smallest edits needed to achieve compliance
while preserving grounding and avoiding invented facts.
End with:
- Next step:  Apply fixes, rerun audit, or stop.

# Greeting
If the user's name is known, begin with:  Hi $username.  Otherwise, begin
with:  Hi.
```markdown
I'll help you draft a M.I.J.O. applicability alert.

To start, please share the dataset to analyze, and optionally a research
question.

It is ideal to provide the dataset paper or documentation, but we can start
with the name.

If no question is provided, I'll analyze the dataset broadly for potential
limitations.

For help, just ask me questions.
```
````

Supplementary Figure S1. M.I.J.O. reference system prompt. Used to transform user-provided data into a M.I.J.O. alert via guided workflows.