World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

Data sources, quality and preprocessing challenges in cancer research: A comprehensive review

Mehdi Zaeifi and Sindhuja Rao Yerra *

Ph. D Student, Department of Data Science and Analytics, University of Oklahoma, Norman 73019.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 1866-1871

Publication history: Received on 09 May 2025; revised on 14 June 2025; accepted on 17 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.1120

Abstract

Cancer is a major global health issue and is responsible for close to one-sixth of all global deaths, states the World Health Organization. With pervasive availability of multi-modal datasets such as genomic profiles, clinical reports, and imaging data, cancer biology and patient care have been advanced greatly. Contemporary data science techniques, from simple description and statistical analyses to advanced machine learning models, are all highly dependent on the quality and reliability of such data. While several studies emphasize working on new models with minimal attention to the fundamental step of data preprocessing, various problems like missing values, measurement error, heterogeneity, and privacy can dramatically degrade the validity of research findings if ignored. Therefore, enhancing data curation and preprocessing standardization and raising awareness is imperative for improving reproducible and influential cancer research. This paper seeks to (i) give an overview of some of the major publicly accessible cancer datasets, (ii) describe some of the shared data quality problems experienced within cancer research, and (iii) make recommendations and synthesize best practices for data preprocessing and management. The target group comprises data scientists and researchers dealing with oncology, bioinformatics, and biomedical informatics. Maintaining high data quality is more than just a technical task it's an ethical responsibility. Inaccurate or poorly handled data can lead to misleading clinical decisions and ultimately impact patient care and outcomes.

Keywords: Cancer Informatics; Data Quality; Preprocessing; Big Data; Batch Effects; Missing Data

1. Introduction

Cancer continues to pose a significant global health burden, accounting for nearly one in every six deaths worldwide according to the World Health Organization. With advances in high-throughput technologies and widespread digitization of medical records, the availability of multi-modal datasets including genomic profiles, clinical records, and medical imaging has transformed cancer research and patient care. These diverse data streams have enabled groundbreaking discoveries in tumor biology, prognostic modeling, and personalized therapy design. However, the reliability and impact of these scientific advancements are deeply tied to the quality and integrity of the underlying data. Poor data quality stemming from missing values, measurement errors, batch effects, and inconsistent coding standards can lead to flawed analyses and misleading conclusions, ultimately affecting clinical decisions and patient outcomes. Despite this, many studies continue to prioritize novel model development while underemphasizing the critical first step of robust data preprocessing and curation.

This review addresses this gap by providing a comprehensive overview of key publicly available cancer datasets, common data quality issues prevalent in cancer research, and practical, evidence-based preprocessing strategies. By synthesizing best practices and highlighting current challenges, we aim to support data scientists, bioinformaticians, and clinical researchers in producing more accurate, reproducible, and ethically responsible cancer research. Ensuring

* Corresponding author: Sindhuja Rao Yerra

high-quality data is not only a technical necessity but an ethical imperative that underpins trustworthy scientific progress and better patient care.

2. Literature Review

2.1. Public Cancer Data Initiatives

Public large-scale consortia have been pivotal to democratizing access to cancer data on an enormous scale. Landmark initiatives such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) have given researchers multi-omics profiles accompanied by extensive clinical metadata (Weinstein et al., 2013; Zhang et al., 2011). These assets have supported tumor classification, signatures of mutational patterns, and therapeutic target discovery (Hoadley et al., 2018). Population registries like the SEER Program and National Cancer Database (NCDB) have enabled epidemiologic surveillance, disparities research, and survival modeling for large, heterogeneous populations (Howlader et al., 2020). Likewise, imaging repositories such as the Cancer Imaging Archive (Clark et al., 2013) and an imaging extension of the UK Biobank have enabled radiomics and imaging biomarker development.

2.2. Data Quality Issues within cancer research

Prior studies have established that technical and biological variability pervasively affect cancer datasets. Leek et al. (2010) highlighted that batch effects can add confounding noise greater than genuine biological variation, particularly for large-throughput omics data. Lazar et al. (2012) reviewed several batch correction methods, stressing routine adjustment necessity. Missing data continue to persist extensively in clinical and genomic datasets (Little & Rubin, 2019). Ad hoc imputation can result in biased point estimates, and although better imputation techniques are available, these are based on specific and often untestable assumptions regarding missingness processes. Data heterogeneity also makes reproducibility and meta-analyses difficult. As an example, local coding standards are commonly applied for clinical records instead of international vocabularies like ICD-10 or SNOMED CT, preventing integration (Hripcsak et al., 2015). In imaging, variations in type of imaging scanner and imaging protocol impact feature reproducibility (Traverso et al., 2018).

2.3. Current Gaps and Need for Standardized Pipelines

While different reviews address data quality for certain modalities, no practical, comprehensive guidance for end-to-end preprocessing for multi-modal cancer datasets exists. Recent literature calls emphasize FAIR data principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016) and sharing preprocessing scripts with publications for reproducibility purposes (Peng, 2011). This work contributes by summarizing practical advice and pointing out typical pitfalls for data scientists and cancer researchers. Yet most research mainly targets novel model designing without giving sufficient attention to the first step of thorough data preprocessing. Missing values, measurement errors, heterogeneity, and privacy problems can largely undermine research findings' validity if these are unresolved. Therefore, enhancing recognition and standardization of data curation and preprocessing processes are key to promoting reproducible and significant cancer research. This article seeks to (i) give an overview of some leading publicly available cancer datasets, (ii) outline some frequent data quality problems faced in cancer research, and (iii) summarize best practices and recommendations for data preprocessing and management. The target audience is researchers and data scientists who work at the crossroads of oncology, bioinformatics, and medical informatics.

3. Methods

For this review, a systematic search was carried out across peer-reviewed literature available in PubMed, Web of Science, and Google Scholar, focusing on studies published between 2010 and 2024. The search used key terms such as “cancer datasets,” “cancer data quality,” “batch effects,” “missing data in oncology,” and “radiomics reproducibility.” Major cancer data repositories were chosen based on how frequently they were cited, their public availability, and their relevance to multi-modal cancer research. Information on data quality challenges and preprocessing approaches was gathered and summarized from various methodological and review articles.

3.1. Overview of Major Public Cancer Data Sources

This paper provides an detailed description of the most commonly utilized data stores, including their scope and type of data and normal applications.

3.2. The Cancer Genome Atlas (TCGA)

TCGA provides comprehensive, multi-layered data including somatic mutations, gene expression profiles, DNA methylation patterns, and copy number changes — for more than 30 types of cancer. Detailed clinical annotations make it a valuable resource for integrative studies and the discovery of molecular subtypes.

3.3. International Cancer Genome Consortium (ICGC)

The ICGC serves as a complement to TCGA by focusing on diverse and often underrepresented populations. Its whole-genome sequencing data allow researchers to explore mutational patterns across populations and identify rare genetic variants.

3.4. Surveillance, Epidemiology, and End Results (SEER) Program

SEER remains a key resource for monitoring cancer trends across the United States. It offers extensive data on cancer incidence, survival rates, and treatment patterns for large and diverse patient populations.

3.5. The Cancer Imaging Archive (TCIA)

TCIA contains thousands of de-identified CT, MRI, and PET scans along with detailed annotations. This resource supports radiomics research, the development of AI models, and the validation of imaging biomarkers.

4. Common Data Quality Challenges

To conduct high-quality and reproducible cancer research, it's crucial to tackle common data quality challenges using suitable preprocessing strategies. Table 1 provides a clear summary of the most frequent issues found in multi-modal cancer datasets like missing data, batch effects, heterogeneity, imbalanced classes, and privacy considerations and lists practical, widely used solutions for each problem. This straightforward overview is intended as a handy reference to help researchers choose the right methods for cleaning and harmonizing their data before moving on to further analyses.

Table 1 Common Data Quality Issues and Solutions

Data Issue	**Solution**
Missing Data	Deletion, mean/median imputation, kNN, MICE
Batch Effects	ComBat, harmonization algorithms, standard protocols
Heterogeneity	Use standard vocabularies, normalization, mapping
Imbalanced Classes	Resampling, synthetic data generation (SMOTE)
Privacy & Ethics	Automated de-identification tools, secure storage, compliance

4.1. Missing Data

Missing data often result from patient dropouts, incomplete follow-ups, or technical glitches. If not handled properly, missing information can introduce bias and weaken the statistical reliability of study results.

4.2. Batch Effects

Differences caused by varying labs, experimental protocols, or collection time points can introduce systematic errors. These batch effects may mask true biological signals and limit how well models perform on new data.

4.3. Heterogeneity

Cancer datasets come in diverse formats, scales, and semantic structures, which makes integrating and reproducing multi-modal analyses challenging.

4.4. Imbalanced Class Distributions

When certain cancer subtypes or clinical outcomes are rare, it becomes difficult to train and validate models effectively, increasing the risk of overfitting.

4.5. Privacy and Ethics

Ensuring patient privacy while maintaining data usefulness requires strong de-identification measures and adherence to privacy regulations like HIPAA, GDPR, and institutional guidelines.

5. Recommended Preprocessing Strategies

Effective preprocessing is crucial for minimizing the impact of low-quality data. This section highlights practical techniques commonly discussed in the literature.

5.1. Handling Missing Data

- Deletion: Suitable only when data are missing completely at random and represent a small fraction of the dataset.
- Simple imputation: Use mean, median, or mode values to fill in missing numerical or categorical data.
- Advanced imputation: Techniques like k-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), or model-based approaches such as Expectation-Maximization can provide more accurate estimates.

5.2. Batch Effect Correction

- ComBat: A widely adopted method for adjusting batch effects in genomic datasets.
- Harmonization algorithms: Used to standardize MRI images obtained from different scanners.
- Calibration techniques: Help maintain consistency across labs and experimental setups.

5.3. Normalization

- Genomics: Methods like log transformation and quantile normalization adjust data distributions for comparability.
- Imaging: Intensity normalization and image resampling ensure consistent image quality and resolution.
- Clinical lab values: Standardizing results to reference ranges improves interpretability across patients.

5.4. Outlier Detection

- Statistical methods: Tools such as z-scores and interquartile range (IQR) rules identify extreme values.
- Machine learning: Algorithms like Isolation Forest and robust clustering detect anomalies in complex datasets.

5.5. De-identification

- Use automated software to remove personal information from DICOM headers and electronic health records.
- Apply secure access controls and data use agreements to protect patient privacy.

6. Case Examples

6.1. Breast Cancer Subtyping

TCGA data have played a pivotal role in identifying molecular subtypes of breast cancer, which inform prognosis and treatment plans. Rigorous preprocessing, including normalization and batch effect correction, was vital to these breakthroughs.

6.2. SEER-based Survival Analyses

SEER datasets have supported extensive research on survival disparities. Properly addressing censored data and missing follow-up records is essential to generate unbiased survival estimates.

6.3. Radiomics in Lung Cancer

Non-small cell lung cancer (NSCLC) data from TCIA highlight the need for strong image preprocessing, systematic feature extraction, and harmonization to produce reliable radiomics biomarkers.

7. Discussion and Future Directions

Despite major strides in cancer data sharing and analytics, the field still lacks standardized preprocessing guidelines. Journals and funding bodies should advocate for the publication of clear workflows and encourage researchers to share preprocessing scripts alongside their analyses. Emerging approaches like federated learning and privacy-preserving computation can help tackle privacy challenges while promoting collaboration between institutions. Additionally, expanding and diversifying data sources will enhance the generalizability of models across various patient populations.

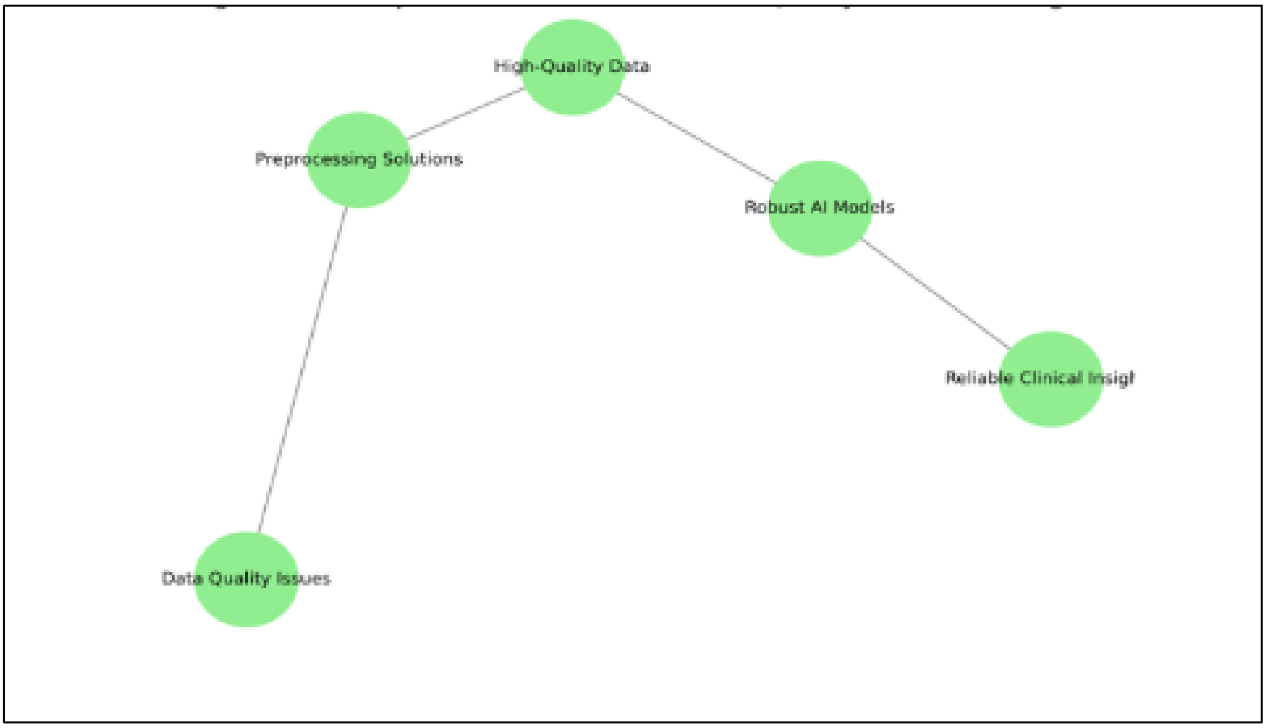

Figure 1 Conceptual Framework from Data Quality to Clinical Insights

Figure 1 shows a conceptual diagram shows the progression from recognizing data quality problems to applying appropriate preprocessing techniques, which ensures clean, high-quality data. This solid foundation enables the development of robust AI models that, in turn, generate trustworthy clinical insights. In addition to robust preprocessing, open data sharing through repositories like Zenodo and Dryad, along with version-controlled preprocessing pipelines hosted on GitHub, can substantially increase transparency and reproducibility.

8. Conclusion

Reliable, thoroughly preprocessed data are fundamental for credible cancer research. This review summarizes key public datasets, common challenges in data quality, and actionable strategies for effective data management. By following best practices and supporting transparency and reproducibility, the cancer research community can continue to drive progress toward better patient care and outcomes.

Emerging Challenges and Opportunities

The rise of single-cell sequencing and spatial transcriptomics has added new layers of complexity to cancer research, creating challenges like greater data sparsity and batch effects at the single-cell level. Likewise, the growth of digital pathology and whole-slide imaging is producing massive petabyte-scale image collections, demanding innovative preprocessing methods and efficient storage solutions. Combining these new data types with traditional clinical and genomic information remains a significant hurdle. Recent progress in artificial intelligence and deep learning offers exciting possibilities for improving cancer diagnosis and prognosis. However, these advanced models are highly sensitive to differences in how data are preprocessed. Without standardized workflows, model accuracy and reliability can fluctuate widely between institutions and patient populations, which raises questions about fairness and real-world applicability. Going forward, researchers should focus on building robust, transparent AI systems trained on well-prepared, consistent datasets. Methods like synthetic data generation and privacy-preserving techniques, such as federated learning, could help tackle privacy concerns while fostering multi-institutional collaboration. To ensure reproducible and fair cancer research globally, the research community must work together to create and share open-source preprocessing pipelines. Another pressing issue is algorithmic bias: when models are trained on poorly curated or demographically unbalanced datasets, they can reinforce existing health disparities. To prevent this, preprocessing should also involve fairness-aware sampling and methods for detecting and addressing bias.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest to disclose.

References

[1] Weinstein, J. N., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120.

[2] Zhang, J., et al. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database, 2011, bar026.

[3] Hoadley, K. A., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, 173(2), 291–304.e6.

[4] Howlader, N., et al. (2020). SEER Cancer Statistics Review, 1975–2017. National Cancer Institute.

[5] Clark, K., et al. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. Journal of Digital Imaging, 26(6), 1045–1057

[6] Leek, J. T., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733–739.

[7] Lazar, C., et al. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. Briefings in Bioinformatics, 14(4), 469–490.

[8] Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data (3rd ed.). Wiley.

[9] Hripcsak, G., et al. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Studies in Health Technology and Informatics, 216, 574–578.

[10] Traverso, A., et al. (2018). Repeatability and reproducibility of radiomic features: A systematic review. International Journal of Radiation Oncology Biology Physics, 102(4), 1143–1158.