23 June 2025 · CC-BY 4.0

Qeios

Peer Review

# Review of: "Kepler: The Pioneer of Data Science and AI"

**Luciano Sathler**[1]

1. Independent researcher

Great article, especially when inviting scientists and educators to expand their possibilities in research based on an interaction with Artificial Intelligence mobilized by good questions, mastery of available techniques, and academic ethics/integrity.

Perhaps one of the biggest challenges for researchers today is keeping up to date with the data published in scientific journals, to identify what is truly scientifically valid and what was eventually produced by AI without true validation of experiments in virtual or real laboratories. Little or no human supervision means that publications run the risk of being populated by scientific articles with data that are not truly valid. In this sense, critical thinking and new validation are what will allow researchers to be sure that they are heading in the right direction in terms of their AI-supported research.

Ensuring data quality and validation in AI-driven research is a cornerstone of scientific rigor and reproducibility. As a high-level researcher, one must adopt a multi-layered approach that integrates methodological precision, domain expertise, and computational robustness. Here are specific strategies and examples:

### *1. Rigorous Data Preprocessing and Cleaning*

Before any modeling, researchers must address missing values, outliers, and inconsistencies. For instance, in a study using electronic health records (EHRs) to predict disease onset, missing timestamps or inconsistent diagnostic codes must be resolved using imputation techniques or domain-specific rules. Tools like **Pandas Profiling** or **Great Expectations** can automate data audits.

### *2. Provenance and Metadata Tracking*

Maintaining detailed metadata (e.g., source, collection method, timestamp, transformations) ensures traceability. For example, in environmental AI models predicting air quality, sensor calibration data and geolocation metadata are essential for contextual accuracy.

### *3. Data Annotation Quality Control*

In supervised learning, label quality is paramount. Researchers should use **inter-annotator agreement metrics** (e.g., Cohen's Kappa) and **active learning** to prioritize ambiguous samples for expert review. In NLP sentiment analysis, for instance, ambiguous texts should be flagged and re-evaluated by multiple annotators.

### *4. Bias and Fairness Audits*

AI models often inherit biases from data. Researchers must conduct **demographic parity** and **equalized odds** tests. For example, in a hiring algorithm, ensuring that prediction accuracy is consistent across gender and ethnic groups is critical. Tools like IBM's **AI Fairness 360** can assist in these audits.

### *5. Cross-Validation and External Validation*

Robust validation includes **k-fold cross-validation** and testing on **external datasets**. In medical imaging, a model trained on one hospital's data should be validated on data from another institution to ensure generalizability.

### *6. Synthetic Data and Data Augmentation*

When data is scarce or imbalanced, synthetic data generation (e.g., using GANs) can help. However, researchers must validate that synthetic data preserves the statistical properties of the original dataset. In fraud detection, synthetic fraudulent transactions can be used to balance training data, but must be validated against real-world patterns.

### *7. Transparent Reporting and Reproducibility*

Publishing **data sheets for datasets** (Gebru et al., 2018) and **model cards** ensures transparency. These documents describe dataset composition, collection process, intended use, and limitations, which are critical for peer review and replication.

### *8. Continuous Monitoring and Drift Detection*

In production environments, data distributions can shift. Implementing **data drift detection** (e.g., using KL divergence or population stability index) ensures that models remain valid over time. For example, in financial forecasting, macroeconomic changes can render historical data less predictive.

## Declarations

**Potential competing interests:** No potential competing interests to declare.