LONGITUDINAL ANALYSIS OF RESEARCH PERFORMANCE USING INTEGRATED INSTITUTIONAL DATA

Martin Ndithi Ndeto

1. Problem Statement

Academic institutions generate large volumes of administrative and scholarly data through grant management systems, publication repositories, and researcher profile platforms. These data sources are typically siloed, maintained independently, and optimized for operational reporting rather than longitudinal analysis. As a result, institutions lack the empirical foundation required to study relationships between funding, research activity, and output over time.

From a research perspective, this setting exhibits characteristics common to institutional analytics problems: fragmented data, inconsistent identifiers, temporal misalignment, and the absence of ground-truth performance measures. These characteristics complicate both descriptive analysis and causal inference.

This memo investigates the following research question:

How can integrated institutional data enable longitudinal and policy-relevant analysis of research performance under real-world data constraints?

The objective is to construct an analytics-ready data foundation that supports empirical study rather than to define or optimize performance metrics.

2. Data Description

2.1 Data Sources

The analysis integrated multiple institutional data sources, including:

- Grant management records capturing applications, awards, funding amounts, and timelines
- Publication metadata linked via ORCID identifiers
- Researcher profiles containing departmental affiliation and role information

Each source differed in structure, update frequency, and data quality.

2.2 Data Characteristics and Constraints

Several challenges shaped the analysis:

- Inconsistent identifiers across systems
- Delays between funding receipt and observable research output
- Incomplete or missing publication records

These constraints required conservative integration strategies and informed the choice of evaluation metrics.

3. Methodology

3.1 Data Integration Strategy

A unified data schema was designed to support longitudinal tracking of grants, researchers, and outputs. ORCID identifiers were used where available to reduce ambiguity in author attribution, consistent with best practices in scholarly data integration.[1]

Data integration prioritized traceability and auditability over completeness, allowing unresolved ambiguities to be explicitly flagged rather than silently resolved.

3.2 Analytics Pipelines

Analytics-ready pipelines were constructed to:

- Normalize funding and publication timelines
- Associate research outputs with funding periods
- Enable cohort-based and temporal analysis

The pipelines were designed to support exploratory analysis and hypothesis generation rather than automated evaluation.

4. Analytical Opportunities

4.1 Descriptive Analysis

The integrated dataset enabled empirical analysis of:

- Research output trajectories following funding awards
- Time lags between funding receipt and publication
- Participation patterns across departments and funding schemes

These analyses were not feasible using siloed systems.

4.2 Exploratory Modeling

The infrastructure supports exploratory analysis of grant success probability, output trajectories, and aggregate funding efficiency metrics. These analyses align with prior work in the science-of-science literature, which emphasizes empirical study of research systems over prescriptive evaluation.[2]

5. Evaluation

5.1 Evaluation Strategy

Given the absence of ground-truth performance measures, evaluation focused on the feasibility and consistency of longitudinal queries enabled by data integration.

5.2 Findings

Data integration substantially improved visibility into research trajectories and revealed systematic delays between funding and measurable output. The unified dataset supported empirical questions that were previously infeasible to address due to data fragmentation.

6. Discussion

6.1 Interpretation

The results demonstrate that integrated institutional data can function as a research instrument for studying research performance dynamics. Importantly, the system enables empirical analysis without imposing normative performance judgments.

6.2 Limitations

Metrics derived from integrated data may encode disciplinary and structural bias, and publication counts alone do not capture research impact. Prior work cautions against the uncritical use of such metrics for evaluation or ranking.[3]

6.3 Research Directions

Future research directions include:

- Causal modeling of funding interventions
- Incorporation of alternative impact measures
- Comparative analysis of funding policies across cohorts and time

7. Concluding Note

This memo illustrates how integrated institutional data can support longitudinal and policy-relevant analysis of research performance under real-world constraints. By emphasizing auditability, interpretability, and empirical grounding, the work provides a foundation for principled study of institutional research systems.

References

1. *Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. Learned Publishing.*
2. *Fortunato, S., et al. (2018). Science of science. Science.*
3. *Hicks, D., et al. (2015). The Leiden Manifesto for research metrics. Nature.*

Status Note

This memo documents exploratory research derived from operational institutional systems and is shared to provide context on problem formulation, methodology, and evaluation under real-world data constraints. The work has not yet been submitted for peer review and continues to evolve.