https://doi.org/10.51408/issi2025_138

Algorithmically Calculated Mentorship: The Netherlands Validation Study

Kathryn O. Weber-Boer[1], Carlos Areia[2], Tamarinde Haven[3]

[1]*k.weberboer@digital-science.com*,[2]*c.areia@digital-science.com*
Cornell University, Digital Science (USA)

[3]*T.L.Haven@tilburguniversity.edu*
Tilburg University (Nederlands)

Abstract

Many efforts to intervene in research practices, with the aim of promoting open science and research integrity, are based on intuitive speculation about what actions might be effective. Research culture involves the norms for registration, research, and publication, but also what responsibility is taken for training, and which behaviours are conventional in collaboration. Efforts to drive shifts in research culture often focus on awareness-raising activities, based on the assumption that a lack of knowledge or familiarity hinders practices of openness and integrity. These activities can be resource intensive, and participants may be self-selecting (where participation is voluntary). The hope is that awareness will spread organically into departments and disciplines. Testing the assumptions upon which these interventions are based provides data-driven evidence to support and strengthen these efforts.

One of the assumptions we are working to test is that open science and integrity practices are related to mentorship, or whether these practices are driven by other forces (e.g., career stage, national or institutional policies). In order to enhance the effectiveness of interventions, we seek to contribute to efforts to quantify the impact of mentorship on open science and research integrity practices. The research in progress presented here takes a first step in this quantification, by testing the foundation of a systematic approach to identifying mentor-mentee pairs. The ability to identify mentorship relationships at scale will enable the analysis of the relationship between mentor and mentee research practices, as well as allow for the assessment of other variables.

This work compares a manually curated dataset of candidates with PhDs awarded from 2021 and 2022 by four Dutch university medical centers and their supervisors (supervisory), to a dataset of pairs of researchers in which a mentor-mentee relationship was algorithmically determined (mentorship). All but one of the supervisory pairs were found in the mentorship dataset, and the strength of mentorship likelihood was largely high or very high. The mentorship dataset further includes informal mentors for the junior researchers. This lays the groundwork for a comparison of the research culture practices of supervisors and supervisees, compared to mentors from formal and informal relationships. This research so far demonstrates high confidence for algorithmically determined mentorship.

Introduction

The broader work of which this is a part aims to investigate the transmission of research culture between supervisors and supervisees. It is essential to be able to qualify and quantify the effect of research policy on research practice, to demonstrate the potential effects of incentives on open science practices. Without knowing whether there is an effect, we are limited in our ability to advocate for training programs, codes of conduct, or other efforts to enhance desirable research practices (Haven, 2025). Interventions in good research practice can be very resource

intensive, and a data-based assessment of the efficacy of these interventions would be valuable to the community.

Scholarly mentorship may play a vital role in shaping the careers of early-stage researchers. However, the impact of mentorship on research culture and practices is relatively under-explored. Correlation between mentorship, research integrity and open science practices remains unknown and there is a need for investigation to quantify the impact of mentorship and identify the factors that contribute to impactful relationships. At Digital Science [REF], we calculated billions of researcher to researcher relationships, including mentorship, however validation is required to test the accuracy and generalisability of this algorithm.

Therefore, the aim of this work is to establish the validity of a mentorship algorithm, by ensuring that manually curated supervisor-PhD pairs are identified in the resulting mentorship dataset, evaluating whether the strength of the relationship correlates with formal supervision, and assessing whether the mentorship dataset also provides likely candidates for informal mentors.

Mentorship is algorithmically determined by drawing upon evidence in publication and grant metadata for collaboration, combined with researcher-specific evidence of seniority. The algorithm produces a dataset of researcher pairs with numeric estimates of the closeness of the relationship and the degree and direction of seniority. The curated list of supervisor-PhD pairs was manually collected as part of a previous research project. This list is used to evaluate the accuracy of the mentorship dataset.

Methods

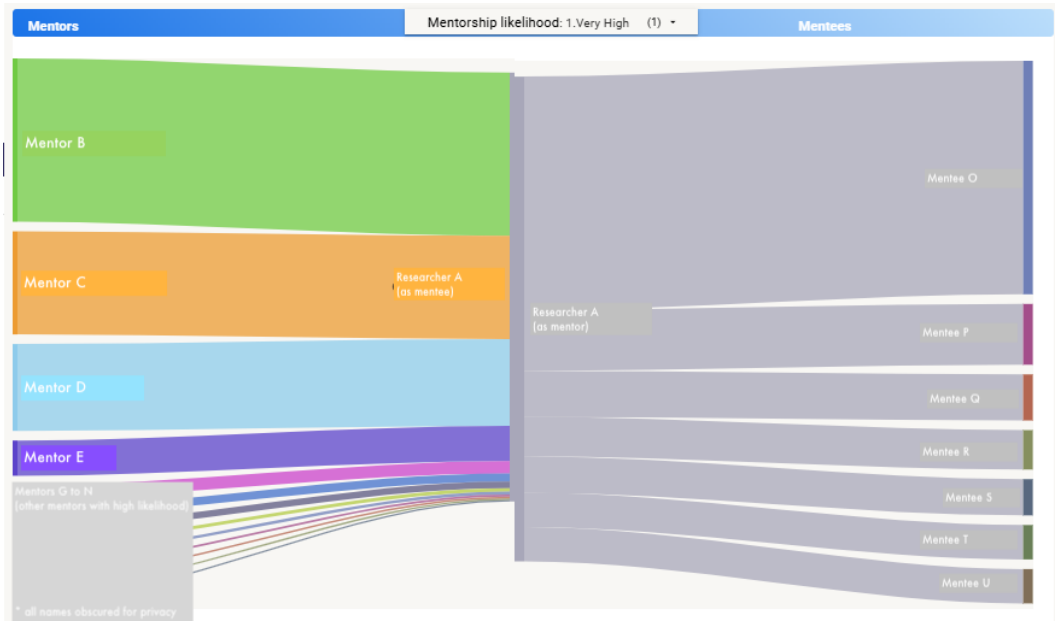
Curated PhD supervisor pairs

The curated PhD-supervisor pair dataset was curated as part of a process that developed new methods for quantifying role modelling of open science practices; the data are publicly available online. (https://github.com/tamarinde/ResponsibleSupervision/tree/main/Pilot-responsible-supervision). The data was manually collected based on a standardized protocol where researchers used PhD thesis metadata to systematically gather data about PhD candidate and main PhD supervisor (in Dutch: *promotor*). We uploaded the tables of PhD-supervisor pairs from four Dutch UMCs (Amsterdam, Groningen, Leiden, Maastricht) into Google BigQuery, and joined them into a final table. The data were cleaned for consistency. The resulting table consists of the PhD candidate and supervisor names, pair ID, and their publication DOIs, amongst other data (all publicly available in GitHub).

Mentorship Dataset

The mentorship database is composed of pairs of co-authors, with a calculated relationship strength and seniority estimation between two researchers: the mentor (Researcher) and the mentee (Co-Researcher). Both Dimensions and Altmetric data are used for this calculation, which considers:

1- Strength of the relationship, that includes data like the number of publications, citations and attention information (according to Altmetric data) of joint publications, grants and clinical trials; number of years sharing research, number of years in the same institution, and publication age.
2- Specific indicators of mentorship, including authorship position on the Researcher's first publications, and investigator roles in the Researcher's first grants.
3- Direction of the relation, using a seniorship score, where a higher score indicates the mentor in the Researcher-Co-Researcher pair.
Based on the above three points we calculated the strength and direction of the mentorship score, which we will refer to as the "Mentometer". A positive Mentometer score indicates that the Researcher is the mentor and the Co-Researcher is the mentee. We also used the Mentometer to calculate a categorical variable of the mentorship likelihood that ranged from "Very Low" to "Very High".



Dimensions data matching

To be able to match Dimensions researcher data to the correct PhD candidate and supervisors, we have followed these steps:

1. Grouped all DOIs available for each researcher
2. Extracted all authors names for each researcher publications
3. Tried to match all original tables PhDs/supervisors with the correct Dimensions researcher ID using the first 2 letters of the first name and the last 2 letters of the last name for the PhD candidates or last 3 letters of the last name for supervisors.
4. Ranked each researcher-Dimension author match automatically and only selected the top match
5. Two independent researchers (CA and KB) manually cross-checked the final PhDs and supervisors matching list, deleting abnormal matches when

multiple matches occurred, so there was only one researcher-Dimensions author match per PhD candidate and supervisor

6. Finally, the matched table includes the pair ID, the PhD candidate name, the supervisor name, subfield, and the thesis year from the curated PhD-supervisor pair dataset, and the supervisor and PhD candidate researcher identifiers from the Dimensions researchers dataset.

Using this linked dataset, we pulled the pairs of researchers from the Mentorship dataset and extracted the likelihood value of the mentor relationship of the supervisor-PhD candidate pairs. We then looked at other mentor candidates to establish whether there were stronger candidates identified in the mentorship algorithm.

One feature of the Dimensions researchers dataset is a tendency to privilege precision over recall. That is, whereas one researcher profile is highly unlikely to contain publications which are not authored by that researcher, it is not unexpected to find multiple profiles per researcher. We selected the strongest mentorship relationship pair, since there were a number of occasions on which multiple mentor-mentee pairs were found (representing the same PhD-supervisor pair). We also alerted the Dimensions support team of any duplicate researcher profiles found, for merging.

Results

This study included 213 distinct supervisors and 213 PhD candidates, all successfully matched to their respective Dimensions researcher IDs.

Table 1. Datasets and the number of supervisor and PhD names and pairs per set.

Dataset	*Supervisor Names*	*PhD Candidate Names*	*Pairs*
Manually curated	219	214	213
Matched Dimensions Researcher profiles	220	220	228
Matched in Mentorship dataset	214 (218 IDs)	213 (218 IDs)	212

Of the 213 PhD-supervisor pairs, 212 were found as pairs in the Mentorship dataset. Because of the additional Dimensions researcher profiles per researcher, there were more mentorship pairs than PhD-Supervisor pairs. Of the mentorship pairs, 188 were classified with a very high likelihood mentorship, and a further 11 had a high likelihood. The remainder of PhD-Supervisor pairs had likelihood of medium, low, or very low. One pair was not identified (Table 2).

Table 2. Mentorship likelihood and pairs matched from manual dataset.

Dataset	*Unique Pairs*
1. Very high	187
2. High	11
3. Medium	7
4. Low	4
5. Very Low	3
not identified	1

Discussion

This study aimed to validate an algorithmic calculation of researchers' mentorship relation. The mentorship score under validation was algorithmically determined by drawing upon evidence in publication and grant metadata for collaboration, combined with researcher-specific evidence of seniority. The algorithm produced a dataset of researcher pairs with numeric estimates of the closeness of the relationship and the degree and direction of seniority. This algorithmically determined mentorship dataset has been previously used by two of the authors (CA and KWB) to explore the transmission of open access publication practices. While the results were promising, suggesting a positive correlation between the open access publishing of the supervisor and the open access publishing rate of the supervisee, three questions required additional investigation: 1) did the relationships identified by the algorithm reflect real-life supervision, and 2) how does the influence of informal mentors on research and publication behavior compare to that of formal supervisors? The research presented in this paper addresses the first question.

Our results support the use of our algorithm in similar populations, as the majority of the manually curated supervisions were identified by our algorithm as having "Very High" likelihood of mentorship. To our knowledge, this is the first study to validate an algorithmic-based mentorship relationship calculation amongst researchers. This validated algorithm will open the door to future exploration of the effect of these relationships on other research practices.

These results also increase our confidence in calculating and using our mentorship algorithm at scale within similar fields included in this dataset, often including millions of mentor-mentee pairs. The authors plan to use these manually curated supervisor relationships and algorithmically determined mentorship relationships to evaluate the role mentorship plays in the transmission of research culture, including open science practices such as ethical approval statements, authorship contribution statements, and data and code sharing. This research also serves as the foundation for other types of analysis, such as geographical mobility and impact related to mentorship, amongst others.

Limitations

In the manually curated dataset, we identified a number of PhD-supervisor pairs where the names of either the supervisor or the PhD candidate varied (e.g., middle initial vs. full middle name). This is a valuable data artefact, as it demonstrates the

limitations of manual curation. Conversely, a major limit of algorithmic identification is the inability to distinguish formal mentorship from informal with certainty.

This work is focused on the biomedical field (specifically researcher pairs from medical centers in the Netherlands). There will be fields for which this approach is less well-suited. Future work will explore these limitations.

Despite encouraging results, we acknowledge that the results of this study may only be generalizable within biomedical and clinical fields, and other validation is required in other fields. For example, we foresee our algorithm performance to be affected in fields where authorship behaviours are different than in medical fields (for example, in mathematics where authorship is usually alphabetical, or the humanities where single authorship is more common).

Acknowledgments

The authors acknowledge the valuable contribution of Susan Abunijla and Nicole Hildebrand, who helped collect, curate, and analyse the manually curated dataset (Haven et al., 2023).

References

Haven, T. (2025). It takes two flints to start a fire: A focus group study into PhD supervision for responsible research. Accountability in Research, 1–24. https://doi.org/10.1080/08989621.2025.2457584

Haven, T.L., Abunijela, S., Hildebrand, N. (2023). Biomedical supervisors' role modeling of open science practices. *eLife*, 12:e83484. https://doi.org/10.7554/eLife.83484

Weber-Boer, K., Areia, C., and Taylor, M. (2024). *Is openness heritable: the transmission of integrity from mentor to mentee*. World Conference on Research Integrity. 3 June 2024.