

# **The effect of mentee and mentor gender on scientific productivity of applicants for NIH training fellowships**

Hao Yu<sup>1,2</sup>, Kristine A. Willis<sup>3</sup>, Aviva Litovitz<sup>1,2</sup>, Robert M. Harriman<sup>1,2</sup>, Matthew T. Davis<sup>1,2</sup>,  
Payam Meyer<sup>1,2</sup>, Brad Busse<sup>1,2</sup>, Rebecca A. Meseroll<sup>1,2</sup>, Hashanthi D. Wijayatilake<sup>1,2</sup>,  
Matthew J. Perkins<sup>1,2</sup>, James M. Anderson<sup>2</sup>, and George M. Santangelo<sup>1,2\*</sup>

<sup>1</sup>Office of Portfolio Analysis, National Institutes of Health, Bethesda, MD 20892

<sup>2</sup>Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of Health, Bethesda, MD 20892

<sup>3</sup>National Cancer Institute, National Institutes of Health, Rockville, MD 20850

\*corresponding author

## 1 **Abstract**

2 Several studies have suggested that women in science are less productive than men, and that this gap  
3 contributes to their under-representation in the ranks of senior researchers. However, few studies have  
4 examined the role of mentoring, and in particular mentor gender, on the productivity of female  
5 scientists early in their careers. Such efforts are limited by the difficulties of unambiguously linking  
6 mentees to their mentors and measuring the research productivity resulting from those relationships.  
7 Here we use our novel author disambiguation solution to investigate the role of self-identified gender in  
8 mentorship of 12,932 trainees who either successfully or unsuccessfully applied to the National  
9 Institutes of Health for research fellowships between fiscal years 2011 and 2017, applying a multi-  
10 dimensional framework to assess productivity. We found that, after normalizing for the funding level of  
11 mentors, the productivity of female and male mentees is indistinguishable; it is also independent of the  
12 gender of the mentor, other than in measures of clinical impact, where women mentored by women  
13 outperform other mentee-mentor dyads.

## 14 **Introduction**

15 It is well established that women who pursue careers in biomedical sciences face formidable barriers.  
16 Gender bias may contribute, since for example university professors given identical resumes headed by  
17 either a male or female name are more likely to view male lab technician candidates as competent and  
18 hireable, and to offer male candidates higher salaries [1]. After controlling for prior productivity and  
19 achievements, reviewers of postdoctoral fellowship applications view male candidates as more  
20 meritorious [2, 3]. Women receive harsher teaching evaluations, are less likely to be judged as stars in  
21 their field by reviewers of R01 applications to the National Institutes of Health (NIH), and after  
22 normalizing for a number of confounding factors, including journal of publication, number of authors,  
23 and seniority, their work accrues fewer citations than that of their male colleagues [4-7]. Though women  
24 in the life sciences represent only slightly more than a third of all tenured or tenure-track professors  
25 employed by universities or four-year colleges, they are awarded roughly half of all doctoral degrees [8].  
26 Importantly, women of color face a double bind that hinders their entry into, and retention and  
27 advancement in, biomedical careers [9].

28 While the overall progress and remaining challenges experienced by women in biomedicine have been  
29 widely discussed, the potential effects of mentorship on their career progress have received relatively  
30 little attention. Recently though, a small number of papers have raised the possibility that mentorship,

31 career progress, and gender interact in important ways. Among the very small and elite group of science  
32 faculty who have funding from the Howard Hughes Medical Institute, have been inducted into the  
33 National Academy of Sciences, and/or have won a Nobel prize, men are significantly more likely to  
34 employ other men as postdoctoral fellows; members of the National Academy of Science, which is 85%  
35 male, train 58% of future faculty [10]. In contrast with these data, which suggest that the careers of  
36 female mentees may be disadvantaged by their exclusion from elite male networks, other work suggests  
37 that having a female mentor is an advantage to mentees; for example, a study of roughly 900 PhD  
38 students at a single university found that on average, doctoral candidates studying biology under female  
39 advisors publish approximately 10% more papers, and tend to publish in more influential venues, than  
40 those with male advisors [11]. In the adjacent field of chemistry, a larger study found that women who  
41 chose a female advisor for their doctoral studies were more productive and more likely to go on to  
42 faculty positions than those who chose a male advisor [12]; however, this work systematically excluded  
43 students with Chinese and Korean names because of the difficulty in assigning them algorithmically to a  
44 gender, significantly weakening its conclusions.

45 These inconsistent findings suggest that a more comprehensive analysis of mentorship and gender  
46 might identify factors that either exacerbate or mitigate the barriers faced by women in science. In  
47 addition to their small sample sizes, previous studies have been further limited by inaccuracy of  
48 assigning mentees to their mentors, difficulty verifying the gender of both, and/or limiting the  
49 measurement of research productivity such as publications to the training experience. To overcome  
50 these drawbacks, we studied mentee-mentor relationships among applicants for individual training  
51 fellowships from the NIH. Most applicants for NIH fellowships choose to self-identify gender, and all are  
52 required to identify their mentor(s). Almost all mentors are also NIH-funded investigators who self-  
53 identify gender in their own applications. We used our novel disambiguation method to accurately  
54 document mentee research productivity. Since NIH fellowships cover topics ranging from computational  
55 biology, synthetic chemistry, and biophysics, to epidemiology and clinical psychology, this allowed the  
56 construction of a large, reliable set of self-reported mentee-mentor pairs spanning a wide range of  
57 scientific disciplines, which we analyze here.

58

59

60

## 61 Results

62 We began our analysis with 18,600 applications for individual fellowships (predoctoral mechanisms F30  
63 and F31, and postdoctoral mechanism F32, K01, K08, K23, and K99) submitted to NIH between fiscal  
64 years (FY) 2011 and 2017. Over this time frame, women and men applied for fellowships in similar  
65 numbers and received awards at the same rate; this is true if applications for pre- or post-doctoral  
66 fellowships are considered either together (two leftmost bars, **Figure 1a**) or separately (two leftmost  
67 bars, **Figure 1b** and two leftmost bars, **Figure 1c**). As mentioned above, a unique feature of this dataset  
68 is that applicants communicated to NIH the names of the independent investigators who would act as  
69 their sponsor(s) or mentor(s). NIH requires that any independent investigator named by a fellowship  
70 applicant in either of these capacities (for simplicity, referred to hereafter as a 'mentor') must  
71 demonstrate an understanding of the candidate's training needs, as well as the ability and commitment  
72 to assist in meeting these needs. Mentors must provide a letter of support as a part of the fellowship  
73 application package, and an evaluation of this statement, as well as evidence of successful outcomes for  
74 the mentor's past mentees, are among the explicit criteria that review panels are instructed to use in  
75 their evaluation. Focusing on mentors identified in fellowship applications allowed us to analyze the  
76 relative impact of gender on mentee productivity, though applicants may have access to other  
77 individuals who provide advice and guidance.

78 Consistent with previously published data on the proportion of women in the R01 applicant pool [13],  
79 approximately 30% of mentors in our dataset are female. A subset of applications (17%) were submitted  
80 by mentees who, either in a single application or in two or more different applications, identified both  
81 male and female independent investigators as mentors. To simplify our analysis and avoid double  
82 counting, we removed those applications, which reduced the proportion of female mentors from 30.4%  
83 to 25.6% but had no effect on mentee award rates (third and fourth bars, **Figure 1a**). Mentees of either  
84 gender who list exclusively female mentors, and those who list exclusively male mentors, have identical  
85 award rates (fifth and sixth bars, **Figure 1a**); further dividing applicants into four dyads based on the  
86 gender of both mentee and mentor also fails to identify any gender-based differences in award rates,  
87 regardless of whether pre- and post-doctoral fellowships are considered together (last four bars, **Figure**  
88 **1a**) or separately (last four bars, **Figure 1b, c**).

89 We next asked if the genders of the mentee/mentor dyads influenced mentee research productivity.  
90 Most analyses of productivity are limited to awardees due to the need to rely on the grant number cited  
91 in the resulting publications to link an investigator to his or her papers. One major drawback of this

92 method is that it is unable to assign papers to unsuccessful applicants. Past studies have attempted to  
93 address this problem by creating a restrictive set of criteria to match grant applicants with the papers  
94 they have authored, such as requiring identical first names, last names, and institutional affiliations.  
95 However, such methods will fail to match authors who change names or institutions, or where  
96 mismatches have been introduced as the result of typos, inconsistent spellings, or the inconsistent use  
97 of a middle initial. To address this problem, we developed a disambiguation solution that used article-  
98 level metadata [14-16] to assign 24.5M unique papers from the PubMed database to 16.0M unique  
99 author names, then used a novel neural network model trained on ORCID identifiers to determine  
100 whether author-publication pairs refer to variant representations of the same person (see **Methods** for  
101 details). For example, our model (**Figure 2**) can determine whether hypothetical records listing Jane  
102 Smith and Jane M. Smith were the same person, or two different people, based on variables that include  
103 institutional affiliation, co-authorship, and article-affiliated Medical Subject Heading (MeSH) terms. We  
104 then matched unambiguously identified author and applicant names. Importantly, the model does not  
105 require last names to match, so women who change their name can be successfully merged. We used  
106 this method to reduce the 16.0M unique author names to 13.3M disambiguated people; the F1 score for  
107 people with at least one NIH application is 0.945, indicating both high precision and high recall.  
108 Disambiguation of the people associated with the fellowship applications in our dataset indicates that  
109 we have captured a large fraction of NIH trainees and their mentors, since the papers of these  
110 unambiguously identified applicants together amount to 57.7% of all publications since 2011 that cite  
111 NIH grant support.

112 Our data show that male-male dyads, and more specifically, male-male post-doctoral dyads, have more  
113 publications prior to the time of their first application (set at time = 0, **Figure 3a-c**). Male-male post-  
114 doctoral dyads also start out with more papers in the top decile of Relative Citation Ratio (RCR) values  
115 (**Figure 3d-f**); RCR is an article-level, field- and time-normalized measure of scholarly influence [17]. This  
116 early advantage in publication and citation metrics is maintained for years after the time of first  
117 application (**Figure 3a-f**), consistent with a hysteretic process. Beginning around the time of application,  
118 the number of highly influential (top decile) papers authored by female-female post-doctoral dyads  
119 begins to diverge from the number published by male-female and female-male dyads (**Figure 3e**). Eight  
120 years after their first application, this difference is outside the confidence interval; it should be noted  
121 that this is roughly correlated with the point at which women are more likely to leave academia [18].  
122 However, median RCR values are indistinguishable for all four types of mentee-mentor dyads over the  
123 entire eighteen-year time frame of our analysis (**Figure 3g-i**), and applicants for pre-doctoral fellowships

124 exhibit no meaningful differences outside the 95% confidence interval for any of these productivity  
125 measures, regardless of mentee or mentor gender, among applicants for pre-doctoral fellowships  
126 **(Figure 3c, f, i)**.

127 Publication and citation metrics are the typical, but not the only, measure of scholarly contribution to  
128 scientific progress [19]. Biomedical research also leads to patentable inventions/technological (tech)  
129 impact, measured by the citation of publications by patents, and clinical impact, measured by the  
130 citations of publications by clinical trials and guidelines. Female-female dyads appear to have less tech  
131 impact **(Figure 3j-l)** and more clinical impact **(Figure 3m-o)**, in both pre- and post-doctoral applicant  
132 populations; since these forms of citations are slower to accrue than citations to peer-reviewed  
133 publications, censoring (the absence of hypothetical future citations; [20]) makes it difficult to  
134 determine whether these differences are maintained. However, clinical impact can also be measured  
135 with APT (Approximate Potential to Translate) scores, which are machine-learning based predictions of  
136 future clinical citations [21]; these predictions are particularly useful because they are less subject to  
137 censoring. Both before and after applying for a fellowship, APT scores are highest for female-female  
138 dyads. Together, the greater number of clinical citations **(Figure 3m-o)** and higher APT scores **(Figure**  
139 **3p-r)** indicate that this dyad generates the highest level of clinical impact.

140 Although small dollar amounts may sometimes be budgeted for a training course or similar expense, NIH  
141 fellowships generally provide salary only. The productivity of the mentee is therefore heavily reliant on  
142 the amount of research funds available to the mentor. Interestingly, for each of the four mentee-mentor  
143 dyad categories, mentors of post-doctoral fellowship applicants have a higher level of median total  
144 costs, adjusted for inflation to 2019 dollars by using the Biomedical Research and Development Price  
145 Index (BRDPI; **Figure 4**). This is unlikely to be explained by the institutional affiliation of mentors, since  
146 pre-doctoral and post-doctoral applications distribute similarly across institutions receiving widely  
147 different levels of NIH support, regardless of whether aggregate inflation-adjusted funding or dollars per  
148 investigator are considered (see **Supplemental Data**). Our data are also consistent with previously  
149 published work [13] showing that on average, women hold fewer awards **(Supplemental Data)** and have  
150 fewer research dollars than men **(Figure 4)**. This disadvantage does not influence the chance of winning  
151 a fellowship award for applicants with female mentors **(Figure 1a)**, but might have an impact on mentee  
152 productivity. We therefore normalized the number of publications, scholarly influence, tech impact, and  
153 clinical impact of mentees to the total amount of NIH funding held by their mentors.

154 Strikingly, adjusting for the funding available to mentors eliminates the advantage of male-male  
155 post-doctoral dyads in number of publications (**Figure 5a, b**), scholarly influence (**Figure 5d, e**), and tech  
156 impact (**Figure 5j, k**). More than simply closing the gap, after adjusting for funding, female-female  
157 post-doctoral dyads have a slightly higher number of papers immediately prior to the time of their first  
158 application for a fellowship (**Figure 5b**). They also have greater clinical impact (**Figure 5m, n**) and higher  
159 APT scores (**Figure 5p, q**). Female-male dyads also have higher APT scores than male-female or  
160 male-male dyads (**Figure 5q**), suggesting that female mentees in general are more successful at  
161 producing clinical impact. Finally, the median RCR values of post-doctoral dyads are largely  
162 indistinguishable (**Figure 5g, h**), and the high signal to noise ratio for the four pre-doctoral dyads could in  
163 part be responsible for the failure to detect differences in those funding-adjusted metrics (**Figure 5c, f, l,**  
164 **l, o, r**).

165

## 166 **Discussion**

167 Previous reports have found that two prominent and potentially interrelated barriers faced by female  
168 scientists are the postdoc-to-faculty transition to independence and the greater difficulty in achieving  
169 higher levels of influence via citation of their research outputs. Our analysis of the productivity to date  
170 of FY11-FY17 applicants for individual NIH fellowships confirmed the latter observation: female mentees  
171 produce fewer papers in the top decile of RCR values, as well as fewer papers overall. However,  
172 normalizing to mentor funding levels eliminates both of those gaps. We also found that the median RCR  
173 for female and male mentees are indistinguishable, and male mentees, especially when they are paired  
174 with male mentors, have lower clinical impact. Taken together, the data indicate that, if there is any  
175 gender-based difference in mentorship at all, it manifests as an advantage of female mentoring of  
176 female mentees in producing clinically relevant research.

177 Our data suggest that the initial appearance of lower productivity of female mentees might be a direct  
178 result of the funding gap between independent female and male investigators that has already been  
179 noted in the literature [13]. In addition to confirming that gap, we found that pre-doctoral mentors in  
180 each of the four dyad categories are less well funded than post-doctoral mentors (**Figure 4**). Since  
181 institutional affiliation makes at most a minor contribution to this difference, the funding gap between  
182 pre- and post-doctoral mentors may indicate a preference on the part of individual independent

183 investigators. If so, less well-funded scientists are more willing or able to sponsor individual pre-doctoral  
184 applications, while well-funded investigators prefer to sponsor post-doctoral fellowships.

185 Among the benefits of data-driven decision making is the potential to inform the pursuit of desired goals  
186 and to dispel misconceptions that might result in unhelpful guidance or policy decisions. Large-scale  
187 metascience uses many different types of information towards this end (e.g. grant applications,  
188 publications, patents, clinical trials), and requires the careful generation of accurate linkages among  
189 these disparate sources of data. Author names, applicant names, affiliations and other metadata are  
190 presented in many different variations (e.g. John D Smith vs. JD Smith vs. John Smith); identification of a  
191 single person, accurately linked to their full publication and application record, requires a rigorous  
192 method of name disambiguation. Without such methodology in hand, attempts at meaningful analysis  
193 are plagued by incomplete records and/or multi-counting errors. Our newly developed high-  
194 performance Artificial Intelligence/Machine Learning (AI/ML)-based disambiguation method improves  
195 on previous attempts [22-24] to address this problem, achieving an F1 score of 0.945 and allowing us to  
196 clean the large datasets of PubMed author names and NIH grant applicants then integrate that  
197 information to create the dataset used in this study.

198 Our disambiguation solution also allows us to overcome several challenges created by the common  
199 practice of linking outputs to funded grants. As we have shown here, it allows publications to be linked  
200 to applicants who have not yet (or ever) received an NIH award. It also allows grants to be linked to  
201 papers if the authors failed to cite their award or cited it in a non-specific way, such as acknowledging  
202 support from the NIH without providing an identifying number. Finally, person-level links solve the  
203 problem posed by publications in journals that lack an acknowledgements section.

204 We have relied on these person-level links, and the availability of self-identified information provided by  
205 NIH training fellowship applicants, to investigate the role of mentee and mentor gender on mentee  
206 productivity. This has removed the inevitable errors associated with a reliance on tenuous assumptions  
207 in defining mentee-mentor relationships. Of course, not all mentoring occurs in the context of a formal  
208 relationship with one or more doctoral or post-doctoral advisors. A variety of scenarios, ranging from  
209 structured, regular meetings with thesis committee members to informal, transient interactions in  
210 which a more senior scientist gives technical or career advice to a junior colleague, may be interpreted  
211 as mentorship. While often critically important, methods capable of fully capturing these networks must  
212 go beyond a simple analysis of co-authorship [25].



213 By identifying the outsized clinical impact of female mentees, which is a previously unappreciated  
214 contribution to biomedicine, our analysis demonstrates the value of using a multifaceted framework for  
215 measuring research productivity. The development of additional metrics that measure other factors  
216 supporting scientific progress, such as rigor/reproducibility and data sharing, should further improve  
217 analyses that can be effective in informing and guiding policy [19]. We are also now poised to go beyond  
218 the current analysis to investigate the role of specific mentor and mentee characteristics (e.g. career  
219 stage, mentoring track record, affiliation, previous publication record) in promoting the successful  
220 transition to productive independent careers. This type of information has the potential to inform  
221 guidance and policy-making that provide robust support for the scientific enterprise, as former mentees  
222 in turn train the next generation of scientists and perpetuate the cycle of progress that has now proved  
223 its worth as a means of advancing knowledge that improves human health.

224

225

## 226 **Methods**

227

### 228 **Author and applicant name disambiguation methodology**

229 Author name disambiguation was carried out in two stages (**Figure 2**). The first stage was  
230 disambiguation of PubMed authors and deduplication of grant applicants. The second stage of the  
231 process involved matching and merging the disambiguated PubMed author records with the  
232 deduplicated applicant records to generate disambiguated author profiles that contain specific linkages  
233 to a person's publications and NIH funding.

234

235 **PubMed author disambiguation:** PubMed author disambiguation was performed in a similar fashion as  
236 described previously [14-16]. Essentially, authors were disambiguated in the same first initial last name  
237 (FILN) block using hierarchical agglomerative clustering algorithm based on pairwise similarity.

238

239 During preprocessing, each individual author on a publication was listed separately to form author-  
240 publication entries. Author and publication metadata (author name, author affiliation, coauthor names,  
241 location, journal name, linked grant numbers, clinical trials, and patents), citations, and content features  
242 (MeSH keywords, title tokens, broad subject terms) were collected or extracted and stored in author-  
243 publication entries. The author-publication entries were then grouped based on the author FILN.

244

245 In order to cluster the author-publication entries within each FILN block to form disambiguated author  
246 records, a fully connected neural network with two hidden layers was trained as a binary classifier to  
247 determine if two author-publication entries belong to the same author using the attributes stored in  
248 author-publication entries. The probability output of the model was considered to reflect the similarity  
249 of the input author-publication entries. Training and test datasets were generated using ORCID  
250 profiles. Hierarchical agglomerative clustering algorithm [26] was used to cluster the author-publication  
251 entries using similarity scores generated by the trained neural network model. The resulting  
252 disambiguated author records were used for the person record merging stage below.

253

254 **Grant applicant deduplication:** In parallel to PubMed author disambiguation, NIH grant applicants were  
255 deduplicated by deduplicating their Principal Investigator IDs (PIIDs). Ideally, PIIDs should map to  
256 applicants in a one-to-one relationship. However, we estimated that 10-15% of all PIIDs were duplicates.  
257 These PIIDs and their associated applications needed to be merged before linking them to the  
258 disambiguated PubMed authors. PIID deduplication was performed as following: In a preprocessing  
259 step, applications were unwound on all PIs listed on the application. The following information was then  
260 extracted: PIID, PubMed IDs (PMIDs) linked from the NIH Scientific Publication Information Retrieval and  
261 Evaluation System (SPIRES; we collected links of match case 3, 4, and 5  
262 and then further screened them with our name matching algorithm), PMIDs resolved from grant  
263 applicants' biosketches, and metadata such as applicant name and grant number. The unwound  
264 applications and all the extracted data associated with the application were aggregated to PIIDs.

265

266 In the deduplication step, PIIDs were combined if they met one of two criteria: 1) Consecutive PIIDs with  
267 matched applicant names; 2) Same grant number with matched applicant names. A few hundred  
268 PIIDs were also manually curated and used for deduplication. The resulting applicant records were used  
269 for the person record merging stage below.

270

271 **Person record merging:** The disambiguated PubMed authors and deduplicated grant applicants from  
272 the first stages were merged based on name matching and publication overlap for records that shared  
273 the same FILN. However, we observed that a small percentage of applicants had name variants with  
274 different FILNs caused by various reasons. The most common reasons include typos in the last names,  
275 re-arranged first/middle names, and name change due to marriage. As a result, their publications were

276 disambiguated in different FILN blocks. To allow combining over-split records of the same applicant from  
277 different FILN blocks, we merged records across FILN if they were associated with the same PIID.

278

279 **Enrichment:** To facilitate downstream analyses, disambiguated person records were enriched by  
280 generating the best name for the person using all the name variants that appeared in their publications  
281 and grant applications, and populating any useful data for each publication and grant application in the  
282 disambiguated record.

283

284 **Assessing performance of the disambiguation method:** We developed an evaluation method using  
285 ORCID profiles which we treated as ground truth. ORCID authors were mapped to disambiguated author  
286 records by name matching and publication overlap. Precision and recall were computed accordingly. If  
287 one ORCID author was mapped to more than one disambiguated author records, the record that had  
288 the highest F1 was designated as the disambiguated author for that ORCID author. Because ORCID data  
289 are largely incomplete in terms of their publication records, we only considered PMIDs that could be  
290 found or resolved in ORCID profiles for precision calculation.

291

292 Since only about 1.8% of all disambiguated authors are associated with NIH grants and this study  
293 concerns grant applicants, we evaluated the performance of our disambiguation process in two groups  
294 of authors: those with grants and those without grants. Micro-precision, micro-recall, and micro-F1 were  
295 computed for random samples of these two groups. For one experiment, 7000-7500 samples for authors  
296 without grants and 450-500 samples for authors with grants were randomly selected to compute the  
297 performance metrics for each group. This experiment was repeated five times and the metrics were  
298 compared using unpaired Student's t-test.

299

300 We found no statistically significant difference between the precision of disambiguated authors with  
301 grants and without grants ( $0.985 \pm 0.003$  for authors without grants vs.  $0.984 \pm 0.007$  for authors with  
302 grants,  $p = 0.86$ ). In contrast, disambiguated authors with grants had significantly higher recall than  
303 disambiguated authors without grants ( $0.783 \pm 0.008$  for authors without grants vs.  $0.908 \pm 0.016$  for  
304 authors with grants,  $p < 0.0001$ ). F1, as a result, showed the same trend as recall ( $0.872 \pm 0.006$  for  
305 authors without grants vs  $0.945 \pm 0.010$  for authors with grants,  $p < 0.0001$ ).

306

307

## 308 **Identifying fellowship applications**

309 Mentees were defined as NIH grant applicants who had applied for pre-doctoral or post-doctoral  
310 training fellowships (F30 and F31 for pre-doctoral; F32, K01, K08, K23, and K99 for post-doctoral) in fiscal  
311 years 2011-2017. 2011 was chosen as the initial analysis year since it was the first year fellowship  
312 applicants were able to self-identify their mentors/sponsors as a “Key Person” during the application  
313 process. This official record of self-identified mentee-mentor relationships was considered critical in  
314 identifying valid and accurate mentee/mentor dyads. 2017 was chosen as the final year of analysis to  
315 give the mentees sufficient time to produce publications after their fellowship  
316 application, while allowing time for reliable publication-related metrics to subsequently accrue.

317  
318 A total of 57,425 applications from 37,918 mentee applicants were first identified. The  
319 following selection criteria was then applied to the dataset: 1. mentors should have a PIID in the Key  
320 Person field or the name/organization search should return a single profile match (for mentors that did  
321 not have a PIID associated with the mentee application, mentor PIIDs were added from the matched  
322 profile; n=35,999 applications from 21,856 applicants were excluded); 2. mentors should have  
323 only one PIID (those with zero or >1 were excluded; n=3,902 applications from 2,853 applicants);  
324 3) mentees should have no more than one PIID (those with >1 PIIDs were excluded; n=414 applications  
325 and 277 applicants).

326  
327 The selection criteria above yielded a dataset of 18,600 unique applications from 12,932  
328 mentee applicants. All these mentees were initially analyzed, including those who had mixed-gender  
329 mentors (Figure 1a). To simplify the analysis, avoid double counting, and avoid conflating effects of  
330 mixed (both gender) mentors, 3,215 applications (17%) from 1,858 mentee applicants who had a  
331 combination of female and male mentors were eliminated for the subsequent analysis. This yielded a  
332 final dataset of n=15,386 applications from 11,074 mentee applicants with single-gender (i.e. female-  
333 only or male-only) mentors.

334  
335 Fellowship award rates are presented as a percentage of total awards.

336

337

338

339

## 340 **Assignment of gender**

341 Gender is self-identified by NIH applicants during the application process. Of the mentee/mentor dyads  
342 in our dataset, 94.1% of mentees and 90.09% of mentors had self-identified records of  
343 gender. Genderize (genderize.io) was used to assign gender to those without self-  
344 identified gender information, as done in previous studies [12, 27, 28]. The agreement between self-  
345 identified gender and Genderize was 98%, allowing confidence in using Genderize to populate the small  
346 portion of non-self-identified records.

347  
348 Statistical significance was calculated using Fisher Tests compared to the rest of the population.

349  
350

## 351 **Analysis of mentee productivity over time**

352 Productivity metrics were analyzed for each mentee, and time-shifted for each mentee-  
353 mentor dyad such that year 0 represented the year of the mentee's first fellowship grant application  
354 (regardless of awarded status). Mentees were split between pre-doctoral mentees who  
355 received only F30 or F31 awards over the course of the analysis period (right columns in Figures 3 and  
356 5), and post-doctoral mentees who received any other type of award listed in the section above (middle  
357 columns in Figures 3 and 5). n=138 mentees had both pre- and post-doctoral fellowship applications and  
358 were counted in the post-doctoral group.

359  
360 In each case four subpopulations of mentee/mentor dyads were measured independently, based on  
361 their respective gender: male mentee/male mentor (MM), male mentee/female mentor (MF),  
362 female mentee/male mentor (FM), female mentee/female mentor (FF). As noted above, mentees with  
363 multiple mentors of both genders were excluded from the analysis. Shaded regions indicate 95%  
364 confidence intervals, determined through bootstrap analysis.

365  
366 Six research productivity metrics were examined, per mentee, per year, before and after the first  
367 fellowship application (**Figures 3 and 5**): Mean number of publications, mean number of high-  
368 influence publications (defined as having an RCR [17] in the top 10% of all NIH publications), median RCR  
369 of all publications, technological impact (fraction of publications which have been cited by at least one  
370 US patent submission), clinical impact (fraction of publications which have been cited by at least one  
371 clinical trial or guideline), Approximate Potential to Translate (APT, [21]) score (predicted clinical impact

372 fraction based on citation trends). Dyads with no mentee publications (n=1,999, or 11.0%) were  
373 excluded from the analysis. Further, to facilitate comparisons with funding information (see below),  
374 dyads with no funding information (e.g. those with mentors entirely funded outside of NIH) were also  
375 excluded from the normalized analysis (n=3,236, or 20.0%).

376

377 The same six mentee productivity metrics noted above were subsequently re-analyzed to normalize  
378 to mentor funding levels (**Figure 5**). Mentor funding levels were defined as the amount of NIH mentor  
379 funding (averaged between mentors if multiple are linked to the same mentee) in the fiscal year of the  
380 mentee's training application, adjusted for inflation to BRDPI 2019 dollars. Time points before the first  
381 training application use the first application's value, and time points after use the most recent  
382 application. Each mentee's productivity metric was divided by their mentors' funding level, and  
383 normalized to productivity per million PI dollars. For example, mean publications per year presented in  
384 Figure 3 are presented as mean publications per million dollars per year in Figure 5.

385

## 386 Figure legends

387 **Figure 1. Fellowship applications and awards by mentee and mentor gender.** NIH fellowship  
388 applications (pre-doctoral: F30 and F31; post-doctoral: F32, K01, K08, K23, and K99) submitted by  
389 females (F) and males (M) in FY11 through FY17. Light blue and orange bars represent unawarded and  
390 awarded applications, respectively; black dots indicate award rates (secondary Y axis). **(a)** The first two  
391 bars represent applications from all mentees with single-gender or mixed-gender mentors; the third and  
392 fourth bars represent applications from mentees with single-gender mentors, separated by mentee  
393 gender; the fifth and sixth bars represent applications from mentees with single-gender mentors,  
394 separated by mentor gender. The remaining four bars show the data analyzed by mentee-mentor dyads  
395 for mentees with single-gender mentors. Mentee gender is presented first, mentor gender second, e.g.,  
396 FM = female mentees with male mentors. **(b and c)** Same data as in (a), analyzed by pre- or post-  
397 doctoral career stage of the mentee applicant, respectively. The first two bars represent applications  
398 from all mentees with single-gender or mixed-gender mentors. The remaining four bars show the data  
399 analyzed by mentee-mentor dyads for mentees with single-gender mentors. There is no statistically  
400 significant difference in award rate for any group.

401

402 **Figure 2. Disambiguating authors and NIH applicants.** Graphical representation of the workflow used to  
403 disambiguate unique author names and link publications and NIH applications to specific authors and  
404 applicants. The process began by assigning 24,453,076 unique publications to 15,985,142 unique author  
405 names, and resulted in 13,324,796 disambiguated people. A fully connected neural network with two  
406 hidden layers, trained on a series of author and publication features, generated pairwise author-  
407 publication entry similarity scores (left side of the illustration; see Methods). Those similarity scores  
408 were used by a hierarchical agglomerative clustering algorithm to merge author-publication entries,  
409 resulting in disambiguated PubMed author records (blue oval). In parallel, preprocessing of NIH  
410 applications and applicant deduplication (see Methods) generated applicant records (red oval).  
411 Subsequent matching and merging of disambiguated PubMed author records with deduplicated  
412 applicant records generated disambiguated author profiles that contain specific linkages to a person's  
413 publications and NIH applications. The disambiguated person records were enriched with the person's  
414 metadata and data for each publication and grant application to facilitate downstream analyses (grey  
415 oval).

416

417 **Figure 3. Mentee productivity over time.** Six different measures of mentee research productivity,  
418 shown per mentee per year, where the first fellowship application is set to time=0 (vertical dashed grey  
419 line): **a-c**, mean number of publications; **d-f**, mean number of high-influence publications (defined as  
420 having an Relative Citation Ratio (RCR) value in the top decile); **g-i**, median RCR; **j-l**, technological impact;  
421 **m-o**, clinical impact; **p-r**, mean Approximate Potential to Translate (APT) score (see Methods for  
422 additional details). First column, all (both post- and pre-doctoral fellowship applicants; second column,  
423 post-doctoral applicants only; third column, pre-doctoral applicants only). Shaded regions indicate 95%  
424 confidence intervals, determined via bootstrap analysis. The dyads are annotated with mentee gender  
425 first, mentor gender second (e.g., FM – female mentees with male mentors)

426

427 **Figure 4. Mentor funding over time.** Mentor funding levels per mentee for gender-based mentee-  
428 mentor dyads in the six years following a mentee's first fellowship application (set to time=0). Left  
429 graph, all (both post- and pre-doctoral fellowship applicants; middle graph, post-doctoral applicants  
430 only; right graph, pre-doctoral applicants only). Shaded regions indicate 95% confidence intervals,  
431 determined via bootstrap analysis. The dyads are annotated with mentee gender first, mentor gender  
432 second (e.g., FM – female mentees with male mentors)

433

434 **Figure. 5. Funding-normalized mentee productivity over time.** In order to account for the effect of  
435 differing mentor resources on mentee productivity, Figure 3 data was re-analyzed to normalize for  
436 mentor funding levels. As in Figure 3, the six different measures of mentee research productivity are  
437 presented for gender-based mentee/mentor dyads, normalized per million mentor Principal Investigator  
438 dollars: Mean number of publications (**a-c**), mean number of high-influence publications (**d-f**), median  
439 RCR of all publications (**g-i**), technological impact (**j-l**), clinical impact (**m-o**), and APT score (**p-r**). Data are  
440 analyzed per mentee per year, grouped by post- or pre-doctoral career stage of the mentee applicant  
441 (columns left to right), and are presented across a time scale of ten years before and after a mentee's  
442 first post-doctoral (middle column) or pre-doctoral (right column) fellowship application. The first  
443 fellowship application is indicated at time=0 (vertical dashed grey line). Shaded regions indicate 95%  
444 confidence intervals, determined via bootstrap analysis. When a mentee is linked to multiple mentors,  
445 their funds are averaged. See Methods for more details. The dyads are annotated with mentee gender  
446 first, mentor gender second (e.g., FM – female mentees with male mentors)

447



448 **References**

- 449 1. Moss-Racusin, C.A., et al., *Science faculty's subtle gender biases favor male students*.  
450 Proceedings of the National Academy of Sciences, 2012. **109**(41): p. 16474.
- 451 2. Holst, S. and S. Hägg, *Positive bias for European men in peer reviewed applications for faculty*  
452 *position at Karolinska Institutet*. F1000Res, 2017. **6**: p. 2145.
- 453 3. Wennerås, C. and A. Wold, *Nepotism and sexism in peer-review*. Nature, 1997. **387**(6631): p.  
454 341-343.
- 455 4. Dworkin, J.D., et al., *The extent and drivers of gender imbalance in neuroscience reference lists*.  
456 Nature Neuroscience, 2020. **23**(8): p. 918-926.
- 457 5. Kaatz, A., et al., *Analysis of National Institutes of Health R01 application critiques, impact, and*  
458 *criteria scores: does the sex of the principal investigator make a difference?* Academic Medicine,  
459 2016. **91**(8): p. 1080-1088.
- 460 6. MacNell, L., A. Driscoll, and A.N. Hunt, *What's in a Name: Exposing Gender Bias in Student*  
461 *Ratings of Teaching*. Innovative Higher Education, 2015. **40**(4): p. 291-303.
- 462 7. Astegiano, J., E. Sebastián-González, and C.d.T. Castanho, *Unravelling the gender productivity*  
463 *gap in science: a meta-analytical review*. Royal Society Open Science, 2019. **6**(6): p. 181566.
- 464 8. *Women, Minorities, and Persons with Disabilities in Science and Engineering*. National Center for  
465 Science and Engineering Statistics (NCSES) 2019.
- 466 9. Malcom, L. and S. Malcom, *The double bind: The next generation*. Harvard Educational Review,  
467 2011. **81**(2): p. 162-172.
- 468 10. Sheltzer, J.M. and J.C. Smith, *Elite male faculty in the life sciences employ fewer women*.  
469 Proceedings of the National Academy of Sciences, 2014. **111**(28): p. 10107-10112.
- 470 11. Pezzoni, M., et al., *Gender and the publication output of graduate students: A case study*. PLoS  
471 One, 2016. **11**(1): p. e0145146.
- 472 12. Gaule, P. and M. Piacentini, *An advisor like me? Advisor gender and post-graduate careers in*  
473 *science*. Research Policy, 2018. **47**(4): p. 805-813.
- 474 13. Hechtman, L.A., et al., *NIH funding longevity by gender*. Proceedings of the National Academy of  
475 Sciences, 2018. **115**(31): p. 7943-7948.
- 476 14. Cota, R.G., et al., *An unsupervised heuristic-based hierarchical method for name disambiguation*  
477 *in bibliographic citations*. Journal of the American Society for Information Science and  
478 Technology, 2010. **61**(9): p. 1853-1870.
- 479 15. Song, Y., et al. *Efficient topic-based unsupervised name disambiguation*. in *Proceedings of the*  
480 *7th ACM/IEEE-CS joint conference on Digital libraries*. 2007.
- 481 16. Song, Y., et al., *Generative models for name disambiguation*, in *Proceedings of the 16th*  
482 *international conference on World Wide Web*. 2007, Association for Computing Machinery:  
483 Banff, Alberta, Canada. p. 1163–1164.
- 484 17. Hutchins, B.I., et al., *Relative Citation Ratio (RCR): A new metric that uses citation rates to*  
485 *measure influence at the article level*. PLoS biology, 2016. **14**(9): p. e1002541.
- 486 18. Lerchenmueller, M.J. and O. Sorenson, *The gender gap in early career transitions in the life*  
487 *sciences*. Research Policy, 2018. **47**(6): p. 1007-1017.
- 488 19. Santangelo, G.M., *Article-level assessment of influence and translation in biomedical research*.  
489 Molecular biology of the cell, 2017. **28**(11): p. 1401-1408.
- 490 20. *The SAGE Encyclopedia of Social Science Research Methods*. 2004.
- 491 21. Hutchins, B.I., et al., *Predicting translational progress in biomedical research*. PLOS Biology,  
492 2019. **17**(10): p. e3000416.

493 22. Liu, W., et al., *Author name disambiguation for PubMed*. Journal of the Association for  
494 Information Science and Technology, 2014. **65**(4): p. 765-781.

495 23. Song, M., E.H.-J. Kim, and H.J. Kim, *Exploring author name disambiguation on PubMed-scale*.  
496 Journal of Informetrics, 2015. **9**(4): p. 924-941.

497 24. Torvik, V.I. and N.R. Smalheiser, *Author name disambiguation in MEDLINE*. ACM Transactions on  
498 Knowledge Discovery from Data (TKDD), 2009. **3**(3): p. 1-29.

499 25. Kirchmeyer, C., *The effects of mentoring on academic careers over time: Testing performance  
500 and political perspectives*. Human Relations, 2005. **58**(5): p. 637-660.

501 26. Eppstein, D., *Fast hierarchical clustering and other applications of dynamic closest pairs*. ACM J.  
502 Exp. Algorithmics, 2001. **5**: p. 1-es.

503 27. Harris, J.K., et al., *Diversify the syllabi: Underrepresentation of female authors in college course  
504 readings*. PLOS ONE, 2020. **15**(10): p. e0239012.

505 28. Xu, R.F., N.H. Varady, and A.F. Chen, *Trends in Gender Disparities in Authorship of Arthroplasty  
506 Research*. JBJS, 2020. **102**(23): p. e131.

507

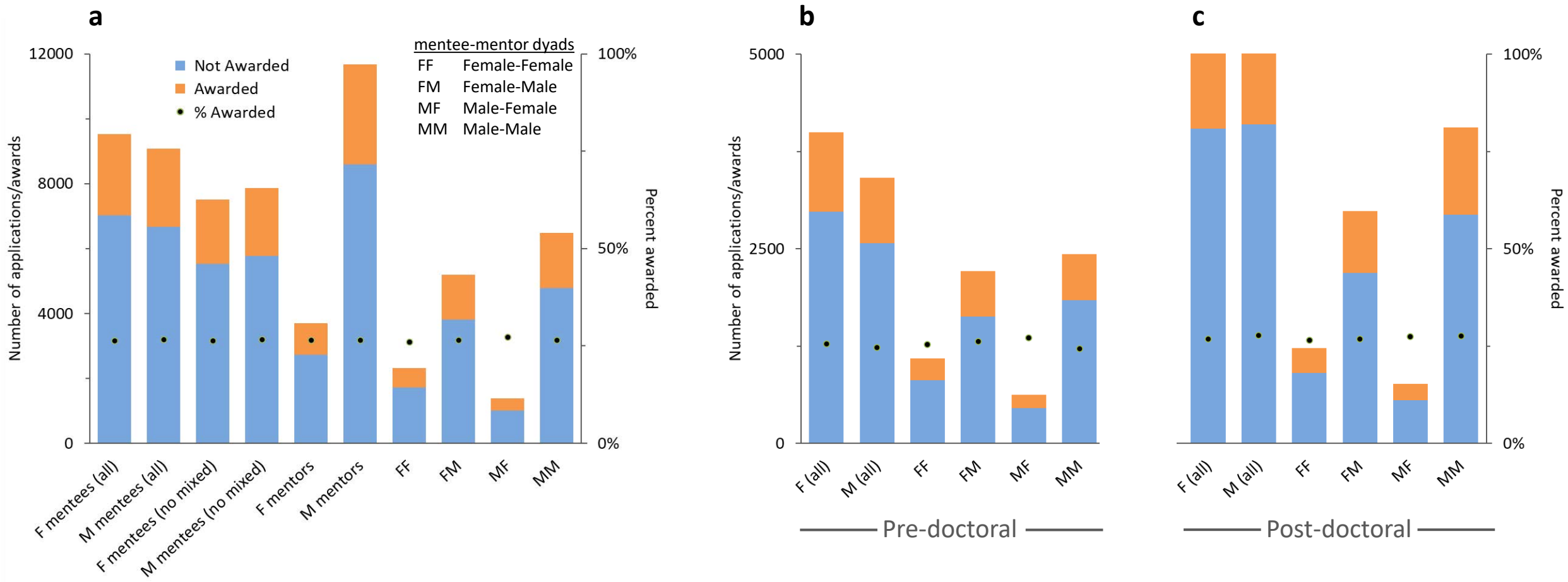


Figure 1

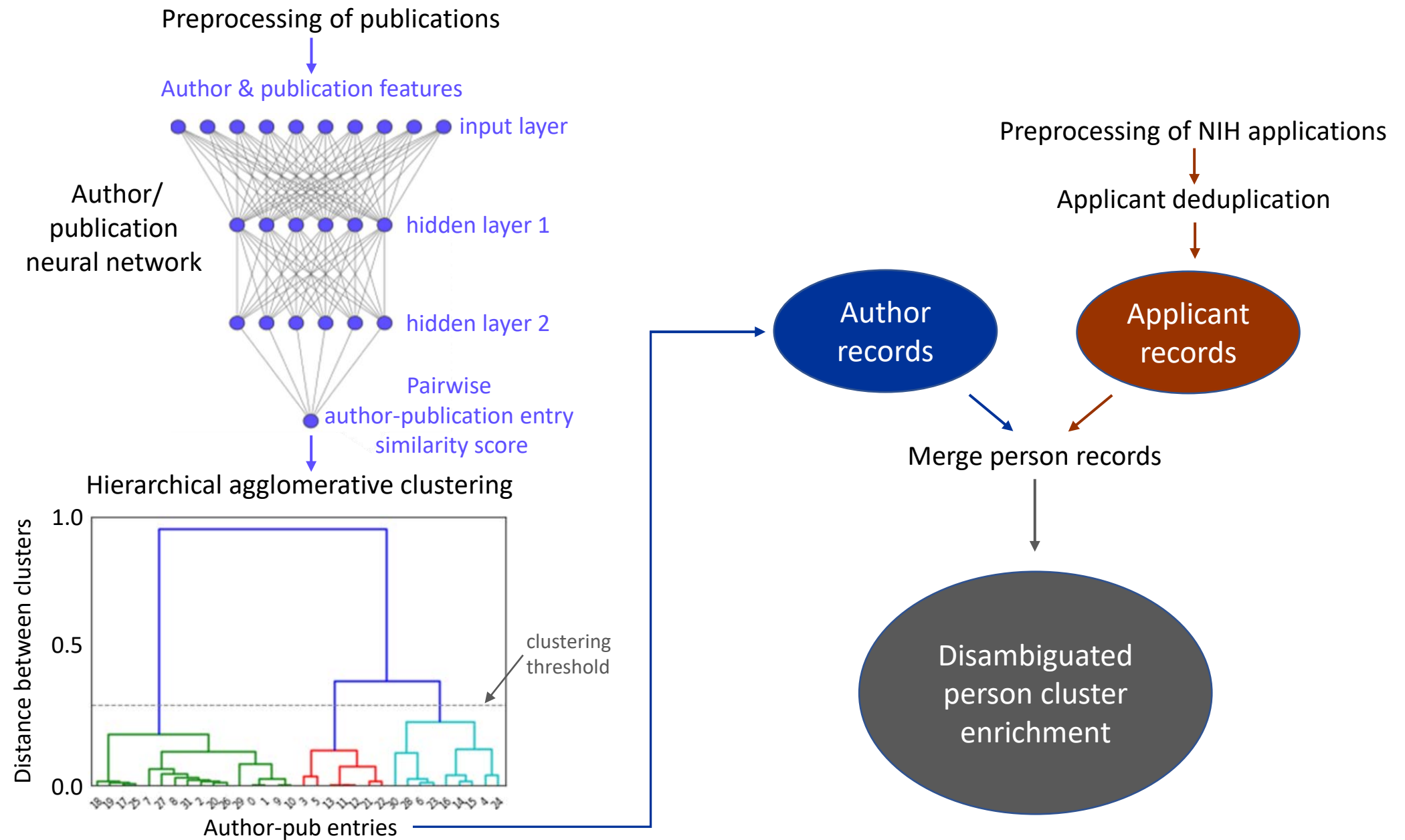


Figure 2

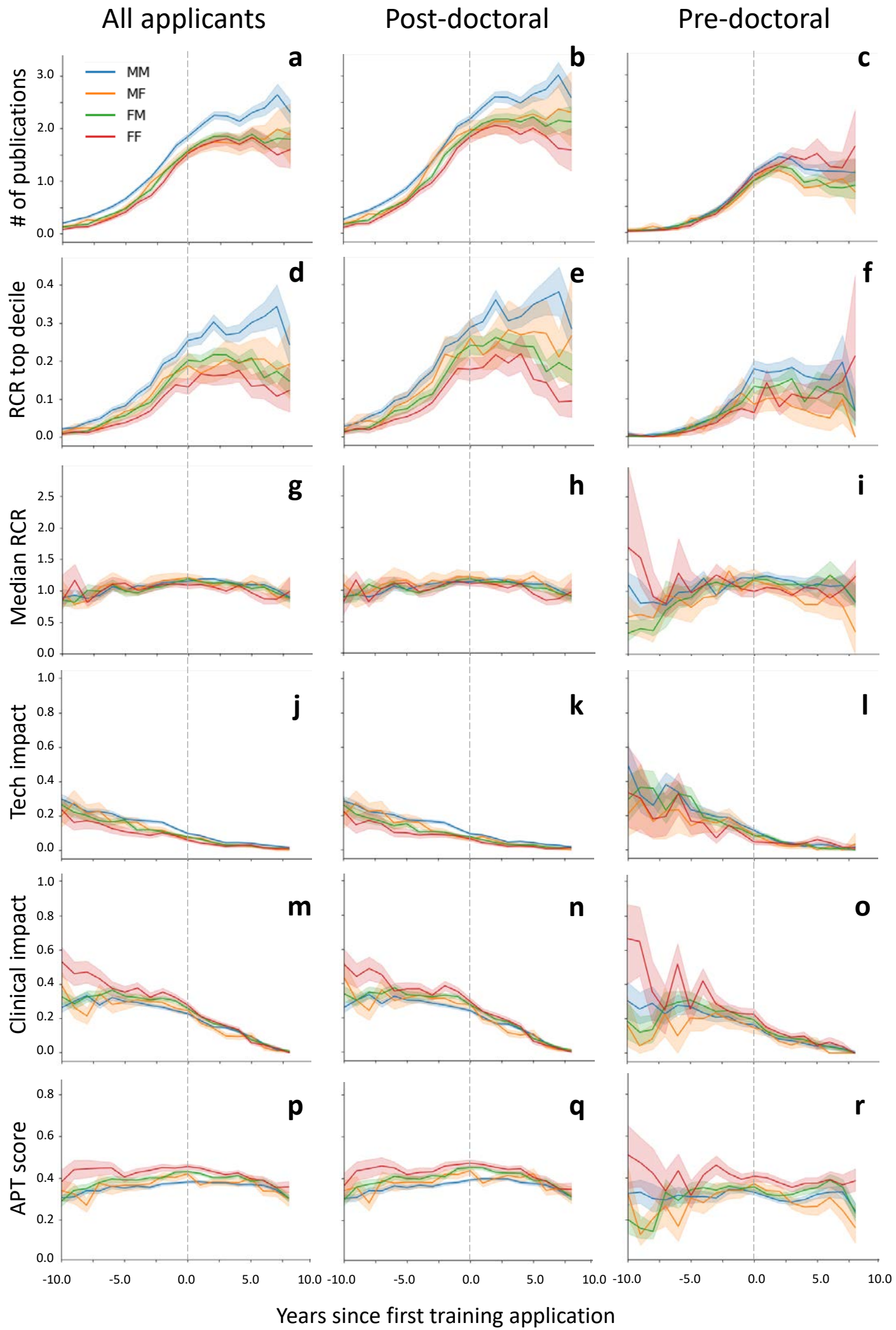


Figure 3

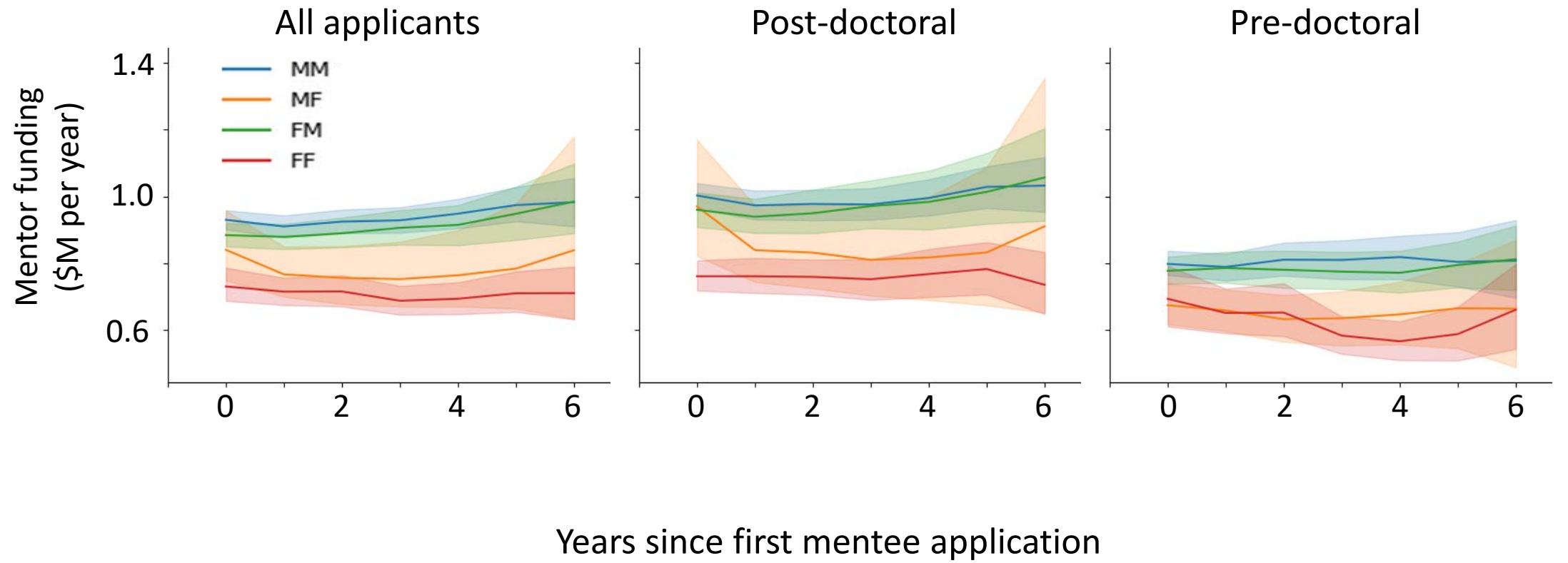


Figure 4

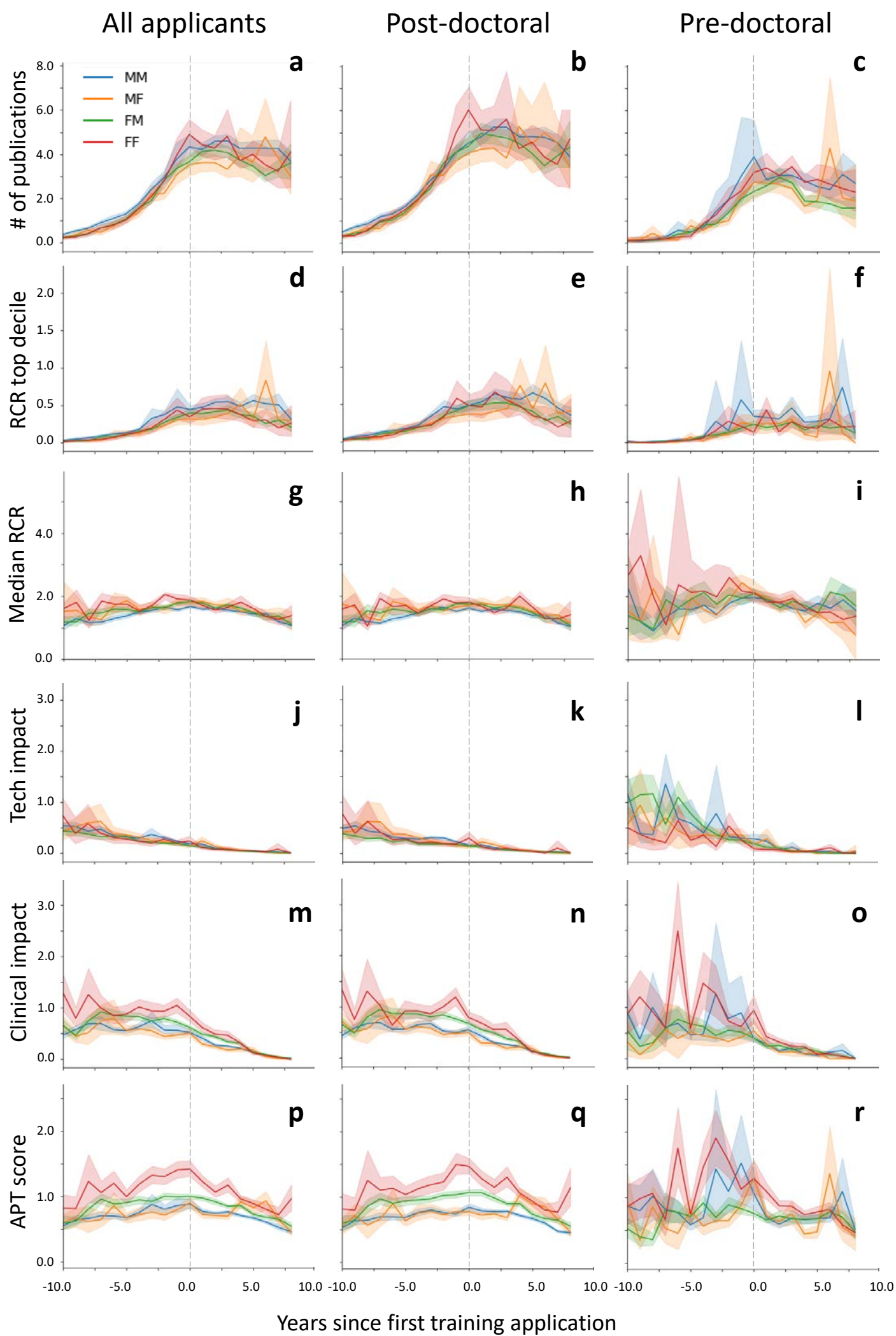
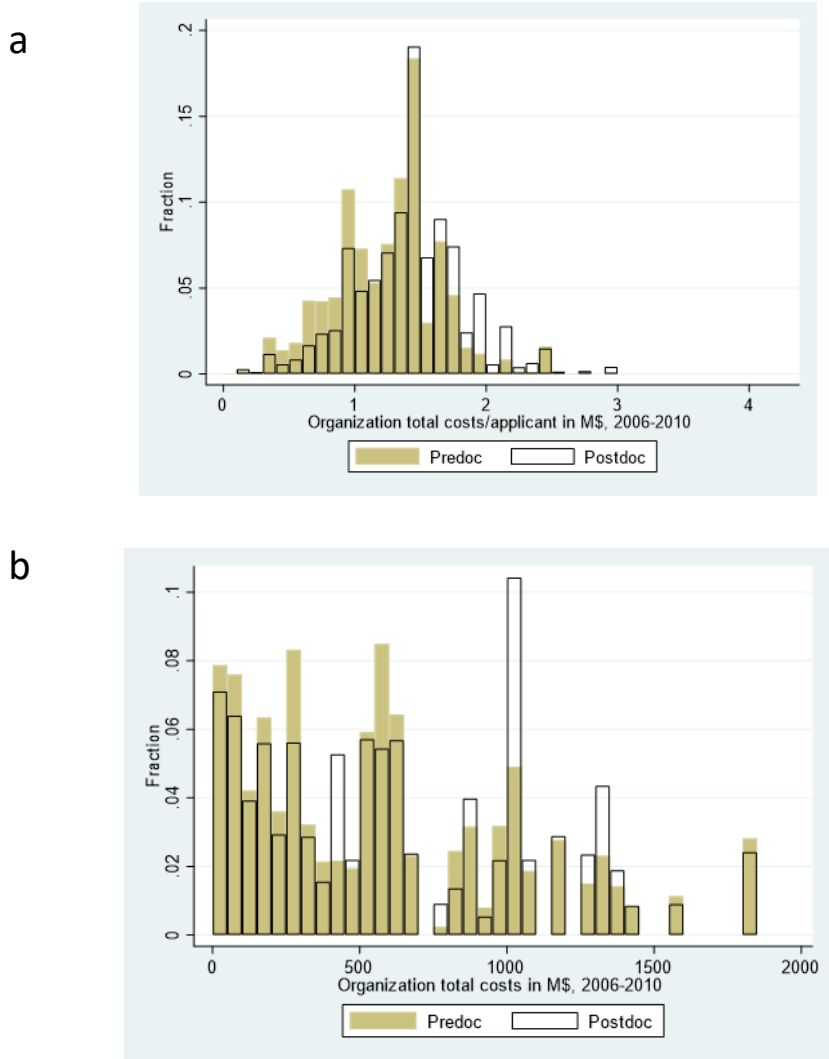


Figure 5

The effect of mentee and mentor gender on scientific productivity of applicants for NIH training fellowships

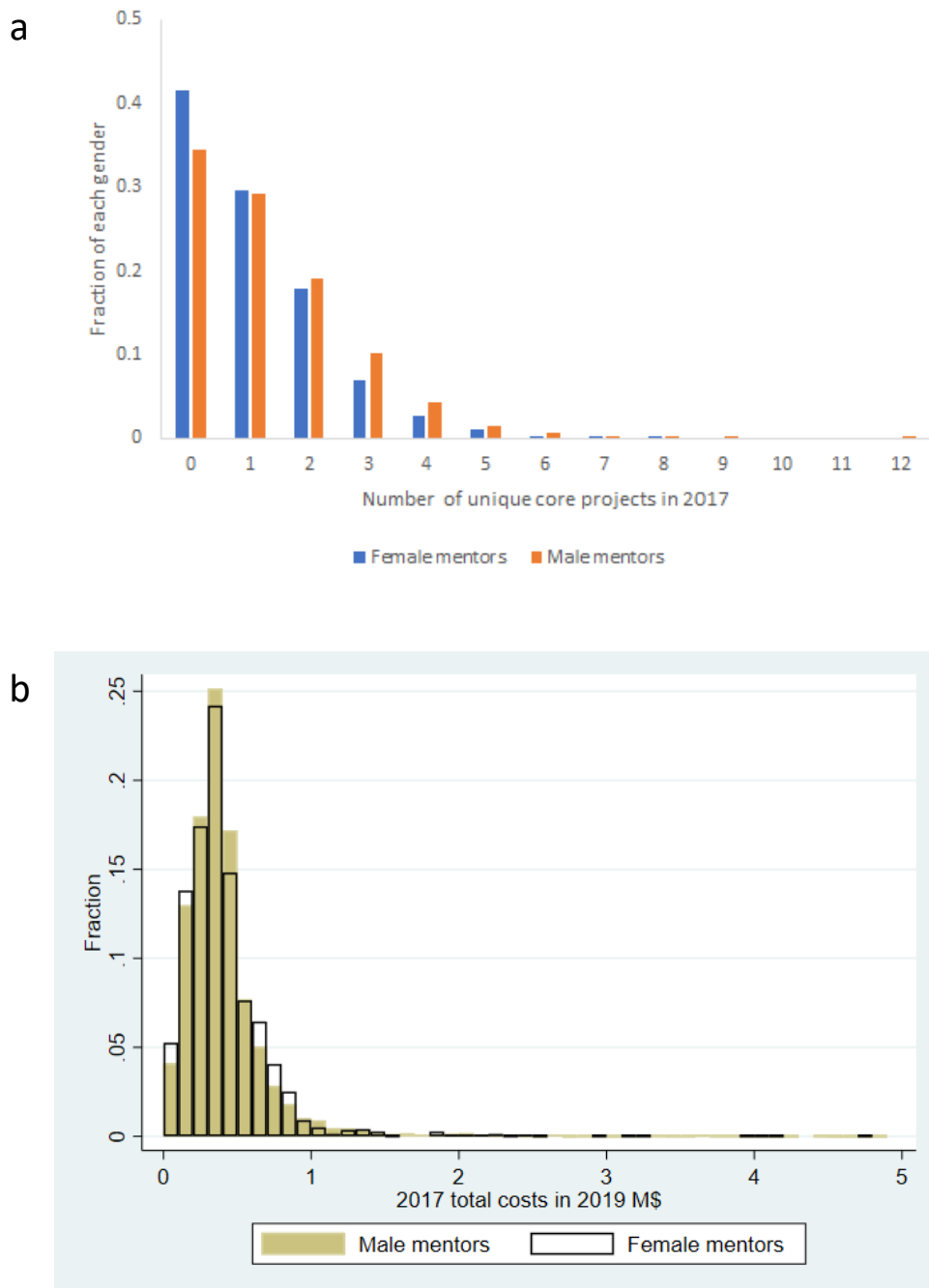
SUPPLEMENTAL INFORMATION



**Figure S1. Distribution of funding to organizations applying for pre- and post-doctoral fellowships.**

**(a)** Funding dollars to organizations listed on mentee fellowship applications are analyzed based on each organization's R01 applications in 2006-2010, and presented per applicant. Total costs to the organizations over the entire period of 2006-2010 are presented in units of million dollars, inflation adjusted to 2019\$ using BRDPI, and then divided by the number of unique Principal Investigator (PI) applicants for R01s from that organization in 2006-2010. This provides the total costs awarded per organizational PI applicant, taking into account organizations that have a relatively small number of PIs but high total costs per PI. **(b)** Same data as in (a) but presented as total costs to the organizations, rather than as a per applicant metric. For both (a) and (b), the fractions are presented within each career stage category, i.e. all clear bars (organizations from mentee post-doctoral fellowship applications) sum to 1; same for brown bars (organizations from mentee pre-doctoral fellowship applications).





**Figure S2. Distribution of mentor award number and award size, by gender.**

In order to understand the effect caused by normalizing to mentor dollar resources (Figure 5), mentor award numbers and sizes were further analyzed by gender. **(a)** The fraction of mentors with the given number of unique awarded projects in 2017. The fractions are presented within each gender, i.e. all the blue bars (female mentors) sum to 1; same for the orange bars (male mentors). Using a Mann-Whitney test, the number of awards for male mentors is significantly higher than for female mentors ( $p$ -value $<0.0001$ ). **(b)** The size of each award received by mentors by mentor gender in 2017. The unit of analysis is the award-mentor with each entry representing the estimated total costs allocated to the

given mentor-PI. The total costs are in units of million dollars, inflation adjusted to 2019\$ using BRDPI. The fractions are presented within each gender, i.e. all clear bars (female mentors) sum to 1; same for brown bars (male mentors). PIs with no funding in 2017 are not included in this figure. Awards above \$5 million have also been excluded. Using a Mann Whitney test, the distributions of award sizes are not significantly different by mentor gender.