Research Evaluation **22** (2013) pp. 272–284 doi:10.1093/reseval/rvt024

Piloting an approach to rapid and automated assessment of a new research initiative: Application to the National Cancer Institute's Provocative Questions initiative

Elizabeth R. Hsu[1,*], Duane E. Williams[2], Leo G. DiJoseph[2],
Joshua D. Schnell[2], Samantha L. Finstad[1], Jerry S. H. Lee[3],
Emily J. Greenspan[3] and James G. Corrigan[1]

[1]Office of Science Planning and Assessment, National Cancer Institute, Bethesda, MD 20892, USA
[2]Thomson Reuters, Rockville, MD 20850, USA and [3]Center for Strategic Scientific Initiatives, National Cancer Institute, Bethesda, MD 20892, USA
**Corresponding author. Email: hsuel@mail.nih.gov*

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint first authors.

Funders of biomedical research are often challenged to understand how a new funding initiative fits within the agency's portfolio and the larger research community. While traditional assessment relies on retrospective review by subject matter experts, it is now feasible to design portfolio assessment and gap analysis tools leveraging administrative and grant application data that can be used for early and continued analysis. We piloted such methods on the National Cancer Institute's Provocative Questions (PQ) initiative to address key questions regarding diversity of applicants; whether applicants were proposing new avenues of research; and whether grant applications were filling portfolio gaps. For the latter two questions, we defined measurements called focus shift and relevance, respectively, based on text similarity scoring. We demonstrate that two types of applicants were attracted by the PQs at rates greater than or on par with the general National Cancer Institute applicant pool: those with clinical degrees and new investigators. Focus shift scores tended to be relatively low, with applicants not straying far from previous research, but the majority of applications were found to be relevant to the PQ the application was addressing. Sensitivity to comparison text and inability to distinguish subtle scientific nuances are the primary limitations of our automated approaches based on text similarity, potentially biasing relevance and focus shift measurements. We also discuss potential uses of the relevance and focus shift measures including the design of outcome evaluations, though further experimentation and refinement are needed for a fuller understanding of these measures before broad application.

1. Introduction

Funders of biomedical research often seek to stimulate research in specific areas by establishing new funding initiatives. These organizations face a number of challenges in the development of such initiatives, including understanding how the new effort fits within the agency's portfolio and other efforts within the larger research community. For newly established initiatives, early use of portfolio assessment can inform initiative refinement and provide baseline information useful for subsequent outcome evaluations. Frequently, analyses are conducted retrospectively and are unavailable to inform program modifications or serve as the basis for a prospectively designed outcome evaluation.

In addition, depending on the size of the initiative, assessment methods relying on manual review by subject matter experts may be cost-prohibitive. Automated approaches leveraging existing administrative and grant application data can be an attractive option for rapid and objective assessment of critical aspects of a new research portfolio. Here, we have piloted such methods on the National Cancer Institute's (NCI) Provocative Questions (PQ) initiative (Varmus and Harlow 2012).

The PQ initiative was conceived in 2011 to challenge the scientific community to creatively think about and propose answers to important but non-obvious or understudied questions in cancer research. PQ research areas were identified through two community dialog processes: (1) workshops across the country with leading researchers, and (2) a public website for submission of comments. The final set of 24 questions was selected by expert judgment complimented by gap analysis of the existing published literature and the funded portfolio of the National Institutes of Health (NIH), the National Science Foundation, the Department of Defense, the Department of Energy, and the Defense Technical Information Center. These questions were included in the PQ request for applications (RFA), which was issued through two mechanisms, the Research Project Grant (R01) and the Exploratory/Developmental Research Grant (R21). The full list of questions, including sections on background, feasibility, and implications for success for each, can be found in the Supplementary Data. A total of 754 grant applications were received in response to the RFA.

Scientific administrators of the PQ initiative envisioned that this initiative would stimulate innovative research applications, pushing researchers beyond the bounds of their traditional lines of thinking. To measure progress against this vision, the scientific administrators were interested in studying characteristics of both the applicant population and the grant applications submitted. In addition, administrators also desired to have a diverse pool of applicants from a range of scientific field and career stages responding to the initiative, introducing new perspectives and approaches to the PQs. Finally, it was important that the grant applications directly address the question posed by the RFA. With these objectives in mind, the analysis was framed around the following key questions:

- *Diversity of applicants*: how diverse is the population of applicants responding to the initiative in terms of their scientific disciplines and experience? Did applicant diversity differ by question?
- *New lines of research*: does the initiative compel applicants to propose new avenues of research, different from their own previous work and also from the larger research field? Are there correlations between the measures of applicant diversity and focus shift?
- *Addressing gaps in the research funder's portfolio*: does the grant application fill gaps in the portfolio, as identified by the PQ initiative? Are there correlations between measures of applicant diversity and relevance?

Manual review of the grant applications by subject matter experts in the fields is generally considered to be the optimal approach to assess the types of questions that were of interest, but was not feasible for the 754 grant applications received. To provide rapid and objective insight into these questions while minimizing manual effort, we applied automated text mining tools and leveraged pre-existing administrative data associated with the grant applications. The success of this type of approach would have the added benefit of being robust and repeatable for the assessment of these characteristics for grant applications received in future funding iterations of the initiative.

In this study the diversity of applicants was addressed using administrative data, focusing on two characteristics of the primary applicant, i.e. the contact principal investigator (PI): scientific discipline and experience. Text mining to compute numeric relevance and focus shift values for the grant applications was used to aid scientific administrators in assessing the success of the initiative in stimulating new approaches to challenging problems. The relevance measurement is intended to assess if the application responds to the PQ. Focus shift is intended to measure the extent to which an application describes a research approach different from previously submitted applications. Both measurements are potential proxies for manual review, intended to provide a quick and objective overview of applications within a funder's portfolio.

Automated text mining techniques that assess document relevance relative to other text has a long history and resulted in an extensive catalog of document similarity scoring algorithms. Manconi (2012) and references within surveyed tools and methods used for text mining in the context of Bioinformatics. The text mining and proxy packages in the R language provide a selection of 48 similarity and dissimilarity algorithms (Lee 1999; Meyer and Buchta 2011; Feiner 2012). One of the most tested and broadly successful algorithms is the Okapai BM25 algorithm whose theoretical foundations are described in Robertson and Zaragoza (2009) and whose performance was assessed in Leveling (2012). This study uses a commercial BM25 implementation to measure relevance scores, rescaled to a 0–1 range and tested by expert review to determine a threshold for a binary relevant/non-relevant classification of the PQ applications.

Novelty detection also has a rich history both generally as a machine learning topic and specifically as a text mining problem. Marsland (2002) summarizes a wide selection of machine learning methods for measuring when a data observation differs substantially from what has been observed previously. Taylor and MacIntyre (1998) proposed a Local Fusion System neural network, which contains the essential 'closest approach' concept we use in our definition of focus shift. In a similar vein, several authors have addressed the problem of detecting documents that are both relevant to a topic or query and also add unique information to the current set of search results. Carbonell and Goldstein (1998) defined a Maximum Marginal Relevance criterion as a weighted average of a relevance score and a penalty determined by the maximum similarity to the current search results, whereas Zhang et al. (2002) studied the first story detection problem using a closest-approach calculation, using several measures such as cosine similarity and a combination of a mixture model and Kullback–Leibler divergence. Finally, the concept of novelty in patents was examined by Gerken and Moehrle (2012) with a goal similar to ours in measuring focus shift in PQ applications, and using an equivalent similarity formula but a different scoring algorithm than we employed.

Although text similarity scoring has been used in other applications for portfolio analysis or research evaluation [e.g. NIH topic maps (http://nihmaps.org) (Herr et al. 2009; Talley et al. 2011), maps of science (http://www.mapofscience.com) (Cambrosio et al. 2006; Boyack et al. 2011; Porter and Zhang 2012), semantic MEDLINE (http://skr3.nlm.nih.gov/SemMedDemo/) (Rindflesch et al. 2011)], our approach differs in that we are applying the analysis to grant applications prior to peer review rather than to successfully funded grant applications or to published literature. This important distinction allows for these tools to be used by scientific administrators and evaluators to quickly and objectively characterize submitted grant applications. Of these, NIH topic maps are the only use of text similarity to grant application text; our implementation differs in incorporating unfunded grant applications prior to peer review and by returning a numerical score as a similarity measurement. In contrast, NIH topic maps provide a visual interpretation of document similarity as compared with other funded NIH grants. In combination with administrative data, methods described here could, with additional validation, provide funders in the future with a means to estimate relevance of grant applications, whether grant applications are proposing new ideas, and whether applicant characteristics are correlated with either of those estimates.

2. Methodology

2.1 Discipline

The NIH IMPAC II grants database was used to obtain the following information on each primary applicant: degree, institution department, primary degree field, and expertise. For most applicants, only a subset of these fields contained information; populated fields were used to assign scientific discipline. In cases where all fields were missing data, manual review of the Biosketch provided by the applicant was used to assign scientific discipline. The Biosketch was also used to assign a 'primary' discipline in cases where applicants could be assigned to multiple discipline categories.

The first level of characterization of discipline was by the degree type. Clinical degrees were assigned to an MD or MD/PhD subcategory. We did not further categorize clinical degrees due to the burden of manually assigning a specialty research area, because the information available tended to be too generic to assign a scientific discipline without reviewing the applicant's previous research (e.g. an applicant with a departmental affiliation of 'surgery' did not provide sufficient information to distinguish the exact research area). All other degrees (primarily PhD or PhD-equivalent) received a second level of characterization based on the department, primary degree field, and expertise information. Each applicant was ultimately assigned to one of the six mutually exclusive disciplines: basic/life sciences; behavioral (including psychology, sociology, social policy, and human behavior); epidemiology; physical science/engineering [including chemistry, organic chemistry, physics, biophysics, mathematics, statistics, all engineering degrees, computational biology, bioinformatics, or imaging (e.g. MRI, contrast agents, optical imaging, radiology, nuclear medicine)]; clinical sciences—MD (including dental and veterinary degrees); and clinical sciences—MD/PhD (including veterinary/PhD degrees).

2.2 Principal investigator experience and stage

Experience was defined in terms of prior funding from the NIH and length of time since obtaining the highest terminal degree. We categorized applicants into one of the three mutually exclusive experience categories: new investigator (NI), early stage investigator (ESI), or experienced investigator (EI). NIH has specific definitions for NI and the NI subset known as ESI (http://grants.nih.gov/grants/new_investigators/). We used designations in the IMPAC II database to identify applicants as NI or ESI. Due to known issues with the NI designation prior to 2008, we verified NI status using the date of the first R01-equivalent grant received by the applicant. In cases where there appeared to be a discrepancy between the NI designation and date of first R01-equivalent grant, we determined NI status by examining the full NIH grant history of the PI. Applicants without an NI or ESI designation were assumed to be EIs.

2.3 Diversity and experience at the question level

To determine whether particular questions attracted a greater diversity of scientific discipline, we defined a measure of applicant diversity using the Gini index of the distribution of application counts over the applicants' discipline categories, defined as:

$$1-\sum_{i=1}^{n}\left(\frac{x_i}{\sum_{i=1}^{n}x_i}\right)^2$$

where $n = 6$ is the number of categories and the x_i are the application counts for each category. The Gini index measures the extent to which applications are concentrated in one or a few discipline categories, rather than dispersed over a larger number of categories. With six discipline categories, this form of the Gini index ranges from 0, if all applications fall into a single discipline category (the least diversity) to 0.833, if the applications are evenly distributed over all six categories (the most diversity).[1] Therefore, higher values for the Gini index indicate questions with a more diverse pool of applicants.

We selected a form of the index proposed by Gini (1912) as a measure of diversity on the basis of its broad use and ease of calculation. However, in future work we will examine alternative calculations based on the general structure described by Stirling (2007) and applied to the question of diversity in published journal subject categories by Porter and Rafols (2009).

To determine whether particular questions attracted a higher proportion of NIs or ESIs, we computed the investigator experience category proportions at the question level. Questions with a higher or lower than average proportion of NIs or ESIs were identified.

2.4 Relevance and focus shift measurement definitions

Using text similarity measurements, we examined the title and abstract of each application submitted in response to one of the 24 PQ questions. Each application was assessed and assigned a single relevance measurement compared with the text used within the public description of the RFA. Two focus shift values were calculated for each PQ application. The first was in comparison with the investigator's own previous work ('by-self'), and the second was in comparison to NIH grant applications received from other investigators ('general'). Relevance of a given PQ application to a given RFA, and the two versions of focus shift of a given PQ relative to the by-self or general previous applications were defined as:

$$\text{Relevance(RFA, PQ)} = \textit{score}(\text{RFA, PQ, corpus1})/\max_{\text{RFA}}$$

$$\begin{aligned}\text{Focus shift}_{\text{by}-\text{self}}&(\text{PQ, Previous})\\ &= \min_{\text{by}-\text{self(PQ)}}(1 - \textit{score}(\text{PQ, previous, corpus2})/\\ &\qquad \max_{\text{PQ, corpus2}-\text{PQ}})\end{aligned}$$

$$\begin{aligned}\text{Focus shift}_{\text{general}}&(\text{PQ, Previous})\\ &= \min_{\text{general(PQ)}}[1 - \textit{score}(\text{PQ, previous, corpus2})/\\ &\qquad \max_{\text{PQ, corpus2}}]\end{aligned}$$

where by-self(PQ) and general(PQ) indicate the subsets of previous applications either from the same or different investigators, respectively, and the other terms are explained below.

In all three formulas, the *score* refers to text similarity scores obtained using the FREETEXTTABLE function in Microsoft® SQL Server™, which is based on the Okapi BM25 algorithm (Microsoft Corporation 2008). Although the details of the FREETEXTTABLE function are proprietary and unknown to the authors, the BM25 approach computes similarity between one document thought of as a search query, and a second document thought of as being considered for retrieval by the search on the basis of relevance, selected from within a corpus of other candidate documents. The BM25 formula, as described by Robertson and Zaragoza (2009) consists of a sum, over all matching terms between the query and the searched document of a product of two factors: (1) a term frequency (tf) factor that increases as the matching term is repeated in searched document but then levels off for terms repeated beyond some saturation point, and (2) an inverse document frequency (idf) factor that is larger for matching terms that appear less frequently across the search corpus. The saturation point for the tf term is determined by tuning parameters and an adjustment for the size of the search document relative to the average size of documents in the search corpus.

We used the default options for the Full Text Search feature in Microsoft® SQL Server™ 2008 and did not create a customized stoplist, word breaker, or stemmer. FREETEXTSCORE returns values in the range from 0 (least similar) to 1000 (most similar). The highest similarity score we observed for any document pair in the study was 764. Corpus1 consisted of PQ applications and other similar grant applications identified as being coincidentally relevant to the PQ RFA based on original gap analysis done to inform the selection of the PQ questions. Corpus2 consisted of the union of the 'by-self' and general subsets of previous NIH grant applications, as well as the PQ applications themselves, which were added to permit calculation of a self-similarity score.

Each formula applies a scaling rule to return a value ranging from 0 (least relevant or least focus shift) to 1 (most relevant or most focus shift). For relevance, all scores for a given RFA question were divided by the largest score (max) observed from any of the applications in corpus1, whether obtained from an application in response to that question, an application responding to a different question, or one of the coincidentally relevant applications. For focus shift, two different scaling rules were used. For focus shift relative to the by-self subset of previous applications, all scores for a given PQ application were divided by the largest score observed from any of the previous applications in corpus2, excluding the PQ application self-comparison score (corpus2-PQ). For focus shift relative to the general subset, scores were divided by the largest score observed using all of corpus2, including the self-comparison score. In all cases, the maximum was found to be the self-comparison score.[2] The decision to use different scaling rules for the two versions of focus shift was based on manual review of a sample of PQ applications compared to the corresponding previous applications.

The scaled focus shift similarity score was subtracted from 1, which means the least similar document pairs would have the highest focus shift. For a given PQ application, the smallest such value found relative to the previous applications in the respective by-self or general subsets was selected as the value of the corresponding focus shift measurement; this was a conservative approach, as only a single previous application could result in a low focus shift score.

2.5 Relevance and focus shift thresholds

Rather than attempting to calibrate the relevance and focus shift measurements against the results of expert manual comparison, we determined a fixed threshold value to classify applications as either relevant or not relevant, and as either shifted in focus or not shifted in focus. Thresholds were chosen by manual review of a subset of the previous applications selected across a range of the relevance and focus shift measures. The value that provided the best concordance with subject matter expert opinion was selected as the threshold. The thresholds were 0.53 for focus shift (both forms) and 0.47 for relevance

3. Results

3.1 Applicant discipline

The disciplinary distribution of PQ applicants is summarized in Table 1. Unsurprisingly, the majority of applicants fell in the basic/life sciences category. However, the PQ initiative did attract MD and MD/PhDs, who made up 35.3% of the applicants (15.3% and 20.0%, respectively). In comparison, in fiscal year (FY) 2012, the number of MD or MD/PhD applicants excluding PQ applicants to NCI R01/R21 grants was ~30%, which was significantly lower than number of PQ MD or MD/PhD applicants (χ^2 test, $\chi^2 = 9.3$, df = 1, $P = 0.002$). In FY 2009, the number of successful NIH applicants with an MD or MD/PhD was ~28% (~17% and 11%, respectively, http://nexus.od.nih.gov/all/2011/06/23/who-are-we/). Not only did the PQ initiative attract a higher than average proportion of MD and MD/PhD applicants, every PQ question received at least one application from an MD or MD/PhD, while two questions (15 and 16) did not receive any basic/life sciences applicants. The small handful of applicants in the epidemiology and behavioral disciplines were concentrated in just a few questions, whereas those in the physical science/engineering disciplines were distributed across the majority of questions.

Table 1. Distribution of applicants across mutually exclusive discipline categories

Applicant Discipline	Percent of PQ applicants
Basic/life sciences	47.7
Behavioral	1.6
Epidemiology	2.1
Physical science/engineering	13.3
Clinical sciences—MD	15.3
Clinical sciences—MD/PhD	20.0

3.2 Principal investigator experience and stage

Table 2 shows the distribution of experience levels of the PQ applicants. Although nearly two-thirds of the applications received were from experienced investigators, the PQ initiative attracted a high percentage of NIs and ESIs, at 20.7% and 15.1%, respectively. For the R01 applications only, the percentage of NIs and ESIs was 19.4% and 10.9%, respectively. This is slightly lower than the percentage of R01 applications received by NCI from NIs (including ESIs) excluding PQ applicants in FY 2012 at ~31% (https://gsspubssl.nci.nih.gov/roller/ncidea/entry/2012_funding_patterns), but the difference is not significant (χ^2 test, $\chi^2 = 0.5$, df = 1, $P = 0.5$). By comparison, the proportion of NIs was ~27% of all competing NIH R01 awardees in FY 2012, down from ~30% in FY 2009. (http://report.nih.gov/NIHDatabook/Charts/Default.aspx?showm = Y&chartId = 273&catId = 22).

3.3 Diversity and experience at the question level

Figure 1 shows the computed Gini index for each question. Question 2, which had applicants from all six discipline categories, had the highest Gini index. A number of questions had a Gini index of 0.691 or higher (questions 10, 19, 24, 20, 14, and 17). Question 11 had the lowest Gini index; although it did have applicants from four discipline categories, it was dominated by those from the basic/life sciences. Question 16, which only had applicants from the MD and MD/PhD disciplines, had the second lowest Gini index.

Table 2. Distribution of applicants across mutually exclusive experience levels

Applicant Experience	Percent of PQ applicants
New investigator (NI)	20.7
Early stage investigator (NI subset)	15.1
Experienced investigator	64.2

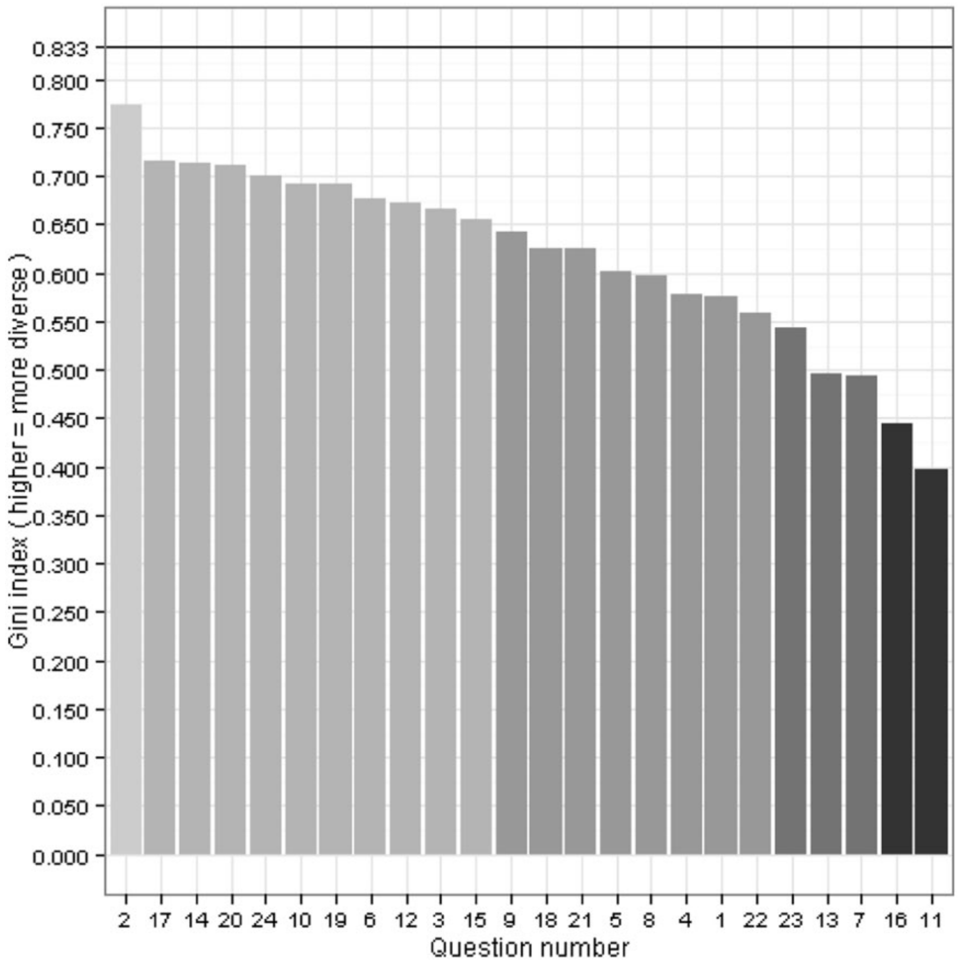

Figure 1. Gini index for each question. Higher values indicate higher diversity in applicant discipline categories. The maximum possible Gini index for a perfectly uniform distribution over six categories is 0.833.

The investigator experience distribution for each question is illustrated in Fig. 2. Questions 2, 8, 9, 14, 15, 20, and 24 had a higher proportion of NI or ESI applicants, whereas questions 3, 7, and 11 had a lower proportion of NI or ESI applicants.

3.4 Relevance

Of the 754 PQ applications, 614 (81.4%) were classified as relevant by the relevance measurement. Box plots of the measured relevance of all PQ applications are shown in Fig. 3 (Wickham 2009; R Development Core Team 2012). The portion of the distribution to the *right* of 0.47 represents the applications that were classified as relevant to the RFA text for each question. The graph shows a high degree of variability in relevance among the applications for particular questions and significantly different distributions across the PQs.

3.5 Focus shift

Of the 754 PQ applications, 39 (5.2%) were classified as shifted in focus relative to the by-self previous subset by the focus-shift measurement and 271 (35.9%) were classified as shifted in focus relative to the general subset by the focus shift measurement. Box plots of the by-self and general forms of the focus shift measurement for all PQ applications are shown in Figs 4 and 5, respectively. The portion of the distribution to the *right* of 0.53 represents applications that had a shift in focus relative to the previous applications for each question.

3.6 Degree of scientific similarity

To better understand the correspondence of the focus shift measurements with actual scientific similarity between two grant applications, we conducted a manual subject matter expert review of a subset of grant applications with very low focus shift by-self measures (focus shift by-self <0.05). Using this criterion, we found that 41% (311/754) of PQ applications were below the 0.05 threshold; 25% (189/754) of PQ applications had similar text to unfunded previous grant applications from any PI on the PQ application; and 12% (88/754) of PQ applications had similar text to

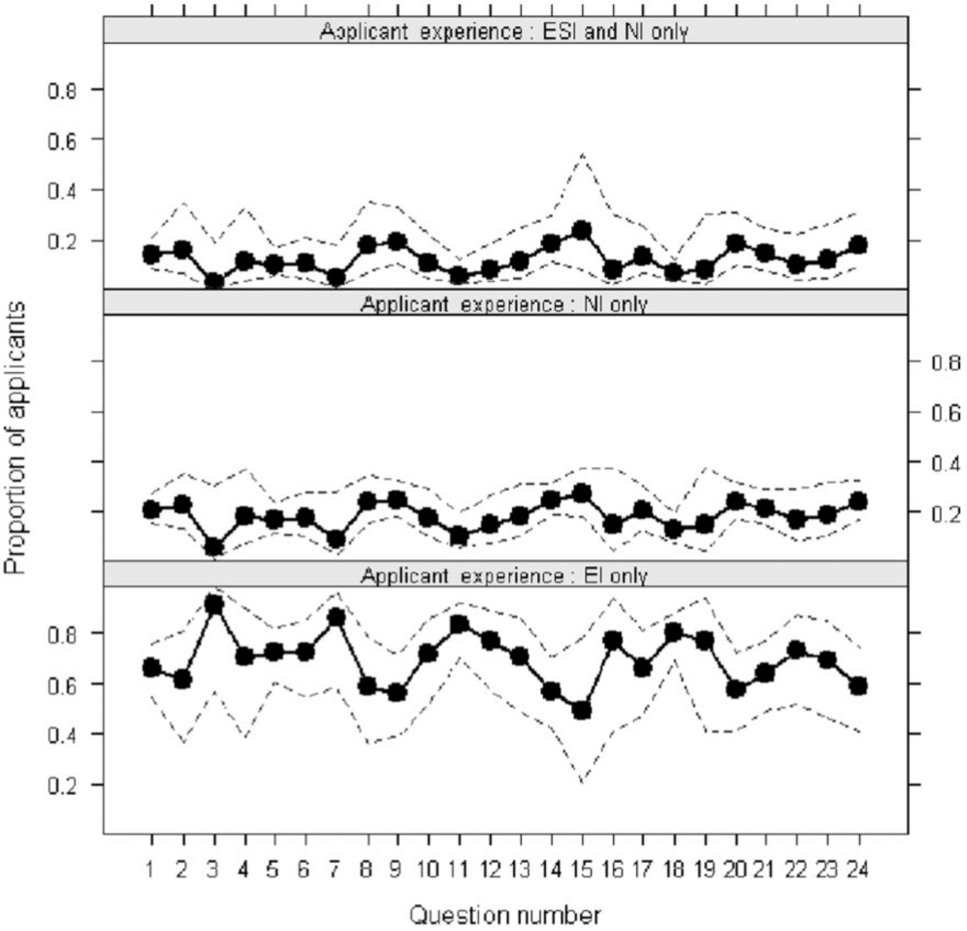

Figure 2. Proportions of the three observed applicant experience levels in the 754 PQ applications aggregated by question number. The dashed lines represent 95% confidence intervals.

funded previous grant applications that resulted in publications from any PI on the PQ application.

Manual review was conducted on 40 PQ applications subdivided into two groups based on the nature of the previous grant application to which they were most similar: PQ applications similar to unfunded previous grant applications from any PI on the PQ application; and PQ applications similar to funded previous grant applications with publications from any PI on the PQ application. The review found that PQ applications with low focus shift by-self measures cannot be assumed to have been reused from previous grant applications (Table 3). Of the PQ applications that were similar to previously funded grant applications with publications from any PI on the PQ application, a larger percentage was found to be an extension of the previous work (45%) than the percentage that appeared to be repurposing previous grant applications (25%). PQ applications that were similar to previous unfunded grant applications from any PI on the PQ application had a greater likelihood of actually repurposing previously submitted grant applications (55%).

3.7 Overall distributions of PQ applications over relevance and focus shift quadrants

The 702 PQ applications with previous by-self applications have two sets of paired values: (focus shift by-self, relevance) and (focus shift general, relevance). The remaining 52 applications only had the focus shift general, relevance pair. In this section, we examine the distribution of these paired values. The two thresholds define four quadrants in which a given application can be found, as illustrated in Fig. 6.

Table 4 shows the overall distribution of PQ applications across the focus shift/relevance quadrants, using the focus shift by-self measurement. This table includes all 754 applications; those with no previous by-self applications were classified into either the (**) or (*R) quadrants depending on whether they were relevant. Table 5 shows the quadrant distribution using the focus shift general measurement.

Finally, we performed χ^2 tests to examine the association between the applicant characteristics and the FsR quadrant outcome for the PQ applications. As illustrated in Table 6, NI and ESIs had a higher than expected

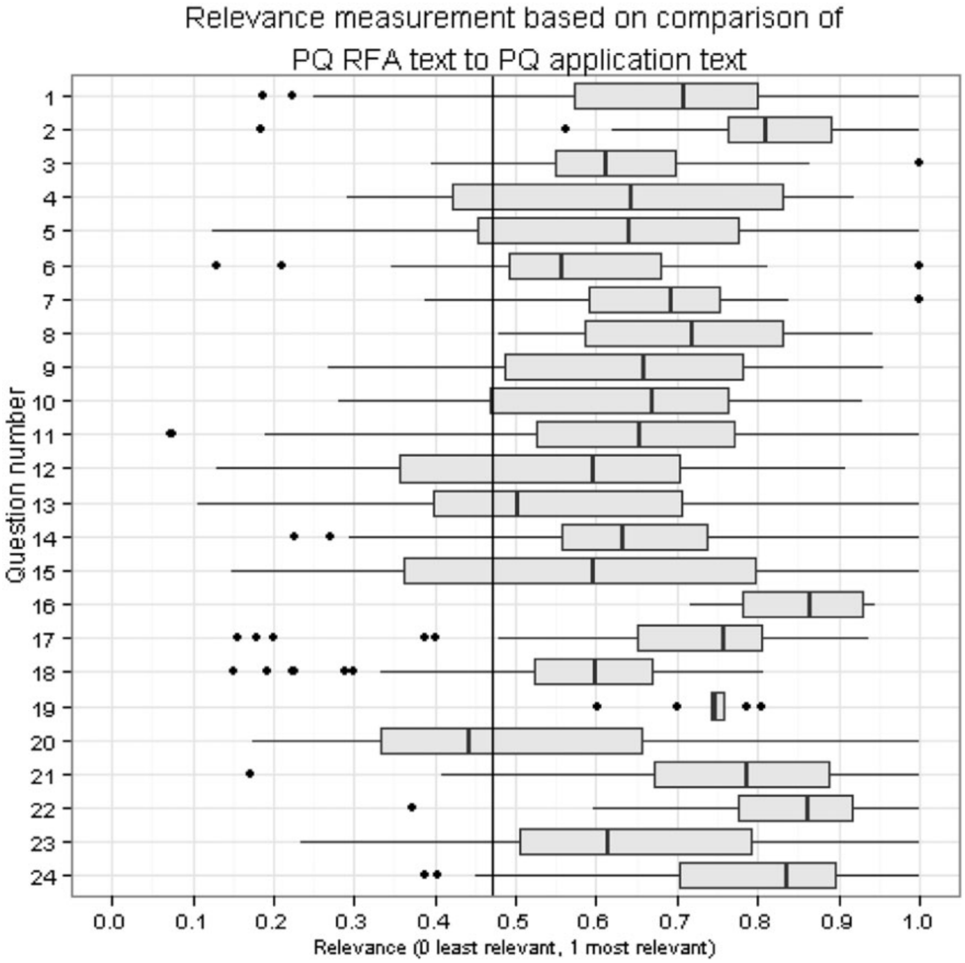

Figure 3. Box plots of relevance of PQ application text to RFA text.

representation in the FsR quadrant, whereas EIs had a higher than expected representation in the *R quadrant. EIs were significantly underrepresented in the optimal FsR quadrant. The differences seen were supported by a linear regression model, which showed a positive effect for both NI and ESIs on the focus shift by-self score ($P < 0.01$). Only marginally significant associations with discipline were found, and they were not supported by the linear regression analysis (Table 7).

4. Discussion

Using the NCI’s PQ initiative, we demonstrated that administrative data associated with grant applications can be used to provide program and evaluation staff early indicators to address key questions that often arise in portfolio assessment and gap analysis. Specifically, we addressed questions regarding the scientific diversity of applicants, proposed new avenues of research, and portfolio gaps. For the latter two questions, we utilized relatively rapid, automated methods based on text similarity scoring in a unique way.

4.1 Discipline

Applicant pool characteristics are often of interest to funders to get a sense of who is attracted to their initiative; in the case of the PQ initiative, there was interest in attracting applicants with diversity in terms of both scientific discipline and experience. Prior to receiving the applications, one concern regarding the PQ initiative had been its apparent focus on basic science. The strong showing of applicants with clinical degrees suggests this concern may not have been valid; however, a limitation of using administrative data for classification of applicant discipline was that more specific categorization of applicants with MDs or MD/PhDs to a particular scientific discipline would have required a labor-intensive manual review of Biosketches. Although it may be more likely that applicants with an MD or MD/PhD degree are conducting more clinically focused research, it is not a given that their research interests are not also in the basic sciences.

In calculating the Gini index to assess the diversity of disciplines at the question level, we chose to keep the possible number of discipline categories for each question constant at six, though some questions may

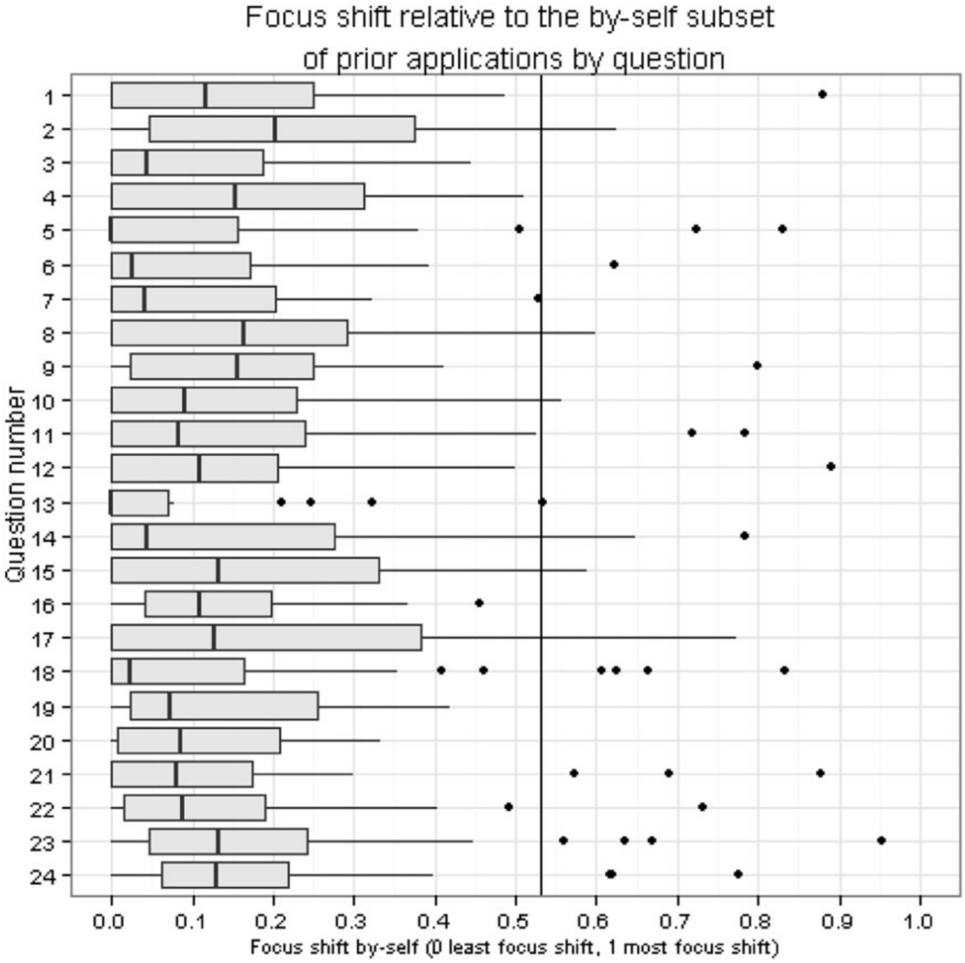

Figure 4. Box plots of PQ application focus shift relative to previous by-self applications.

have been more amenable to some disciplines than others. However, as the intent of the PQ initiative was to generate innovative approaches to addressing questions, an applicant outside of an 'expected' discipline category might be desirable. Applied in this way, the Gini index appears to balance both the number of categories and distribution among categories. Questions with applicants from multiple discipline categories and also with a relatively even distribution across the categories tended to have the highest Gini index. Questions with applicants from multiple discipline categories but with a single discipline dominating tended to be penalized with a lower Gini index. For example, while question 1 had applicants in five discipline categories, 60.7% of the applicants were in the basic/life sciences category (compared with 47.7% of all applicants), resulting in a lower Gini index of 0.577.

4.2 Principal investigator experience and stage

It is important to the NIH mission to attract and support young investigators for multiple reasons: sustaining a critical mass of investigators over time, bringing fresh perspectives and ideas to approaching research problems, and incorporating the use of newly developed technologies and methods. We have shown that the proportion of NI and ESI applications to the PQ initiative was on par with the proportion of NI and ESI applicants to all NCI R01s and slightly higher than the proportion of NI and ESI awardees on competing NIH R01s. Although we identified which questions had a higher or lower than average probability of NI or ESI applicants, the underlying factors driving a higher or lower probability are unclear.

4.3 Relevance and focus shift

Recognizing whether an applicant is proposing a distinctly new research can be difficult in manual review, depending on how familiar the reviewer is with the applicant's previous research and the broader field in general, and would be labor-intensive for a large initiative. As investigators tend to carry over ideas from previous research, achieving a focus shift classification in comparison with one's own previous applications, that is, the by-self subset, was expected to pose a challenge. Conversely, it

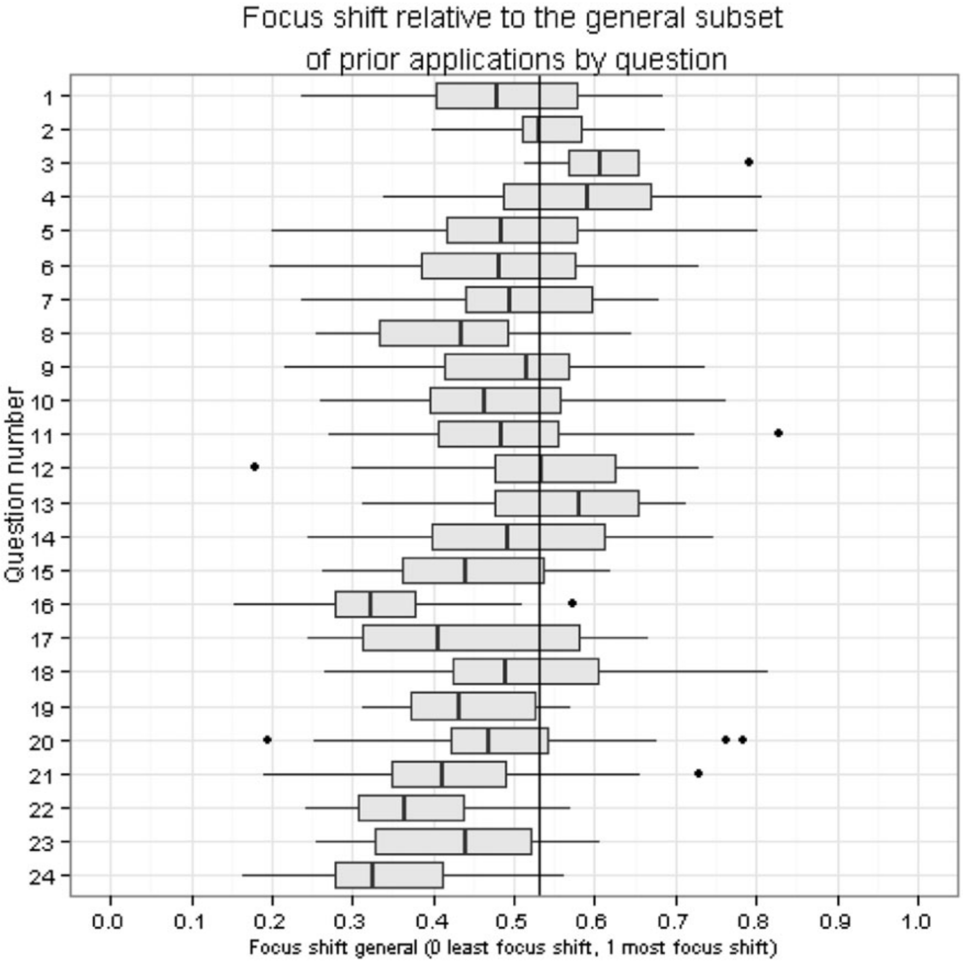

Figure 5. Box plots of PQ application focus shift versus previous NIH general applications.

Table 3. Results from manual review of 40 PQ applications with very low focus shift by-self measurements (<0.05)

Classification	Similar to unfunded grant applications (%)	Similar to funded grant applications with publications (%)
Repurposed previous grant application	55	25
Reused background/stage setting, scientific approach substantially different	30	30
Extensions of previous work	15	45

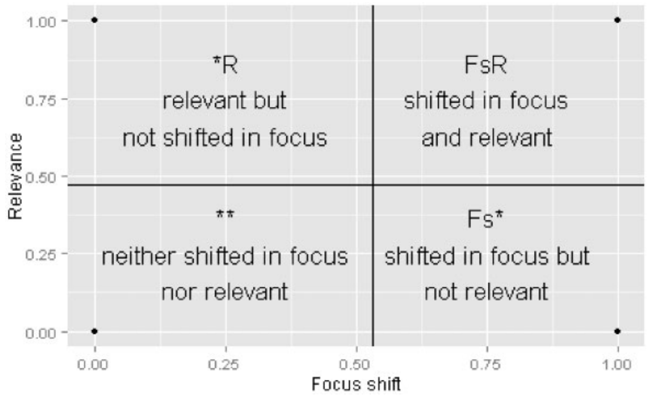

Figure 6. Focus shift /relevance quadrants.

was expected that finding a very similar scientific approach within a general previous application (which excludes the investigator’s own applications) would be less likely. These ideas were supported by our findings, as only 5.2% of applications were classified as shifted in focus relative to their own previous applications, and 35.9% were classified as shifted in focus relative to a comparison cohort of NIH applications. The high proportion of applications in the second focus shift/relevance quadrant suggests the intuitive result: generally, researchers did not stray far from previous work, but they did respond to the questions.

The χ^2 tests examining the association between applicant experience and focus shift/relevance quadrants seem

Table 4. Counts and percentages of PQ applications in the focus shift by-self/relevance quadrants

Focus shift by-self/ relevance quadrant	Description	PQ application count	Percentage of applications (n = 754)
FsR	Shifted in focus and relevant	26	3.4
*R	Relevant but not shifted in focus	588	78.0
Fs*	Shifted in focus but not relevant	13	1.7
**	Neither shifted in focus nor relevant	127	16.8

Table 5. Counts and percentages of PQ applications in the focus shift general/relevance quadrants

Focus shift general/ relevance quadrant	Description	PQ application count	Percentage of applications (n = 754)
FsR	Shifted in focus and relevant	182	24.1
*R	Relevant but not shifted in focus	432	57.3
Fs*	Shifted in focus but not relevant	89	11.8
**	Neither shifted in focus nor relevant	51	6.8

Table 6. Standardized residuals from a χ-square test of focus shift by-self/relevance quadrant and applicant experience

Applicant Experience	FsR	*R	Fs*	**
EI	**−4.5**	**2.3**	−2.0	0.3
ESI	**2.8**	0.8	−0.8	−2.0
NI	**2.8**	**−3.4**	**3.0**	1.4

Values in bold (>|2|) indicate significant over- or under-representation in a given focus shift/relevance quadrant.

Table 7. Standardized residuals from a χ^2 test of focus shift general/relevance quadrant and applicant discipline

Applicant discipline	FsR	*R	Fs*	**
Basic/life sciences	−1.5	2.0	−0.1	−1.3
Behavioral	**2.8**	**−3.5**	1.4	0.2
Epidemiology	1.3	0.4	−1.5	−1.1
Physical science/engineering	−0.3	−0.7	1.1	0.5
Clinical sciences—MD	0.3	−1.0	1.1	0.1
Clinical sciences—MD/PhD	0.5	−0.1	−1.6	1.4

Values in bold (>|2|) indicate significant over- or under-representation in a given focus shift/relevance quadrant.

to suggest that experienced investigators were less likely to stray far from previous work. However, the contrast may also be an artifact of new investigators having fewer previous applications available for comparison.

4.4 Factors affecting text similarity measurements

The relevance and focus shift measurements are both based on text similarity measurements, and are therefore limited by factors affecting text similarity measurements. Results from a subject matter expert review of a sample of applications suggest that the current approach for measuring relevance and focus shift using text similarity could be enhanced by a more sophisticated method that appropriately accounts for semantic differences within the text. For example, for funded grant applications with publications, low focus shift scores were often attributable to an inability of the text similarity algorithm to distinguish subtle differences in scientific nuance rather than a true lack of new ideas by the applicant. Nearly half of these applications were found to be extensions of previous work; in further development of the focus shift measure, it will be important to ensure that applicants proposing tangents to existing lines of research are correctly classified as shifted in focus by the text similarity algorithm.

One important consideration when using text similarity measurements for comparisons that the two bits of text being compared must be sufficiently similar in intent; the method may be limited by misalignment of the data sources relative to the study question being addressed. In this case, we found that these tools were generally effective in measuring focus shift because the comparison involved similar types of text—titles and abstracts of two grant applications. Additionally, the text similarity measurement is sensitive to numerous factors, including the text chosen for comparison. For example, application text that incorporates the comparison text may inflate similarity scores. Improvements to the measurement, in particular the focus shift measurement, may be obtained by using more text from the grant applications beyond the titles and abstracts. For example, the specific aims may contain more scientific nuances than the abstract and may better distinguish between applications with subtle differences in scientific approach; evaluating whether including the specific aims section of the grant applications improves performance will be an important next step. Other factors that may affect the text similarity measurement are cohort size and scaling parameters. We found that using too small a comparison cohort resulted in PQ applications incorrectly classified as shifted in focus—that is, expanding the comparison cohort revealed applications that were found to be similar to the PQ applications that had previously been missed. In conducting the manual assessment to choose the appropriate threshold for the relevance and focus shift measurements, we discovered that the measurements needed to be scaled differently in order to allow for adequate spread near zero.

5. Conclusion

Although we have not proposed any new extensions to text mining methods for similarity measurements, we have demonstrated a unique implementation of such measurements to grant application text. We have also demonstrated combining these measurements with applicant characterization information for the purposes of evaluating grant applications prior to peer review and to funding. With further refinement, these methods may provide a less burdensome alternative to subject matter expert review and qualitative data collection to guide both evaluators of research programs and research funders.

Two possible applications of the relevance and focus shift measurements are for design of an outcome evaluation and as an adjunct tool for expert review methods. Outcome evaluations of biomedical research initiatives are often designed retrospectively, commonly when initiative impact needs to be demonstrated for a variety of reasons (e.g. renewal of funding). Early analyses conducted at the application stage could generate baseline data in preparation for a future outcome evaluation. An outcome evaluation of an initiative like the PQs might ask if the research resulting from the funding was truly innovative or effectively addressed gaps in the scientific portfolio. Collecting measurements at the application stage allows for the construction of questions such as: is there a correlation between relevance measurements at the application stage and how well the funded research addresses the PQ question? Do focus shift measurements at the application stage serve as an indicator of innovative research? Or was progress in the understudied PQ research areas more likely to be made by applications classified as relevant or demonstrating a shift in focus at the application stage? Comparing similar measurements of the outputs of the funded research (e.g. publications) to the focus shift and relevance measurements of the applications might also provide a better indicator of how the overall field has shifted over time, not just the research funded by the initiative. In addition to serving as a baseline for comparison to measurements of the outputs of the funded research, application-based measures might also serve as early predictors of research trajectory. The feasibility of using relevance and focus shift measures as early indicators of research trajectory needs further exploration as the outputs of the PQ-funded projects are generated in coming years.

In addition to these uses, funders could potentially utilize relevance measures to guide reviewers and highlight grant applications that may not have addressed the questions as intended by the initiative. Similarly, focus shift measurements could serve to alert reviewers to proposed research that is potentially similar to something already being supported by the funding agency. We do not suggest that automated assessment methods can replace peer review by subject matter experts, but instead that these methods could potentially help to guide review. For example, these automated methods could act as an adjunct to manual review as part of a hybrid system in which peer review would be informed by the automated analysis, highlighting items for closer inspection by reviewers. It is also possible that not every funding initiative could benefit from such an assessment—the PQ initiative was unusual in that it was asking applicants to be innovative in their approaches and covered a wide breadth of subjects, making review particularly challenging. More standard funding initiatives might have reviewers who are more familiar with the breadth of the subject, or the initiatives may not ask applicants to be truly innovative in the approach but to instead fill a recognized unmet research need.

A natural extension of this analysis is to examine how relevance and focus shift scores correlate with review scores and likelihood of being funded. Questions that could be addressed in such an analysis are whether relevance and focus shift correlate with any of the existing review measures or represent independent measures. If they represent independent measures from existing review scoring criteria, that might further justify their examination as a complement to the peer review system. It would also be interesting to examine whether a model could be developed incorporating the relevance and focus shift measures that could predict likelihood of being funded.

The difference in structure of the PQ initiative compared with traditional funding initiatives notwithstanding, the broader utility of our approach is that it may provide automated assessments of questions that might otherwise not be feasible when using purely manual review. Future work will include further refinement of our methodology to allow incorporation of additional text from the grant application such as specific aims or research strategy and to account for semantic differences, providing better estimates of relevance and focus shift that can be used to inform assessments of research initiatives.

Acknowledgements

The authors would like to thank Yvette Seger for her work in the manual review for the 'degree of scientific similarity' analysis.

Funding

This work was supported in part by contract funding from the National Cancer Institute at the National Institutes of Health.

Notes

1. For just two categories, the maximum (most diverse) Gini index is 0.5, the maximum Gini index approaches 1 from below as the number of categories increases to infinity.
2. A score larger than a self-score can be obtained if another document has the same number of term matches but is shorter in length.

References

Boyack, K. W. *et al*. (2011) 'Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches', *PLoS One*, 6/3: e18029.

Cambrosio, A. *et al*. (2006) 'Mapping the Emergence and Development of Translational Cancer Research', *European Journal of Cancer*, 42/18: 3140–8.

Carbonell, J. and Goldstein, J. (1998) 'The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries', *SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–6. New York, NY, USA: ACM.

Feinerer, I., Hornik, K. and Meyer, D. (2008) 'Text Mining Infrastructure in R', *Journal of Statistical Software*, 25/5: 1–54.

Gerken, J. M. and Moehrle, M. G. (2012) 'A New Instrument for Technology Monitoring: Novelty in patents measured by semantic patent analysis', *Scientometrics*, 91/3: 645–70.

Gini, C. (1912). 'Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche' [Variability and mutability: contribution to the study of distributions and statistical relationships]. (Bologna: Tipogr. di P. Cuppini).

Herr, B. W., 2nd *et al*. (2009) 'The NIH Visual Browser: An Ineractive Visualziation of Biomedical Research', *IEEE International Conference Information Visualisation*, pp. 505–9.

Lee, L. (1999) 'Measures of Distributional Similarity', *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*, pp. 25–32. Stroudsburg, PA, USA.

Leveling, J. *et al*. (2012) 'DCU@TRECMed 2012: Using Ad-hoc Baselines for Domain-Specific Retrieval', *The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings*, SP 500-298: NIST Special Publication.

Manconi, A. *et al*. (2012) 'Literature Retrieval and Mining in Bioinformatics: State of the Art and Challenges', *Advances in Bioinformatics*, 2012: 1–10.

Marsland, S. (2002) 'Novelty Detection in Learning Systems', *Neural Computing Surveys*, 3: 1–29.

Meyer, D. and Buchta, C. (2011) 'Proxy: Distance and Similarity Measures', (R package version 0.4-7) http://cran.r-project.org/web/packages/proxy/.

Microsoft Corporation. (2008) '*How Search Query Results are Ranked (Full-Text Search)*', (SQL Server 2008 R2) <http://msdn.microsoft.com/en-us/library/ms142524> accessed 9 October 2013.

Porter, A. and Rafols, I. (2009) 'Is Science Becoming More Interdisciplinary? Measuring and Mapping Six Research Fields Over Time', *Scientometrics*, 81/3: 719–945.

Porter, Alan L. and Zhang, Yi (2012) 'Text Clumping for Technical Intelligence'. In: Sakurai, Shigeaki (ed.) *Theory and Applications for Advanced Text Mining*, InTech: Rijeka.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rindflesch, Thomas C. *et al*. (2011) 'Semantic MEDLINE: An Advanced Information Management Application for Biomedicine', *Information Services & Use*, 31: 15–21.

Robertson, S. and Zaragoza, H. (2009) 'The Probabilistic Relevance Framework: BM25 and Beyond', *Foundations and Trends in Information Retrieval*, 3/4: 333–89.

Stirling, A. (2007) 'A general framework for analysing diversity in science, technology and society', *Journal of the Royal Society Interface*, 4/15: 707–19.

Talley, E. M. *et al*. (2011) 'Database of NIH Grants Using Machine-Learned Categories and Graphical Clustering', *Nature Methods*, 8/6: 443–4.

Taylor, O. and MacIntyre, J. (1998) 'Adaptive Local Fusion Systems for Novelty Detection and Diagnostics in Condition Monitoring'. In: Dasarathy, B. V. (ed.) *Sensor Fusion: Architectures, Algorithms and Applications II*, pp. 210–8. Orlando, FL: SPIE.

Varmus, H. and Harlow, E. (2012) 'Science Funding: Provocative Questions in Cancer Research', *Nature*, 481/7382: 436–7.

Wickham, Hadley (2009) *ggplot2: Elegant Graphics for Data Analysis*, 213. New York: Springer.

Zhang, Y., Callan, J. and Minka, T. (2002) 'Novelty and Redundancy Detection in Adaptive Filtering', *SIGIR Forum, ACM Special Interest Group on Information Retrieval*, 81–8.