Research Policy 53 (2024) 105075

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Short communication

Unexpectedness in medical research

Yasemin Aslan [a], Ohid Yaqub [a,*], Bhaven N. Sampat [c], Daniele Rotolo [a,b]

[a] SPRU (Science Policy Research Unit), University of Sussex Business School, UK
[b] DMMM (Department of Mechanics, Mathematics and Management), Polytechnic University of Bari, Italy
[c] Arizona State University, USA

ARTICLE INFO

Keywords:
Research funding
Research direction
Priority setting
Research outputs
Spillovers
Medical research
Machine learning

ABSTRACT

Whether research funding is targetable is one of the central unresolved questions of science policy. A particular question is how often research aimed at understanding one disease or problem spills over to others. This has been a perennial topic of debate at the world's largest single funding body of biomedical research, the U.S. National Institutes of Health (NIH). Critics of the agency's priority-setting process have repeatedly called for better alignment between funding and disease burden, and patient advocates for specific diseases for more funding for their causes. In response, opponents of planning have argued that research in one area frequently leads to advances in others. In this study, we provide new evidence to inform these debates by examining the extent to which research funding (grants) in one scientific or disease area leads to research findings (publications) in another. We used the NIH's Research, Condition, and Disease Categorization (RCDC) to identify categories for NIH grants awarded between 2008 and 2016. We applied machine-learning to map text to these categories and use this model to categorize publications resulting from these grants. We categorized over 1.2 million publications, resulting from over 90,000 grants. We found that 70 % of the publications have at least one RCDC category not in its grant, which we termed 'unexpected' categories. On average, 40 % of categories assigned to a publication were unexpected. After adjusting for similarity across some of the RCDC categories by empirically clustering the categories, we found 58 % of the publications had at least one unexpected category and, on average, 33 % of publication categories were unexpected. Our results suggest that disease-orientation and clinical research were less likely to be associated with spillovers. Grants resulting from targeted requests for applications were more likely to result in publications with unexpected categories, though the magnitude of the differences was relatively small.

1. Introduction

Over 75 years ago in *Science, The Endless Frontier*, sometimes considered the blueprint for postwar U.S. science policy, Vannevar Bush made the case that science cannot be neatly planned or tied to social goals, since much progress comes from unexpected sources (Mowery, 1997; Nelson, 1997). This view contrasted with that of many New Deal liberals, most prominently Harley Kilgore (D-W.Va) who argued for a postwar science policy more targeted to specific social and economic goals, and for political considerations (not just scientific ones) to guide the allocation of federal funds for science (Kevles, 1977).

On the issue of medical research, the "Bush report" (as Endless Frontier came to be known) asserted "Discoveries pertinent to medical progress have often come from remote and unexpected sources, and it is certain that this will be true in the future." This idea has continued to be an important one in postwar biomedical research policy. In 1946, the first director of the grants program at the National Institutes of Health (NIH) argued against targeting of biomedical research because "in the normal course of scientific investigation, however, the bypaths often lead to more important findings than do the roads from which they branch" (Van Slyke, 1946). In debates over the War on Cancer, spokesmen for the NIH argued that cancer should not be targeted exclusively since cancer research benefits from that aimed at other diseases (Rettig, 1977).

More recently, in deflecting criticism from Congress and others that the agency's funding patterns were misaligned with U.S. disease burden, NIH Director Harold Varmus suggested that priority setting and planning may be misguided because "research aimed in one direction frequently provides benefits in an unexpected direction" (Varmus, 1997). Disease advocates lobbying for funding for 'their' disease (Best,

* Corresponding author.
E-mail address: o.yaqub@sussex.ac.uk (O. Yaqub).

https://doi.org/10.1016/j.respol.2024.105075
Received 20 March 2023; Received in revised form 3 July 2024; Accepted 9 July 2024
Available online 22 July 2024

2012) implicitly take the opposite view on targetability: that funding in one area is likely to generate local advances. As do members of Congress who have tried to stimulate research on specific diseases through creation of new categorical institutes and centers at the NIH, typically over the objections of NIH leaders and the scientific community (Varmus, 2001).

The NIH budget has grown substantially since the time Bush and Van Slyke wrote, rising to about $45 billion today in fiscal year 2023. However, there remains limited large-sample empirical evidence to inform these debates on the magnitude of cross-category knowledge spillovers, or their correlates. One reason for this is that, until relatively recently, there were no formal reporting categories for NIH funding. This changed with the 2008 "Research, Condition, and Disease Categorization" (RCDC) effort, through which the agency (at the request of Congress) developed standardized methodologies to classify funds by area (Sampat et al., 2013). The NIH notes, "RCDC provides consistent and transparent information to the public about NIH-funded research. For the first time, a complete list of all NIH-funded projects related to each category is available" (NIH, 2011, cited in Sampat et al., 2013). The NIH began with 212 categories starting in 2008, and currently has 293 categories covering a range of diseases, conditions, or research areas. NIH grants awarded from 2008 onward are matched to RCDC categories by NIH, and this information is publicly available in the NIH RePORTER database.

The RCDC provides the categories for grants but not for their corresponding outputs. To answer the question of whether the funded research delivers outputs in unexpected categories, we need to measure and categorize the outputs – i.e., the publications resulting from grants as reported in the NIH RePORTER. Accordingly, we applied a machine-learning model that maps publication-text into RCDC categories. The model is trained on grant abstracts and their assigned RCDC categories. We then used the model to categorize over 1.2 million publications, resulting from over 90,000 grants, amounting to $94.1 billion in research funding.

We found that 70 % of publications had at least one RCDC category not in its grant, which we termed 'unexpected' categories. On average, 40 % of categories assigned to a publication were unexpected. After adjusting for similarity across some of the RCDC categories by empirically clustering the categories, we found 58 % of the publications had at least one unexpected category and, on average, 33 % of publication categories were unexpected. We also found that the propensity to yield unexpected RCDC categories differed between 'basic' and 'applied' medical research, and investigator-initiated and funder-initiated research, though the magnitudes were small.

This paper is organized as follows. Section 2 describes our empirical approach; Section 3 presents the results; and Section 4 concludes the paper with a discussion of policy implications and suggestions for future research.

2. Empirical approach

2.1. Data sample: NIH grants and publications

We began by gathering all NIH funded grants and abstracts, and lists of their resulting publications, from the NIH RePORTER database.[1] This database (through the ExPORTER tool) provides unique identifiers (PubMed identifiers or PMIDs) for all publications resulting from NIH grants (also termed projects). Using PMIDs, publication abstracts were then retrieved from NIH's National Library of Medicine through PubMed.

We started with 742,339 RePORTER records from FY (fiscal year) 2008 to FY 2016. We started with FY 2008 since that is the year when the RCDC was introduced by NIH, and ended in 2016 to allow an ample window to observe publication outputs from the grants. We excluded 132,794 records where RCDC codes were not assigned. We also excluded 28,676 records without abstracts or with very short (<250 words) or very long (>5000 words) abstracts, since our pilot analyses suggested the machine-learning algorithm would not work on these cases, and for similar reasons 1164 additional records with more than fifteen RCDC categories. This yields a sample of 579,705 RePORTER records. Further, we dropped any records with subprojects (e.g., center grants with subcontracts) since we want to focus on projects with a unique abstract we can categorize. This left 472,457 records, accounting for 64 % of grants and 65 % of funding over the period.

The unit of analysis in RePORTER is the application number. In the database, multiple application numbers can be associated with one grant, including for renewals, extensions, revisions etc. Since publications are recorded at the "core" project level,[2] project applications and their publications were therefore aggregated into their corresponding core projects (for example, R01CA012345-01 and R01CA012345-02, corresponding to different grant years, were aggregated to the core project R01CA012345). The 472,457 records were associated with 157,871 core project numbers.

We then collected publication data for these projects by using the corresponding PMIDs as available in RePORTER. For each, we collected abstracts from PubMed, the National Library of Medicine publication database, by querying PubMed with PMIDs. As with grants, we dropped publications without abstracts or with abstracts that were very short (<250 words) or very long (>5000) words. In cases where a grant yielded no publications, or where none of the publications could be reliably categorized, we dropped the underlying grant. With these drops, we were left with 134,964 core project numbers.

As a final step, we further excluded 39,818 core projects whose first funding preceded 2008, and 4184 intramural core projects leaving us with a final sample of 90,962 core projects and 1,251,826 corresponding publications (of which 715,961 are unique). The final set of grants accounts for 58 % of all grants (as records) and 48 % of all funding from all grants that began after 2008, over the 2008–2016 period. With these data, our unit of analysis was grant-publication pairs or dyads.

2.2. RCDC-based categorization of publications

As noted, RCDC categories are assigned to NIH grants, not publications. In order to examine category crossing, we need to assign categories to grants and publications to the same categorization scheme. To link publications to RCDC categories, we applied and compared 11 machine learning algorithms trained on NIH grant abstracts and their RCDC assignments. To assess the performance of the selected algorithms, we set aside 10 % of project abstracts – not used to train the algorithms – and compared the NIH's assignment of RCDC to projects with that of the algorithm. Among the different machine learning algorithms, the linear support vector classifier with a term frequency-inverse document frequency (tf-idf) vectorizer returned the highest values on the two performance indicators we describe below: Jaccard similarity, and adjusted cosine similarity (see Appendix A). This is consistent with support vector machines often performing well in theoretical and empirical evaluations of text categorization (Joachims, 1998).

We calculated the macro-average to F1 score to enable comparison. The macro-average F1 score – which accounts for *precision* and *recall*

[1] Data were downloaded from the NIH's ExPORTER tool (https://exporter.nih.gov) on August 17, 2022.

[2] Core project numbers remain fixed over the course of a grant, including for renewals after initial funding. Core projects can also include different subprojects, in cases where there are subcontracts beyond the main funded institution.

performance – achieved a value of 74 % for our machine learning model, which compares favourably with notable benchmarks. For example, the Medical Text Indexer, an algorithm that assigns Medical Subject Headings (MeSH) terms to text, shows a macro-average F1 score of 56 % to 65 % (Jimeno-Yepes et al., 2013), while Madjarov et al. (2012) reported an average macro-average F1 score of 35 % for text-classification when applying 12 machine learning methods for multilabel learning over a variety of benchmark datasets.

However, traditional performance metrics that involve a combination of true/false positives/negatives offer an incomplete approach to evaluating classification in our setting, since we have a dataset with multilabel classes and with non-mutually exclusive categories. This is because, with multilabel classification, a classification may not necessarily be wholly correct or incorrect. For example, if only a subset of the actual categories is predicted correctly, then this is better than predicting no categories correctly. Predicting a higher proportion of the subset correctly is preferred. Moreover, a misclassification may be more or less incorrect depending on its similarity to the actual category. A misclassification with a category that is similar to the correct category is preferred. We address these two issues respectively by using the two performance indicators: Jaccard similarity and adjusted cosine similarity. Our machine learning model achieved a Jaccard similarity of 80 %, and a cosine-adjusted similarity of 91 % (Appendix A).

After selecting our machine-learning algorithm, we deployed the algorithm to assign RCDC categories to publication abstracts.

2.3. Spillover measures

We focused on two main outcome measures for a given grant-publication dyad. First, we defined a dummy variable – *Unexpected Category* – that assumed a value of one if *any* of the publication's categories were unexpected (i.e., they are not assigned to the corresponding grant) and a value of zero otherwise. Second, we measured the *proportion* of a publication's categories that were unexpected – *Share of Unexpected Categories*. These measures give us an indication of whether, and how far, the research went beyond its expected target.

To evaluate the validity of these RCDC-based measures as indicia of cross-category spillovers, we compared the results to those obtained by manual classification, asking how the RCDC-based spillover measures compare to one developed through hand-coding. For this validation, we focused on disease-oriented grants only; i.e., those that were assigned one or more of the RCDC disease categories. We developed the list of RCDC disease-related categories manually, starting with the list of 74 categories that report a burden of disease measure, and adding a further 97 categories after manual review (Appendix B provides the list and further details).

For validation purposes, we randomly sampled 40 disease-focused grants and one publication from each. We then took the grant and publication abstracts, scrambled their ordering, and had two expert reviewers independently assign each to categories in the International Classification of Disease (ICD) system, 10th edition. They assigned both broad ICD-10 chapters (e.g., Chapter 1, Certain Infectious and Parasitic Diseases) and narrower ICD-10 codes (e.g., B55, Leishmaniasis) – see Appendix C for more information on coding. Inter-rater reliability scores were "substantial" (0.68) at the ICD code level, and "almost perfect" (0.89) at the ICD chapter level (Landis and Koch, 1977; Gwet, 2008). Differences between the two reviewers were resolved by discussion. The agreed set of ICD codes assigned to the abstracts were then sent for comparison with RCDC codes. One grant abstract and one publication abstract were not classified into any ICD codes, so these along with their two paired abstracts were discarded from the comparison. This left a total of 38 pairs. We reassociated each grant with its publication, and asked whether the ICD disease category of the publication was the same as that for the grant. If not, we called the publication 'unexpected'. We then compared this measure to the unexpectedness measured using the RCDC approach. We found that only 4 out of 19 (21 %) pairs without unexpected RCDC disease categories had unexpected manually coded ICD categories, and that 18 out of 19 (95 %) of pairs with unexpected RCDC disease categories also had unexpected manually coded ICD categories.

From this validation exercise, we concluded that RCDC-based measures do generally capture cross-category spillovers, and decided to use RCDC-based measures in the analyses that follow.

2.4. Clustering similar RCDC categories

A potential issue with RCDC is that some categories are quite similar (e.g., "tobacco" and "tobacco smoke and health"). This feature of classification may lead to cases of spillover that are actually a reflection of the similarity between certain RCDC categories. To address this, we used cluster analysis to aggregate RCDC categories assigned to grants. We examined the co-occurrence (cosine similarity) network of RCDC categories and identified 32 clusters that maximize similarity within each cluster while minimizing similarity between the clusters (see Appendix D for further details). We then defined the *Unexpected Category – Clustered* variable assuming value of one if *any* of the publication's clustered categories were unexpected, zero otherwise; and the *Share of Unexpected Categories – Clustered* variable representing the *proportion* of a publication's clustered categories that were unexpected. In most of the analyses below, we show results for both clustered and unclustered categories.

2.5. Categorizing basic versus applied grants

In addition to measuring whether funded research delivers outputs in unexpected categories (through the two measures *Unexpected Category* and *Share of Unexpected Categories*), we also related the scores on these measures to specific characteristics of the grants. One policy-relevant question is whether basic research is more likely to yield unexpected results than applied research. This is the argument that was made by Bush and Van Slyke, quoted in the introduction, and has been a mainstay in biomedical research policy debates ever since. One difficulty in assessing this is that the labels "basic" versus "applied" may imply a false dichotomy, as Stokes (1997), among others, has pointed out. Rather than engaging in this age-old debate, here we follow previous research in examining different dimensions of "appliedness". Li et al. (2017) investigated how likely basic/applied grants were to be linked to patents. As indicia of applied grants, they examined whether individual grants were focused on diseases, whether they were patient-oriented, and whether they were solicited through requests for applications "used to direct research at particular diseases or problems" (Li et al., 2017).

Here, we generated similar measures from the RCDC and NIH administrative data. We categorized a grant as "disease-oriented" if at least one of the RCDC categories was a disease category (see Appendix B for details on the construction of this set). We categorized a grant as "clinical" if the NIH has assigned it to the "Clinical Research" RCDC category. To determine grants emanating from requests for applications (RFAs), we consulted RePORTER's Funding Opportunity Announcement (FOA) field to see if it contained the string "RFA". RFAs are "a narrowly defined area for which one or more NIH institutes have set aside funds for awarding grants" and are typically reviewed by NIH Institutes and Centers, rather than through standard peer-review mechanisms.[3] RFAs are often the way NIH responds to Congressional directives to fund specific types of research or, are an effort by agency officials to solicit

[3] See https://grants.nih.gov/grants/guide/description.htm Last accessed February 20, 2023.

research that addresses specific problems or is in a specific field, or both (Sampat, 2012; Hegde and Sampat, 2015). These attempts by the NIH to steer research by the NIH have been controversial. And RFAs are growing: the percentage of NIH grants resulting from RFAs increased from 2 % in 1985 to over 25 % by 2015, and the corresponding share of NIH funding from RFAs increased from 5 % to 33 % (NIH RePORTER, 2023). Like Li et al. (2017), here we use RFA grants as a third measure of more targeted research.

3. Results

3.1. Descriptive statistics

Table 1 shows descriptive statistics at the grant level for each of the 90,962 grants in our dataset. The grants span the 2008–2016 period, by construction. About 43 % of the grants were R01 or equivalent, the main investigator-initiated grant mechanism at the NIH. On average, grants were assigned about six RCDC categories (range: 1–20) and were funded for about $1.03 million. Of the grants in our sample, about 13 % were from RFAs, 38 % were for clinical research, and 82 % were disease-focused. Table 2 shows the descriptive statistics for grant-publication dyads, while Table 3 shows the correlation matrix.[4]

Across the more than 1.2 million grant-publication pairs in this study, 70 % of publications were assigned at least one RCDC category that differed from the categories assigned to the associated grant. Not surprisingly, if we focus on the "clustered" categories we created to consolidate very similar RCDC terms, this percentage is lower, at 58 % (Fig. 1A and B). Across publications, an average of 40 % of the categories were not present in the underlying grant, which was reduced to 33 % when focusing on clustered categories (Fig. 1A and B). In the empirical analyses we also looked at the timing on cross-category spillovers (*Publication Lag*). On average, the publications were published 3.4 years after the start of the grant (in a small number of cases where the publication pre-dated the grant start year, we set the lag to zero). Fig. 1C and D shows that the distribution of the proportion of a publication's categories that were unexpected is bimodal, with most publications having zero unexpected categories but a substantial portion (13 % without clustering, 11 % with) where none of the publication's categories were assigned to the funding grant.

3.2. Regression analysis

We explored grant-level variables associated with cross-category spillovers.[5] We estimated linear probability models regressing whether

Table 1
Descriptive statistics (core project-level).

Variable	Observations	Mean	Std. Dev.	Min.	25th Perc.	Median	75th Perc.	Max.
RFA	89,754	0.134	0.341	0	0	0	0	1
Disease Oriented	90,962	0.815	0.388	0	1	1	1	1
Clinical Research	90,962	0.379	0.485	0	0	0	1	1
R01 (and equivalent)	90,962	0.434	0.496	0	0	0	1	1
Research Grants (R Series)	90,962	0.732	0.443	0	0	1	1	1
Career Development Grants (K Series)	90,962	0.086	0.281	0	0	0	0	1
Research Training Fellowships (T and F Series)	90,962	0.112	0.315	0	0	0	0	1
Program Center Grants (P Series)	90,962	0.007	0.086	0	0	0	0	1
Cooperative Agreements (U Series)	90,962	0.043	0.204	0	0	0	0	1
Total Cost (million $)	90,962	1.035	2.129	1e-6	0.278	0.533	1.383	165.299
Number of Grant Categories	90,962	5.857	3.048	1	4	6	8	20
FY Start	90,962	2012	2.628	2008	2010	2012	2014	2016

Note: R01 (and equivalent) includes R01, DP1, DP2, DP5, R37, R56, RF1, RL1 and U01 (https://grants.nih.gov/grants/funding/funding_program.htm).

Table 2
Descriptive statistics (grant-publication dyad-level).

Variable	Observations	Mean	Std. Dev.	Min	25th Perc.	Median	75th Perc.	Max
Unexpected Category	1,251,826	0.699	0.459	0	0	1	1	1
Unexpected Category – Clustered	1,251,826	0.584	0.493	0	0	1	1	1
Share of Unexpected Categories	1,251,826	0.396	0.349	0	0	0.333	0.667	1
Share of Unexpected Categories – Clustered	1,251,826	0.329	0.341	0	0	0.25	0.5	1
RFA	1,236,270	0.250	0.433	0	0	0	0	1
Disease Oriented	1,251,826	0.820	0.384	0	1	1	1	1
Clinical Research	1,251,826	0.524	0.499	0	0	1	1	1
R01 (and equivalent)	1,251,826	0.515	0.500	0	0	1	1	1
Research Grants (R Series)	1,251,826	0.607	0.488	0	0	1	1	1
Career Development Grants (K Series)	1,251,826	0.100	0.300	0	0	0	0	1
Research Training Fellowships (T and F Series)	1,251,826	0.051	0.219	0	0	0	0	1
Program Center Grants (P Series)	1,251,826	0.051	0.220	0	0	0	0	1
Cooperative Agreements (U Series)	1,251,826	0.171	0.377	0	0	0	0	1
Total Cost (million $)	1,251,826	3.861	9.904	1e-6	0.537	1.394	2.629	165.299
Number of Grant Categories	1,251,826	6.043	3.319	1	3	6	8	20
FY Start	1,251,826	2012	2.476	2008	2010	2012	2014	2016
Publication Lag	1,046,870	3.405	2.315	0	2	3	5	11

Note: R01 (and equivalent) includes R01, DP1, DP2, DP5, R37, R56, RF1, RL1 and U01.

[4] We found no evidence of major collinearity issues that may affect the econometric estimation presented later in the paper.

[5] We used OLS estimation for all our models. Results related to the unexpected categories variable (dummy) were qualitatively the same when using a logit model.

Table 3
Correlation matrix (grant-publication dyad-level).

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Unexpected Category															
2. Unexpected Category - Clustered	0.777														
3. Share of Unexpected Categories	0.747	0.717													
4. Share of Unexpected Categories – Clustered	0.633	0.815	0.889												
5. RFA	0.118	0.132	0.182	0.170											
6. Disease Oriented	−0.036	−0.072	−0.154	−0.162	0.013										
7. Clinical Research	0.060	0.046	0.001	−0.014	0.251	0.097									
8. Total Cost (million $)	0.093	0.112	0.169	0.161	0.373	0.071	0.229								
9. Number of Grant Categories	−0.136	−0.204	−0.386	−0.382	−0.050	0.370	0.260	−0.028							
10. Publication Lag	0.010	0.002	−0.019	−0.022	−0.003	−0.002	0.058	0.025	0.084						
11. R01 (and equivalent)	−0.134	−0.148	−0.203	−0.189	−0.240	0.046	−0.218	−0.182	0.129	0.090					
12. Research Grants (R Series)	−0.152	−0.167	−0.225	−0.209	−0.381	0.028	−0.336	−0.287	0.068	0.037	0.619				
13. Career Development Grants (K Series)	0.021	0.011	−0.024	−0.029	−0.086	0.056	0.215	−0.099	0.098	−0.011	−0.353	−0.437			
14. Research Training Fellowships (T and F Series)	0.027	0.040	0.067	0.073	−0.082	−0.073	−0.072	−0.075	−0.125	−0.007	−0.234	−0.290	−0.077		
15. Program Center Grants (P Series)	0.064	0.066	0.123	0.106	0.051	−0.084	0.063	−0.003	−0.083	−0.021	−0.228	−0.282	−0.075	−0.050	
16. Cooperative Agreements (U Series)	0.120	0.140	0.183	0.177	0.584	0.027	0.297	0.515	−0.028	−0.038	−0.229	−0.561	−0.149	−0.099	−0.096

a publication showed unexpected categories on our main measures of interest (whether it was disease-oriented research, clinical research, and for whether it resulted from an RFA) and a range of grant-level controls, as well as a publication-level variable (the lag from grant to publication). Table 4 shows the results. We examined this overall in Model 1 and for clustered categories in Model 2; Models 3 and 4 show the results when we focused on the proportion of categories assigned to a publication that were not in the underlying grant. All models include fixed effects indicating which of the 27 NIH Institutes or Centers funded the grants and for the starting year of the project (FY Start). Robust standard errors are clustered on the core grant number and reported in the parentheses.

Across all specifications, publications from larger grants were more likely to have unexpected categories, as were articles published longer after

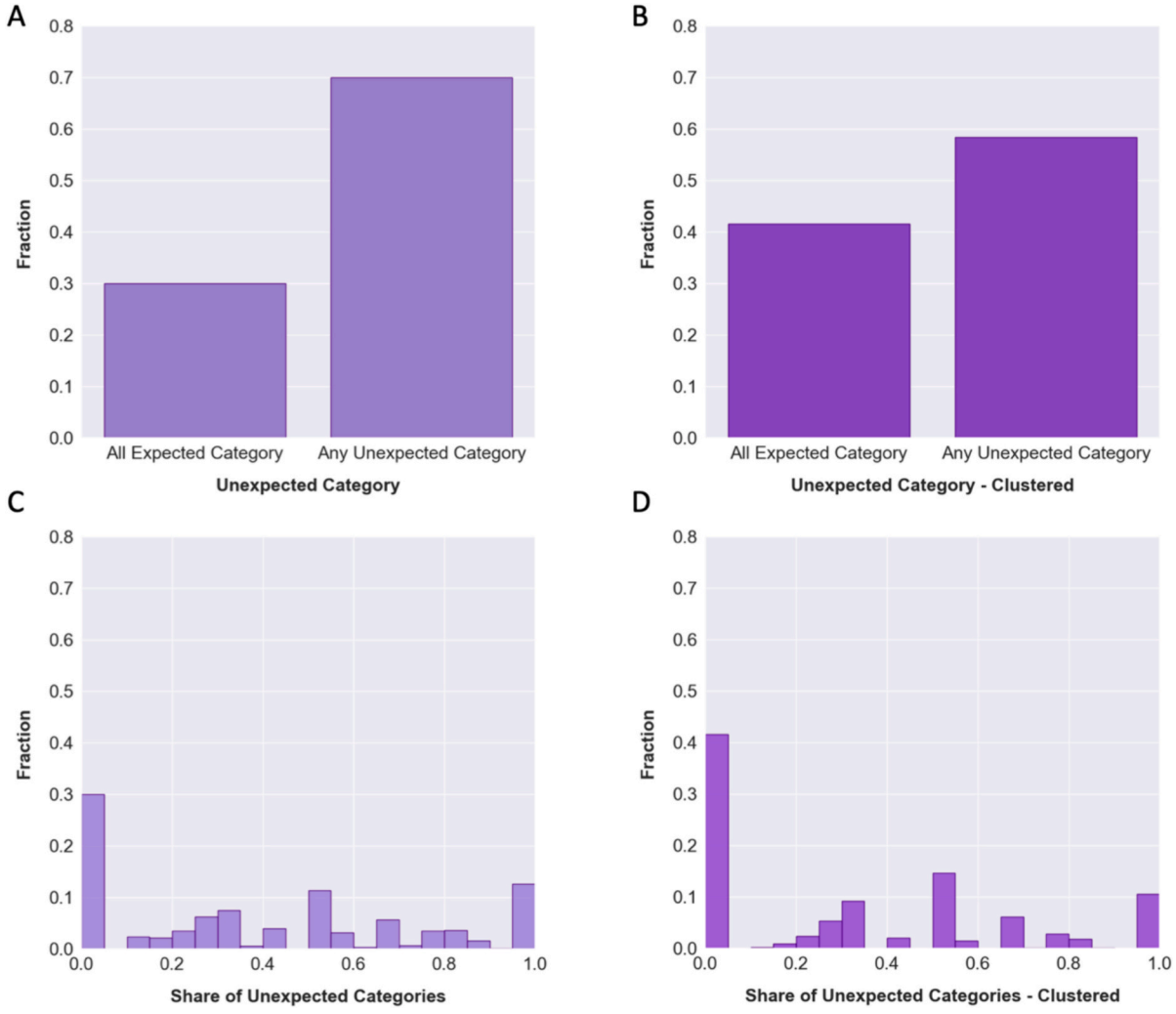

Fig. 1. Proportion of publications with (A) one or more unexpected categories or (B) one or more unexpected categories (clustered); distribution of (C) unexpected category proportions and (D) unexpected category proportions (clustered).

Table 4
OLS regression results including R01 (and equivalent) as a control variable.

	(1) Unexpected Category	(2) Unexpected Category – Clustered	(3) Share of Unexpected Categories	(4) Share of Unexpected Categories – Clustered
RFA	0.0435***	0.0473***	0.0350***	0.0299***
	(0.0035)	(0.0042)	(0.0042)	(0.0041)
Disease Oriented	−0.0002	−0.0123*	−0.0330***	−0.0390***
	(0.0045)	(0.0054)	(0.0051)	(0.0054)
Clinical Research	0.0258***	0.0205***	−0.0091**	−0.0182***
	(0.0031)	(0.0038)	(0.0035)	(0.0035)
Total Cost (million $)	0.0008***	0.0013***	0.0011***	0.0011***
	(0.0002)	(0.0003)	(0.0003)	(0.0003)
Number of Grant Categories	−0.0174***	−0.0269***	−0.0305***	−0.0291***
	(0.0005)	(0.0007)	(0.0006)	(0.0006)
Publication Lag	0.0087***	0.0095***	0.0063***	0.0058***
	(0.0004)	(0.0005)	(0.0004)	(0.0004)
R01 (and equivalent)	−0.0624***	−0.0663***	−0.0542***	−0.0478***
	(0.0029)	(0.0035)	(0.0031)	(0.0029)
Intercept	0.8970***	0.8599***	0.6828***	0.6011***
	(0.0125)	(0.0154)	(0.0140)	(0.0150)
Research Institutes	Included	Included	Included	Included
FY Start	Included	Included	Included	Included
Adjusted R-squared	0.0569	0.0879	0.2382	0.2273
Observations	1,032,629	1,032,629	1,032,629	1,032,629

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; Heteroskedasticity-robust standard errors, clustered on the originating grant, are reported in parentheses; R01 (and equivalent) includes R01, DP1, DP2, DP5, R37, R56, RF1, RL1 and U01.

the start of funding. Across the board, if a grant was associated with more categories, then there was a lower likelihood of any unexpected categories in the resulting publications, and a lower proportion of unexpected categories. Interestingly, across all specifications, R01 grants and their equivalents were less likely than other grants to yield publications with unexpected categories.

Turning next to our indicators for more applied research, the sign of the coefficient on clinical research varied by specification. For both clustered and unclustered RCDC categories, clinical grants were more likely to lead to publications with at least one unexpected category (Models 1 and 2). But, when the share of unexpected categories is considered, clinical research led to publications with slightly fewer unexpected categories (Models 3 and 4). Disease-oriented research also led to fewer unexpected categories across all models, though significance levels vary (Models 1 and 2). Surprisingly, grants resulting from RFAs were more likely to result in publications with unexpected categories, across the board.

Table 5 shows that the results were similar after including indicators for the major grant types. Table 6 shows that the results were

Table 5
OLS regression results with 1-label level activities.

	(5) Unexpected Category	(6) Unexpected Category – Clustered	(7) Share of Unexpected Categories	(8) Share of Unexpected Categories – Clustered
RFA	0.0278***	0.0261***	0.0212***	0.0157***
	(0.0038)	(0.0045)	(0.0042)	(0.0044)
Disease Oriented	0.0011	−0.0100	−0.0312***	−0.0367***
	(0.0042)	(0.0052)	(0.0049)	(0.0051)
Clinical Research	0.0143***	0.0070	−0.0188***	−0.0270***
	(0.0031)	(0.0037)	(0.0032)	(0.0032)
Total Cost (million $)	0.0005*	0.0009**	0.0009**	0.0008*
	(0.0002)	(0.0003)	(0.0003)	(0.0003)
Number of Grant Categories	−0.0175***	−0.0270***	−0.0299***	−0.0286***
	(0.0005)	(0.0006)	(0.0005)	(0.0005)
Publication Lag	0.0080***	0.0088***	0.0056***	0.0052***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Research Grants (R Series)	−0.0363***	−0.0171	−0.0471***	−0.0168
	(0.0086)	(0.0106)	(0.0108)	(0.0112)
Career Development Grants (K Series)	0.0268**	0.0496***	−0.0076	0.0164
	(0.0091)	(0.0112)	(0.0112)	(0.0115)
Research Training Fellowships (T and F Series)	0.0282**	0.0632***	0.0367**	0.0683***
	(0.0097)	(0.0118)	(0.0121)	(0.0125)
Program Center Grants (P Series)	0.1251***	0.1543***	0.1465***	0.1469***
	(0.0114)	(0.0143)	(0.0154)	(0.0153)
Cooperative Agreements (U Series)	0.0414***	0.0813***	0.0122	0.0440***
	(0.0097)	(0.0120)	(0.0122)	(0.0128)
Intercept	0.9181***	0.8685***	0.7088***	0.6076***
	(0.0144)	(0.0173)	(0.0161)	(0.0164)
Research Institutes	Included	Included	Included	Included
FY Start	Included	Included	Included	Included
Adjusted R-squared	0.0598	0.0914	0.2472	0.2347
Observations	1,032,629	1,032,629	1,032,629	1,032,629

Note: *p < 0.05, **p < 0.01, ***p < 0.001; Heteroskedasticity-robust standard errors, clustered on the originating grant, are reported in parentheses; R01 (and equivalent) includes R01, DP1, DP2, DP5, R37, R56, RF1, RL1 and U01.

Table 6
OLS regression results with R01 (and equivalent) grants only.

	(9) Unexpected Category	(10) Unexpected Category – Clustered	(11) Share of Unexpected Categories	(12) Share of Unexpected Categories – Clustered
RFA	0.0352***	0.0308***	0.0245***	0.0187***
	(0.0045)	(0.0050)	(0.0037)	(0.0034)
Disease Oriented	0.0179**	0.0117	−0.0066	−0.0064
	(0.0056)	(0.0065)	(0.0054)	(0.0050)
Clinical Research	−0.0019	−0.0157***	−0.0348***	−0.0457***
	(0.0035)	(0.0039)	(0.0028)	(0.0026)
Total Cost (million $)	0.0022**	0.0031**	0.0011	0.0011
	(0.0008)	(0.0010)	(0.0007)	(0.0006)
Number of Grant Categories	−0.0141***	−0.0223***	−0.0246***	−0.0234***
	(0.0006)	(0.0007)	(0.0006)	(0.0005)
Publication Lag	0.0071***	0.0076***	0.0035***	0.0034***
	(0.0004)	(0.0005)	(0.0004)	(0.0004)
Intercept	0.8252***	0.7192***	0.5707***	0.4831***
	(0.0287)	(0.0435)	(0.0307)	(0.0294)
Research Institutes	Included	Included	Included	Included
FY Start	Included	Included	Included	Included
Adjusted R-squared	0.0152	0.0272	0.0859	0.0887
Observations	533,930	533,930	533,930	533,930

Note: *p < 0.05, **p < 0.01, ***p < 0.001; Heteroskedasticity-robust standard errors, clustered on the originating grant, are reported in parentheses; R01 (and equivalent) includes R01, DP1, DP2, DP5, R37, R56, RF1, RL1 and U01.

qualitatively similar when focusing on R01 and equivalent grants only, the main NIH grant mechanism. Of note: for these grants, RFAs continued to be more likely to produce publications associated with unexpected categories, though the magnitudes were relatively small – about half of the magnitudes reported in the overall sample.

4. Conclusions

The extent to which biomedical research is targetable has been at the center of longstanding debates in science policy. We aimed to provide new evidence to inform these debates. We used machine learning approaches to develop new measures of how often NIH funding results in publications with categories not in the original grant. Our results indicate that the majority of publications exhibit at least some degree of unexpectedness (70 %, and 58 % with clustered categories) and that, for the average publication, a substantial proportion of all its assigned categories are unexpected (40 %, and 33 % with clustered categories). This does not just reflect noise in classification. Our simulations[6] suggest that, after allowing for noise, publications still show one or more unexpected categories (66 %, and 55 % with clustered categories) and a considerable average proportion of unexpected categories (35 %, and 30 % with clustered categories). When restricting our sample to only the top 10 %, top 5 %, and top 1 % cited publications, our results change by a mere 1.4 % at most.[7]

Whether one considers the percentages we outlined above to be high or low indications of unexpectedness in research would depend on one's priors. For example, Azoulay et al. (2019) estimated the causal impact of NIH-funding on private sector patenting. They found that over half of the patents were not in the original disease area, with "disease area" broadly defined on the basis of the funding NIH institute/center. Our results – using publications, not patents, and fine-grained RCDC categories – also suggest substantial spillovers.

The finding that 70 % of publications have one or more unexpected categories assigned to them would seem to lend support to assertions by Bush, Van Slyke and others about spillovers in research. However, publications can fit multiple categories, some expected and others unexpected (as shown in Fig. 1, lower panel). Thus, in parallel, our results also support the ideas of Kilgore, disease advocates, and others arguing for more targeted and strategic allocation of research funding towards specific diseases. In particular, the case for research-targeting finds some support in our finding that most publications reflected at least one category from their underlying grant, with only 13 % of publications having only unexpected categories (11 % when clustering). This dual interpretation is a reflection of multilabel classification, in which most publications are assigned multiple categories. More generally, any individual research project can have on-target and off-target outcomes.

[6] The performance of our machine learning model varies at the RCDC category-level. Specifically, the model is more likely to correctly assign certain categories than others to publication records. This performance relates to the feature extraction step (see Appendix A) and how some categories may be more readily associated with specific terms that are clearly extractable from the training dataset. We therefore examined the extent to which the potential misallocation of categories (false positives) affects the percentage of all publications with any unexpected categories and the proportions of unexpected categories in publications. To estimate this, we first calculated the misallocation error of each category as one minus the corresponding category-level precision of our model (or as the ratio between false positives and the sum of true and false positives). We then considered all grant-publication dyads with at least one unexpected category, dropped each unexpected category with a probability equal to the corresponding category-level misallocation error, and recalculated the two dependent variables. We repeated this operation 100 times and estimated the resulting distribution of our dependent variables. Publications with one or more unexpected categories reduced from 70 % (58 % for clustered) to 66 % (55 % for clustered); and, the average proportion of unexpected categories in a publication reduced from 40 % (33 % for clustered) to 35 % (30 % for clustered), all averaged across the 100 iterations and all with a standard deviation 0.02 % or less. The approach is conservative in the sense that it did not consider false negatives i.e., missed unexpected categories that should have been, but were not, allocated to publication records by our model.

[7] We matched the publication sample with OpenAlex data (Priem et al., 2022). We searched for all PMIDs associated with our publications by querying OpenAlex through its API and the R package *openalexR* (Aria et al., 2024). Citation data were retrieved for 99.3 % of the publications and normalised by year and OpenAlex's subfields (https://help.openalex.org/how-it-works/topics, last accessed May 13, 2024). This allowed us to identify publications in the sample that are highly cited – top 10 %, top 5 %, and top 1 % – in the corresponding year and subfield(s). Our main measures remain robust to restricting to only top 10 %, top 5 % and top 1 % cited publications. When adding a control variable for the top 10 %, top 5 % and top 1 % cited publications, our regression results also changed little from our main regression results (these three sets of regression results are available on request).

Turning next to the correlates of spillovers, our results suggest that two measures of applied research (disease-orientation and clinical research) are less likely to be associated with spillovers, consistent with the longstanding idea that basic research is more likely to result in unexpected results (Nelson, 1959; Arrow, 1962). However, across the board, research targeted via RFAs was more likely to be associated with spillovers. Although the magnitude of this effect was small (as with clinical- and disease-oriented research), our results do seem to rule out that RFA-funded grants are less likely than investigator-initiated grants to generate spillovers. This resonates with previous work (e.g., Myers, 2020), which found that RFAs are three times as productive in terms of its publication rate, than investigator-initiated grants. We use the RFA indicator as a broad measure of "targeted" research. In future work, it may be useful to distinguish between RFAs resulting from the activities of disease-advocates and Congress (to target research on specific diseases) versus those from NIH-program managers (to target science that the NIH-peer review process penalizes under the normal bottom-up investigator-initiated model, including potential "breakthrough" research).

The measures we have developed and used in this study, implemented via machine learning run across a large sample of publications, may be helpful in future research. For example, they could be useful in analyses relating NIH funding to disease burden (Gross et al., 1999). Is it possible, for example, that diseases that are underfunded relative to disease burden generate higher cross-disease spillovers? The measures may also be useful in exploring the characteristics of researchers who are willing to go off-target and pursue unexpected findings, and any penalties associated with such pivoting (Hill et al., 2021). One possible comparison could be between extramural and intramural researchers, although such comparisons would need to account for the fact that the program is small (less than 10 % of total NIH funding) and that its aims and approaches could be substantially different from the extramural program. Additionally, one could explore the feasibility of using RCDC to examine broader spillovers in clinical trials, patents, start-up enterprises, and pharmaceuticals (see for example, Blume-Kohout (2012) and Choi et al. (2022)).

Our results here are descriptive, not causal. In future work, researchers might also exploit natural experiments in the history of the NIH to identify the effects of policies and institutional factors on spillovers. Another set of limitations relates to categorization. We used the NIH's RCDC to define cross-category movements, using machine-learning to extend a system originally developed to classify research funding to instead classify research publications. Although RCDC is an official categorization, it is imperfect (like all categorizations, see for example Bowker and Star, 1999). A major issue is similarity across some of the categories, which we tried to address via clustering. We also validated RCDC-based measures of spillovers versus our own manual categorization using another classification scheme (the International Classification of Diseases), and found high overlap. It would be useful to see if our general results, on the proportion of grants with spillovers and the correlates of these spillovers, replicate using other categorization systems, such as MeSH. There is also a deeper issue: what we are calling "unexpected" on the basis of text in the grants and publications may not in fact be unexpected for the investigators, or others in the field. Future research could make progress on these issues by using interviews and surveys to validate text-based measures of unexpected spillovers.

CRediT authorship contribution statement

Yasemin Aslan: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Ohid Yaqub:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Bhaven N. Sampat:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Daniele Rotolo:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We thank Kyle Myers for additional checks on some of our descriptive statistics, Julie Weeds for advice on algorithm development, Duncan Moore for excellent research assistance, and participants at the 2023 Atlanta Conference on Science and Innovation Policy and the 2023 EuSPRI Annual Conference for their feedback. All errors belong to us. We acknowledge funding from ERC grant 759897.

Appendix A. Machine-learning to assign RCDC categories to publications

The NIH assigns RCDC codes to grants starting from FY 2008, but RCDC codes are not assigned to the corresponding publications. We trained a range of machine-learning algorithms on the grant abstracts (2008–2016), using 90 % of the data. We used the remaining 10 % to compare the performance of the algorithms.

Our data are characterised by multilabel classes that are non-mutually exclusive, i.e. each grant record can have more than one RCDC category. This implies that different grants may have the same RCDC category, and/or similar and dissimilar/less similar RCDC categories (there is a certain level of similarity among certain pairs of RCDC categories, see Appendix D). Therefore, we applied a multilabel classification by adopting 11 different algorithms with different vectorizers and parameters. We used the "sklearn" package in Python. We then tested for the accuracy of each algorithm, as described later in this Appendix.

Text pre-processing. The abstracts of grants were tokenized – i.e., text data were broken down into words, symbols, or other meaningful elements. The text was then cleaned – i.e., punctuation, stopwords, and generic terms not representing the topic of a grant (e.g., "description", "provided by applicant") were removed. We used the "nltk.corpus" module in Python to delineate the lexicon of stopwords. Data on RCDC category labels were also cleaned to remove extremely minor variations in labelling. The NIH's downloadable files listed 355 categories (unique text strings) when we queried the RePORTER database on Aug 17, 2022. These were manually matched with the categories listed on the NIH website. For example, we unified "smoking and health" (taken from database) and "Tobacco Smoke and Health" (taken from website) as "Tobacco Smoke and Health". As another example, we unified "Nanotechnology - Nanoscale devices and systems" (taken from database) and "Nanotechnology" (taken from website) as

"Nanotechnology". This cleaning step resulted in 283 categories allocated to our grant records to the years 2008–2016.

Feature extraction. We encoded text data (abstracts of grant records) as a vector of numerical values by using three different vectorizers as feature extraction: a count vectorizer (CV), a hashing vectorizer (HV) and a term frequency-inverse document frequency (tf-idf) vectorizer. CV counts how frequently each word appears in an abstract while tf-idf normalizes these frequencies by decreasing the effect of common words across the text corpus (Kowsari et al., 2019). Thus, tf-idf lessens the effect of the terms appearing very often in all abstracts and having little information for a specific abstract. HV is also a variation of CV and produces results faster with a lower memory, and is more scalable to large datasets by storing the tokens as numerical indexes instead of strings with a hashing function (see Kaur et al., 2020). This data processing was performed using "CountVectorizer", "TfidfVectorizer", "HashingVectorizer" from the "sklearn.feature_extraction.text" module in Python.

Classification. We applied two traditional algorithms of text categorization – multinomial naïve Bayesian and logistic regression – as probabilistic classifiers (Rogers and Girolami, 2016; Kowsari et al., 2019). We also applied a support vector machine (SVM), specifically the linear support vector classifier, as a non-probabilistic classifier for stronger empirical performance (Rogers and Girolami, 2016; Kowsari et al., 2019). We further applied stochastic gradient descent as another example for linear models, because linear classifiers work well for text classification (Kowsari et al., 2019; Kaur et al., 2020) – for a detailed description of these methods, see Rogers and Girolami (2016), Kowsari et al. (2019), and Kaur et al. (2020). In Python, we used "MultinomialNB" from the "sklearn.naive_bayes" module, "LogisticRegression" and "SGDClassifier" from the "sklearn.linear_model" module, and "LinearSVC" from the "sklearn.svm" module; while, for modelling purposes, we used "MultiOutputClassifier" from the "sklearn.multioutput" module.

Performance of classification. We used traditional performance-evaluation metrics such as recall, precision and F-measure with macro-averaging (Rogers and Girolami, 2016; Kowsari et al., 2019), which are based on combinations of true/false positives/negatives. These metrics, however, provide an incomplete approach to evaluating the classification performance of a machine learning model for our dataset that features multilabel classes and non-mutually exclusive categories. In this context, a classification may not necessarily be wholly correct or incorrect. For example, if only a subset of the actual categories is predicted correctly, then this is better than predicting no categories correctly. Predicting a higher proportion of the subset correctly is preferred. Moreover, a misclassification may be more or less incorrect based on its similarity to the actual category. A misclassification with a category that is similar to the correct category is preferred. We address these two issues respectively by using two non-traditional performance indicators. First, we used Jaccard similarity. This metric aims to assess the extent to which machine-allocated categories match with the actual categories in the grant:

$$Jacard\ similarity = \frac{count(actual\ categories \cap predicted\ categories)}{count(actual\ categories \cup predicted\ categories)}$$

The Jaccard similarity metric has several properties that are useful to our classification challenge: (i) the metric increases with the increase of correctly predicted categories and ranges from a minimum value of 0 (no category correctly predicted) to a maximum value of 1 (full match between actual and predicted categories); (ii) the metric is independent from the number of categories since it normalizes the matching value with the number of maximum unique categories.

Second, we used the adjusted cosine similarity. While some RCDC categories are dissimilar (e.g. "tobacco" and "lyme disease"), others are highly similar (e.g. "tobacco" and "tobacco and health") or somewhat similar (e.g. "breast cancer" and "cancer"). More traditional metrics would consider similar categories as different categories. On this basis, those cases where the machine learning model allocates a category such as "breast cancer" to a grant record with an actual category "cancer" would be considered as not a match by traditional metrics. The adjusted cosine similarity aims to address this limitation by also considering the similarity between categories and it works as follows:

1. Defining the sets of RCDC categories:
   1.1. $A = \{a_1, a_2, \ldots, a_i, \ldots, a_M\}$ represents the set of RCDC categories that are assigned by the NIH to a grant with $M > 0$;
   1.2. $P = \{p_1, p_2, \ldots, p_j, \ldots, p_N\}$ represents the set of RCDC categories that are assigned to a grant by machine learning with $N > 0$.

Computing distances between actual and predicted RCDC categories: We compute the matrix of cosine similarities (squared) C between each pair of categories in A and P. This matrix has M rows and N columns.

2. Identifying the closest RCDC category pairs:
   2.1. Identify those pairs of RCDC categories (one from set A and one from set P) with the highest cosine similarity values. Remove these matched categories from both sets.
   2.2. Repeat Steps 2 and 3.1 until no more RCDC categories can be matched between A and P.
   2.3. If RCDC categories still remain in A or P, consider all categories in the other set, and repeat Steps from 2 to 3.2.
3. **Computing the adjusted cosine similarity score**: Add up all the identified cosine similarity values (squared) and divide this sum by the maximum length between A and P.

To clarify the two metrics, we present an example. We consider a grant record with categories "breast cancer" and "cancer" (i.e. actual categories) to which are machine learning-allocated the categories are "breast cancer", "cancer", and "clinical research" (i.e. predicted categories). We assume that the cosine similarity (squared) values between pairs of categories are those reported in Table A1 – cosine similarities between RCDC categories are assessed on the basis of the co-occurrence of categories as assigned to grant records.

Table A1
Example of cosine similarity (squared) matrix.

Category	Breast cancer	Cancer	Clinical research
Breast cancer	1.00	0.68	0.20
Cancer	0.68	1.00	0.30
Clinical research	0.20	0.30	1.00

For such a case, the Jaccard similarity is 0.67, while the adjusted cosine similarity is 0.77 as detailed below:

$$Jacard\ similarity = \frac{count(breast\ cancer, cancer)}{count(breast\ cancer, cancer, clinical\ research)} = \frac{2}{3} = 0.67$$

$$Adjusted\ cosine\ similarity == \frac{2(matches : breast\ cancer, cancer) + 0.30(clinal\ research\ vs.cancer)}{max(2,3)} = 0.77$$

The joint use of the Jaccard similarity and adjusted cosine similarity metrics is helpful for evaluating the overall performance of machine learning model. Although these metrics are correlated, they are not identical. Table A2 provides a range of possibilities, and varying performance metrics.

Table A2
Examples of Jaccard similarity and adjusted cosine similarity metrics.

Grant	Actual categories	Predicted categories	Jaccard similarity	Adjusted cosine similarity
Grant 1	Breast cancer; cancer	Breast cancer; cancer; clinical research	0.67	0.77
Grant 2	Breast cancer; cancer; clinical research	Breast cancer; cancer	0.67	0.77
Grant 3	Breast cancer; cancer; HIV/AIDS	Breast cancer; cancer	0.67	0.72
Grant 4	Breast cancer; cancer; genetics	Breast cancer; cancer	0.67	0.81
Grant 5	Biotechnology; breast cancer; cancer	Biotechnology; breast cancer; cancer; genetics	0.75	0.92
Grant 6	Cancer	Breast cancer; cancer	0.50	0.84
Grant 7	Breast cancer; cancer; genetics; prevention	Breast cancer; cancer	0.50	0.68
Grant 8	Breast cancer; cancer; complementary and alternative medicine; nutrition	Breast cancer; cancer	0.50	0.60
Grant 9	Breast cancer; cancer	Biotechnology; breast cancer; cancer; estrogen; genetics	0.40	0.62
Grant 10	Breast cancer; cancer	Biotechnology; cancer	0.33	0.61
Grant 11	Cancer	Breast cancer; cancer; genetics	0.33	0.70
Grant 12	Breast cancer; cancer	Breast cancer; cancer	1.00	1.00

As noted, we applied 11 different algorithms, with various combinations of CV, HV and tf-idf for vectorizing the tokens of abstracts as feature extraction, and multinomial naïve Bayesian, logistic regression, linear support vector and stochastic gradient descent for classification. Table A3 compares the performance of the different algorithms, using Jaccard and adjusted cosine similarity metrics.

We selected the linear support vector classifier with tf-idf vectorizer (with parameters C = 4 and class_weight = None) since this machine learning model achieved the highest values for predicting the test dataset – for example, it can predict RCDC categories assigned to grant records with an accuracy of 91 % according the median values of the adjusted cosine similarity metric. The accuracy varies across grant records (Figs. A1 and A2) and RCDC categories (Fig. A3).

Table A3
Comparison of the results of different machine learning algorithms.

Algorithm	Comparison between trained and actual data			
	Jaccard similarity (median)	Jaccard similarity (mean)	Adjusted Cosine similarity (median)	Adjusted Cosine similarity (mean)
SGD, TF-IDF	0.778	0.743	0.910	0.865
SGD, HV	0.750	0.725	0.899	0.855
SGD, CV	0.750	0.717	0.898	0.847
Linear SVC, TF-IDF	0.800	0.745	0.912	0.866
Linear SVC, HV	0.778	0.735	0.906	0.860
Linear SVC, CV	0.750	0.729	0.902	0.855
Logistic Regression, TF-IDF	0.800	0.742	0.911	0.863
Logistic Regression, CV	0.778	0.731	0.904	0.857
Logistic Regression, HV	0.714	0.696	0.881	0.835
Multinomial NB, TF-IDF	0.500	0.502	0.762	0.718
Multinomial NB, CV	0.450	0.459	0.747	0.718

Note: HV: hashing vectorizer, CV: count vectorizer, TF-IDF: term frequency-inverse document frequency; SGD: stochastic gradient descent, SVC: support vector classifier, NB: naïve Bayesian.

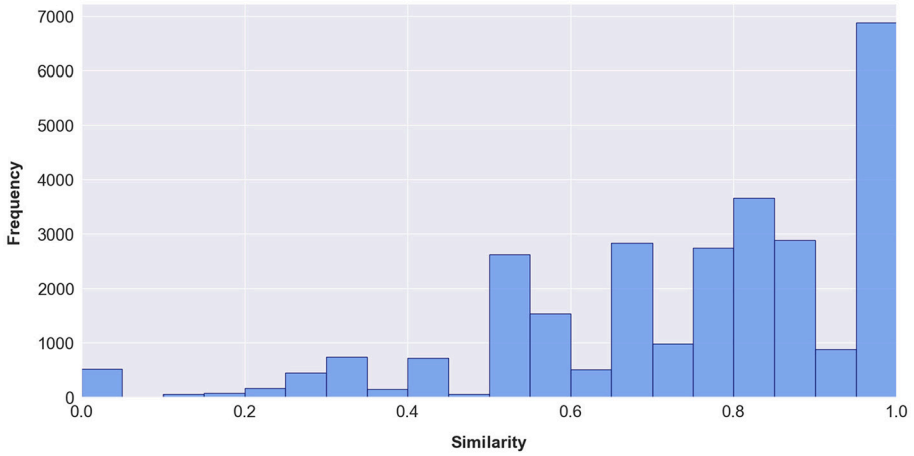

Fig. A1. Distribution of the Jaccard similarity (actual vs. predicted RCDC categories) across grant records.

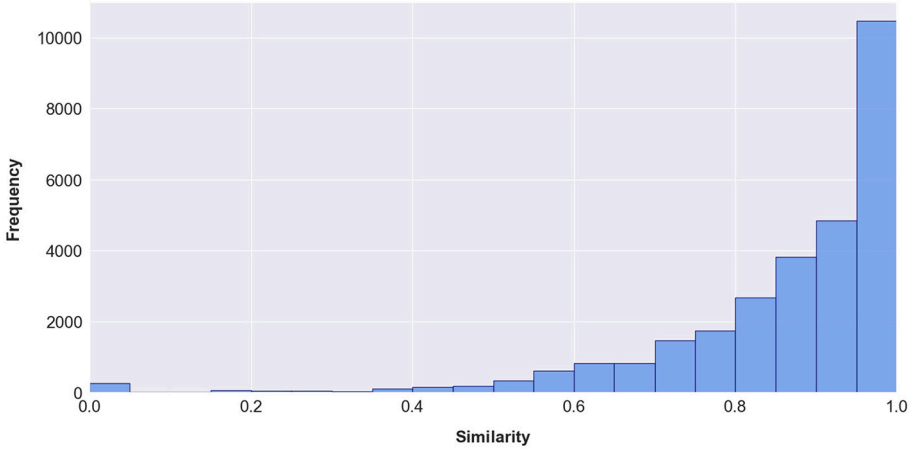

Fig. A2. Distribution of the adjusted cosine similarity (actual vs. predicted RCDC categories) across grant records.

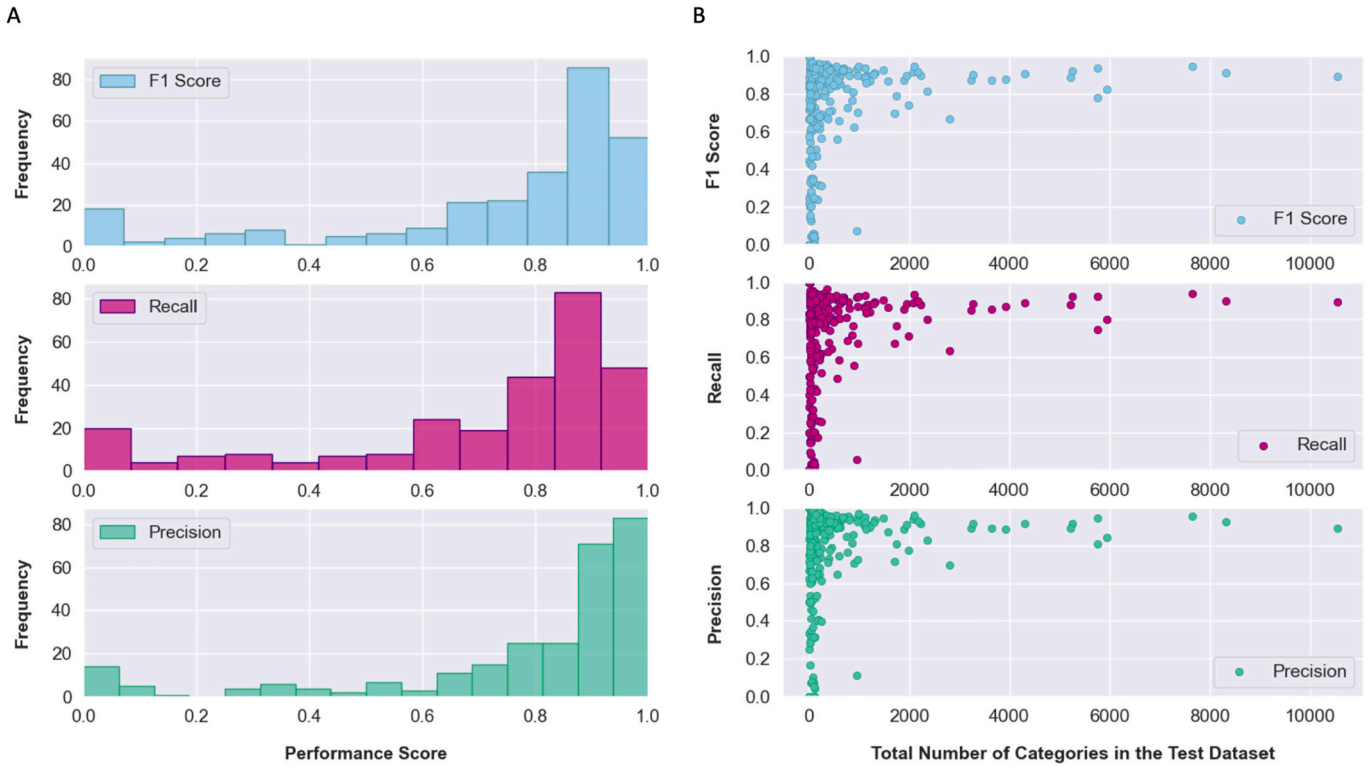

Fig. A3. Distribution of F1 score, recall, and precision (actual vs. predicted RCDC categories) across RCDC categories.

Appendix B. Disease-related RCDC categories

We delineated a number of RCDC categories associated with diseases to examine the extent to which our results of unexpectedness may depend on more 'applied' research. We first identified categories for which the NIH provides information on disease burden (i.e. 74 categories). Two reviewers then independently hand-coded RCDC categories and added further disease-related categories – discrepancies between the reviewers (5 categories) were resolved by a third reviewer. This resulted in 171 disease-related RCDC categories (Table B1).

Table B1
List of identified disease-related RCDC categories.

Disease-related RCDC categories		
• Acquired cognitive impairment	• Endometriosis	• Neurofibromatosis
• Acute respiratory distress syndrome	• Epilepsy	• Obesity
• Alcoholism, alcohol use and health	• Eye disease and disorders of vision	• Osteoarthritis
• Allergic rhinitis (hay fever)	• Facioscapulohumeral muscular dystrophy	• Osteogenesis imperfecta
• Als	• Fetal alcohol syndrome	• Osteoporosis
• Alzheimer's disease	• Fibroid tumors (uterine)	• Otitis media
• Alzheimer's disease including Alzheimer's disease related dementias (AD/ADRD)	• Fibromyalgia	• Ovarian cancer
• Alzheimer's disease related dementias (ADRD)	• Food allergies	• Paget's disease
• Anorexia	• Foodborne illness	• Pancreatic cancer
• Anxiety disorders	• Fragile x syndrome	• Parkinson's disease
• Aphasia	• Frontotemporal dementia (FTD)	• Pediatric aids
• Arthritis	• Headaches	• Pediatric cancer
• Asthma	• Heart disease	• Pediatric cardiomyopathy
• Ataxia telangiectasia	• Heart disease - coronary heart disease	• Pelvic inflammatory disease
• Atherosclerosis	• Hepatitis	• Perinatal period - conditions originating in perinatal period
• Attention deficit disorder (ADD)	• Hepatitis - a	• Peripheral neuropathy
• Autism	• Hepatitis - b	• Pick's disease
• Autoimmune disease	• Hepatitis - c	• Pneumonia
• Back pain	• HIV/AIDS	• Pneumonia & influenza
• Batten disease	• Hodgkin's disease	• Polycystic kidney disease
• Bipolar disorder	• Huntington's disease	• Post-traumatic stress disorder (PTSD)
• Brain cancer	• Hydrocephalus	• Prescription drug abuse
• Brain disorders	• Hypertension	• Preterm, low birth weight and health of the newborn
• Breast cancer	• Infant mortality	• Prostate cancer
• Cachexia	• Infectious diseases	• Psoriasis
• Cancer	• Infertility	• Rare diseases
• Cerebral palsy	• Inflammatory bowel disease	• Rett syndrome
• Cervical cancer	• Influenza	• Reye's syndrome
• Charcot-Marie-tooth disease	• Injury - childhood injuries	• Rheumatoid arthritis
• Childhood leukemia	• Injury - trauma - (head and spine)	• Sarcopenia
• Chronic fatigue syndrome (ME/CFS)	• Injury - traumatic brain injury	• Schizophrenia
• Chronic liver disease and cirrhosis	• Injury - unintentional childhood injury	• Scleroderma
• Chronic obstructive pulmonary disease	• Injury (total) accidents/adverse effects	• Sepsis
• Chronic pain	• Intellectual and developmental disabilities (IDD)	• Serious mental illness
• Colo-rectal cancer	• Interstitial cystitis	• Sexually transmitted infections
• Conditions affecting the embryonic and fetal periods	• Kidney disease	• Sickle cell disease
• Congenital heart disease	• Lead poisoning	• Small pox
• Congenital muscular dystrophy	• Lewy body dementia	• Spina bifida
• Congenital structural anomalies	• Liver cancer	• Spinal cord injury
• Cooley's anemia	• Liver disease	• Spinal muscular atrophy
• Creutzfeldt-Jakob	• Lung cancer	• Stroke
• Crohn's disease	• Lupus	• Substance abuse
• Cystic fibrosis	• Lyme disease	• Sudden infant death syndrome
• Dementia	• Lymphoma	• Temporomandibular muscle/joint disorder (TMJD)
• Dental/oral and craniofacial disease	• Macular degeneration	• Tourette syndrome
• Depression	• Major depressive disorder	• Transmissible spongiform encephalopathy (TSE)
• Diabetes	• Malaria	• Tuberculosis
• Digestive diseases	• Mental illness	• Tuberous sclerosis
• Digestive diseases - (gallbladder)	• Migraines	• Urologic diseases
• Digestive diseases - (peptic ulcer)	• Mucopolysaccharidoses (MPS)	• Usher syndrome
• Down syndrome	• Multiple sclerosis	• Uterine cancer
• Duchenne/Becker muscular dystrophy	• Muscular dystrophy	• Vaginal cancer
• Dystonia	• Myasthenia gravis	• Valley fever
• Eating disorders	• Myotonic dystrophy	• Vascular cognitive impairment/dementia
• Eczema/atopic dermatitis	• Neonatal respiratory distress	• Vector-borne diseases
• Emerging infectious diseases	• Neuroblastoma	• Vulvodynia
• Emphysema	• Neurodegenerative	• West Nile virus

Appendix C. Assigning International Classification of Diseases (ICD) codes

We validated our RCDC-based measures of unexpectedness for a set of disease-oriented grants. To do so, we first randomly sampled a set of disease-oriented grants (see Appendix B). We then asked two reviewers to classify these grants and the resulting publications in the categories of the International Classification of Diseases, 10th edition (ICD10). The reviewers relied on the ICD10 browser (https://icd10cmtool.cdc.gov/?fy=FY2022, accessed Dec 21, 2022.) The coding guide provided to the reviewers is reproduced below:

We want to assign ICD10 codes to each text. Use an ICD10 browser to help assign codes. The text to be categorized will consist of a title and abstract and will have an ID code. The text may relate to a grant or a publication, though in either case they should be treated the same and categorized in the same way, as follows:

1. For each ID row, read the title and abstract. Then, select which top level ICD10 codes best describe what the content is about and enter them into the ICD10 chapter column (e.g., chapter XI; XIV). Replace roman numerals with digits (e.g., 11; 14). If multiple codes are applicable, then separate them with semicolons. If none are, then leave this blank.
2. Then, select which two-digit level ICD10 codes best describe what the content is about and enter them into the ICD10 column (e.g., N18; K74). If multiple codes are applicable, then separate them with semicolons. If none are, then leave this blank.
3. Enter the two-digit level ICD10 code (e.g., K74) that most applies into ICD10 priority. Priority codes should designate the main disease that the text is oriented towards. For example, this might be obtainable from the title alone, or from specific statements about past or planned work. Ideally, this should be one code but multiple codes are acceptable if this is not possible. Leave blank if none apply.
4. The ICD10 browser enables you to search with key terms and retrieves suggested codes for you to select from. Search terms entered into the browser could be disease names, or other terms taken from the context. These could be terms that indicate that the direct outputs of a project are expected to substantially relate to a disease, or that the results being reported and discussed from the work undertaken substantially relate to a disease.
5. If you use any other information outside of the title and abstract, such as full texts of publications or further project information, please indicate this in the notes field.
6. Repeat for all rows.

Appendix D. Clustering similar RCDC categories

Some categories are quite similar (e.g., “tobacco” and “tobacco smoke and health”). As a result, some cases of spillover could be a reflection of the similarity between certain RCDC categories. We address aggregating similar RCDC categories into clustered RCDC categories – in particular, we identified 32 clustered RCDC categories. To do so, we first generated the co-occurrence network of RCDC categories in grants. We then produced the corresponding cosine similarity (squared) network. We relied on the *igraph* package in R (Kolaczyk and Csárdi, 2014). This resulted in the network depicted in Fig. D1. This network includes 283 nodes (RCDC categories) and 39,902 edges (cosine similarity (squared) between RCDC categories). To delineate clusters of densely connected nodes, we applied the Louvain method for community detection to the sub-network having edges with weight (cosine similarity (squared)) higher than 0.2 – we performed this analysis in Gephi (https://gephi.org). With a resolution of 0.44, the algorithm returned 32 clusters with a modularity value of 0.129 (see Fig. D1 and Table D1).

Although we performed the analysis by adopting different thresholds for the selection of edges (i.e. cosine similarity (squared) higher than 0.1, 0.2, 0.3, and 0.5) and different clustering resolutions (e.g. 0.6, 0.5, 0.4), the above clustering results generate more meaningful clusters in terms of size and meaning of RCDC categories (i.e. they reflect a good compromise between maximizing similarity within each cluster while minimizing similarity between the clusters) and retain a large number of nodes and edges. For example, a cluster analysis based on a sub-network with edges whose weight is higher than 0.5 returns a higher modularity value (0.78). However, this solution retains only 1.5 % of the initial sample of edges and delineates a large number of small clusters (i.e. 63 clusters many of which with one node). A cluster analysis based on a sub-network with edges whose weight is higher than 0.2 with a resolution of 1.0 also returns a higher modularity value (0.51) but only 7 large clusters of RCDC categories.

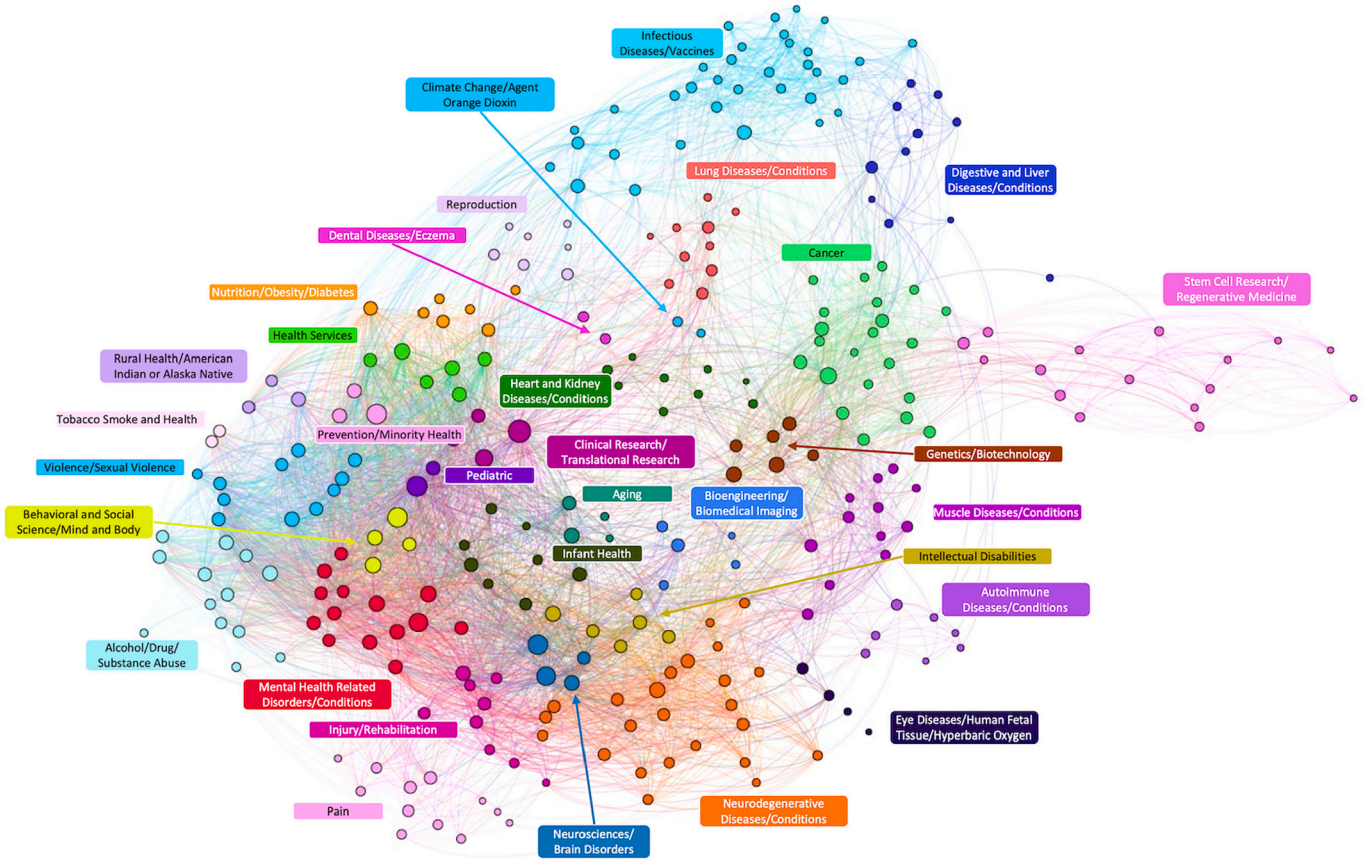

Fig. D1. Similarity network of RCDC categories (2008–2016). Nodes represent RCDC categories (the size of each node is proportional to its degree centrality), edges represent the cosine similarity (squared) between pairs of RCDC categories (edges with cosine similarity (squared) higher than 0.2 are depicted), and colours correspond to the resulting clusters of RCDC categories.

Table D1
Correspondence table: clustered and unclustered RCDC categories.

Clustered RCDC categories	RCDC categories
Aging	Aging; caregiving research; osteoporosis; sarcopenia
Alcohol/Drug/Substance Abuse	Alcoholism, alcohol use and health; cannabidiol research; cannabinoid research; drug abuse (NIDA only); endocannabinoid system research; homelessness; methamphetamine; prescription drug abuse; screening and brief intervention for substance abuse; substance abuse; substance abuse prevention; therapeutic cannabinoid research; underage drinking
Autoimmune Diseases Conditions	Arthritis; autoimmune disease; lupus; myasthenia gravis; psoriasis; rheumatoid arthritis; scleroderma
Behavioral and Social Science/Mind and Body	Basic behavioral and social science; behavioral and social science; mind and body; sleep research
Bioengineering/Biomedical Imaging	Assistive technology; bioengineering; biomedical imaging; nanotechnology; networking and information technology R&D
Cancer	Brain cancer; breast cancer; cachexia; cancer; cancer genomics; childhood leukemia; colo-rectal cancer; Hodgkin's disease; lung cancer; lymphoma; neuroblastoma; neurofibromatosis; orphan drug; ovarian cancer; pancreatic cancer; pediatric cancer; precision medicine; prostate cancer; radiation oncology; rare diseases; urologic diseases; uterine cancer
Climate Change/Agent Orange Dioxin	Agent orange & dioxin; climate change
Clinical Research/Translational Research	Clinical research; clinical trials and supportive activities; effectiveness research; translational research
Dental Diseases/Eczema	Dental/oral and craniofacial disease; eczema/atopic dermatitis
Digestive and Liver Diseases/Conditions	Chronic liver disease and cirrhosis; Crohn's disease; digestive diseases; digestive diseases - (gallbladder); hepatitis; hepatitis - b; hepatitis - c; inflammatory bowel disease; liver cancer; liver disease; organ transplantation
Eye Diseases/Human Fetal Tissue/ Hyperbaric Oxygen	Chemical preparedness and decontamination activities; eye disease and disorders of vision; human fetal tissue; hyperbaric oxygen
Genetics/Biotechnology	Biotechnology; gene therapy; gene therapy clinical trials; genetic testing; genetics; human genome
Health Services	Burden of illness; comparative effectiveness research; cost effectiveness research; emergency care; health services; patient safety
Heart and Kidney Diseases/Conditions	Atherosclerosis; cardiovascular; congenital heart disease; heart disease; heart disease - coronary heart disease; hypertension; kidney disease; pediatric cardiomyopathy; polycystic kidney disease
Infant Health	Cerebral palsy; conditions affecting the embryonic and fetal periods; fetal alcohol syndrome; infant mortality; neonatal respiratory distress; perinatal period - conditions originating in perinatal period; preterm, low birth weight and health of the newborn; sudden infant death syndrome
Infectious Diseases/Vaccines	Anthrax; antimicrobial resistance; biodefense; cervical cancer; digestive diseases - (peptic ulcer); emerging infectious diseases; food defense; foodborne illness; hepatitis - a; HIV/AIDS; HPV and/or cervical cancer vaccines; immunization; infectious diseases; influenza; Lyme disease; malaria; malaria vaccine; otitis media; pelvic inflammatory disease; pneumonia; pneumonia & influenza; Reye's syndrome; sepsis; sexually transmitted infections; small pox; topical microbicides; tuberculosis; tuberculosis vaccine; vaccine related; vaccine related (AIDS); vaginal cancer; valley fever; vector-borne diseases; West Nile virus

(*continued on next page*)

Table D1 (*continued*)

Clustered RCDC categories	RCDC categories
Injury/Rehabilitation	Aphasia; injury - trauma - (head and spine); injury - traumatic brain injury; injury - unintentional childhood injury; injury (total) accidents/adverse effects; physical rehabilitation; rehabilitation; spinal cord injury; stroke
Intellectual Disabilities	Autism; down syndrome; fragile x syndrome; intellectual and developmental disabilities (IDD); Rett syndrome; tuberous sclerosis
Lung Diseases/Conditions	Acute respiratory distress syndrome; allergic rhinitis (hay fever); asthma; chronic obstructive pulmonary disease; climate-related exposures and conditions; cystic fibrosis; emphysema; health effects of household energy combustion; health effects of indoor air pollution; lung
Mental Health Related Disorders/ Conditions	Anorexia; anxiety disorders; attention deficit disorder (add); bipolar disorder; depression; eating disorders; major depressive disorder; mental health; mental illness; post-traumatic stress disorder (PTSD); schizophrenia; serious mental illness; suicide; suicide prevention
Muscle Diseases/Conditions	congenital muscular dystrophy; congenital structural anomalies; Duchenne/Becker muscular dystrophy; facioscapulohumeral muscular dystrophy; mucopolysaccharidoses (MPS); muscular dystrophy; myotonic dystrophy; osteogenesis imperfecta; Paget's disease; spina bifida; spinal muscular atrophy; usher syndrome
Neurodegenerative Diseases/Conditions	Acquired cognitive impairment; als; Alzheimer's disease; Alzheimer's disease including Alzheimer's disease related dementias (AD/ ADRD); Alzheimer's disease related dementias (ADRD); ataxia telangiectasia; batten disease; cerebrovascular; Charcot-Marie-tooth disease; Creutzfeldt-Jakob; dementia; dystonia; epilepsy; frontotemporal dementia (FTD); Huntington's disease; Lewy body dementia; macular degeneration; multiple sclerosis; Parkinson's disease; neurodegenerative; peripheral neuropathy; Pick's disease; transmissible spongiform encephalopathy (TSE); vascular cognitive impairment/dementia
Neurosciences/Brain Disorders	Brain disorders; hydrocephalus; neurosciences; Tourette syndrome
Nutrition/Obesity/Diabetes	Complementary and alternative medicine; diabetes; dietary supplements; food allergies; nutrition; obesity; physical activity
Pain	Back pain; chronic fatigue syndrome (ME/CFS); chronic pain; fibromyalgia; headaches; interstitial cystitis; migraines; neck pain; osteoarthritis; pain research; temporomandibular muscle/joint disorder (TMJD); vulvodynia
Pediatric	lead poisoning; pediatric
Prevention/Minority Health	health disparities; minority health; prevention
Reproduction	Contraception/reproduction; diethylstilbestrol (DES); endocrine disruptors; endometriosis; estrogen; fibroid tumors (uterine); infertility
Rural Health/American Indian or Alaska Native	American Indian or Alaska Native; Arctic; rural health
Stem Cell Research/Regenerative Medicine	Cooley's anemia; hematology; regenerative medicine; sickle cell disease; stem cell research; stem cell research - embryonic - human; stem cell research - embryonic - non-human; stem cell research - induced pluripotent stem cell; stem cell research - induced pluripotent stem cell - human; stem cell research - induced pluripotent stem cell - non-human; stem cell research - nonembryonic - human; stem cell research - nonembryonic - non-human; stem cell research - umbilical cord blood/placenta; stem cell research - umbilical cord blood/ placenta - human; stem cell research - umbilical cord blood/placenta - non-human; transplantation
Tobacco Smoke and Health	Tobacco; tobacco smoke and health
Violence/Sexual Violence	Adolescent sexual activity; child abuse and neglect research; homicide and legal interventions; injury - childhood injuries; pediatric aids; sexual and gender minorities (SGM/LGBT*); teenage pregnancy; violence against women; violence research; youth violence; youth violence prevention

References

Aria, M., Le, T., Cuccurullo, C., Belfiore, A., Choe, J., 2024. openalexR: an R-tool for collecting bibliometric data from OpenAlex. R J. 15, 167–180.

Arrow, K.J., 1962. Economic welfare and the allocation of resources for invention. In: Nelson, R.R. (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton University Press, Princeton, pp. 609–625.

Azoulay, P., Graff Zivin, J.S., Li, D., Sampat, B., 2019. Public R&D investments and private-sector patenting: evidence from NIH funding rules. Rev. Econ. Stud. 86, 117–152.

Best, R.K., 2012. Disease politics and medical research funding: three ways advocacy shapes policy. Am. Sociol. Rev. 77, 780–803.

Blume-Kohout, M.E., 2012. Does targeted, disease-specific public research funding influence pharmaceutical innovation? J. Policy Anal. Manag. 31, 641–660.

Bowker, G., Star, S., 1999. Sorting Things Out: Classification and Its Consequences. MIT Press.

Choi, H., Yoon, H., Siegel, D., Waldman, D.A., Mitchell, M.S., 2022. Assessing differences between university and federal laboratory postdoctoral scientists in technology transfer. Res. Policy 51, 104456.

Gross, C.P., Anderson, G.F., Powe, N.R., 1999. The relation between funding by the National Institutes of Health and the Burden of Disease. N. Engl. J. Med. 340, 1881–1887.

Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61, 29–48.

Hegde, D., Sampat, B., 2015. Can private money buy public science? Disease group lobbying and federal funding for biomedical research. Manag. Sci. 61, 2281–2298.

Hill, R., Yin, Y., Stein, C., Wang, D., Jones, B.F., 2021. Adaptability and the Pivot Penalty in Science arXiv preprint arXiv:2107.06476.

Jimeno-Yepes, A.J., Plaza, L., Mork, J.G., Aronson, A.R., Díaz, A., 2013. MeSH indexing based on automatically generated summaries. BMC Bioinform. 14, 208.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (Eds.), Machine Learning: ECML-98. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 137–142.

Kaur, S., Kumar, P., Kumaraguru, P., 2020. Automating fake news detection system using multi-level voting model. Soft. Comput. 24, 9049–9069.

Kevles, D.J., 1977. The National Science Foundation and the debate over postwar research policy, 1942–1945: a political interpretation of science–the endless frontier. Isis 68, 5–26.

Kolaczyk, E., Csárdi, G., 2014. Statistical Analysis of Network Data With R. Springer, New York.

Kowsari, K., Meimandi, Kiana Jafari, Heidarysafa, Mojtaba, Mendu, Sanjana, Barnes, Laura, Brown, Donald, 2019. Text classification algorithms: a survey. Information 10, 150.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 159–174.

Li, D., Azoulay, P., Sampat, B., 2017. The applied value of public investments in biomedical research. Science 356, 78–81.

Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S., 2012. An extensive experimental comparison of methods for multi-label learning. Pattern Recogn. 45, 3084–3104.

Mowery, D.C., 1997. The bush report after fifty years—blueprint or relic. In: Barfield, C. E. (Ed.), Science for the Twenty-First Century: The Bush Report Revisited. American Enterprise Institute, Washington.

Myers, K., 2020. The elasticity of science. Am. Econ. J. Appl. Econ. 12, 103–134.

Nelson, R.R., 1959. The simple economics of basic scientific research. J. Polit. Econ. 67, 297–306.

Nelson, R.R., 1997. Why the bush report has hindered an effective civilian technology policy. In: Barfield, C.E. (Ed.), Science for the Twenty-First Century: The Bush Report Revisited. American Enterprise Institute, Washington.

Priem, J., Piwowar, H., Orr, R., 2022. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts, ArXiv. https://arxiv.org/abs/2205.01833.

Rettig, R., 1977. Cancer Crusade: The Story of the National Cancer Act of 1971. Princeton University Press, Princeton, New Jersey.

Rogers, S., Girolami, M., 2016. A First Course in Machine Learning. Taylor and Francis Group. https://doi.org/10.1201/9781315382159.

Sampat, B., 2012. Mission-oriented biomedical research at the NIH. Res. Policy 41, 1729–1741.

Sampat, B., Buterbaugh, K., Perl, M., 2013. New evidence on the allocation of NIH funds across diseases. Milbank Q. 91, 163–185.

Stokes, D., 1997. Pasteur's Quadrant: Basic Science and Technological Innovation. Brookings Institution Press, Washington DC.

Van Slyke, C.J., 1946. New horizons in medical research. Science 104, 559–567.

Varmus, H., 1997. Testimony on Setting Research Priorities at NIH. Before the Subcommittee on Public Health and Safety Committee on Labor and Human Resources. United States Senate.

Varmus, H., 2001. Proliferation of National Institutes of Health. Science 291, 1903–1905.