Using ontology-based annotation to profile disease research

Yi Liu,1 Adrien Coulet,1,2 Paea LePendu,1 Nigam H Shah1

ABSTRACT
Background Profiling the allocation and trend of research activity is of interest to funding agencies, administrators, and researchers. However, the lack of a common classification system hinders the comprehensive and systematic profiling of research activities. This study introduces ontology-based annotation as a method to overcome this difficulty. Analyzing over a decade of funding data and publication data, the trends of disease research are profiled across topics, across institutions, and over time.

Results This study introduces and explores the notions of research sponsorship and allocation and shows that leaders of research activity can be identified within specific disease areas of interest, such as those with high mortality or high sponsorship. The funding profiles of disease topics readily cluster themselves in agreement with the ontology hierarchy and closely mirror the funding agency priorities. Finally, four temporal trends are identified among research topics.

Conclusions This work utilizes disease ontology (DO)-based annotation to effectively the landscape of biomedical research activity. By using DO in this manner a use-case driven mechanism is also proposed to evaluate the utility of classification hierarchies.

Scientists from varied fields are interested in profiling research activity4–6 and in identifying innovation hotspots.4–6 In bibliometrics, research activity is profiled by analyzing citation and publication patterns. Journals are quantified for impact, using the impact factor, which measures the mean citations received per paper, for papers published in the two most recent years.7 The use of citation counts to evaluate impact is common beyond the evaluation of journals. For example, metrics like the h-index are designed to measure the research impact of an individual researcher, and are calculated using the citations of a researcher’s publications.5–10

In the biomedical sciences, profiling of funding and publication is common,11–15 and researchers profile funding and publications, across research topics, across institutions, over time, or a combination of these. For example, Argawal and Serls6 profiled demand and supply factors in the drug discovery cycle by analyzing relationships between funding, publications, and disease burden, per therapeutic area. They concluded that the analysis of literature can support decision-making for the drug discovery cycle.

It is essential—even if at a crude level—to profile research activity as well as innovation because in the light of limited resources, funding agencies must prioritize among competing projects. At the National Institutes of Health (NIH), such profiling—also called portfolio analysis—is supported by the Office of Portfolio Analysis16 within the Division of Program Coordination, Planning, and Strategic Initiatives. Several existing initiatives catalog and analyze grants, publications, and citations: the Thomson Reuters Web of Knowledge focuses primarily on publications and the NIH research portfolio online reporting tool (RePORT) catalogs grants and publications that are linked to grants.17

However, most initiatives, including RePORT, do not yet enable integrated mining of publications and grants within a single framework because they do not share a common terminology for indexing grants and publications. This disconnect, which we will address in this paper, arises because efforts tracking research funding and publications have different objectives and thus use different vocabularies. For example, NIH grants in RePORT are tagged with terms from the research, condition, and disease categorization. PubMed publications are tagged with terms from medical subject headings (MeSH). This disconnect also affects the integration and interpretation of machine learned topic maps such as the NIH visual browser18 because the topics identified in grants may not necessarily correspond exactly to topics learnt from publications.

To compare activity across research topics, the publication and funding data need to be tagged using a common hierarchy of research topics. In the domain of biomedical research, human disease hierarchies provide a plausible categorization that can provide the necessary, shared frame of reference to analyze publications and grants jointly. One solution then is to tag both publications and grants with a common hierarchy of human diseases—ie, a disease ontology (DO). Based on the mentions of disease topics found in grant summaries and publication abstracts, we can tag (or annotate) these documents with disease terms. Such annotation is similar in spirit to gene ontology annotation, but is done automatically. After tagging with terms from a DO, we can profile research activity on a per-topic basis using the diseases terms as the ‘joining keys’ to integrate the grants and publications datasets.

Over 250 bio-ontologies are available in the BioPortal maintained by the National Center for Biomedical Ontology (NCBO),19 including the human DO, which is a manually curated ontology of disease concepts derived from the UMLS metathesaurus.20 The NCBO also creates annotation tools that allow users to annotate textual metadata with ontology terms and integrate disparate data resources.19 21

1 Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA
2 LORIA—INRIA Nancy, Vandoeuvre-lès-Nancy Cedex, France

Received 10 October 2011 Accepted 8 March 2012 Published Online First 11 April 2012

Correspondence to
Yi Liu, Stanford University, Stanford Center for Biomedical Informatics, 1265 Welch Road, Medical School Office Building, Room X215 Mail Code 5479, Stanford, CA 94305, USA; liuyipei@stanford.edu

Additional materials are published online only. To view these files please visit the journal online (www.jamia.bmj.com/content/19/e1.toc).
In this work, our objective is to demonstrate that the profiling of research activity in a domain of interest can be conducted using existing ontologies, without requiring manual assignment of terms. We use the human DO and annotate publications and grants as a proof of concept for the ability to perform profiling of research activity using automated annotation. We profile research activity spanning grants and publications in terms of sponsorship, allocation, co-funding, and trends over time. We explore the interplay between funding and publications and the allocation of research activity with respect to disease burden using an existing DO as a reference.

METHODS

Data

We obtained a courtesy export of a database of grants from 1997 to 2007 for 83 funding institutions, listed in supplementary table S1 (available online only), from Research Crossroads (RC).\(^22\) RC is a company, which aims to increase the transparency of publicly funded research. They collect and clean up grant data from NIH, National Science Foundation, Health Resource and Service Administration, and a number of not-for-profit organizations.\(^22\) We use their data to avoid redoing the tedious clean up. This dataset was then supplemented with an export of the Scholarly Database (SDB) from Katy Börner’s group at the University of Indiana.\(^23\)

The coverage of funding institutions and range of time are not identical between the data received from RC and the SDB. While RC-provided data dump goes up to 2007, the SDB contains data up to 2010. The SDB does not contain data from Health Resource and Service Administration, Food and Drug Administration, National Aeronautic and Space Administration, Centers for Disease Control and Prevention (CDC), and Department of Defense. In contrast, the export from RC contains funding information for these funding institutions. We decided to restrict our analysis to the end of 2007 in order to be consistent and comprehensive across funding institutions. Our resulting database contains the grant sources, amounts, fiscal years, and recipients (supplementary figure S1, available online only). Between 1997 and 2007, our dataset contains funding data on US$327 billion across 81 858 grants. US$157 billion of the funding correspond to grants that are annotated with at least one disease term. Among the US$157 billion, US$101 billion are allocated to the top 110 institutions.

We use publication data from PubMed, which organize publications into types such as review, journal articles, meta-analysis and editorials. Journal articles include the majority of research results and clinical trials—and we restrict our analysis to those. Temporally, we restrict our analysis to the time range between 1997 and 2007. Finally, to limit the scope of our analysis to research activities in the USA, we consider only articles that are affiliated with US research institutions.

Given the publication type, temporal, and geographical constraints, we use approximately 2.4 million journal articles. As the names for a given institution appear in varying representations, we normalize them through the use of regular expressions. The regular expressions are described in supplementary table S1 (available online only). We ignore department level granularity, which is not consistently available. Among journal articles affiliated with research institutions in the USA, approximately 994 000 of them are tagged with a disease term; and 441 000 of these articles are affiliated with the top 110 funding recipient institutions in the USA (supplementary figure S2, available online only).

The names of research institutions are not used consistently across the funding and publication data; therefore, to map research institutions across the grants and publications, we normalize alternative representations for the top 110 recipient institutions (ranked by funding) by a combination of manual inspection and regular expressions. Among these institutions, we distinguish highly related organizations (eg, Harvard University versus Brigham and Women’s Hospital). In these cases, the publication and funding can appear unbalanced if the academic institution attributed for a grant is different from the institution attributed for the corresponding publication.

Ideally, there should be a temporal lag between funding times and publication times. However, because this lag is difficult to quantify across journals, research topics and across a decade, we directly align the funding and publication data with no time offset.

Annotation workflow

The annotator web service recognizes ontology terms in user-submitted text and returns the recognized terms, which we call annotations, back to the user.\(^21\) For example, each experimental series in the gene expression omnibus has a title and a summary describing the data, analysis, and conclusion that can be used to annotate the gene expression omnibus.\(^22\) By systematically tagging elements from multiple biomedical resources, the NCBO created the NCBO resource index,\(^23\) in which elements from biomedical resources are linked to over 5 million terms via 16.4 billion annotations. Using the same workflow, we annotate our database of grants and PubMed articles with terms from the DO. The free text metadata for a grant consist of the grant’s title and summary, while that of a publication consist of the publication’s title and abstract.

Ontologies in biomedicine may refer to hierarchical controlled vocabularies, information models, or repositories of axiomatic logic.\(^26\) The DO provides a subsumption hierarchy of ‘is-a’ relationships between terms. Each term has a preferred name, and usually a set of synonyms. When the preferred name or a synonym of a term appears in the free text metadata of a document, the document is subsequently tagged with that term and all of the term’s ancestors in the ontology. We interpret each term and its corresponding synonyms as a research topic.

For example, ‘Alzheimer’s disease’ is a term in the DO and it has ‘Alzheimer’s dementia’ as a synonym. When the phrase ‘Alzheimer’s disease’ or ‘Alzheimer’s dementia’ is mentioned in a publication abstract, we tag the given abstract with the DO term ‘Alzheimer’s disease’ (DOID:10652). The DO hierarchy specifies that ‘Alzheimer’s disease’ is both a type of ‘taupathy’ and a type of ‘dementia’; therefore, we tag the abstract with the terms ‘dementia’ and ‘taupathy’. We recursively follow the subsumption hierarchy, until reaching the root term, and tag the given abstract with all terms encountered on the path to the root (figure 1A).

We tag publication abstracts and grant summaries with terms from the DO to create ontology-based annotations.\(^27\) \(^28\) While numerous ontologies provide disease hierarchies, we chose the DO for several reasons. First, the DO has a very permissive license—it is licensed under a creative commons attribution 3.0 unported license.\(^29\) The DO is mapped back to UMLS concepts in BioPortal, which enumerates many plural forms, acronyms and alternative names.\(^30\) We treat these additional strings as synonyms in our annotation workflow. The DO borrows heavily from MeSH and systematized nomenclature of medicine—clinical terms (SNOMED CT),\(^20\) \(^31\) and contains more disease terms than MeSH.\(^20\) Compared with SNOMED CT, the DO is more tractable to perform computation on because the DO
excludes terms that are rarely used and thus is significantly smaller in size. For the purpose of this study, the DO provides sufficient granularity.

While analyzing grants, we track funding quantities because the amount of the funds provided, rather than the count of grants, is a better measure of research activity. Analogously, when analyzing annotations on publications, we track the impact factors of the publications’ journals. Each individual article is published in a specific journal, and thus has a corresponding impact factor value. The impact factor-weighted count of a list of journal articles is computed by summing the impact factor values of each article in the list. Using these weighted counts as opposed to raw counts emphasizes publications in high impact journals over those in lower impact journals and is a better representation of scientific activity in a research area.

Figure 1B,C shows our annotation workflow, which uses a dictionary-based term recognizer called Mgrep developed at the University of Michigan. For recognizing disease terms, Mgrep is more accurate and more efficient than MetaMap—Mgrep had 0.77 precision when annotating publication abstracts with disease terms. Mgrep’s performance depends on the ontology used to create the dictionary and the source of the textual data annotated—the findings from five studies are described in supplementary table S2 (available online only). Overall precision is reported to be between 60% and 95% and recall between 79% and 93%.

Our workflow uses BioPortal mappings to retrieve additional synonyms for each DO term from the UMLS. For example, the British spelling ‘haemoglobinopathy’ is absent in the DO; nevertheless BioPortal recognizes it as a synonym of a term in SNOMED CT, which in turn is mapped to ‘hemoglobinopathy’ in the DO. The alternative spelling of ‘haemoglobinopathy’ thus maps to the DO term ‘hemoglobinopathy’.

One drawback of this approach is that the inclusion of certain short acronyms out of context may lead to spurious annotations. To alleviate this drawback, acronyms with length three or less are not included in our analysis. For example, aortic stenosis has the acronym AS, which is a short acronym. We use version 1.238 of DO, which has 9438 terms.

Analysis
After tagging each grant with ontology terms, we produce a table of funding by research topic, with information on year,
funding institution, and research institution. This table allows us to profile research funding quantitatively. By applying an analogous process to publications, we produce an analogous table for publications. The two tables share common keys—which are the research topic, research institution, and year—permitting a quantitative profiling of the interplay between publications and funding.

Sponsorship is one way of quantifying the interplay between publications and funding. For a topic, we define sponsorship as the US dollar funding amount divided by the impact factor weighted publication count for that topic. For each disease topic, we calculate its annual funding, its annual impact factor weighted publications, and subsequently compute a sponsorship level.

In addition to sponsorship, we also profile the allocation of research activity in relation to the problem size—which we quantify using disease mortality rates. After computing per-topic funding and per-topic publication levels, we cross-reference these research activity levels to mortality rates on a per-disease basis. This comparison quantifies the allocation of research resources against mortality rates (which are a surrogate for the ‘size of the problem’). The mortality rates are obtained from external sources: the CDC published a report for 2007 causes of mortality within the USA,36 and the WHO published a similar report for 2004 causes of mortality worldwide.37 We use the 2004 WHO data and the 2007 CDC data because more recent datasets are not yet finalized. Supplementary table S5 (available online only) shows the mapping from the causes of mortality to DO terms.

Our database of grants identifies the funding institution of each grant, allowing us to profile individual funding agencies. For each research topic, and each funding agency, we calculate the total funding amount. As shown in figure 2, these quantities can be represented in a matrix in which rows correspond to research topics and columns correspond to funding institutions. Analysis of sponsorship is meaningful for broad, well-funded research topics; thus we restrict our analysis to topics that received at least US$100 000 of funding from 1997 to 2007. By reducing the number of concepts examined, this restriction also improves the tractability of our workflow. Two thousand nine hundred and eighty-five topics pass this constraint, so the topic by funding institution matrix is of size $2985 \times 33$ (figure 2A).

To visualize the similarities between the 33 funding institutions’ research portfolios, we represent their Pearson correlations as a heatmap. Correlation between funding institutions is computed by comparing profiles of the institutions across each research topic. Profiles are numeric vectors representing US dollars allocated, and the correlations between these vectors are Pearson correlations. When measuring similarity between numeric vectors representing the profiles, Pearson’s correlation is more appropriate because it scales in proportion to the actual numeric quantities in the profiles. Upon this correlation matrix, we apply MATLAB’s implementation of unweighted pair group method with arithmetic mean using Euclidean distances. This produces a heatmap in which clusters of granting institutions visually emerge as highly correlated blocks on the diagonal (figure 2B). Analogously, we apply the same process to the rows of the topic by funding institution matrix to generate a $2985 \times 2985$ heatmap of Pearson correlations between research topics (figure 2C).

Finally, we identify and summarize clusters of research topics defined by their temporal funding trends (figure 2D). Each research topic has a temporal funding trend, and we compute distance between trends as one minus their Pearson correlation.

For the temporal analysis, we only plot topics with a mean annual funding of US$1 million; 1613 topics pass this additional constraint. As we track the topics over 11 years, the time by topic matrix is of size $11 \times 1613$. To determine the ideal number of clusters in a systematic manner, we apply the gap statistic38 to determine that four clusters of research topics are present in our data. We then apply k-means clustering to determine the members of these four clusters.

**RESULTS**

We demonstrate that by using the DO as a common vocabulary for annotation, we are able to profile funding and publications on a common framework. We identify and profile research topics of high sponsorship along with their leading research institutions. We cross-reference leading causes of mortality with the funding and publication data to understand the allocation of research activity across institutions. Furthermore, we profile similarities between research topics and between funding agencies. Finally, we identify clusters of trends in research funding. These profiles provide a unique disease-centric view of the landscape of research activity by combining publication and funding in a common framework. We argue that such an integrated analyses is possible primarily because of the use of a common vocabulary—namely the DO—to annotate both funding and publication data in a consistent manner. Such profiles of research activity can help policy-makers understand the impacts of the research they fund.
Degree of sponsorship of research topics

To study research activity, we track funding amounts and impact factor weighted publication counts across research topics. For each disease topic, we divide the funding by weighted publication counts for the topic area to create a measure, which we refer to as sponsorship (described in the Methods section). Supplementary figure S3 (available online only) presents a plot of logarithmically binned mean annual sponsorship across DO topics with at least 100 annual impact factor weighted publications and US$10 000 of annual funding between 1997 and 2007. We make these constraints on funding and publications because the analysis is more reliable for topics with higher funding and more publications. Notably, the two extreme tails of the distribution differ by two orders of magnitude. Table 1 presents the top 10 highly sponsored research topics. Supplementary figure S4 (available online only) represents this hierarchically. Drug abuse, Alzheimer’s disease, retroviridae infectious disease, and pervasive development disorder are all highly sponsored research topics. It is also interesting that cancer is not among the disease topics with the highest levels of sponsorship. Compared with research topics with lower sponsorship, like cancer or heart disease, the highly sponsored topics in table 1 attract more attention from funding agencies despite appearing infrequently in the abstracts of high impact publications. High-sponsorship disease topics are interesting because they have extreme levels of skew between funding and publications for some reason. Sometimes, the discrepancy can be explained by differences in the cost of doing research (eg, studying Alzheimer’s disease is more expensive than informatics research). These research areas may be very new and thus do not have the volume of publications. Alternatively, they may be scientifically ‘stuck’, and in spite of significant effort no major breakthroughs are occurring. Although our current analysis is unable to discern the differences between these cases, it is nevertheless worthwhile to identify the high-sponsorship topics.

Figure 3 shows research activity in high-sponsorship diseases. From figure 3, we see that Yale and Baylor College of Medicine lead in pervasive development disorder research. In research of retroviral infectious diseases, Johns Hopkins and University of California San Francisco lead. For drug abuse research, Yale leads in publications while Johns Hopkins leads in funding. Finally, in the research of Alzheimer’s disease, University of Pennsylvania and University of California San Diego lead.

Research topics with top sponsorship

Table 1 lists the top research topics ordered by sponsorship. Only topics with at least 100 annual impact factor weighted publications and over US$50 million of annual funding between 1997 and 2007 are included in this overview table. This constraint is applied to avoid repeating variants of the same topic at varying granularities of the DO.

Allocation of research versus disease burden

While it is interesting to examine high-sponsorship diseases topics, sponsorship cannot determine whether the research attention corresponds to the size of the problem. For diseases, one way to quantify the size of the problem is by measures of the disease burden, an umbrella concept encompassing the overall impact of a health problem. Disease burden may be quantified by numerous measurements, including mortality, morbidity, years lost to disability, quality-adjusted life-years, and disability-adjusted life-years. We use mortality as the measurement because there is little ambiguity in its measurement.

The WHO published a list of top causes of mortality worldwide for 2004. This is the most recent and comprehensive in their publication series on the global burden of disease. The causes of mortality across diseases serve as surrogates for disease burden. To profile research allocation, we compare the proportions of mortality caused by diseases against research activities in those disease areas. Figure 4A plots funding and publication activities for top causes of mortality worldwide. Severe worldwide problems like cerebrovascular diseases receive only moderate levels of research. Diarrheal diseases appear under-researched because other disease topics (eg, nephritic syndromes, malaria, and tuberculosis) have similar or higher levels of research activity but are responsible for fewer deaths. A caveat to interpreting such trends is that the hierarchy used by the WHO to aggregate mortality data may lead to bias. For example, cholera is a ‘diarrheal disease’, which we mapped to ‘diarrhea’; the DO lists cholera under infectious diseases and thus misses cholera-related grants when counting diarrhea grants. However, we expect most of the top causes of mortality such as cancer, heart diseases, and cerebrovascular diseases—which are also high-level terms in the DO—to agree (see supplementary table S2, available online only).

Analogous to the WHO dataset for worldwide mortality, the CDC tracks mortality rates within the USA. The CDC dataset on US mortality provides an alternative to the WHO worldwide data. Although 2009 data are also available, they are preliminary and we decided to use the most recent finalized data from 2007. Based on figure 4B, heart disease is a huge burden in the USA. When compared with cancer, heart disease has higher mortality rates, but lower research activity. Such a skew might also be caused by the advances in cancer treatments—due to the higher
research activity—which result in lower mortality; making cancer appear like a ‘smaller problem’ when measured purely based on mortality.

Given these major areas of disease burden, it is of interest to identify leaders in these research areas. Supplementary figures S5 and S6 (available online only) identify leaders of research activities in the major causes of mortality for the USA and the world, respectively. Harvard, Johns Hopkins, and Mayo Clinic are among the leaders for research activity in most of the leading causes of mortality in the USA and worldwide. Supplementary figure S5 (available online only) shows that in cancers in particular, the University of Texas MD Anderson Cancer Centre leads far ahead of competitors in both funding and publications. Supplementary figure S6 (available online only) shows that Johns Hopkins University, Mayo Clinic, Brigham and Women’s Hospital, and University of Washington are consistently among

Figure 3  Research activity in high-sponsorship areas. Research activity in high-sponsorship research topics selected from table 1 is presented in scatter plots of publications against funding. We select the more general topics in disease ontology when multiple overlapping topics appear in table 1. For example, alcohol abuse is a specific kind of drug abuse, so we only plot drug abuse. We plot only institutions ranking within either the top five funding or the top five for publications.

Figure 4  Funding, publications, and disease load. The left panel (A) shows annual US dollar funding and annual impact factor weighted publications on a per-disease basis. The sizes of the bubbles correspond to the relative disease burdens worldwide. Disease burden is characterized by worldwide mortality statistics provided by the WHO for 2004. The right panel (B) shows annual US dollar funding and annual impact factor weighted publications on a per-disease basis. The sizes of the bubbles correspond to the relative disease burdens within the USA. Disease burden is characterized by US mortality statistics provided by the Centers for Disease Control and Prevention for 2007.
the leading research institutes for worldwide leading causes of mortality. For research in cardiovascular disease, which is the top killer worldwide, Duke University, Columbia University, University of Pittsburgh, University of Pennsylvania, and Stanford University are also among the leaders in research activity.

Clusters and trends in research funding
We visualize the topic by topic Pearson correlation matrix (described in the Analysis section) as a heatmap shown in figure 5. The topic by topic heatmap shows that research topics organize themselves into clusters based on their funding sources. The clustering in topic—topic heatmap tends to follow the DO is-a hierarchy. While partly due to the aggregation of our workflow, this correspondence is also sensible in the interpretation that similar diseases have similar sources of funding.

Inter-cluster co-funding relationships also emerge in this visualization, represented by highly correlated submatrices away from the diagonal. The smaller heatmap in figure 5 shows heavily co-funded topics from two different disease groups. The first group is highly coherent, consisting primarily of dementia and its various forms, like vascular dementia, Alzheimer’s disease, tauopathy, and Lewy body disease. The second group of diseases relate to the brain. However, these diseases are less coherent, ranging from neuromuscular diseases to Down’s syndrome to prion diseases. Despite the differences, these two disease groups have heavily co-funded profiles.

We also profile similarities between funding institutions. Our data contain 33 funding institutions. Figure 2B and supplementary figure S7 (available online only) show a $33 \times 33$ funding institution by funding institution heatmap colored by the Pearson correlations between their respective funding profiles—which show the degree to which they co-fund topics. National Institute on Drug Abuse (NIDA) and National Institute on Alcohol Abuse and Alcoholism (NIAAA) form a tight

![Figure 5](https://academic.oup.com/jamia/article-abstract/19/e1/e177/728255)
As both institutions deal with substance abuse of some sort, this similarity builds confidence in our approach. In fact, there is a proposition to merge the two institutes.39 We also see a large general cluster at the lower left corner, which includes National Center for Research Resources (NCRR), National Library of Medicine (NLM), National Institute of Nursing Research (NINR), National Human Genome Research Institute (NHGRI) and National Institute of General Medical Sciences (NIGMS) indicating synergistic co-funding between these institutions.

Finally, we profile research topics in terms of their temporal patterns (figures 2D and figure 6). We characterize each research topic as a vector of funding over time. As described in the Methods section, there are four clusters in the data based on the gap statistic (see supplementary figure S8, available online only). Each cluster’s progression describes a simple story. The largest cluster (figure 6, cluster 1) contains the most general root node of our DO, the ‘disease’ topic. This general cluster corresponds to the overall funding levels of the NIH from 1997 to 2007. In contrast, the smallest cluster is a group of topics whose funding levels peaked around 1999, but has practically bottomed out by 2007. Top funded examples include metabolic brain disease and papillary epithelial neoplasm. Among the two remaining intermediate clusters, the smaller one is also slowing down, but its peak was near 2002. Examples include herpessviridae infectious disease, connective tissue neoplasm, and embryonal cancer. Finally, the larger of the two intermediate clusters is just starting to experience exponential growth. Nutrition disorder, parasitic infections, hypersensitivity, and pervasive development disorder are all part of this cluster. Supplementary table S4 (available online only) provides a list of topics in each cluster.

DISCUSSION
Existing work
Several studies link grant data to individual articles; usually the studies focus on specific domains40—42 and on analyzing impact.12 15 40 41 In previous work researchers have extracted the funding data from the acknowledgments in journal articles.43 Lewison et al44 used citation data from the Science Citation Indexes and acknowledgment data from the UK Research Outputs Database to study national level impacts in gastroenterology. Boyack and Börner44 linked grants to individual articles through common author/principal investigator and institution using data from the National Institute on Aging, and found that impact increased with grant amount.45 Zhao42 analyzed a set of 266 publications, and reported that articles acknowledging grant funding were cited over 40% more than those without a grant acknowledgment

Boyack and Jordan46 have also linked grants in NIH RePORTER and articles in PubMed based on the funding acknowledgments and the article-to-grant linkage data provided by NIH for recent years. Based on the funding attributions, they quantified the impact of funding source upon a publication’s citations and the synergy between NIH institutions. In a separate work, they clustered publications based on the lexical similarity of the abstracts.47 Their work provides excellent insight into the interplay between grants and publications as well as the impact (in terms of citation counts) of NIH-funded research. In a highly complementary analysis, we present results at the research topic level granularity and compare funding and publication activity with the disease burden (ie, the size of the problem). LaRowe et al23 have published on the visualization of knowledge domains, and their research explores the interplay between disease burden and the size of the problem.
between publications, citations, and authors. Complementary to the previous work in this field, the use of the DO gives us the unique ability to integrate across datasets, aggregate along a hierarchy of research topics across funding and publications, correlate research activity with disease burden and visualize funding trends over time on a per-topic level.

By analyzing funding and publication rates over time, Gilhus and Sivertsen showed that publications correlate with future funding rates in neurology at a per-research institution level. Their analysis spanned four Norwegian research institutions over 4 years. Our approach differs in scale: our work examines a decade of funding and publication data, across over 9000 research topics and across 53 funding agencies. Our annotation workflow enables high-throughput analysis of the trends in research over a larger and richer space of time, topics, and institutions.

Contributions
This work demonstrates that the profiling of research activity in a domain of interest can be conducted using existing ontologies, without requiring manual assignment of index terms. As a proof of concept, we profile disease research over 10 years based on sponsorship, allocation, co-funding, and trends over time. We demonstrate that ontology-based annotation can enable comprehensive profiling of research activity spanning grants and publications. We profile disease research activity across topics, across institutions, and over time, and explore the allocation of research activity with respect to disease burden. We also identify and characterize emergent patterns that arise between research topics over time. In our work, the ontology provides a common ‘key’ between the two datasets, and a predefined hierarchy along which topics aggregate. In methods in which topics are learnt from the data, the topic sets and their hierarchy change when new data get added, making comparison across years difficult. Finally, our work demonstrates the applicability of aggregating diseases along the subsumption hierarchy of ontologies when profiling research activities.

Our work also has novel uses—such as for the evaluation of the coverage of ontologies. Figure 5 shows that the major clusters of research topics correspond to major branches of the DO hierarchy. This correspondence demonstrates that the DO classifies diseases in a consistent manner without multiple inheritances and that the classification asserted by the DO matches with the observed differences in funding priorities of different funding institutions.

A different ontology might not produce the same level of agreement between the ontology’s hierarchy and the empirically produced clusters and ontology branches corresponding to the areas of disagreement can be flagged for further review. In addition, by analyzing the items that remain unannotated, we enable identification of terms that are currently missing from the ontology used for annotation.

Limitations
Our work has certain limitations. The data we analyzed end in 2007. We used 2009 impact factors to weight all publications; a finer grained analysis might use yearly impact factors. We only examined the top 110 recipient institutions receiving the bulk of the funding because it was intractable to craft regular expression for all the recipient organizations. We are unable to have department level granularity when identifying leaders of research, because although department level data were available for most publications, they are not available for grants. The sponsorship measure that we introduced does not evaluate efficiency. Our work does not attempt to distinguish between the many possible reasons leading to extreme levels of sponsorship and cannot be interpreted as a surrogate for efficiency.

Our analysis ignores grant information reported in MEDLINE as it is not always available. In annotating the grants, we had access to the summaries, which include titles and abstracts, but not specific aims. In aiming for high precision of the annotations, we sacrificed on recall of Mgrep. We did not perform negation detection during annotation. However, the prevalence of negation in sentences in publication abstracts is estimated to be approximately 10%, and for the majority of these sentences, the negation applies to a relationship among two entities mentioned in the sentence (discussed further in supplementary text, available online only). Finally, we chose mortality as a surrogate for disease burden; one can argue that for chronic diseases—such as diabetes—the prevalence might be a better measure of disease burden.

CONCLUSION
Through a novel application of an ontology-based annotation workflow, this work provides a proof of concept for automated profiling of disease research activity using an existing ontology. We profiled disease research across topics, across institutions, and over time. We identified research topics of high sponsorship and characterized research activity in those domains. We cross-referenced research activity against US and worldwide mortality rates, as a method to assess funding allocation by comparing with the disease burden. We performed similarity-based profiling of research institutions and research topics. Finally, we identified and characterized temporal trends among research topics. Our workflow generalizes well in the sense that an alternative ontology other than the DO can be swapped in the DO’s place—enabling a use-case driven mechanism to evaluate the utility of classification hierarchies.

In previous work, we have shown that annotation corpora, such as the NCBO resource index, can provide integrated search on biomedical data. As a further instantiation of NCBO’s mission to annotate publically available resources, our work demonstrates that ontology-based annotation can play a role in broad questions such as profiling the landscape of biomedical research activities.

Contributors YL implemented the final annotation workflow, integrated data, performed analysis, generated the results, and wrote the manuscript. AC collected the grants and publications data, and implemented a proof of concept for the annotation workflow. PL provided significant expertise in optimizing the annotation workflow and generating figures. NHS conceived the study, provided scientific guidance, and edited the final manuscript. All authors contributed to the project direction through discussions. All authors read and approved the manuscript.

Funding This study received support from the NIH grant U54 HG004028 for the National Center for Biomedical Ontology.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data are available upon request.

REFERENCES
Research and applications


42. Zhao D. Characteristics and impact of grant-funded research: a case study of the library and information science field. Scientometrics 2010; 84:283–306.


