## SBIR PHASE II STATEMENT OF WORK

### TITLE: Creation of an accurate model of the topical structure of PubMed and associated indicators

### Background Information and Objectives

1. Background Information – The ability to identify emerging topics is one of the most important tasks in developing a research strategy. The ability to determine if a topic is no longer attractive and should therefore receive less funding is of comparable importance for portfolio analysis. Overall, the accuracy by which one identifies topics, and the validity of any indicator of topic attractiveness, is a necessary first step towards research evaluation and planning that is evidence-based.

This need is not adequately addressed in a cost-efficient manner by any existing product. The current market for these types of indicators is being served by the three companies with citation databases: Clarivate (Web of Science), Elsevier (Scopus) and Springer Nature (Dimensions). The ultimate objective of this project is to develop an accurate and viable open source alternative to citation-based indicators of science and technology that can be used for research assessment and funding decisions.

This SBIR will produce a highly detailed and accurate model of the PubMed database, where each paper is assigned to one of about 50,000 topics. It will also produce new topic-level metrics of both impact and virtuous research based on open source (rather than citation) data and will validate these metrics against traditional citation-based metrics, thus providing an alternative to citation-based metrics.

B. Phase II Technical Objectives – The key objectives of this project are to:

1. Increase the accuracy of the PubMed model successfully created in Phase I. New models of PubMed data containing features not tested in Phase I will be created and their accuracy will be compared to the best Phase I model using our established methods. Specifically, we plan to test features based on novel (or atypical) combinations of features such as chemicals, equipment, diseases, methods, etc. into the model.
2. Develop a methodology to accurately update the PubMed model at specified intervals (e.g., every six or twelve months) with new data while maintaining a stable solution and identifying new emerging topics. To be useful for decision makers, models of science must be stable over a period of time. In this context, stability means that the majority of papers and clusters must remain intact when new data are added to the model. However, in addition to stability, to be realistic, the model must include provision for creating new emerging topics. Thus, this task will test different methods of updating a model while maintaining stability, preserving accuracy, and nominating new emerging topics.
3. Increase the accuracy of the indicators of scientific and economic impact at the topic level that were created in Phase I. Larger tests of the feature space will be done in Phase II to identify those features that correlate positively with citation counts and funding amounts, and to identify which combinations of those features will be the best predictors of topic growth, impact and funding.
4. Assess virtue of research indicators and their correlation to measures of impact in samples of papers published in 2015-2018 and create an observatory of biomedical research that combines optimal field partitioning based on open sources and a suite of impact, virtue, quality, and translational indicators. We will extend our previous work on virtue indicators to capture information on papers published in 2015-2018 and will continue to explore their correlation and complementarity against other types of impact and other measures that can be generated from open source information. We will also explore the capturing and validation of additional indicators of quality, methods and translational potential. This comprehensive information can generate an observatory of the quality, impact, virtue, and potential of biomedical research.
5. **Services to be Performed.**
6. General Requirements
7. The contractor shall independently perform all work and furnish all labor, materials, supplies, equipment, and services (except as otherwise specified in the contract)
8. The Contracting Officer’s Representative (COR) specified in Section G of the contract will monitor the work.
9. Specific requirements – The work plan is comprised of the following tasks. The lead institution for each task is noted.
10. *Create a more accurate model of the PubMed database by creating and testing multiple new models, measure the accuracy of each model, and select the best model to use going forward (SciTech).*

In Phase I we found that a PubMed model based on related article (RA) scores is just as accurate as a model based on citation features. For this task we will create and test multiple models that add different features to our Phase I model using the same process established in Phase I. Specifically, we will, at a minimum, create and test the following models:

* Phase I model + novel chemical combinations
* Phase I model + novel disease combinations
* Phase I model + novel equipment combinations
* Phase I model + open source direct citation information

All models will be created using up-to-date data. The accuracies of each model will be calculated using the multiple metric process established in Phase I, and the most accurate model will be used as the baseline for the additional tasks going forward.

1. *Propose, test, and validate approaches for accurately updating the PubMed model with new data while maintaining stability and accuracy, and nominating new emerging topics (SciTech).*

Models of science need to be stable over periods of several years to be most useful to stakeholders. In addition, as new topics emerge in science, they need to be added to the model. Calculating a new model each time new papers are added to the database would solve the emerging topic problem, but would create large instabilities in existing topics. Thus, we propose to develop an accurate approach to add papers to the existing model and to also identify emerging topics. Two main possibilities exist, each of which will be tested.

1. Add new articles to the model using the best similarity method resulting from task A. Once this has been done, analyze the largest clusters to determine if they can be reasonably split into an older decaying component and a younger emerging component, and split (a small number of) clusters accordingly.
2. Compare the feature space of papers in the youngest clusters with those in older clusters to identify features that might be associated with emerging topics. Use this result to identify the new papers containing these features, and create a number of new clusters using these papers. Once these papers have been clustered, then add new articles to the model using the best similarity method.

Choosing between these methods may be difficult. We propose to choose between them by creating an entirely new model, identifying the emerging topics and their papers in the new model, and determining which of the stability-preserving approaches does the best job of creating emerging topics. We expect this task to require several rounds of iteration where we progressively learn which features are best suited to the task of creating accurate emerging topics as we add data to models.

1. *Create and test updated and more accurate measures of scientific, clinical and economic impact at the topic level (SciTech).*

This task will continue to develop measures of scientific, clinical and economic impact at the topic level that can be used for decision making. A large number of features was tested in Phase I. The list of features to be tested will be expanded, particularly using new indicators of our own making, such as:

* Presence of discovery papers, as indicated by full text analysis
* Presence of methods papers, as indicated by full text analysis
* Presence of language indicating uncertainty, as indicated by full text analysis, and using abstracts.

Note that each of these new metrics is something that was not mentioned in our proposal, but that we have begun researching since the proposal was written. In each case, we have promising results (unpublished) in terms of identifying individual papers, but have not yet incorporated these features into models at the topic level. We expect indicators based on these features to be significant at the topic level. We will also do research to understand exactly what each of these indicators will mean in practice. For example, an indicator of uncertainty is unlikely to relate to scholarly impact. However, it may relate to social value. Hypotheses such as these will be formulated and tested.

1. *Assess virtue of research indicators and their correlation to measures of impact in samples of papers published in 2015-2018 (Stanford).*

In previous work, we evaluated a random sample of 500 papers from PubMed published in 2000-2014 to code them for virtuous indicators. This methodology was used in Phase I of this project to code an additional 100 papers from 2015-2016. For this task, we will now extend this assessment to an additional 500 papers published in 2015-2018 to get a more comprehensive insight of the most current status of these measures of the virtue of the research using the same processes detailed in our Phase I report.

The Phase I study showed us that the virtuous indicators we seek (e.g., reproducibility, novelty, translatability, etc.) are more prevalent in recent papers than in earlier papers. The study of an additional 500 papers in this task will tell us if this trend is likely to continue, and will help us to formulate strategies that might be used later on to develop algorithmic approaches to extracting such features from abstracts and full text documents.

In addition, using extracted data from the full set of recent papers (2015-2018), we will examine whether there is a correlation between measures of impact and hot science and indicators of virtue, as described in our technical proposal, Phase I report, and in our previous publications. The correlational analyses will be done across the biomedical literature, and also across different fields defined with different levels of granularity and employing both open-source data and all-data options for comparison of performance and consistency of results. Even if the open-source classifications do not perform as optimally as we expect in classifying different fields, they may still provide reliable results when such correlational analyses are performed. The availability of all-source data for comparison will allow us to investigate whether results with different approaches are similar or not.

We will also examine whether measures of impact/hot research can be combined with indicators of virtue for papers published in 2015-2018 as outlined also in phase I for papers published in 2000-2014. As we have stated, it is very likely that, while different metrics of impact/hot research are relatively highly correlated with each other, their correlation with indicators of quality, transparency, and reproducibility may not be necessarily that high.

1. *Create an observatory of biomedical research that combines optimal field partitioning based on open sources, measures of virtue of research on expanded samples of PubMed articles and additional indicators of quality of methods, translational potential, and novel linkages (Stanford).*

The principles expanded in steps described above can be expanded and scaled up to include up to 1,500 articles per year (2015-2018). We will focus on indicators that are easier and most straightforward and unambiguous/accurate to extract manually, including availability of raw data, full protocols, funding, and conflicts of interest disclosures. Random sampling will be used with in-depth data extraction and much larger samples of articles where information on virtue indicators may be collected with text-mining from abstracts and full-text articles. We have extensive experience on using text-mining in samples as large as the entire biomedical literature. For example, our METRICS team at Stanford published an in-depth evaluation of the use of p-values across all PubMed abstracts published in 1990-2015 and also of over 800,000 full-text articles from PubMed Central. We will explore different text mining approaches with different text term searches to identify those that may come closer to results obtained from in-depth manual extraction of these indicators. Text mining may be particularly useful to use for indicators that are difficult to extract manually, e.g. whether a study is a replication and whether it has been used in a systematic review/meta-analysis.

Moreover, we will explore additional indicators that will try to capture the use of statistical methods and the potential translational potential. Based on our experience, we are able to capture in large-scale information on the use and reporting of specific methods, e.g. frequentist methods using p-values as well as other metrics that reflect use of specific statistical tests as well as their misuse, e.g. reporting of p-values that do not correspond to other statistics reported in the same spot (e.g. t-statistics and degrees of freedom), a manifestation of questionable research practices and addition of spin and exaggeration in the reported results.

The addition of up to 1,500 articles per year examined in-depth manually and/or with validated text-mining (up to 6,000 articles examined) and the potential addition of extra information from text-mining approaches will allow the creation of an ongoing observatory of impact and virtue in biomedical research. We will operationalize this observatory so as to be maximally useful to stakeholders who are interested to track the course of biomedical research as a whole and of specific fields in terms of impact and virtue indicators. This may include funders, scientists, institutions and other entities. We will seek feedback from surveys and communication with major funding agencies of biomedical research and with several leading institutions as well as many top scientists. METRICS is increasingly recognized as a key hub for meta-research and defining this field at large. We have an extensive network of colleagues who work with us or are interested in meta-research questions and we will mobilize this network to ask for feedback on how this observatory may become even more useful. We expect that this will be an iterative process and will allow developing an observatory that is attractive to attract further support so as to be maintained on a more long-term basis.

In addition to this detailed article-level work, we will also develop indicators of translational potential by building on the previous work of Boyack et al. They found that each paper could be accurately assigned to a research level along the basic-to-applied continuum using titles and abstracts. Using research levels, the translational potential of authors and research groups can be identified using publication profiles. Translational potential is highest when an author or small group of authors publish significant fractions of their papers (at least 30%) in both basic and applied topics. In other words, this method identifies those researchers and groups that speak multiple scientific languages that allow them to personally move research along the translational pathway. We will extend the methodology to identify those topics (clusters of papers) that are involved in translational research.

Finally, although novelty is not listed in the RPF as an indicator of virtuous research, it is an additional dimension that would be worth exploring both on its own terms as well as in terms of its correlation with virtuous indicators. Foster, Rzhetsky and Evans analyzed chemical data from millions of PubMed papers, finding that those publications that proposed new linkages between chemicals were both more likely to be ignored and more likely to achieve high impact than publications that did not. In other words, these publications were riskier, and some of that risk was rewarded. They also suggest that risk-taking is essential to scientific progress and should be encouraged. This research is intriguing in that it suggests that novelty can be inferred from non-citation data. We will expand upon this research in our search for virtuous indicators. Specifically, we will calculate multiple types of linkages using data from PubMed, to include diseases, chemicals and equipment. All linkage types will be explored including linkages between items of the same feature type (e.g., chemical-chemical) and linkages between items of different types (e.g., disease-chemical). Indicators of novelty based on new linkages will be correlated with citation data and with the other indicators mentioned above.

1. *Using results from tasks 4 and 5, develop preliminary approaches to automatically extract features related to virtuous indicators from full text data, and use machine learning to determine if these extracted data are suitable for indicator development (Stanford).*

Using the coded data from tasks 4 and 5, preliminary approaches (e.g., natural language search scripts, etc.) will be created to attempt to automatically identify features related to these data within full text articles. Using the subset of coded articles that are in the PubMed Central Open Access (PMCOA) subset, and for which we thus have full text, these scripts and processes will be applied to the full text articles to identify virtuous features in each article. These results will then be compared with those identified using hand-coding to determine precision and recall for each feature to determine its suitability for automation. Although we are hopeful that some features will be amenable to automatic extraction, we are also practical in stating that we may find many features that cannot be automatically extracted with any fidelity. In our experience, the hand-coding of such information is very much tied to tacit knowledge of subject areas and to synthesis of the meaning of various passages in full text articles, and this is very difficult to duplicate algorithmically. Thus, we are hopeful, and yet do not expect that all features will be amenable to algorithmic extraction.

1. *To demonstrate the usefulness of the resulting model and indicators for stakeholders, we will create workflows designed to answer questions that are likely to be asked (SciTech).*

In preparation for commercialization of our technology, we will determine the types of questions that are likely to be of most interest to stakeholders such as researchers, research administrators, and program officers. We have experience working with such stakeholders, and anticipate questions such as:

* 1. Researcher: What topics are most related to my research? What are the key papers and performers in those topics? What other topics are similar that might be of higher impact? What is the translational potential of my set of topics? If I want to make my research more translatable, which topics (and thus researchers) should I link to in the future?
  2. Researcher: I’m writing a proposal. Which topic(s) are most related to my proposed idea, and are these topics well-funded or not? Are there nearby topics with more funding available?
  3. Program officer: Which topics comprise my portfolio? Are there key players in those topics of whom I’m not aware? Which topics are in emerging and mature/declining phases? What is the translational potential of my topics? What are the natural translational pathways between my topics and others? Are the studies in a particular topic reproducible, or have they been reproduced? What kind of data sources are available in my topic space? … and many others.

Workflows (i.e., approaches to querying the data and obtaining results) will be created to answer each of these questions. Examples will be provided.

1. *Compilations of data will be created and made available for sharing (SciTech).*

At both the midpoint and end of the project, the best available PubMed model will be shared with the scientific community. In addition, metadata descriptions of each topic will be made available. Selected indicators will also be shared along with the model and metadata. However, not all indicators will be shared. We expect to share only samples of those indicators that are deemed to be of the most commercial value (e.g., the scientific and economic impact indicators)

In addition to those things mentioned in the task descriptions above, we anticipate the following end results from the project:

1. An accurate, validated, up-to-date model of the PubMed literature will be created that will be suitable for research planning in the biomedical research space by researchers and agency. This model will be comprised of approximately 50,000 topics (clusters of articles), in which each PubMed article will be assigned to a single topic. Topics will be explicitly linked to historical funding data.
2. Indicators of scientific, clinical and economic impact at the topic level will be created from non-citation data and compiled for use. These indicators will have been compared to best-in-class citation-based indicators to establish their validity and usefulness. General methodologies for each indicator, along with validation experiments, thus facilitating their acceptance by the community.
3. Indicators of virtuous research, including indicators of transparency, reproducibility, translational potential, and novelty will be created and compiled for use. Where enough data exist, these indicators will be linked to topics. These indicators will have been validated using a combination of comparison with other metrics and expert opinion. General methodologies for each indicator, along with validation experiments, thus facilitating their acceptance by the community.
4. Workflows will be created for stakeholders (e.g., researchers, program officers) to query the model to provide information that will help them questions such as those mentioned in conjunction with Task 7.

Reports & Deliverables – Phase I has established that this project can be successful. A detailed technical report was submitted to NIH. We have not, however, published the Phase I model or metric results yet to maintain a competitive advantage using this information. However, we do expect to eventually publish the Phase I results, and those from Phase II as well, at the correct time. This will be done to establish credibility for the approaches and indicators being developed.

Note that while we will publish the general bases and methodologies for our indicators, detailed parameters that would enable full replication of the methods will not be published, thus preserving the potential to commercialize the technology.

Deliverables to NIH will include the full model and a full set of indicators at the topic level that are current as of 1) the mid-term of the project, and 2) the conclusion of the project. This model and set of indicators can be used by NIH internally in any way desired. Subsequent updates to the model will, however, not be freely provided. A detailed project report containing methodologies and results will also be submitted to NIH. We anticipate making progress reports as frequently as it is desirable for NIH.