ProvCaRe: Characterizing Scientific Reproducibility of Biomedical Research Studies using Semantic Provenance Metadata

Satya S. Sahoo[1*], Joshua Valdez[1], Matthew Kim[2], Michael Rueschman[2], Susan Redline[2]

[1]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

[2]Department of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

Keywords: Scientific Reproducibility, Provenance Metadata, W3C PROV Specifications, ProvCaRe ontology, S3 Model, Provenance-based Ranking, ProvCaRe Knowledge Repository

Word Count: 5032

*Corresponding author: Satya S. Sahoo, 2103 Cornell Road, Wolstein Research Building, Cleveland, OH, 44106, USA. satya.sahoo@case.edu. Telephone: 216-368-3286, Fax: 216-368-0207.

ABSTRACT

Objective: Reproducibility of research studies is key to advancing biomedical science by building on sound results and reducing inconsistencies between published results and study data. We propose that the available data from research studies combined with provenance metadata provide a framework for evaluating scientific reproducibility. We developed the ProvCaRe platform to model, extract, and query semantic provenance information from 435, 248 published articles.

Methods: The ProvCaRe platform consists of: (1) the S3 model and a formal ontology; (2) a provenance-focused text processing workflow to generate provenance triples consisting of *subject*, *predicate*, and *object* using metadata extracted from articles; and (3) the ProvCaRe knowledge repository that supports "provenance-aware" hypothesis-driven search queries. A new provenance-based ranking algorithm is used to rank the articles in the search query results.

Results: The ProvCaRe knowledge repository contains 48.9 million provenance triples. Seven research hypotheses were used as search queries for evaluation and the resulting provenance triples were analyzed using five categories of provenance terms. The highest number of terms (34%) described provenance related to population cohort followed by 29% of terms describing statistical data analysis methods, and only 5% of the terms described the measurement instruments used in a study. In addition, the analysis showed that some articles included higher number of provenance terms across multiple provenance categories suggesting a higher potential for reproducibility of these research studies.

Conclusion: The ProvCaRe knowledge repository (https://provcare.case.edu/) is one of the largest provenance resources for biomedical research studies that combines intuitive search functionality with a new provenance-based ranking feature to list articles related to a search query.

1. INTRODUCTION

A key component of biomedical research is transparency in reporting of studies with clear description of design, data collection, analysis, and methodology to support scientific reproducibility and accurate interpretation of research findings [1-3]. However, a recent survey of 1,576 researchers found that 70% of the researchers were unable to reproduce a study conducted by others and 50% of the researchers were unable to reproduce results from their own experiments [2]. Similar studies in a variety of disciplines have shown that a significant number of research studies cannot be replicated. For example, a number of the spinal cord injury studies funded by the US National Institute of Neurological Disorder and Stroke (NINDS) could not be replicated [4] and an analysis of 67 drug target discovery projects found inconsistencies between published and study data in two-thirds of the projects [5]. The lack of reproducibility in biomedical research is an important concern for academic and industry research communities, public and private funding agencies, and patients [1, 6]. In particular, irreproducible studies result in misdirection of research funding from appropriate studies, waste of limited resources, and potential suffering of participants in clinical or preclinical studies.

In addition, the increasing availability of data from research studies as part of several data sharing initiatives led by funding agencies including the National Heart, Lung, and Blood Institute (NHLBI)-funded National Sleep Research Resource (NSRR) [7], the Cancer

Genome Atlas (TCGA) [8], and the National Institute of Mental Health (NIMH) Data Archive (NDA) [9] highlights the need to make the associated contextual metadata available to support scientific reproducibility. The NSRR project is creating the largest repository of sleep medicine study data from more than 40,000 polysomnograms (sleep studies) involving more than 36,000 participants. Researchers can access and download the study data from NSRR after receiving approval from their institutional review board and completing a data access and user agreement. Similarly, the NDA shares de-identified human subject data from the National Database for Autism Research (NDAR), the National Database for Clinical Trials Related to Mental Illness (NDCT), the Research Domain Criteria Database (RDoCdb), and the Adolescent Brain Cognitive Development (ABCD) study. However, these public data repositories form only one of the two core components of scientific reproducibility. Without the availability of appropriate contextual metadata, for example the inclusion and exclusion criteria for a study population, randomization technique used, or statistical analysis methods used in a study, the reproducibility of research studies is extremely challenging.

To address this challenge, there is increasing focus on developing guidelines and best practices to enhance reproducibility of research studies, for example the "Rigor and Reproducibility" guidelines defined by the National Institutes of Health (NIH) [6]. The NIH Rigor and Reproducibility guidelines focus on enhancing transparent reporting of study details, such as use of blinding techniques, methods used to estimate sample size, and instruments used to record data in an effort to facilitate scientific reproducibility. They also require the study authors to provide complete details of the statistical methods used to analyze the data, the procedure used for validation, and suggest the use of

existing domain-specific data and reporting standards [10]. Similar to the Rigor and Reproducibility guidelines, the Transparency and Openness Promotion (TOP) guidelines focus on eight standards for transparent reporting of scientific studies, for example analytical methods, research materials, and description of data used in a study [11]. A common feature underlying these guidelines and other related best practices is their focus on contextual metadata called *provenance*, which describe the history or origin of data.

Provenance metadata has long been used in computer science to trace the origin and the intermediate data processing steps that generate final results, which enables verification of data quality, security, and soundness of results [12-14]. In particular, research in data management systems has focused on "database provenance" to identify the source of a value in a database (called "Where provenance"), the reasons for presence of a value in a database query result (called "Why provenance"), and the specific values associated with a result (called "How provenance") [15-17]. Scientific workflows systems, such as Taverna [18], Kepler [19], and Trident [20], are widely used to automate scientific data processing, integration, and analysis. Therefore, "workflow provenance" is used to keep track of data in a workflow system to ensure systematic identification of computing steps that lead to errors and support scientific reproducibility (a review of workflow provenance is presented in [21]). In 2013, the World Wide Web Consortium (W3C), which is the Web technology standards organization, recommended the PROV specifications to serve as a common model with associated constraints for modeling and managing provenance metadata [14, 22, 23]. The PROV Data Model (PROV-DM) and PROV ontology (PROV-O) define a minimal set of provenance metadata elements that can be extended to represent provenance in a variety of domains.

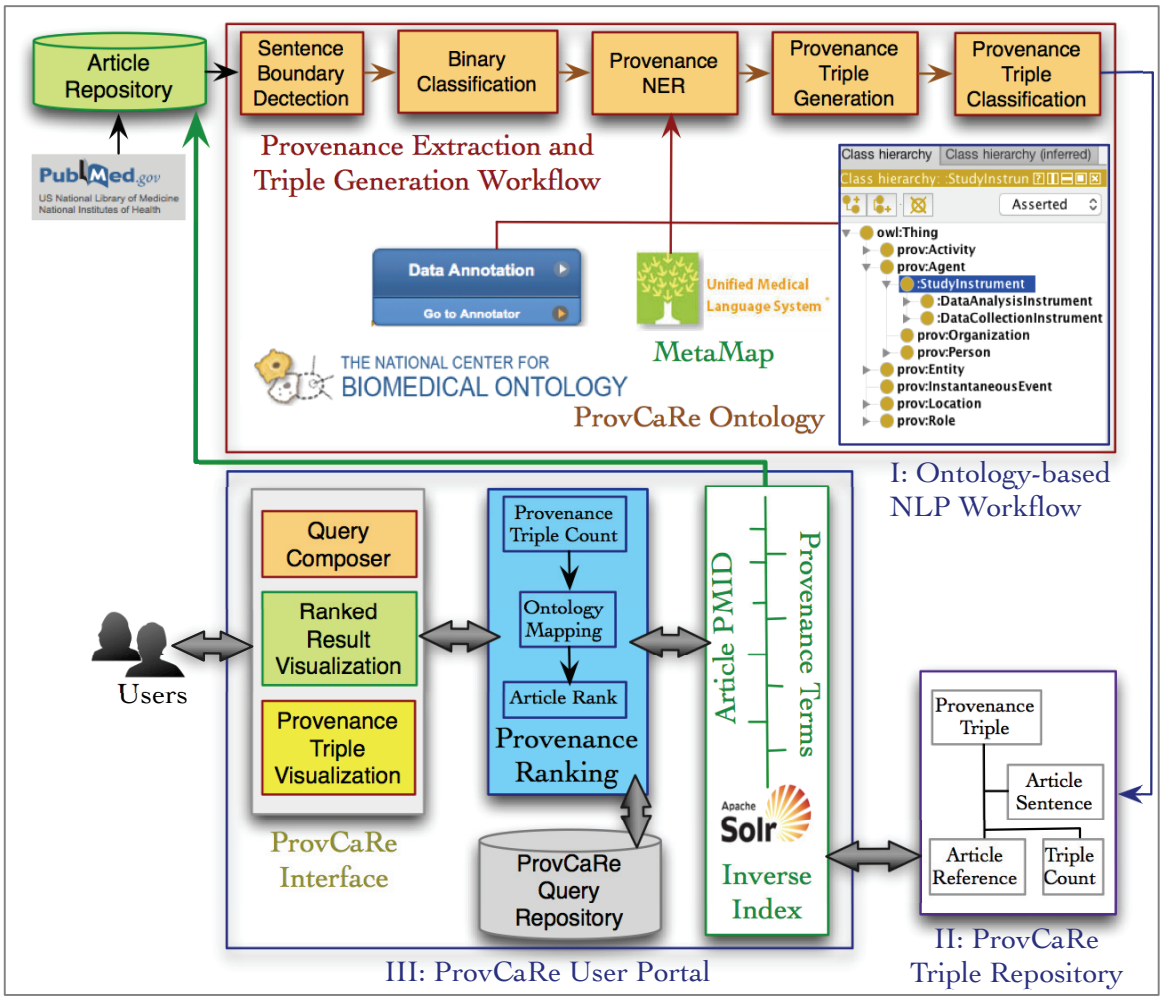

Figure 1. The architecture of the ProvCaRe platform with three components: (1) Ontology-based NLP workflow; (2) Triple repository for provenance metadata; and (3) ProvCaRe user portal. The users use the ProvCaRe interface to perform hypothesis-driven query and exploration of provenance metadata associated with research studies.

The PROV specifications are well-suited to develop and implement a provenance-based framework for supporting reproducibility in biomedical research with two objectives: (1) to systematically characterize provenance metadata available from published research studies, and (2) to provide a practical tool to implement and advance the objectives of initiatives such as the NIH "Rigor and Reproducibility" guidelines. To the best of our knowledge, existing tools and methods for reproducibility of clinical and health research have limited support for modeling, storing, and analyzing provenance metadata using the PROV specifications. In this paper, we describe the development of the Provenance for

Clinical and Health Research (ProvCaRe) framework consisting of three components to extract, analyze, and characterize provenance metadata associated with biomedical research studies (Figure 1 illustrates the overall architecture of ProvCaRe). The first component of the ProvCaRe framework is the ProvCaRe ontology, which is a formal model of provenance terms associated with the design and analysis of research studies that extends the W3C PROV Ontology [22]. The ProvCaRe ontology is used in the biomedical text processing and provenance extraction pipeline to generate provenance graphs from published articles available from the National Center for Biomedical Information (NCBI) PubMed citation database. The second component is the ProvCaRe knowledge repository consisting of semantic provenance metadata extracted from 435, 248 biomedical research articles (focused on sleep medicine as an exemplar). The third component of the framework is the hypothesis-driven search capability and with a new provenance-based ranking feature that allows users to locate research studies with higher likelihood of reproducibility.

1.2 Related Work: Guidelines and Best Practices for Scientific Reproducibility

Reproducibility has been a focus of many community-based initiatives in clinical and basic science research studies that have resulted in multiple guidelines and best practices [1-3, 6]. The Problem/Population, Intervention, Comparison, Outcome and Time (PICOT) model has been extensively used to formulate clinical questions in evidence-based medicine (EBM), which allows physicians to clearly structure clinical problems that leads to better study results and use these characteristics for systematic literature reviews [24]. The PICO(T) model is effective in modeling clinical therapy related studies [25]; however, it has limited features to model provenance information required for

reproducibility. Similar to PICO(T), the Consolidated Standards of Reporting Trials (CONSORT) guidelines consist of 25 terms for reporting of randomized control trials with a focus on individually randomized, two group, and parallel trials [26]. The CONSORT guidelines have been extended to include additional terms to describe the design of trials for transparent reporting of clinical trials. The CONSORT guidelines have been widely adopted with more than 400 journals recommending the use of these guidelines for reporting clinical trials [26].

The Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines have been developed by the National Center for the Replacement, Refinement, and Reduction of Animals in Research (NC3Rs), a government funded organization in the United Kingdom, to enhance transparency in reporting animal research [27]. The ARRIVE guidelines are similar to the CONSORT guidelines and include 20 terms to describe animal research studies. As discussed earlier, the NIH guidelines for Rigor and Reproducibility were published in 2013 to facilitate greater transparency in reporting of biomedical research studies for improved reproducibility [6]. In addition to guidelines and best practices, the Ontology for Clinical Research (OCRe) project has developed a formal model of clinical study protocols to represent study design and eligibility criteria of a study [28]. The OCRe project developed the Eligibility Rule Grammar and Ontology (ERGO) annotation workflow to represent and annotate eligibility criteria in clinical statements. The ProvCaRe framework builds on these guidelines, best practices, and formal model of research studies to extend the W3C PROV specifications for extracting, modeling, and analyzing provenance metadata required to support scientific reproducibility [29]. The ProvCaRe framework includes: (1) the ProvCaRe ontology as a

formal knowledge model of provenance metadata for reproducibility with a compositional grammar syntax to support post-coordinated class expressions [30], a provenance-focused ontology-based text processing pipeline to extract provenance metadata from biomedical literature [31], and a user portal for search and query of provenance metadata [29]. To the best of our knowledge, the ProvCaRe knowledge repository is the largest repository of biomedical semantic provenance metadata with more than 48.6 million "provenance triples" available to the research community for characterizing and evaluating scientific reproducibility.

2. MATERIALS AND METHODS

The current version of the ProvCaRe platform was developed in collaboration with sleep medicine researchers involved in the development of the NHLBI-funded NSRR project [7]. We identified 435, 248 articles related to sleep medicine using a bootstrap approach that used two keywords of "sleep" and "sleep disorder" as input to an ontology-driven term lookup API supported by the National Center for Biomedical Ontologies (NCBO). This search resulted in 2083 unique terms that were manually reviewed for relevance. These 2083 terms were used as input to the NCBI E-Utils APIs to identify relevant PubMed identifiers (PMID) of sleep medicine related articles [32]. The E-Utils tool identified 1,132,528 PMIDs that were used to download 435,248 full-text articles (697,280 PMIDs corresponded to abstracts that were discarded). These articles were processed to extract provenance metadata corresponding to the ProvCaRe S3 model, which is described in the next section.

2.1 ProvCaRe S3: A Provenance Metadata Model for Scientific Reproducibility

Using the NIH Rigor and Reproducibility guidelines and the W3C PROV specifications we identified three core categories of provenance metadata required for supporting reproducibility of research studies: (a) *study method*, (b) *study instrument*, and (c) *study data*. The *study method* describes how a research study was conducted, including criteria used for selecting study cohort, data collection and analysis method. The *study instrument* describes what instruments were used to collect and analyze data, including statistical models. The *study data* represents the data collected and analyzed in a research study, including valid range of the data and threshold used to retain or discard data during analysis. These three categories of provenance metadata terms constitute the core terms of the ProvCaRe S3 model and they are modeled by extending the three core PROV Ontology classes of `prov:Entity`, `prov:Activity`, and `prov:Agent` [22] (`prov` represents the W3C PROV namespace, *http://www.w3.org/ns/prov#*). The S3 model is formally represented in the ProvCaRe Ontology using the Web Ontology Language (OWL2) [33]. These three S3 terms are extended to model provenance terms corresponding to different biomedical domains, including sleep medicine research and neuroscience. For example, the *provcare:StudyInstrument* is extended to model electrophysiological signal data recording instruments (e.g., *provcare:ElectroencephalogramInstrument, provcare:ScalpElectrode*) and sleep questionnaire (e.g., *provcare:ObstructiveSleepApnea18*), where *provcare* refers to the namespace: *http://www.case.edu/provcare#*). The ProvCaRe ontology uses OWL constructs to compose class expressions, for example *provcare:Electroencephalograph* (a subclass of *provcare:StudyData*) has a class-level existential restriction on OWL object

property *provcare:hadDataCollectionMethod* with restriction filler *provcare:ElectroencephalographProcedure* class [33].

A key feature of the ProvCaRe ontology is the extensive re-use of terms modeled in existing biomedical ontologies and creation of mappings between ontology classes. For example, the ProvCaRe ontology classes representing biochemical assays are mapped to existing classes in the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [34] using the `rdfs:seeAlso` property. In addition to ontology class mappings, the `rdfs:label` annotation property is used to represent acronyms, synonyms, and other human readable terms of an ontology class. For example, laboratory test data for Interleukin-1 level is mapped to its different labels such as *IL-1*, *LAF*, and *LEM*. In addition to pre-coordinated class expression, which are "built-in" in an ontology before deployment, we have developed a post-coordinated compositional grammar syntax that is used to represent new class expressions based on requirements of specific disciplines of biomedical research [30]. An important role of the ProvCaRe ontology is to support the generation of "provenance triples" consisting of *subject* → *predicate* → *object*, for example *cross sectional analysis* → *included* → *6132 participants*. These provenance triple structures are aggregated to form to a provenance graph. In the next section, we describe the development of a text processing workflow for provenance extraction and triple generation from published full-text articles.

2.2 Provenance Graph Generation: Structured Metadata Extraction from Published Articles

Provenance metadata terms representing contextual information of a research study are linked together by multiple relations or predicates to form a provenance graph [35].

Therefore, the ProvCaRe project uses graph structure to represent provenance of research studies that can be queried and analyzed using graph traversal and query techniques [36]. Figure 2 illustrates a segment of the provenance graph corresponding to a prospective cohort study that evaluated the correlation between sleep disordered breathing and incident hypertension [37]. This provenance graph can be used to query contextual information of a research study, for example `provcare:recordedUsing` to retrieve the Compumedics PS-2 system used to record polysomnogram data. To generate provenance graphs using structured metadata information from published articles we extended the open source clinical Text Analysis Knowledge Extraction System (cTAKES) [38] with additional functionalities.

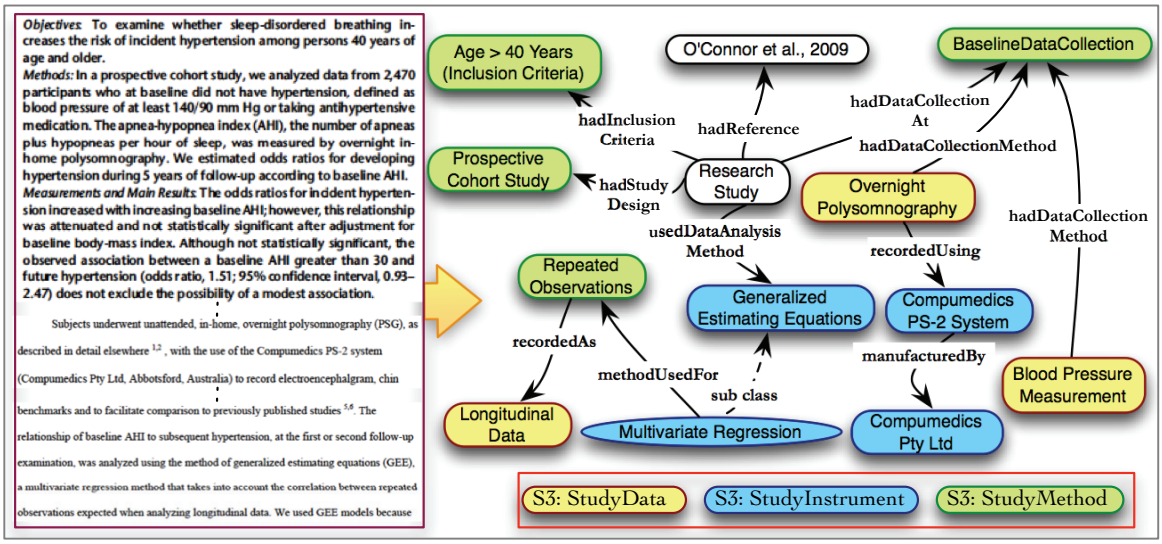

Figure 2: A provenance graph consisting of provenance triples describing three aspects of a research study corresponding to Study Method, Study Data, and Study Tool extracted from different sections of an article by O'Connor et al.

Extension of cTAKES in the ProvCaRe NLP Workflow. The cTAKES modules for sentence boundary detection, tokenization, morphological normalization, and part-of-speech (POS) tagger are used to initially process the published articles, which is followed by provenance-specific processing and extraction of terms (a detailed description of the

extended cTAKES modules is provided in our previous publication [31]). We developed a two-stage process for provenance entity recognition that uses a combination of techniques for sentence classification followed by entity recognition. In the first stage, we use a binary classifier based on Google Tensorflow Long Term Short Memory (LSTM) network library [39] to categorize and identify sentences that contain provenance metadata by extending a sentence classification architecture developed by Kim et al. [40, 41]. In our previous work, we discuss that the LSTM network has the highest classification accuracy of 86% as compared to other deep learning approaches [42]. In the second stage, the provenance Named Entity Recognition (NER) module uses a combination of techniques, including the ProvCaRe ontology as reference knowledge model, which is parsed using the OWLAPI [43], the MetaMap tool [44], and the NCBO Open Biomedical Annotator (OBA) tool [45]. In our previous work, a comparative evaluation of three techniques showed that each of the three approaches successfully identified entities in different categories of the S3 model, therefore using a combined approach improved the performance of the provenance NER module [29].

Generation of Provenance Triples. The NER annotated provenance-related sentences are parsed using the Stanford dependency parser and Semantic Role Labeling (SRL) parser [46] [40, 47] to generate a constituent parse tree and SRL labeled predicates. Figure 3 illustrates the parse trees of the sentence from a research study [37] and the corresponding provenance triple generated from the parse trees using a new ProvCaRe triple generation algorithm. This ProvCaRe algorithm selects noun subject (*nsubj*) of the sentence in the parse tree and performs a depth first search to identify the parent noun phrase (*NP*) containing the *nsubj* node to generate the *subject*. The algorithm uses

heuristics, such as the presence of coordinating conjunction and anaphora resolution in the *NP*, to generate more than one triple, while the remaining components of the *NP* are joined with the identified *nsubj* to create the *subject* of the provenance triple. The *predicate* of a provenance triple is identified by concatenating the root verb with related *verb phrases* in the context of the subject and object of the sentence. The *object* of the triple is generated by extracting the objects of the predicate phrase and identifying its parent structure. A total of 48,916,832 provenance triples were generated from 435, 248 articles, which are available to users for querying and analysis in the ProvCaRe knowledge repository.

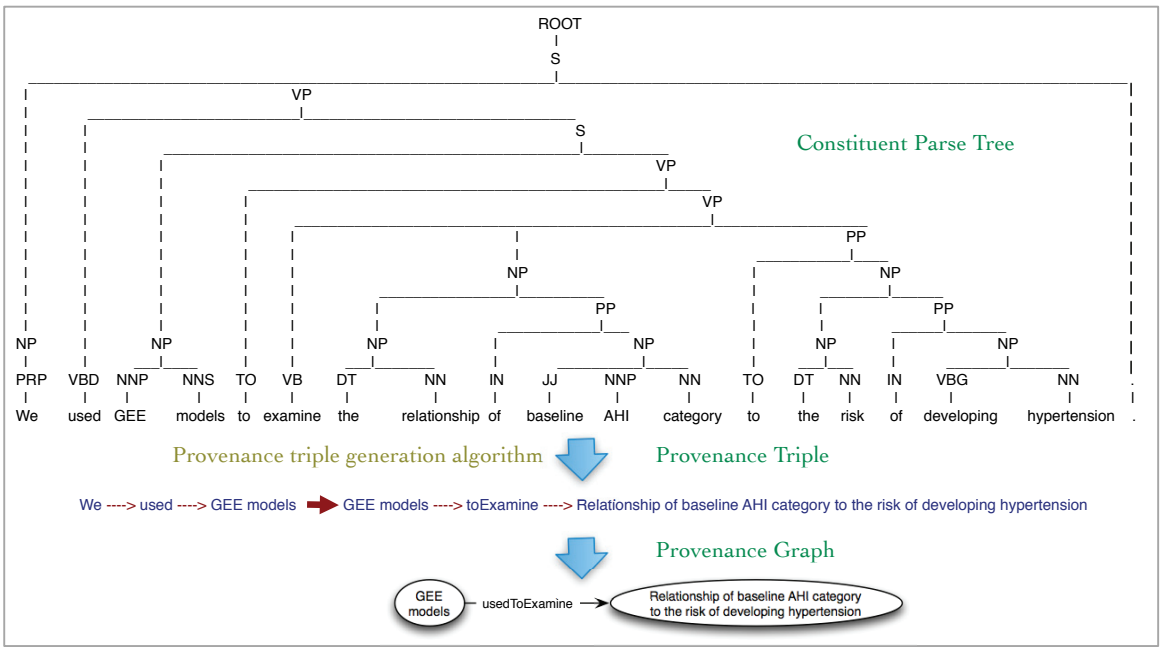

Figure 3: A workflow demonstrating the generation of provenance triples from an example sentence using semantic role labeler and a heuristic-based algorithm. The provenance triples are aggregated to form a provenance graph.

2.3 ProvCaRe Knowledge Repository: Provenance Search and Query Functions to Characterize Reproducibility of Research Studies

The ProvCaRe knowledge repository aims to facilitate the adoption of a "provenance-aware" literature survey and enable translation of existing guidelines for reproducibility

into practice. The ProvCaRe knowledge repository complements data sharing initiatives such as NIH-funded NSRR and NDAR projects by allowing users to access both study data and provenance metadata. The repository is available as a Web-accessible query and search platform (http://www.provcare.case.edu) with several user-focused features. The ProvCaRe interface features query composition functionality that is similar to popular Web search engines with "auto-complete" feature supported by the open source Apache Solr inverse document indexing application [48] to perform quick search over a large volume of provenance triples (Figure 4 shows the rate of increase of provenance triples generated from articles in the ProvCaRe knowledge repository). The articles identified to be relevant to a search query (*M*) are ranked using a new two-step provenance-based ranking algorithm to generate a "reproducibility rank" for each article in the query result.

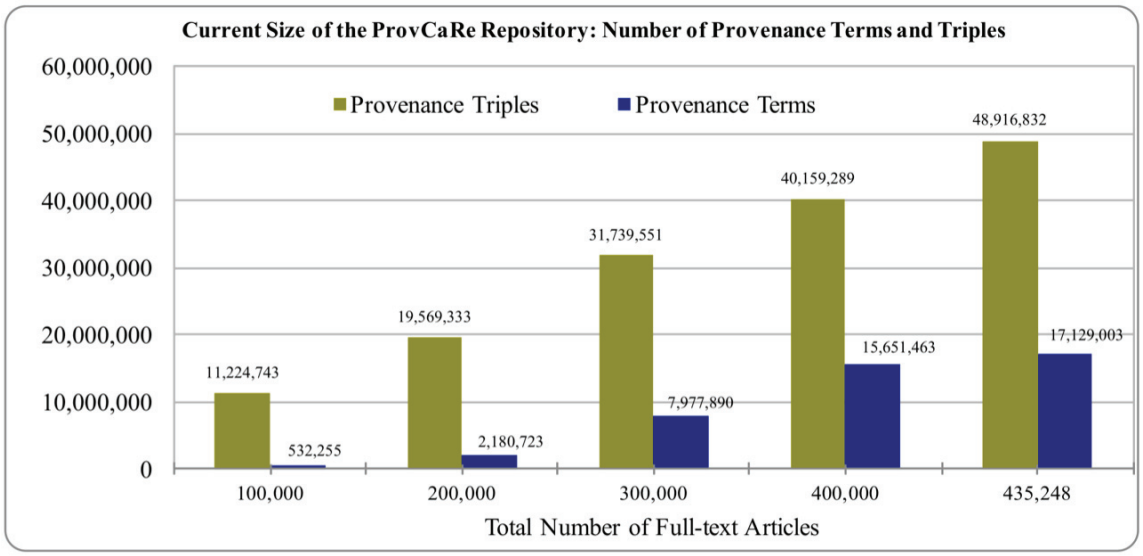

Figure 4: The ProvCaRe repository contains more than 48 million provenance triples with a linear rate of increase in the number of provenance triples extracted from published articles accessible from the PubMed repository.

In the first step, each article is assigned a weight (*pt*) corresponding to the total number of provenance triples extracted from the article. In the second step, the terms in the *subject*, *predicate*, and *object* of a provenance triple are mapped to ProvCaRe ontology terms.

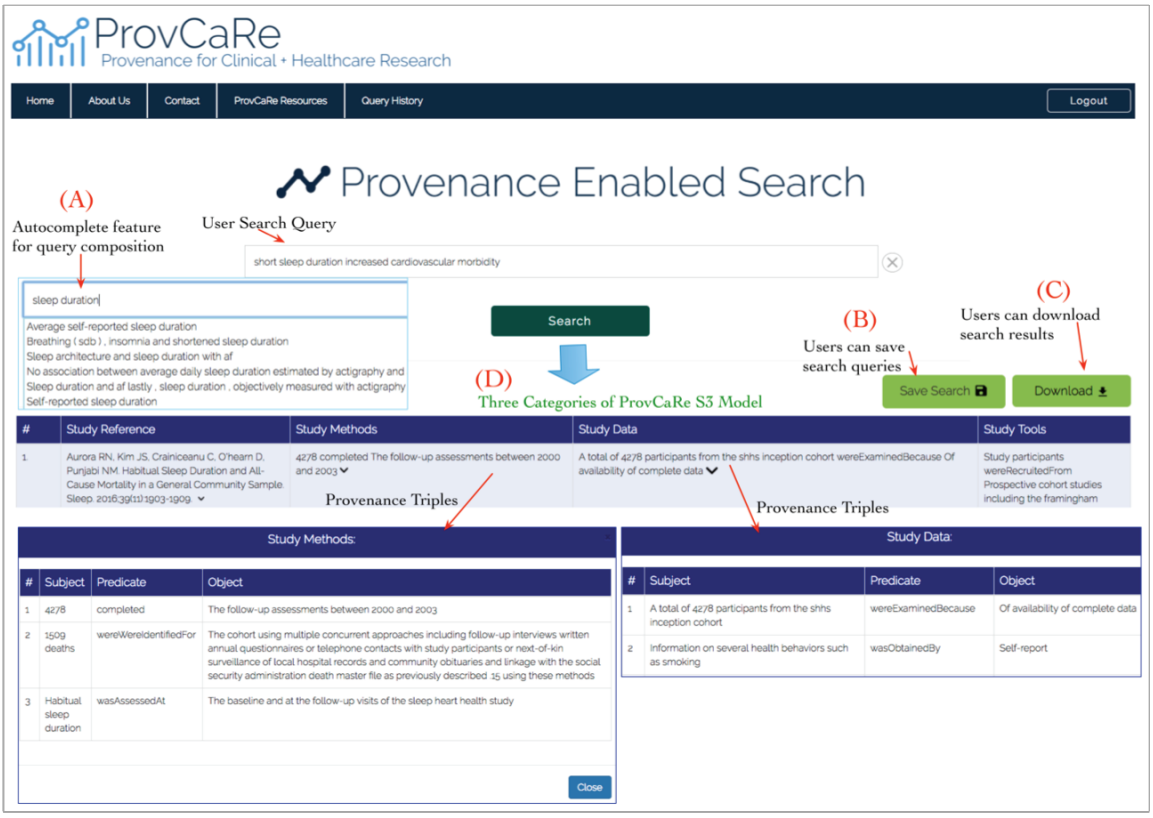

Figure 5: Screenshots of the ProvCaRe repository user interface with: (A) autocomplete feature to facilitate query composition, (B) a save search feature to allow users to store search queries, and (C) a download feature to save the results of a query for further analysis. The results of a search query list articles are ranked using a new provenance-based ranking algorithm.

Using the total number of these ontology mappings for each article ($prov_i$) and the total number of terms in provenance triples of an article ($totalprov_i$), we compute a ratio of $\frac{prov_i}{totalprov_i}$ for each article i. The final rank of an article ($ArticleRank_i$) is defined as the harmonic mean of pt and $\frac{prov_i}{totalprov_i}$, which is used to compute the $ArticleRank_i = 2 \times \frac{pt \times \frac{prov_i}{totalprov_i}}{pt + \frac{prov_i}{totalprov_i}}$. This measure is similar to the F-measure used in information retrieval applications [49]. The provenance-based article rank corresponds to a higher likelihood that the results reported in the article are reproducible. Figure 5 illustrates the results of a query with table-based visualization feature. Users can easily download the provenance triples corresponding to their query results for subsequent analysis. In the next section,

we analyze the characteristics of provenance information in the ProvCaRe knowledge repository.

3. RESULTS

In the following sections, we evaluate the ProvCaRe knowledge repository, the provenance ranking algorithm, and characterize the attributes of the provenance metadata extracted from published articles using seven sleep medicine related research hypotheses.

3.1 Analysis of Provenance Metadata using Hypothesis-driven Search Queries

The seven hypotheses used to evaluate the ProvCaRe knowledge repository explore different aspects of sleep disorders, such as the prevalence of sleep disordered breathing among racial/ethnic groups and the association between heart rate variability and sleep disordered breathing. The provenance triples extracted from the articles corresponding to each hypothesis were analyzed using five categories of provenance terms derived from the ProvCaRe S3 model: (a) statistical data analysis techniques; (b) instruments used in study; (c) software tools; (d) information describing the population cohort, and (e) design of the research studies. The five categories of provenance terms were identified through discussions with members of the NSRR project (co-authors MK and MR). The seven hypotheses used in the evaluation were created by co-author MK and are listed below:

- Hypothesis 1: *The prevalence of sleep disordered breathing varies among different racial/ethnic groups.*
- Hypothesis 2: *The prevalence of periodic limb movement associated with sleep disordered breathing varies among different racial/ethnic groups.*
- Hypothesis 3: *The prevalence of periodic limb movement associated with sleep disordered breathing varies with aging.*
- Hypothesis 4: *Clinical parameters can be used to predict the severity of sleep disordered breathing.*
- Hypothesis 5: *There is a relationship between heart rate variability and sleep disordered breathing.*

- Hypothesis 6: *There is a relationship between sleep disordered breathing and all-cause mortality.*
- Hypothesis 7: *There is a relationship between Vitamin D levels and sleep disordered breathing.*

The results of the search queries are presented in Table 1, which lists the top three articles (in terms of the count of provenance terms) for the seven hypotheses across all five categories of provenance metadata. The results in Table 1 show that provenance triples corresponding to hypothesis 2 related article (PMID: 25325500) have the highest number of terms describing statistical data analysis techniques, including terms such as "t-test" and "p-values". Provenance terms describing the design of studies occur with highest frequency in provenance triples extracted from articles related to hypothesis 1. Similarly, provenance triples extracted from article (PMID: 28835227) related to hypothesis 1 have the highest number of terms describing the study population. Table 1 also lists the number of provenance terms describing software tools used in research studies with the lowest number of terms extracted from article (PMID: 26861778) related to hypothesis 7 and the highest number of provenance terms extracted from article (PMID: 27920726) related to hypothesis 5.

Table 1: The provenance triples extracted from articles related to seven sleep medicine related hypotheses were analyzed using five categories of provenance terms with top three articles (based on number of provenance terms) are listed below.

	Statistical Data Analysis Technique (article ID, provenance terms)		Instruments (article ID, provenance terms)		Software Tools (article ID, provenance terms)		Population Cohort (article ID, provenance terms)		Research Design (article ID, provenance terms)	
1. The Prevalence of sleep disordered breathing varies among different racial/ethnic groups.	PMID: 27768852	16	PMID: 27339289	6	PMID: 28422847	4	PMID: 28835227	40	PMID: 28835227	38
	PMID: 27568910	14	PMID: 28835227	5	PMID: 28848496	4	PMID: 27450684	37	PMID: 27541580	22
	PMID: 28081171	11	PMID: 28081171	5	PMID: 28457559	3	PMID: 28081171	33	PMID: 28457559	17
2. The prevalence of periodic limb movement associated with sleep disordered breathing varies among different racial/ethnic groups.	PMID: 25325500	34	PMID: 26106238	6	PMID: 26197315	5	PMID: 25489744	35	PMID: 25348124	25
	PMID: 25348124	32	PMID: 25325464	5	PMID: 25348124	3	PMID: 27250807	32	PMID: 25325464	22
	PMID: 25489744	24	PMID: 26210395	3	PMID: 25845698	3	PMID: 26106238	28	PMID: 25489744	19
3. There is a relationship between environmental factors and sleep disordered breathing.	PMID: 27070139	32	PMID: 27070639	5	PMID: 26414899	4	PMID: 27070639	29	PMID: 26845389	21
	PMID: 27768852	28	PMID: 26904263	3	PMID: 27091520	3	PMID: 27646537	24	PMID: 27091520	19
	PMID: 26509676	21	PMID: 27314230	2	PMID: 27070139	3	PMID: 27314230	18	PMID: 27810258	18
4. Clinical parameters can be used to predict the severity of sleep disordered breathing.	PMID: 26857052	26	PMID: 28510598	18	PMID: 26280546	4	PMID: 26658438	36	PMID: 26897500	21
	PMID: 26658438	24	PMID: 26414902	13	PMID: 26291487	3	PMID: 28510598	28	PMID: 26291487	16
	PMID: 26280546	23	PMID: 26897500	7	PMID: 26857052	3	PMID: 26857059	16	PMID: 28472141	14
5. There is a relationship between heart	PMID: 11549537	25	PMID: 27920726	15	PMID: 27920726	22	PMID: 25480401	30	PMID: 25480401	21
	PMID:	21	PMID:	5	PMID:	5	PMID:	22	PMID:	17

rate variability and sleep disordered breathing.	25480401		27826247		25555635		25555635		26463420	
	PMID: 25860587	18	PMID: 28899529	5	PMID: 25634206	5	PMID: 27999786	18	PMID: 28118872	12
6. There is a relationship between Vitamin D stores and sleep disordered breathing.	PMID: 26414899	26	PMID: 27707440	2	PMID: 26414899	3	PMID: 26350605	24	PMID: 26845389	19
	PMID: 26845389	22	PMID: 25766695	2	PMID: 24684979	2	PMID: 26414899	17	PMID: 27707440	13
	PMID: 28686746	19	PMID: 26414899	2	PMID: 25580607	2	PMID: 25669179	16	PMID: 25766695	13
7. There is a relationship between sleep disordered breathing and all-cause mortality.	PMID: 26886528	24	PMID: 28146212	3	PMID: 27307401	2	PMID: 27450684	31	PMID: 26856225	28
	PMID: 27307401	21	PMID: 27655449	3	PMID: 29193576	2	PMID: 27105053	28	PMID: 26886528	18
	PMID: 28146212	21	PMID: 25633255	3	PMID: 26861778	1	PMID: 26038534	28	PMID: 26316620	17

The results highlight that articles include more provenance terms for population cohort category as compared to statistical data analysis techniques category (for top three articles across all seven hypotheses). We note that the NIH Rigor and Reproducibility guidelines emphasize the inclusion of detailed description about statistical methods used in a study to support reproducibility. It is interesting to note that some articles have consistently more provenance-related terms across the five categories of provenance terms analyzed. For example, PMID 26414899 is among the top three articles for hypothesis 6 across four of the five provenance categories. Similarly, two articles (PMID: 28081171 and PMID: 25348124) for hypothesis 1 and 2 describe provenance terms for three categories. The multiple occurrences of articles across different categories of provenance metadata may indicate that some articles include more provenance metadata information compared to other articles. In the next section, we describe the results of the new two-step provenance ranking algorithm that allows ranking-based listing of query

results, which often include large number of articles (e.g., 1986 articles for hypothesis 4 and 1311 for hypothesis 5).

3.2 Provenance-based Ranking of Articles

Provenance-based ranking of articles allows users to quickly find articles that provide more provenance metadata information with a corresponding higher likelihood of reproducibility as compared to other articles. We use the seven hypotheses-based search queries to demonstrate the computation of a reproducibility rank for each article in Table 2 with intermediate values corresponding to the total number of provenance triples in an article and the total number of ontology mappings. It is interesting to note that although the number of provenance triples extracted from the top three articles for hypothesis 6 (323 triples) is higher than the number of provenance triples extracted from articles for hypothesis 7 (283 triples), the average reproducibility rank of articles for hypothesis 7 (0.61) is higher than articles for hypothesis 6 (0.53). These results show that the reproducibility rank of an article reflects the quality of provenance terms (mapped to the ProvCaRe ontology terms) instead of only a count of provenance triples extracted from an article. We note that the average provenance rank of the top three articles for hypothesis 3 is the highest (0.66) and it is lowest (0.53) for hypothesis 6 among the seven hypotheses.

Hypothesis 3 has the highest number of ontology mappings for the top three articles whereas hypothesis 2 has the lowest number of ontology mappings for top three articles. A review of the highest number of ontology mappings (for article PMID: 28686746) showed that the maximum number of provenance terms were mapped to ProvCaRe ontology class `provcare:ResearchStudy` followed by mappings to classes

`provcare:StudyPopulation` and `provcare:StudyOutcome` respectively. It is interesting to note that the ratio of the total number of provenance triples and the total number of ontology mappings is highest for an article by Migacz et al. (PMID: 28877768) related to hypothesis 4 and that an article by Xie et al. (PMID: 29275335) related to hypothesis 7 has the lowest number of ontology mappings. In future, it may be helpful to manually validate the reproducibility of studies listed in Table 2 to improve the function of the provenance-based ranking algorithm. In the following section, we describe this and related issues the role of the ProvCaRe platform in scientific reproducibility.

Table 2: Given the large number of articles retrieved as a result of search queries, a new provenance-based ranking algorithm is used to rank and list articles in the ProvCaRe repository. The table lists the results of the intermediate steps of the algorithm that is used to compute the final rank of an article.

Hypothesis	Total Papers	Total Number of Provenance Triples Articles		Total Number of Ontology Mappings	Reproducibility Rank
1. The Prevalence of sleep disordered breathing varies among different racial/ethnic groups.	749	PMID:28848496	111	156	.58
		PMID:28686746	91	118	.56
		PMID:28457559	82	113	.55
2. The prevalence of periodic limb movement associated with sleep disordered breathing varies among different racial/ethnic groups.	328	PMID:28100870	112	131	.62
		PMID:28822017	98	119	.61
		PMID:26725017	108	126	.58
3. There is a relationship between environmental factors and sleep disordered breathing.	471	PMID:28716800	126	158	.69
		PMID:28852230	108	149	.65
		PMID:27878796	115	160	.64
4. Clinical parameters can be used to predict the	1,986	PMID:29060229	101	138	.70
		PMID:29016682	120	147	.64

severity of sleep disordered breathing.		PMID:28877768	105	122	.63
5. There is a relationship between heart rate variability and sleep disordered breathing.	1,311	PMID:28862662	119	162	.68
		PMID:28534047	111	150	.66
		PMID:28445548	107	142	.63
6. There is a relationship between Vitamin D stores and sleep disordered breathing.	21	PMID:28686746	132	179	.55
		PMID:27684795	90	126	.53
		PMID:24684979	101	142	.52
7. There is a relationship between sleep disordered breathing and all-cause mortality.	1,129	PMID:29214822	104	147	.66
		PMID:29070017	97	119	.61
		PMID:29275335	82	116	.58

4. DISCUSSION

Although the current ProvCaRe knowledge repository represents a unique resource with large volume of provenance metadata, there is a clear need to include additional sources to improve the coverage and comprehensiveness of the ProvCaRe repository.

Additional Sources of Research Study Provenance. Although peer-reviewed articles are the primary source of provenance metadata for research studies, there are multiple additional sources of research study provenance including supplementary material and Web-accessible data repositories. Therefore, we are implementing a new functionality in the ProvCaRe pipeline to locate and process information in supplementary material of published articles. For example, we processed the supplementary material associated with research study by O'Connor et al. to extract 33 additional provenance triples [37]. In addition, research studies in some domains such as protein structure determination publish details of their experiment method in Web-accessible databases.

The Protein Structure Initiative (PSI) Structural Biology Knowledgebase (SBKB) provides provenance metadata describing the experimental history of each protein target along with the protocols used for production and structure determination of the protein [50]. We are exploring approaches to extract and include provenance information from these databases in the ProvCaRe repository. There is a clear need to incentivize investigators to share both data and provenance metadata with attribution, for example through the use of the Findable, Accessible, Interoperable and Re-usable (FAIR) principles for data management [51]. Therefore, sharing of research study data as well as provenance metadata requires the biomedical research community to also address social and administrative challenges.

Provenance Ranking and Role of Provenance Graph Properties. The current provenance-based ranking technique in ProvCaRe relies on the count of provenance triples of an article and the relevance of the terms in provenance triples in the context of scientific reproducibility (using the ProvCaRe ontology as reference model). However, this approach does not use the properties of the provenance graphs associated with each article, for example network density, average degree of nodes, and centrality of nodes in a provenance graph. Integrating appropriate network properties into the ProvCaRe ranking algorithm may allow identification and use of provenance terms with higher value of eigenvector centrality, which measures the influence of a node in a graph. This ranking approach is expected to more accurately reflect the provenance information associated with an article and can be potentially used to derive a measure for scientific reproducibility.

This provenance graph-based ranking approach extends existing approaches that have used attributes of a provenance graph, for example the number of links between nodes together with graph traversal techniques such as "random walk algorithm" [52]. We note that extraction of provenance graph properties requires the use of a query language for traversing provenance graph structure, for example ProQL [53] and Query Language for Provenance (QLP) [36]. As part of our ongoing work, we propose to integrate a provenance graph query language in ProvCaRe to support computation of provenance ranking.

5. CONCLUSIONS

There is a clear need to improve the availability of provenance metadata to support scientific reproducibility and complement the increasing availability of research study datasets from NSRR, TCGA, and NDAR projects. In this paper, we presented the ProvCaRe platform that extends the W3C PROV specification to model, extract, and analyze provenance metadata from PubMed articles. The ProvCaRe S3 model consists of Study Method, Study Data, and Study Tool terms to represent provenance metadata. We developed a provenance-focused text processing workflow by extending the cTAKES pipeline to extract provenance metadata from published articles related to sleep medicine research. We analyzed the resulting 48.9 million provenance triples stored in the ProvCaRe knowledge repository to characterize the occurrence of five categories of provenance metadata in published articles. The ProvCaRe knowledge repository (https://provcare.case.edu/) is one of the largest provenance resource for biomedical research studies that combines intuitive search functionality with a new provenance-based ranking feature to list articles related a search query.

Funding

This work is supported in part by the NIH-NIBIB Big Data to Knowledge (BD2K) 1U01EB020955 grant, NSF grant# 1636850, and the NIH-NHLBI R24HL114473 grant.

Contributors

SSS and JV led the development and implementation of the text processing pipeline, user interface, and analysis of provenance metadata. MK, MR, SR, SSS, and JV contributed to the development of the S3 model and the ProvCaRe repository interface. MK and MR contributed to the development of a gold standard used to train the classifier. All authors contributed to the writing and editing of the paper.

Competing Interests

The authors have no competing interests to declare.

Acknowledgements

We thank Sara Arabyaarmohammadi for contributing to the code used to identify and extract articles used in the study from PubMed.

References

1. Landis SC, Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H.E., Golub, R.M., Goudreau, J.L., Gross, R.A., Gubitz, A.K., Hesterlee, S.E., Howells, D.W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S.E., Levine, M.S., Macleod, M.R., McCall, J.M., Moxley, R.T. 3rd, Narasimhan, K., Noble, L.J., Perrin, S., Porter, J.D., Steward, O., Unger, E., Utz, U., Silberberg, S.D.: A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 2012, 490(7419):187-191.
2. Baker M: 1,500 scientists lift the lid on reproducibility. *Nature* 2016, 533(7604):452-454.

3. Munafò MR, Nosek, B. A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E., Ware, J.J., Ioannidis, J.P.A.: A manifesto for reproducible science. *Nature Human Behavior* 2017, 1:0021 EP-.
4. Steward O, Popovich, P.G., Dietrich, W.D., Kleitman, N.: Replication and reproducibility in spinal cord injury research. *Experimental Neurology* 2012, 233(2):597-605.
5. Prinz F, Schlange, T., Asadullah, K.: Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 2011, 10(9):712.
6. Collins FS, Tabak, L.A.: Policy: NIH plans to enhance reproducibility. *Nature* 2014, 505:612-613.
7. Dean DA, Goldberger, A.L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., Sahoo, S.S., Jayapandian, C.P., Cui, L., Morrical, M.G., Surovec, S., Zhang, G.Q., Redline, S.: Scaling up scientific discovery in sleep medicine: the National Sleep Research Resource. *SLEEP* 2016, 39(5):1151–1164.
8. Collins FS, Barker, A.D.: Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Scientific American* 2007, 296(3):50-57.
9. The National Institute of Mental Health Data Archive (NDA) [https://data-archive.nimh.nih.gov/], Retrieved on January 24, 2018
10. Principles and Guidelines for Reporting Preclinical Research [https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research], Retrieved on January 24, 2018
11. Nosek BA, Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D.P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T.A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E.L., Simonsohn, U., Soderberg, C., Spellman, B.A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.J., Wilson, R., Yarkoni, T.: Promoting an open research culture. *Science* 2015, 348(6242):1422-1425.
12. Goble C: Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics. In: *Workshop on Data Derivation and Provenance: 2002; Chicago*.
13. Sahoo SS, Sheth, A., Henson, C.: Semantic Provenance for eScience: Managing the Deluge of Scientific Data. *IEEE Internet Computing* 2008, 12(4):46-54.
14. Moreau L, Missier, P.: PROV Data Model (PROV-DM). *W3C Recommendation.* World Wide Web Consortium W3C; 2013.
15. Buneman P, Davidson, S.: Data provenance - the foundation of data quality. 2010 [*http://www. sei. cmu. edu/measurement/research/upload/Davidson. pdf*], Retrieved on January 24, 2018.
16. Cheney J, Chiticariu, L., Tan, W. C.: Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases* 2009, 1(4):379-474.
17. Green TJ, Karvounarakis, G. ,Tannen, V.: Provenance Semirings. In: *ACMSIGMOD-SIGACTSIGART Symposium on Principles of database systems (PODS): 2007*. 675–686.

18. Wolstencroft K, Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., Nieva de la Hidalga, A., Balcazar Vargas, M.P., Sufi, S., Goble, C.: The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* 2013, 41:W557-561.
19. Ludascher B, Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the Kepler system: Research Articles. *Concurr Comput : Pract Exper* 2006, 18(10):1039-1065.
20. Barga R, Jackson, J., Araujo, N., Guo, D., Gautam, N., Simmhan, Y.: The trident scientific workflow workbench. In: *IEEE Fourth International Conference on eScience: 2008; Bloomington IN*. IEEE: 317-318.
21. Simmhan YL, Plale, A.B., Gannon, A. D.: A survey of data provenance in e-science *SIGMOD Rec* 2005, 34(3):31 - 36
22. Lebo T, Sahoo, S.S., McGuinness, D.: PROV-O: The PROV Ontology. In: *W3C Recommendation.* World Wide Web Consortium W3C; 2013.
23. Cheney J, Missier, P., Moreau, L.: Constraints of the PROV Data Model. In: *W3C Recommendation.* World Wide Web Consortium W3C; 2013.
24. Richardson WS, Wilson, M.C., Nishikawa, J., Hayward, R.S.: The well-built clinical question: a key to evidence-based decisions. *ACP J Club* 1995, 123(3):A12-13.
25. Huang X, Lin, J., Demner-Fushman, D.: Evaluation of PICO as a knowledge representation for clinical questions. In: *AMIA Annual Symposium Proceedings*. 2006: 359-363.
26. Schulz KF, Altman, D.G., Moher, D., CONSORT Group.: CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology* 2010, 63(8):834-840.
27. Kilkenny C, Browne, W.J., Cuthill, I.C., Emerson, M., Altman, D.G.: Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology* 2010, 8(6):e1000412.
28. Sim I, Tu, S.W., Carini, S., Lehmann, H.P., Pollock, B.H., Peleg, M., Wittkowski, K.M.: The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. *Journal of Biomedical Informatics* 2014, 52:78-91.
29. Valdez J, Kim, M., Rueschman, M., Socrates, V., Redline, S., Sahoo, S.S.: ProvCaRe Semantic Provenance Knowledgebase: Evaluating Scientific Reproducibility of Research Studies In: *American Medical Informatics Association (AMIA) Annual Conference; Washington DC*. AMIA 2017.
30. Valdez J, Rueschman, M., Kim, M., Arabyarmohammadi, S., Redline, S., Sahoo, S.S.: An Extensible Ontology Modeling Approach Using Post Coordinated Expressions for Semantic Provenance in Biomedical Research. In: *The 16th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE): 2017; Rhodes, Greece*.
31. Valdez J, Rueschman, M., Kim, M., Redline, S., Sahoo, S.S.: An Ontology-Enabled Natural Language Processing Pipeline for Provenance Metadata

Extraction from Biomedical Text. In: *15th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)* 2016: 699-708.
32. Entrez Programming Utilities (EUtils) [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html], Retrieved on January 24, 2018
33. Hitzler P, Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S.: OWL 2 Web Ontology Language Primer. In: *W3C Recommendation.* World Wide Web Consortium W3C; 2009.
34. SNOMED Clinical Terms [http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html], Retrieved on January 24, 2018
35. Gil Y, Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., Pinheiro da Silva, P., Coppens, S., Garijo, D., Manuel Gomez, J., Missier, P., Myers, J., Sahoo, S.S., Zhao, J.: Provenance xg final report. In: *W3C Technical Report.* W3C; 2010.
36. Anand MK, Bowers, S., Ludäscher, B.: Techniques for efficiently querying scientific workflow provenance graphs. . In: *Proceedings of the 13th international Conference on Extending Database Technology: 2010; Lausanne, Switzerland*. ACM, New York, NY: 287-298.
37. O'Connor GT, Caffo, B., Newman, A.B., Quan, S.F., Rapoport, D.M., Redline, S., Resnick, H.E., Samet, J., Shahar, E.: Prospective study of sleep-disordered breathing and hypertension: the Sleep Heart Health Study. *Am J Respir Crit Care Med* 2009, 179(12):1159-1164.
38. Savova GK, Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010, 17(5):507-513.
39. TensorFlow [https://www.tensorflow.org/], Retrieved on January 24, 2018
40. Collobert R, Weston, J., Bottou, L., Karlen, M., Kavukcuglu, K., Kuksa, P.: Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 2011, 12:2493–2537.
41. Kim Y: Convolutional neural networks for sentence classification. In: *arXiv preprint* arXiv:1408.5882; 2014.
42. Valdez J, Kim, M., Rueschman, M., Redline, S., Sahoo, S.S.: Classification of Provenance Triples for Scientific Reproducibility: A Comparative Evaluation of Deep Learning Models in the ProvCaRe Project. In: *International Provenance Annotation Workshop (IPAW): 2018; London, UK*. Springer.
43. Horridge M, Bechhofer, S.: The OWL API: A Java API for OWL Ontologies. *Semantic Web Journal* 2011, 2(1):11-21.
44. Aronson AR: MetaMap: Mapping Text to the UMLS Metathesaurus. In.: US NLM; 2006.
45. Jonquet C, Shah, N.M., Musen, M.A.: The Open Biomedical Annotator. In: *AMIA Summit on Translat Bioinformatics; San Francisco*. AMIA 2009: 56-60.
46. Jurafsky D, Martin, J.H.: Speech and Language Processing.: Prentice Hall; 2015.
47. Dahlmeier D, Ng, H.T.: Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics* 2010, 26(8):1098-1104.
48. Apache Solr [lucene.apache.org/solr/], Retrieved on January 24, 2018

49. Van Rijsbergen CJ: Information Retrieval. MA, USA: Butterworth-Heinemann Newton; 1979.
50. Gabanyi MJ, Adams, P. D., Arnold, K., et al.: The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *Journal of Structural and Functional Genomics* 2011, 12(2):45–54.
51. Wilkinson MD, Dumontier, M., Aalbersberg, I.J., et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016, 3(160018).
52. Ives ZG, Haeberlen, A., Feng, T.: Querying Provenance for Ranking and Recommending In: *4th USENIX Workshop on Theory and Practice of Provenance; Boston, MA*. USENIX 2012.
53. Karvounarakis G, Ives, Z. G., Tannen, V.: Querying Data Provenance In: *Proceedings of the 2010 international Conference on Management of Data: 2010; Indianapolis, Indiana, USA*. ACM, New York, NY: 951-962.