Building Bridges: Making Statistical Issues Accessible to the Biomedical and Translational Researcher

Taylor R. Pressler[1*], Philip F. Binkley[2], and Philip R.O. Payne[1]
[1]The Ohio State University Medical Center, Department of Biomedical Informatics
[2]The Ohio State University Medical Center, Department of Internal Medicine

Abstract: As biomedical and translational research expands, the need for effective communication between statisticians and scientists presents a major challenge. The communication issues between these two fields affect the quality of research. Research design, inferential statements and outcomes, and the use of proper statistical methodologies are all areas in which statisticians can offer expertise to further scientific research.

One approach to increasing communication is to build an electronic knowledge-base describing statistical methodologies and issues that is accessible to biomedical researchers. This knowledge base also includes a decision support system that guides users to information and literature that relate to their project. The goal of this project is to design, implement, and validate a decision support system (DSS) platform that targets a gap in knowledge related to statistical literacy as applied to biomedical and translational research.

This paper will present the preliminary results of the Statistical Information for Biomedical Research engine (SIBRe) project. The project described in this manuscript consists of three complementary stages, the first stage using literature and end users to identify important concepts and components of such a system. A representation of the knowledge and concepts identified is created during Phase 2,and Phase 3 implements a computer software platform that facilitates scientific interaction.

Key Words: Statistical literacy, biomedical research, translational research, biomedical informatics, decision support system.

1. Introduction

It has become increasingly clear that scientific research has become more interdisciplinary and advanced in nature[1-3]. Commensurately, statistical techniques are becoming more complex as the field of statistics has advanced over the past 50 years. In a comparative study of scientific articles from the journal *Pediatrics*, Hellems et al found that the rate of articles using only descriptive statistics is declining dramatically and the top ten types of statistical procedures found in the articles examined in the study include logistic regression, ANOVA, non-parametric tests, and procedures typically not covered in an elementary biostatistics course[4]. Many biomedical and clinical researchers may take a basic biostatistics course during their course of study, but they also learn from evidence based settings such as conferences and journal clubs [4, 5]. As the statistics within journal articles becomes more advanced and the types of statistical methods and inferential procedures becomes more sophisticated, there is a widening gap of statistical

[*] Corresponding author: *taylor.pressler@osumc.edu*

knowledge for researchers. They must not only be able to read and absorb information from scientific literature, but also find the most appropriate methods for their own prospective research. Such a gap is particularly important, given the need for biomedical and translational researchers to have a functional level of statistical knowledge. Statistical literacy is defined on three levels [6, 7] as follows:

- understanding of statistical methods.
- the ability to understand the effects of statistics on research design and analysis.
- a working vocabulary of statistical terms.

Ideally, all researchers should have a functional level of statistical literacy. As biomedical research uses more advanced techniques, researchers may struggle more with finding the information necessary to keep up with current methods.

2. Background

There are gaps in knowledge that can lead to the mis-use of statistical techniques or the mis-interpretation of statistical results. Statisticians do not always have a clear understanding of the scientific data with which they are working and analyzing. Likewise, scientists do not have a clear understanding of research design methods and the differences and underlying assumptions that statistical methods carry. There are two current areas of research and practice that exist which can be utilized and contribute to the formation of a solution to such barriers, namely:

- Statistical literacy.
- Knowledge engineering and knowledge base systems.

In the work described later in this manuscript, the preceding two fields of study are combined in a novel way to create a platform that can assist in increasing the statistical literacy of biomedical and translational scientists.

2.1 Statistical Literacy Research

Statistical literacy research spans many disciplines. Many research fields have written extensively about the advances in statistical methodologies and the lack of equal progress in statistical literacy. As previously stated, the advancement of more sophisticated statistical methods is reflected in current literature. Hellems states that the proportion of articles in the journal *Pediatrics* that used inferential statistical procedures was approximately 48% in 1982 and increased to approximately 89% in 2005 [4]. However, it has been documented that many clinicians and biomedical researchers have difficulty in interpreting the medical data that is found in such articles [4, 8-11]. Thus, it is necessary to understand where the gaps in knowledge exist for researchers in order to plan solutions for filling them.

Ludbrook and Dudley give an example of a lack of communication between biomedical researchers and statisticians and the effect this misunderstanding has on the resultant analysis that is used for a study, even when working in collaboration with statisticians. The authors reviewed 252 articles from four journals and found that only 4% of the articled used a random sampling method, while the other articles used randomization of the units [12]. Of the studies that used a randomized sample design, 84% used the t-test or F-test to analyze data instead of the more appropriate non-parametric techniques [12]. They use this example to highlight that a major problem in current research occurs when statisticians believe that most researchers are using random sampling techniques when they are instead more often using randomized design and that the researchers are either not aware of the difference in the methods or are unclear of the assumptions of the F and t-tests [12]. Anderson and Warnapala also write that researchers can often misinterpret data or omit measures in the analysis of the data that are critical [13]. Many biomedical studies also have small sample sizes, which can also effect inferential outcomes, but the effects of which are rarely discussed in the published results [12].

2.2 Knowledge Engineering and Knowledge Base Systems

Knowledge engineering and the use of knowledge based systems are techniques that are commonly found in the area of biomedical informatics. The knowledge engineering process looks to identify and process three types of knowledge into a computable form[1]. The knowledge engineering process has three steps: (1) Knowledge acquisition, (2) knowledge representation, and (3) Implementation. Refinement of the engineered knowledge systems then takes place and a decision support system can then be created from the engineered knowledge. A decision support system is a system which provides reference information, reminders, guidelines, and alerts concerning a domain, leveraging knowledge collections or bases established using the preceding knowledge engineering process [14]. Such reference material can be active-provided up front by the decision support system- or it can be passive-sought out by the user given the input of certain criteria [14]. Decision support systems are most effective when there is enough expert input during the collection of knowledge items and representation of terms [15].

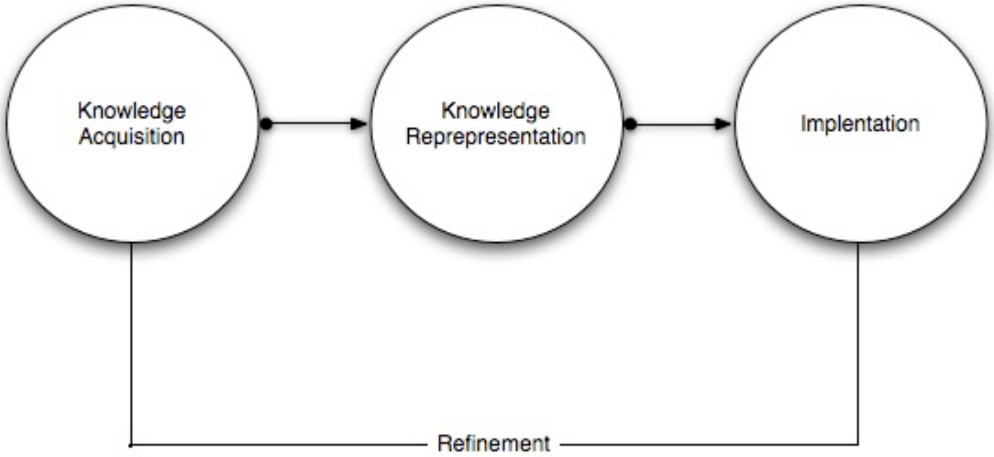

Figure 1: The knowledge engineering process starts with the acquisition of knowledge within a domain. The acquired knowledge is then represented in a computable form, and this form is then refined and used for implementation of a decision support system.

2.2.1 Acquisition of knowledge

Knowledge acquisition is the process by which domain knowledge is acquired for subsequent computation representation and reasoning. It involves identifying different types of knowledge, eliciting the knowledge, and then verifying and validating the knowledge [1, 16]. There are three types of knowledge that should be identified in the knowledge acquisition process. The first type of knowledge is conceptual knowledge, which can be defined as units of information and the relationships between units[1, 16]. The second type of knowledge is procedural knowledge, which is the “process-oriented understanding[1]” of a problem. The third type of knowledge is strategic knowledge, which spans both procedural and conceptual knowledge and could be characterized as understanding the relationships between objects within a domain and knowing how to use them to solve some set problem in that domain.

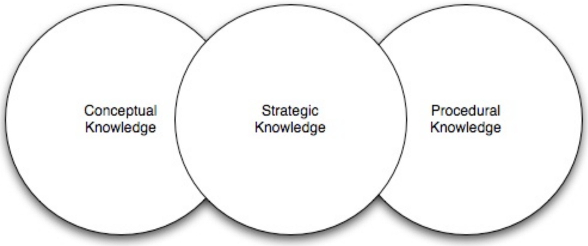

Figure 2: Three types of knowledge. Adapted from R. McCormick and P. Payne [1, 16].

Knowledge can be acquired through multiple sources, including experts, databases, and literature [1]. For the purpose of this project, the domain of interest is statistics, which includes related information in statistical literacy research. Statistical texts, research literature, biomedical and translational researchers, and statisticians are all sources of knowledge that will be used in this knowledge acquisition process of this study.

2.2.2 Knowledge Representation

Knowledge representation can take many forms. In this particular project, the representation of knowledge will use a construct known as an ontology, which can provide a basis for integrating and processing information in a decision support system [17]. An ontology is built by creating a collection of: 1) atomic conceptual entities (such as objects or tasks); and 2) semantic and hierarchical relationships that serve to define the interrelationships between such entities [18]. The first step in creating an ontology is to specify the concepts within the targeted domain that should be included. This is usually achieved during the previously defined knowledge acquisition phase of the overall knowledge engineering process, and is often continued during the refinement of the resulting knowledge-based system or knowledge collection. Subsequently, the interrelationships between the included entities are defined and assigned using logical assertions, usually via a collaborative process involving subject matter experts and knowledge engineering practioners [19]. The resulting ontology can then serve as the foundation for the implementation of a knowledge-based system.

2.2.3 Implementation

In this particular study, the knowledge-base being created is intended to inform the implementation of a decision support system (DSS). In a broad sense, such systems allow users to input context- or problem-specific information such that the DSS may reason upon that information using an underlying knowledge base and business logic (for example, leveraging an ontology) in order to support the decision making activities of the end-user who is interacting with the system. Such platform can be either active (e.g. prospectively reasoning on information and issuing alerts through a variety of mediums) or passive (e.g., end-users access and interact with the system on an on-demand basis). One example of a prototypical clinical decision support system is MYCIN. The MYCIN system assists clinicians in selecting appropriate antimicrobial therapy [20]. This system allows a user to input information relevant to the patient and MYCIN will return the user with antimicrobial therapy choices as well as links to primary literature through Medline [20].

3. Methods

The primary objective of this study is to implement and evaluate the efficacy of a DSS platform targeting the previously defined gap in knowledge relative to clinical/translational research related statistical literacy. We have labelled this DSS platform as the Statistical Information for Biomedical Research engine (SIBRe). The SIBRe platform is intended to serve an electronic resource and an experimental decision support system for clinical and translational researchers. The design of SIBRe has followed a multi-modal approach, as described in the following section. The design and implementation of SIBRe is motivated in part by end user requirements articulated by members of the OSU Center for Clinical and Translational Science (CCTS), which emphasize the need to help a researcher to answer two questions: 1) What is the best methodology for my experiment/research? 2) How can I find an expert for consultation/collaboration?

There overall implementation process for the SIBRe project is summarized in Figure 1. There are three main phases of this implementation process. Phase 1 consists of the knowledge acquisition in the domain of statistics, phase 2 consists of formalizing the acquired knowledge generated during phase 1 using knowledge engineer best practices in order to create an ontology construct, and phase 3 deals with the verification and validation of the ensuing knowledge base. Based upon the output of these three phases, the SIBRe platform will being implemented

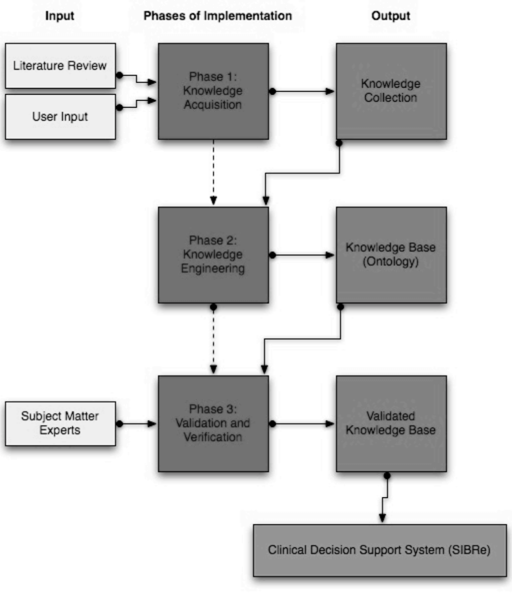
Input
Phases of Implementation
Output
Literature Review
User Input
Phase 1: Knowledge Acquisition
Knowledge Collection
Phase 2: Knowledge Engineering
Knowledge Base (Ontology)
Subject Matter Experts
Phase 3: Validation and Verification
Validated Knowledge Base
Clinical Decision Support System (SIBRe)

Figure 3: Phases of Implementation for the SIBRe tool.

3.1 Phase 1: Knowledge Acquisition

The knowledge acquisition phase of the SIBRe implementation is based upon two sources of information: literature and potential end users of SIBRe. The literature review examines the role of statistics in current published research and serves as a basis for which the ontology created. Input from potential end users of SIBRe, in this case biomedical and translational researchers, is also a valuable source of knowledge.

3.1.1 Literature Review

The literature review begins with a comparative study on recently published articles. In order to carry out this study, a list of current biomedical researchers from the OSU CCTS was retrieved. The published research articles from 2008,2007, and 2006 where then compiled for each CCTS researcher. Approximately 1,082 articles were collected and then a simple random sample of size 100 was taken from the sampled articles. Each article was surveyed and the statistical methodology used in the research was recorded.

3.1.2 End User Verification and Validation

End users are very important to the implementation of the SIBRe system. In order to recognize the needs of the research community, initial input from researchers is essential. It also serves to verify the knowledge that will then be represented in Phase II. Data is collected through two mechanisms: survey data and card sorts.

The survey is an electronic volunteer response survey that is sent to researchers associated with the Center for Clinical and Translational Science (CCTS, ccts.osu.edu) at Ohio State University (OSU). In the survey, the participants are asked to identify their level of knowledge of various statistical techniques, the use of various statistical techniques in their field of expertise, and the use of various statistical techniques in their own research. They are also asked to identify hurdles or setbacks that they feel hinder their research with respect to statistics.

The card sorting techniques are carried out in a focus group setting. During the focus group, participants are asked to sort a set of cards that include the names of different statistical techniques and different types of data. The researchers are asked to sort the cards so that the appropriate statistical methods are groups with the appropriate types of data. They are again asked to repeat the exercise such that types of data are groups with statistical methods. The collected data is then analyzed using cluster analysis. Card sorting exercises are widely used in conceptual knowledge acquisition studies and are considered highly appropriate for statistical analysis[21].

3.2 Phase 2: Knowledge Representation

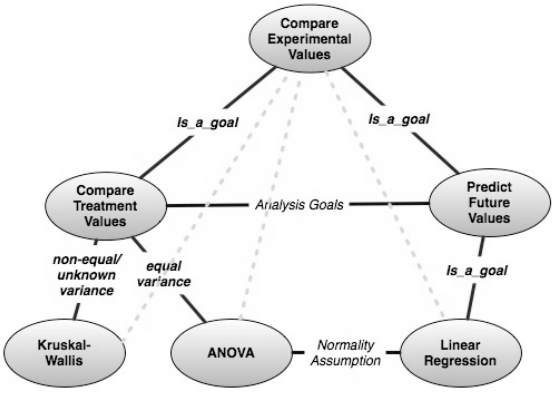

Figure 4: A sample from the SIBRe ontology showing a simple example.

The knowledge representation scheme being used for this project is in an ontology. Specifically, the information and knowledge identified and collected during Phase 1 is aggregated and represented an ontology that will map the concepts and relationships between the items in the knowledge base, as illustrated in Figure 4. Protégé OWL[22] is used as the primary environment for authoring the ontology. Phase 2 runs concurrently with Phase 1, so as more information is gathered, that information is then represented in the ontology. Such an approach is commensurate with the cyclical and iterative nature of prevailing knowledge engineering best practices in the biomedical domain.

3.2 Phase 3: Implementation

After Phases 2 and 3 are completed, a software platform is built using the ontology from Phase 2 as the foundational knowledge base. The design and implementation of the software platform is performed by a biomedical informatics developer, using standards-based rules-engine technologies (e.g., leveraging the Semantic Web Rules Language (SWRL) and Jess rules-engine, both of which are tightly integrated with the Protégé OWL knowledge authoring environment).

The planned implementing SIBRe includes and iterative and ongoing process of validating that the software works the way it is intended. In order to validate the DSS, experts from both statistics and informatics are engaged in a two hour session in which they will be asked to complete two series of tasks using the SIBRe software system and a comparable software. A random selection of half of the participants will complete the first series of tasks using the SIBRe using Google as a search engine and the second series of tasks using the SIBRe system. The other half of the participants will be asked to complete the first series of tasks on SIBRe and then on Google. The use of the MORAE [23] software suite will be used to analyze the outcome of this study. The MORAE software suite allows a researcher to keep track of screen clicks, screen transitions, as well as video to capture cognitive behaviours exhibited by the users. The information and data from the MORAE software is then analyzed quantitatively and thematically such that researchers can identify whether the tasks were able to be completed using the targeted tools and methods, where users became frustrated by the tools/methods, and how the tools/methods compared in efficiency and usability for the user.

4. Results

Partial results for Phases 1 are available at the time of submission, and are summarized below:

4.1 Phase 1

The literature review showed a vast array of statistical methods and inferential techniques that are being used currently by researchers in biomedical and translational publications. The research areas that were included in the literature sample include Pharmacology, Biochemistry, Hematology, Informatics, Cardiology, Internal Medicine, Genetics, Psychiatry, and others. Of these articles in the literature sample, 24% of them included no statistics or inferential procedures. These articles were typically review papers or structural studies. Descriptive statistics were the most common type of statistics to be used in the articles. Descriptive statistics include the use of mean, median, standard error, standard deviation, minimum, and maximum. It was common for both parametric and non-parametric techniques to be used by the authors. There were four papers in which the authors developed their own statistical and/or mathematical model. There were several papers that included more sophisticated techniques, including Bayesian criteria, mixed linear models, Holm's procedure.

The data from the card sorting techniques is not yet available at the time this publication was written, but this data should be able to give some insight to the process that researchers use to choose the statistical methods that are used in their respective research.

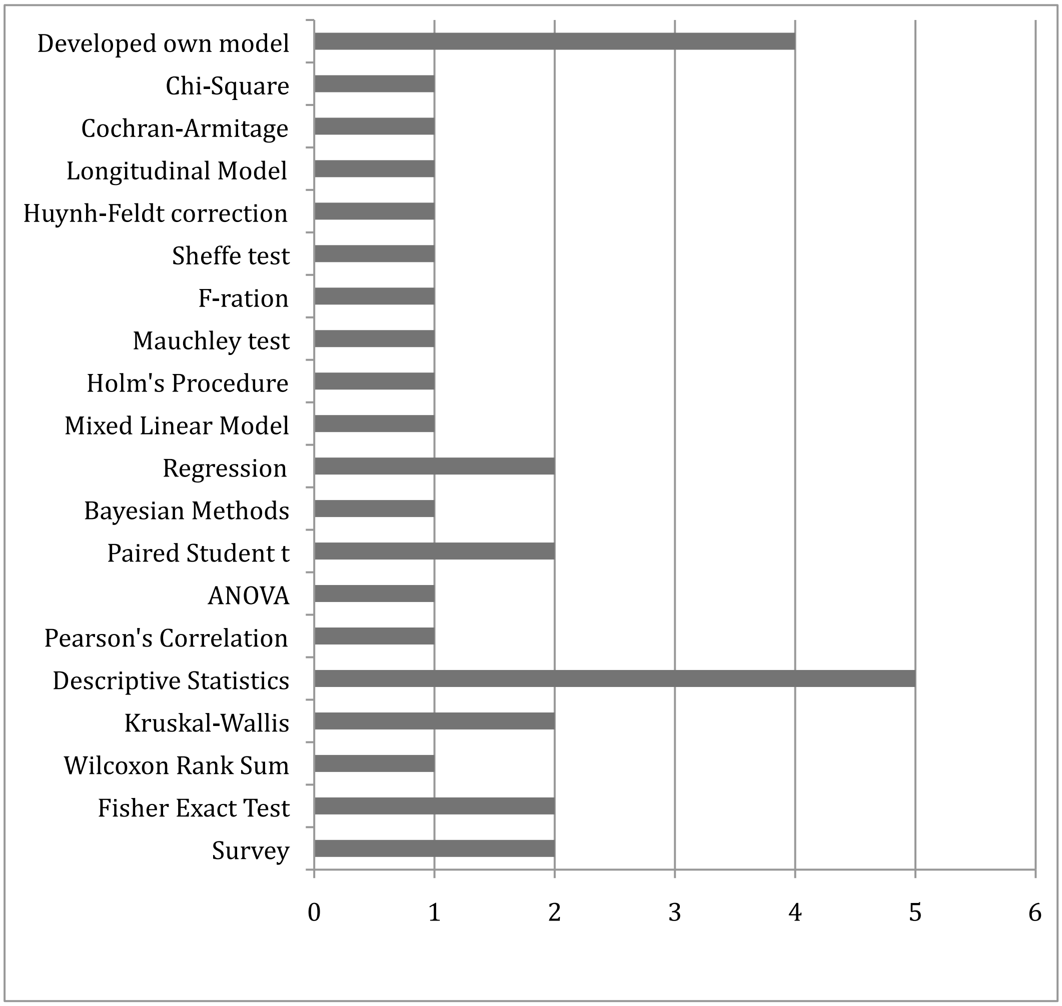

Figure 5: The frequency of statistical methods and inferential procedures is shown above.

4.2 Phase 2 and Phase 3

Phase 2 is not subject to a direct verification or validation procedure. The ontology is changed and updated as results from the verification and validation processes of Phase 1 and Phase 3 become available. Phase 3 cannot be completed until Phase 1 and Phase 2 are completed.

5. Discussion

5.1 Limitation and Future Work

There are several limitations that currently exist in this study and the design to date. The first limitation is due to a limited literature search size. As phase I and II progress, the literature search will extend as the verification and validation process proceeds. There is only one person who is engineering the knowledge involved in the creation of the ontology. After the ontology is validated during phases I and II, it will be open sourced, where other informaticians and knowledge engineers can add and edit the ontology. At this point in the study, there is limited validation and verification available of the ensuing knowledge collection. The results of the verification and validation will be available as the study progresses and will be published at a later date. There are also potential human-computer interaction challenges when implementing SIBRe.

5.2 Implication of SIBRe

This is a very comprehensive project that also includes a very organic process. As more data become available through each phase of the project, refinements can be made to the knowledge base and incorporated into the software platform. This project began by including basic statistical techniques into the knowledge base. However, when literature was reviewed, it became quickly apparent at how wide the range of techniques used in research has become. There has been a fair level of discussion at this point whether to leave SIBRe with information about commonly used and easily executed techniques or to include more advanced and technical types of analysis into the system. It is agreed that while the goal of SIBRe is to educate researchers about statistics and to instruct them in the use of proper methodologies for research. The blueprint for SIBRe is proceeding to include information about more advanced methodologies, but also including several resources that would allow a researcher to get an understanding of a technique and its appropriateness to a particular problem or type of data, but to also give them the information needed to seek out a professional expert to assist with the execution of the technique. Annotated links to articles that summarize a particular method or describe its use in a certain domain as well as contact information to local statistical experts or consulting groups would be included in the resources that SIBRe would return to a user. It is hoped that the end user will get enough information to gain an understanding of what technique(s) is most appropriate and information on whom to contact to help the user execute his/her data analysis.

The card sorting techniques and the survey in Phase 1 will also help to answer the question of how much a researcher does or does not know about statistics and the application of statistical techniques to data. Do researchers continue to use the same statistical technique over and over because they feel comfortable with it, even if it may not be appropriate? If a researcher sees reads an article that analyzes data using a new or different technique, how likely are they to go and find more information about it? What is the biggest hurdle for a researcher when it comes to statistics? This information will also be extremely valuable in the creation of SIBRe, and should also help to direct us as to whether or not more advanced techniques are necessary and useful in the SIBRe system.

Since the foundation of the decision support system in SIBRe is an ontology, it will not be a linear type of decision tree. A user can describe his/her data through several factors such as experimental goals, types of variables, number of variables, etc. so that SIBRe can more accurately address the needs of the researcher. As a user progresses through the decision support system, questions and explanations of terms will remain as non-technical as possible.

6. Conclusion

The process of the creation of SIBRe is still underway. It is hoped that SIBRe will allow for better communication between researchers and statisticians and to encourage more researchers to seek out statistical experts for experimental design and data analysis. Through an integrated platform incorporating a decision support system, annotated literature and reference library, a researcher should be able to find the information he or she will need to strengthen the integrity of his/her area of expertise.

Acknowledgements

This publication was supported in part by an Institutional Clinical and Translational Science Award, NIH/NCRR Grant Number UL1-RR025755. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. Payne, P.R., et al., *Conceptual knowledge acquisition in biomedicine: A methodological review.* J Biomed Inform, 2007. **40**(5): p. 582-602.
2. Payne, P., S. Johnson, J. Starren, H. Tilson, and D. Dowdy, *Breaking the Translational Barriers: The Value of Integrating Biomedical Informatics and Translational Research.* J Investig Med, 2005. **53**(4): p. 192-200.
3. Cech, T.R., *Fostering innovation and discovery in biomedical research.* JAMA, 2005. **294**(11): p. 1390-3.
4. Hellems, M.A., M.J. Gurka, and G.F. Hayden, *Statistical literacy for readers of Pediatrics: a moving target.* Pediatrics, 2007. **119**(6): p. 1083-8.
5. Edwards, K.S., P.K. Woolf, and T. Hetzler, *Pediatric residents as learners and teachers of evidence-based medicine.* Acad Med, 2002. **77**(7): p. 748.
6. Applegate, K.E. and P.E. Crewson, *Statistical literacy.* Radiology, 2004. **230**(3): p. 613-4.
7. Mossman, K.L., *Nuclear literacy.* Health Phys, 1990. **58**(5): p. 639-43.
8. Friedman, S.B. and S. Phillips, *What's the difference? Pediatric residents and their inaccurate concepts regarding statistics.* Pediatrics, 1981. **68**(5): p. 644-6.
9. Berwick, D.M., H.V. Fineberg, and M.C. Weinstein, *When doctors meet numbers.* Am J Med, 1981. **71**(6): p. 991-8.
10. Rickard, C.M., *Statistics for clinical nursing practice: an introduction.* Aust Crit Care, 2008. **21**(4): p. 216-9.
11. Scales, C.D., Jr., B. Peterson, and P. Dahm, *Interpreting statistics in the urological literature.* J Urol, 2006. **176**(5): p. 1938-45.

12. Ludbrook, J.a.H.D., *Why Permutation tests are superior to t and F tests in biomedical research.* American Statistician, 1998. **52**(2): p. 127-133.

13. Anderson, D.C.a.Y.W., *Tests of signifigance and effect size: meaningful interpretation of statistical data in the health sciences.* Essays in Education, 2008. **23**: p. 142-158.

14. Bates, D.W., et al., *Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality.* J Am Med Inform Assoc, 2003. **10**(6): p. 523-30.

15. Kuperman, G.J., et al., *Medication-related clinical decision support in computerized provider order entry systems: a review.* J Am Med Inform Assoc, 2007. **14**(1): p. 29-40.

16. McCormick, R., *Conceptual and Procedural Knowledge.* Int J Technol Design Edu, 1997. **7**: p. 141-159.

17. Hoehndorf, R., et al., *Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies.* BMC Bioinformatics, 2007. **8**: p. 377.

18. Tu, S.W., et al., *Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTEGE-II to protocol-based decision support.* Artif Intell Med, 1995. **7**(3): p. 257-89.

19. Sadeghi, S., A. Barzi, and J.W. Smith, *Ontology Driven Construction of a Knowledgebase for Bayesian Decision Models Based on UMLS.* Stud Health Technol Inform, 2005. **116**: p. 223-8.

20. Kim, D.K., et al., *MYCIN II: design and implementation of a therapy reference with complex content-based indexing.* Proc AMIA Symp, 1998: p. 175-9.

21. McGeorge, G.R.a.P., *The Sorting techniques: a tutorial paper on card sorts and item sorts.* Expert Systems, 1997. **14**(2): p. 13.

22. Medicine, S.C.f.B.I.R.a.S.U.S.o., *Protégé OWL*, in *Protégé*. 2009.

23. in *MORAE*, TechSmith.