Published in final edited form as:
Med One. 2020 January 10; 5: . doi:10.20900/mo20200001.

Artificial Intelligence Pipeline to Bridge the Gap between Bench Researchers and Clinical Researchers in Precision Medicine

Lewis J. Frey[1,2,*], **Douglas A. Talbert**[3,*]

[1]Department of Public Health Science, Biomedical Informatics Center, Hollings Cancer Center, Medical University of South Carolina (MUSC), 135 Cannon St, Charleston, SC 29425, USA

[2]Health Equity and Rural Outreach Innovation Center (HEROIC), Ralph H. Johnson Veteran Affairs Medical Center, Charleston, SC 29401, USA

[3]Department of Computer Science, Tennessee Tech University (TTU), 1 William L Jones Dr, Cookeville, TN 38505, USA

Abstract

Precision medicine informatics is a field of research that incorporates learning systems that generate new knowledge to improve individualized treatments using integrated data sets and models. Given the ever-increasing volumes of data that are relevant to patient care, artificial intelligence (AI) pipelines need to be a central component of such research to speed discovery. Applying AI methodology to complex multidisciplinary information retrieval can support efforts to discover bridging concepts within collaborating communities. This dovetails with precision medicine research, given the information rich multi-omic data that are used in precision medicine analysis pipelines. In this perspective article we define a prototype AI pipeline to facilitate discovering research connections between bioinformatics and clinical researchers. We propose building knowledge representations that are iteratively improved through AI and human-informed learning feedback loops supported through crowdsourcing. To illustrate this, we will explore the specific use case of nonalcoholic fatty liver disease, a growing health care problem. We will examine AI pipeline construction and utilization in relation to bench-to-bedside bridging concepts with interconnecting knowledge representations applicable to bioinformatics researchers and clinicians.

Keywords

translational science; precision medicine; bioinformatics; artificial intelligence; knowledge graph; liver disease; collaboration

[*]Correspondence: Lewis J. Frey, Frey@musc.edu; Douglas A. Talbert, dtalbert@tntech.edu.
AUTHOR CONTRIBUTIONS
Both authors contributed equally to the manuscript and wrote the paper together.

CONFLICTS OF INTEREST
The authors have declared no conflicts of interest.

INTRODUCTION

The following quote by Herbet A. Simon, an artificial intelligence (AI) visionary, clearly articulates the extant dilemma that biomedical researchers face with a plethora of data measured at multiple scales.

> “In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence, a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.” [1]

This new information-rich reality motivates the use of systems that reduce attentional overload instead of contribute to it. An example of increasing attentional overload is alert fatigue where clinicians stop devoting attention to warning messages because of the abundance of unnecessary alerts [2]. Translational precision medicine would benefit from AI systems that glean relevant and useful knowledge from multiple sources in both automated and semi-automated ways through autonomous integration or crowdsourced augmentation to address our attentional limitations. Some areas of precision medicine research, such as translational research with large and diverse research repositories, have an even greater need for AI supported research and collaboration. Using a community driven collaborative information retrieval framework, we will discuss discovery processes relevant to precision medicine that can generalize beyond precision medicine, but is particularly applicable given the exponential growth of multi-omic data that can be leveraged in constructing knowledge representations [3].

Translational research focuses on moving discoveries made at the bench to clinical care and from clinical care to the bench [4]. However, in practice, translational research is very hard to achieve [5]. This is in part because translational research is multidisciplinary [6]. Here is where an AI system can serve as a bridge between concept models from different disciplines [7]. The specifics of an AI pipeline approach include the integration of knowledge sources, knowledge bases, knowledge linkages, and knowledge extraction that can be crowdsourced by humans and/or AI learners to improve the performance and predictive capabilities of the learned perspectives and knowledge [8]. The sources of knowledge reside in domain experts, research journals, guidelines, and data repositories, and the AI learner could be used to assist in the search for and organization of information to lower the attentional load on researchers. This could build on extant work that extracts meaningful information from clinical practice guidelines using question answering systems such as Watson [9]. Our proposed system is not intended to be a question answer system such as IBM Watson or solely an information retrieval system as described by Sparck [10]. We build on concepts from never ending learning systems [11,12] by combining AI learning with crowdsourcing to identifying bridging concepts among researchers to facilitate collaborations. It is not a general AI approach, but is instead focused on building knowledge representations specific to collaborative research teams.

The goal of the proposed system is to provide a multidisciplinary forum that grows with the knowledge of experts and AI learning components, who together collaboratively build knowledge graphs in specific research areas. A knowledge graph is a representation of concepts as nodes that are connected by edges representing relationships between concepts. Clinicians and bench researchers implicitly create knowledge graphs when they construct best practice clinical guidelines or mechanistic models of interacting components in cells or model systems. Knowledge graphs can be used to discover bridging concepts among collaborators in the project. Because there are data rich repositories of high-throughput experiments along with the papers that document them, the task of biomarker discovery in precision medicine is well matched to the challenge of finding bridging concepts using knowledge graph representations. Given that ontological representations converge more quickly for information rich knowledge spaces [13], the density of high-throughput data in the precision medicine space, especially in cancer, lends itself to the task of conceptual ontology space construction. The collaborative information retrieval task combined with converting high quality data into well supported knowledge graphs will be enhanced through the combined efforts of experts in specific domain areas and AI algorithms scaling with the size of the growing data resources through the use of crowdsourcing.

The use of crowdsourcing technology coupled with human and AI expertise supports scalable solutions, even when the AI does not have the complexity or subtlety to comprehend aspects of the problem space. The strength of crowdsourcing enables concepts of word and acronym ambiguity to be resolved through expertise that exists in the community. The idea is that a community would use the pipeline to grow a knowledge graph that is highly relevant to their area of expertise and interest. The approach is not geared to build a single knowledge graph with general knowledge, but instead many specific knowledge graphs that have active community bases that support a living repository of information.

Because there is a risk that the AI could overwhelm a knowledge graph with irrelevant and low quality data and papers, crowdsourcing with experts that rank and prioritize the encoding of information provides a check and balance to the information retrieved by the AI. Low quality data and information would result in a downgrade of the knowledge graph by experts in the field, and a reinforcement component built into the AI learner could support the AI learning from the domain experts’ feedback. The net result would be improved AI support for the collaboration, and the human interaction input improves the ability of the AI to integrate relevant information.

In this paper we describe a roadmap for researchers to build such a system to reduce the attentional strain in the application area of precision medicine translational science research. First, we describe how translational scientists can represent their research questions in a computable knowledge representation. Second, we outline a prototype AI pipeline that addresses attentional overload through computational analysis of knowledge representations. Third, we describe how the three target communities (i.e., bioinformaticians, clinicians, and AI researchers) can be engaged and contribute to its success. As an illustration, we apply these concepts to the domain of nonalcoholic fatty liver disease (NAFLD) from both a bench as well as a clinical researcher perspective [14,15]. NAFLD is related to metabolic syndrome with a constellation of comorbidities including obesity, type 2 diabetes mellitus, hypertension, and dyslipidemia [16–19]. If not properly managed NAFLD can progress to liver fibrosis and cirrhosis with outcomes including hepatocellular carcinoma [20]. For bench research to attract the attention of clinicians and clinical researchers their questions need to align with issues that are associated with improving the care of patients.

Bench Researcher Perspective

Our NAFLD use case is a researcher studying the proteomic signature of extracellular matrix (ECM) interaction in fibrotic tissues. Changes in ECM can be observed in liver tissue disease progression that occurs in NAFLD and in epithelial-mesenchymal transition occurring in an aggressive cancer [21,22]. Bench researchers analyze measurements collected from biopsied tissue and, in the case of proteomics, could use mass spectrometry matrix-assisted laser desorption/ionization with liquid chromatography (LC-MALDI) to obtain collagen and peptide signatures from formalin-fixed paraffin-embedded (FFPE) tissue [23–25]. The difficulty faced by translational bench researchers is knowing how to frame their preclinical questions so they are clinically relevant. This often comes down to selecting phenotypes to examine along with which experimental conditions to investigate by drawing from their own and their collaborator's experiences, all while prioritizing time and resources. Given exponentially expanding data repositories, however, the analysis and investigation can be done without awareness of the full set of highly relevant resources. Hence, there are gaps in utilizing the vast amount of informational resources that continue to accrue daily.

Clinical Researcher Perspective

Continuing our NAFLD example, a clinical researcher is interested in factors associated with patient outcomes that include fibrotic tissue growth in the liver to the point of cirrhosis and potentially liver cancer. The guideline for diabetes care now recommends NAFLD assessment for diabetic patients with elevated liver enzymes such as alanine aminotransferase (ALT) [26]. The data sources associated with fibrosis progression are liver radiology imaging reports, liver biopsy reports, non-invasive liver fibrosis estimates (e.g., fibrosis 4 (FIB4) scores), race, ethnicity, and body mass index (BMI). The problem for the clinical researcher is how to improve their predictive risk models through non-invasive measures. A strong incentive to move forward with a translational collaboration is co-developing experiments and measures that can be validated to improve the prediction of fibrosis progression in NAFLD. Below we explore this scenario using an AI pipeline approach to facilitate the identification of such connections using a formalized knowledge graph approach.

OVERVIEW OF AI PIPELINE

An AI pipeline can bridge collaborations among bench and clinical researchers by identifing concepts that connect the collaborators and seed the common interests among them. To achieve the adaptive knowledge management scenario described in this paper, we envision AI support for engaged research communities that create, use, and share knowledge to collaborate and extend knowledge in precision medicine. This support would combine crowdsourcing with credential and trustworthiness measures with AI to iteratively refine (1) collaborative information retrieval [10,13], (2) knowledge extraction and organization, and (3) concept connection identification. In this way, the AI pipeline combines the interpretive strength of experts and practitioners in the specific areas of interest while enhancing their models with data-informed AI medical models.

While the use of such a tool might not be limited to precision medicine, we believe precision medicine exhibits a number of properties that make it an ideal area in which to apply AI to help initiate collaborative research: (1) need for translational research, (2) challenge of bridging different ways of thinking (patient-centric vs. bench-centric), (3) particularly high volume of literature, (4) presence of well-developed and organized data repositories, and (5) availability of well-recognized and useful vocabularies and ontologies that can support concept identification and extraction.

INTELLIGENT PRECISION MEDICINE PIPELINE

Our proposed framework, which we call the *Intelligent Precision Medicine Pipeline* (IPMP), enables human researchers to collaborate with AI to help organize research questions at different scales of the problem. In our example, the scale of the research question for the clinical research is at the level of the patient, where the focus is identifying actionable predictive models to improve care decisions. For the bench researcher, the scale of the problem is at the cellular level examining fibrosis in the ECM.

One of the challenging tasks that IPMP needs to perform is to identify bridging concepts across the different scales of the research problem. Figure 1 provides a high level diagram of the components involved in a translational human/AI collaborative pipeline. The processes for each researcher are represented in the boxes on the right and left side of the diagram. After developing their research questions, the researchers generate conceptual keywords and identify other sources of knowledge (e.g., data repositories, papers, guidelines) for extraction. There are a number of tools (e.g., Protege, owlready)[27,28] that support manual or programmatic construction and refinement of machine readable knowledge graphs that can be used in the initial knowledge extraction from the collaborating researchers [29]. Given an initial set of small human-generated knowledge graphs, IPMP can utilize the information as input to start the *human/AI collaboration*.

Once the collaboration begins, the AI can assist in knowledge graph refinement to more precisely capture the relevant research interests. A knowledge graph can be pruned or grown as the collaboration provides feedback about its correctness. As the knowledge graph guides the retrieval of information, the researchers can provide feedback on the relevance of selected retrieved items. When retrieved items are marked as irrelevant, the aspect of the knowledge graph responsible for its retrieval is marked for possible refinement. Then the relevant and irrelevant items associated with each knowledge graph element can be analyzed for more precise representations that include the relevant items and exclude the irrelevant ones. Thus, the knowledge graph's precision is improved over time by learning from examples, and IPMP's ability to correctly retrieve relevant literature continually improves.

Depending on the desired level of intelligence, IPMP could employ active learning strategies to target knowledge graph refinement [30].

The pipeline is then tasked with identifying the connecting concepts between these refined perspectives and deriving a common collaborative space of extracted concepts/knowledge. To do this, knowledge is extracted from the literature and converted to a knowledge representation (knowledge graph) with edges linking the concepts to each other. One challenge during knowledge extraction will be the recognition of synonyms and abbreviations. A variety of techniques can be used to address this. Where possible, IPMP can leverage existing resources that map terms to concepts (e.g., the Unified Medical Language System (UMLS) metathesaurus) [31]. Where that is not possible, IPMP can learn to map terms to each other either by being shown that a mapping exists through crowdsourcing or by using machine learning to discover such mappings [32]. For IPMP, this practice needs to be made explicit and represented in a machine readable format. A graph-based approach has shown some success in linking concepts across papers and finding complementary (i.e., cross-specialty) literature [33–35].

For our NAFLD example, the researchers can initiate this by listing keywords and subsets of publications relevant to their joint NAFLD interest. For example, the clinician would have keywords such as *NAFLD*, *fibrosis*, and *BMI* while the bench researcher would have keywords such as *LC-MALDI*, *proteomics*, *collagen*, and *ECM*. Figure 2 illustrates the creation of a knowledge representation, created using Protege/owlready [27,28], of the keywords for bench (Left) and clinical (Right) researchers.

By using the python package owlready2 we constructed the word graphs in Figure 2 for words associated with bench researchers examining ECM remodeling and clinical researchers examining risk models of fibrosis progression in NAFLD. The list of keywords for each research area establishes a position in the conceptual space. For the purposes of explanation, we have made the space tightly related. The separate knowledge representations can be converted into a joint knowledge representation by running them simultaneously in owlready2 and visualizing the result with Protege (see Figure 3).

The two graphs are connected through the concept of *Cancer* in the words used by each group. There are places in the example where the keywords could be different but related (e.g., *bio_marker* at the bench and *risk_model* at the bedside) and hence, may be important to link. In Figure 3, related concepts have been positioned close to each other. Proximity could be used by the researchers to communicate similarity to the AI while being visually informative to the human researchers. Another example is *Fibrosis* and *Fibroblast* since fibroblast cells generate connective tissue associated with fibrosis. The AI can map into related concepts to find a link between them with human assessed accuracy being a measure of the quality of the AI reasoning [36]. Existing terminologies and coding standards such as those in the UMLS [31] could be used to connect concepts through the AI graph search or reasoning algorithms. Bioinformatics ontologies such as the Gene Ontology could also be integrated into the reasoning process to connect and bridge concepts at the molecular and cellular scale [37]. Alternatively the researchers could construct it manually and feed it as input to the AI. For example, they might map predictive *bio_marker* and predictive *risk_model* to a bridge concept like *predictive marker*. Crowdsource participants could also provide links between concepts [8]. Approaches using machine learning and crowdsourcing on real time analysis of tweets during disaster recovery provides examples of the feasibility of supporting crowdsourced collaborative information retrieval using the python package pybossa [38]. The use of pybossa for this realtime task provides a template for supporting multidisciplinary teams performing collaborative information retrieval focused on specific research areas.

In creating the small knowledge graphs integrated in Figure 3, a paper was identified that could be used to pursue future research entitled, “Identifying Nonalcoholic Fatty Liver Disease Patients with Active Fibrosis by Measuring Extracellular Matrix Remodeling Rates in Tissue and Blood” [21]. The ability to use blood measurement on EMC in liver disease creates a potential translational bridge concept for less invasive measures of liver fibrosis progression in NAFLD, a current active topic of research. Having IPMP further investigate other potentially related papers would provide value to the ongoing research. IPMP could also be used to mark areas of research to follow or flag when related discoveries are identified.

IPMP’S COLLABORATIVE SEARCH

Once the process is initiated by entering the initial keyword knowledge graph(s), the volume of information that IPMP needs to process and search requires the power and speed of AI, but the nature and complexity of the task might also necessitate human input, review, and interpretation. Thus, the success of IPMP depends on effective human/AI collaboration, as envisioned in the following steps for an IPMP search:

1. The researchers’ specify their questions and/or areas of interest in the form of connected keywords (i.e., simple knowledge graphs) along with other (optional) relevant material (e.g., research papers, guidelines, or datasets). The AI could provide assistance in this step by capturing this information through an interactive dialog that guides the researchers through the specification of their interest.
2. IPMP extracts knowledge from that input and uses it as a seed to identify other sources of knowledge that it uses to grow its knowledge base.
3. IPMP uses feedback regarding selected retrieved items to learn improved understandings of the researchers’ areas of interest and to improve the precision of its knowledge base.
4. As it grows its knowledge base, IPMP is continually searching for bridges that connect the questions/areas of interest. Any identified bridges are ranked using estimates of confidence and interest (e.g., relevance scores or path length).
5. If crowdsourcing is enabled, IPMP is sharing its discovered knowledge with interested members of appropriate communities. Through a process that incorporates credentialing and trustworthy measures, discoveries can be confirmed or refuted by human experts and directions of potential interest can be

suggested to IPMP through manually added links and through the submission of additional knowledge graphs, papers, or other material.

6. The current status of the search can be observed by any of the contributing researchers, and likewise the researchers can interact with IPMP at anytime during the search. In addition to any of the actions described above with crowdsourcing, the initiating researchers can review and redirect IPMP based on their interests and priorities. For example, they could indicate that one of the connections identified as promising by IPMP is not of interest to them. IPMP would then quit searching for additional information surrounding that connection.

7. IPMP would also seek to learn through the process. As it extracts knowledge and interacts with researchers, it builds a better understanding of areas of interest and of the researchers' intentions and can, in turn, become better at identifying and extracting relevant knowledge and at finding interesting bridges the connect key concepts.

A summary of the steps is visualized in Figure 4.

Thus, collaboration is integrated throughout IPMP, which is intended to provide both human-support of AI (e.g., crowdsourced or researcher-provided feedback) and AI-support of humans (e.g., automated exploration of area of interest for relevant papers, datasets, and guidelines).

IPMP's AI would work as autonomously as it is able. It is human experts, however, who drive the information flow in the pipeline, and thus the inputs to and outputs from the AI are visible and can be reviewed, manipulated, and critiqued by human experts. This feedback is an important component in the AI's learning process and is similar to the human feedback component in Mitchell's NELL project [11]. Additionally, IPMP will learn from the crowdsourced content that humans input, and just as they can help out the AI, human interactions with the system (e.g., entering keywords and constructing knowledge graphs) are supported and mediated by the AI.

The strength of IPMP will be to organize a large volume of information relevant to the researchers' questions and make the connection understandable to them. The extent to which IPMP succeeds in providing relevant information will increase the chances of facilitating a collaboration among the researchers involved. Achieving this ambitious goal will require not only collaboration between biomedical researchers and IPMP, but also collaboration between biomedical researchers and AI researchers.

From the above description, it is clear that clinical and bench researchers are integral to the success of IPMP's search. They need to be able to clearly specify their questions and must be able to logically critique IPMP's outputs and provide timely, meaningful, and insightful feedback to shape the system's understanding of their needs and to contribute to IPMP's continual learning process. In contrast, AI researchers take no active role in the actual search. Their role is the design and implementation, in concert with bench and clinical researchers, of a collaborative AI that can (A) capture and extract information through dialog with human experts; (B) extract and organize knowledge from natural language narratives, structured guidelines, datasets, and terminologies; (C) understand, search, and connect concepts across the different scales in precision medicine (i.e., genes and proteins to organ systems and patients) in ways that make sense to human experts; and (D) continually learn how to get better at doing (A–C).

Available and Needed Technologies

Many of the identified capabilities and gaps are, at least partially, addressed by existing technologies. Details of these technologies are beyond the scope of this paper. There are many approaches to *extracting knowledge from free text,* including named-entity extraction [39–41], topic modeling [42–44], and automatic text summarization [45–47]. Techniques such as clustering [48,49], frequent pattern identification [50,51], and rule extraction [52–54] have been used to *extract knowledge from data*. IPMP could build *additional knowledge representations* such as ontologies [55–58], knowledge graphs [12,59,60] and word embeddings [61–63]. Once knowledge is extracted and represented, a variety of search and information retrieval techniques can be adapted to help find relevant connections [64–67]. The python module pybossa keeps track of user contributions and can provide statistics on the activities of the authenticated and anonymous users on the project: top contributors, time to completion of tasks and other metrics. It potentially could be used in conjunction with a version control system to manage provenance of knowledge graph contributions [38]. Highlighting the existence of these technologies is not meant to trivialize the development of the IPMP. On the contrary, we anticipate that significant work will be needed to realize this AI pipeline. The work highlighted here does, however, suggest that AI researchers have a strong foundation on which to build when developing this AI pipeline.

CONCLUSION

A key component of precision medicine informatics is knowledge generation using learning systems applied to larger data sets [68]. Repositories of biological information have been growing at exponential rates since the rate of cost reduction for genomic data is faster than Moore’s Law [7]. To manage the accumulation of data that has discoverable patterns, we propose IPMP, a human and AI collaboration, that creates hybrid knowledge structures that benefit from combined knowledge generation intrinsic to human and AI learners. Collaborations between bioinformatics and clinical researchers are complex and difficult to achieve given the differences in the nature and use of information in the two communities. In bioinformatics research, large sets of highly complex multi-omic data are often used to better understand the mechanistic behavior of model systems. Clinicians tend to use highly defined measures (e.g., images, labs) to answer specific questions pertaining to the health and survival of patients. The different emphases and risks involved in both fields result in different training regimes and different perspectives when collecting and analyzing information that affect the systems of interest.

Given the gap that exists between the two fields there is a need for approaches that identify knowledge that is relevant to both communities to increase the productivity of collaborations. Such an approach will need to be able to start from the perspective of either community and build bridges that link the disparate perspectives into a unified whole that motivates each researcher to invest resources in building a collective understanding that advances both fields. To achieve this desirable outcome there is a role for AI researchers to adapt or invent approaches and methodologies that spur on the enterprise of discovery at a faster rate with more successful outcomes for all communities involved. It is our goal in the paper to advocate for a crowdsourced multidisciplinary collaborative information retrieval framework based on AI that would enable a true collaboration involving clinical, bench, and AI researchers all working together to improve the efficacy and precision of medical care.

Acknowledgments

FUNDING SOURCES

The work of LJF was supported in part by VA HSR&D NAFLD grant HX002700–01A1, National Institutes of Health (NIH) grant U54-MD010706 and Health Equity and Rural Outreach Innovation Center grant CIN 13–418.

ABBREVIATIONS

NAFLD	nonalcoholic fatty liver disease
ECM	extracellular matrix
LC-MALDI	mass spectrometry matrix-assisted laser desorption/ionization with liquid chromatography
FFPE	formalin-fixed paraffin-embedded
ALT	alanine aminotransferase
FIB4	fibrosis 4 scores
BMI	body mass index
IPMP	Intelligent Precision Medicine Pipeline
UMLS	Unified Medical Language System

REFERENCES

1. Simon HA. Designing organizations for an information-rich world. Lecture at Brookings Institute; 1969 9 1; Washington, DC, USA Available from: http://zeus.zeit.de/2007/39/simon.pdf. Accessed 2019 Oct 15.
2. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce “alert fatigue” while still minimizing the risk of litigation. Health Aff. 2011;30(12):2310–7.
3. Frey LJ, Piccolo SR, Edgerton ME. Multiplicity: an organizing principle for cancers and somatic mutations. BMC Med Genomics. 2011 6 29;4(1):52. [PubMed: 21714919]
4. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L. Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc. 2011 7;18(4):354–7. [PubMed: 21561873]
5. Sabroe I, Dockrell DH, Vogel SN, Renshaw SA, Whyte MKB, Dower SK. Opinion: Identifying and hurdling obstacles to translational research. Nat Rev Immunol. 2007;7(1):77. [PubMed: 17186032]

6. Collins FS. Reengineering Translational Science: The Time Is Right. Sci Transl Med. 2011 7 6;3(90):90cm17.

7. Frey LJ. Artificial Intelligence and Integrated Genotype–Phenotype Identification. Genes. 2018 12 28;10(1):18.

8. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017 1 31;49:170. [PubMed: 28138153]

9. Salvi E, Parimbelli E, Basadonne A, Viani N, Cavallini A, Micieli G, et al. Exploring IBM Watson to Extract Meaningful Information from the List of References of a Clinical Practice Guideline In: ten Teije A, Popow C, Holmes JH, Sacchi L, editors. Artificial Intelligence in Medicine. Cham (Switzerland): Springer International Publishing; 2017 p. 193–7.

10. Sparck JK. Information retrieval and artificial intelligence. Artif Intell. 1999 10 1;114(1):257–81.

11. Mitchell T, Cohen W, Hruschka E, Talukdar P, Yang B, Betteridge J, et al. Never-ending Learning. Commun ACM. 2018 4;61(5):103–15.

12. Nickel M, Murphy K, Tresp V, Gabrilovich E. A Review of Relational Machine Learning for Knowledge Graphs. Proc IEEE. 2016;104(1):11–33. doi: 10.1109/jproc.2015.2483592

13. Jung JJ. Ontological framework based on contextual mediation for collaborative information retrieval. Inf Retr Boston. 2006 9 9;10(1):85.

14. Day CP. Non-alcoholic fatty liver disease: a massive problem. Clin Med. 2011 4;11(2):176–8.

15. Roberts EA. Pediatric nonalcoholic fatty liver disease (NAFLD): a “growing” problem? J Hepatol. 2007 6;46(6):1133–42. [PubMed: 17445934]

16. Singh S, Allen AM, Wang Z, Prokop LJ, Murad MH, Loomba R. Fibrosis Progression in Nonalcoholic Fatty Liver vs Nonalcoholic Steatohepatitis: A Systematic Review and Meta-analysis of Paired-Biopsy Studies. Clin Gastroenterol Hepatol. 2015 4 1;13(4):643–54.e9. [PubMed: 24768810]

17. Patel YA, Gifford EJ, Glass LM, McNeil R, Turner MJ, Han B, et al. Risk factors for biopsy-proven advanced non-alcoholic fatty liver disease in the Veterans Health Administration. Aliment Pharmacol Ther. 2018;47(2):268–78. [PubMed: 29115682]

18. Williams CD, Stengel J, Asike MI, Torres DM, Shaw J, Contreras M, et al. Prevalence of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis Among a Largely Middle-Aged Population Utilizing Ultrasound and Liver Biopsy: A Prospective Study. Gastroenterology. 2011 1 1;140(1):124–31. [PubMed: 20858492]

19. Kanwal F, Kramer JR, Duan Z, Yu X, White D, El-Serag HB. Trends in the Burden of Nonalcoholic Fatty Liver Disease in a United States Cohort of Veterans. Clin Gastroenterol Hepatol. 2016 2;14(2):301–8.e1–2. [PubMed: 26291667]

20. Beste LA, Leipertz SL, Green PK, Dominitz JA, Ross D, Ioannou GN. Trends in Burden of Cirrhosis and Hepatocellular Carcinoma by Underlying Liver Disease in US Veterans, 2001–2013. Gastroenterology. 2015 11 1;149(6):147–182.e5.

21. Decaris ML, Li KW, Emson CL, Gatmaitan M, Liu S, Wang Y, et al. Identifying nonalcoholic fatty liver disease patients with active fibrosis by measuring extracellular matrix remodeling rates in tissue and blood. Hepatology. 2017 1 15;65(1):78–88. [PubMed: 27706836]

22. Théret N Increased extracellular matrix remodeling is associated with tumor progression in human hepatocellular carcinomas. Hepatology. 2001 7;34(1):82–8. [PubMed: 11431737]

23. Neubert H, Bonnert TP, Rumpel K, Hunt BT, Henle ES, James IT. Label-free detection of differential protein expression by LC/MALDI mass spectrometry. J Proteome Res. 2008 6;7(6):2270–9. [PubMed: 18412385]

24. Angel PM, Comte-Walters S, Ball LE, Talbot K, Mehta A, Brockbank KGM, et al. Mapping Extracellular Matrix Proteins in Formalin-Fixed, Paraffin-Embedded Tissues by MALDI Imaging Mass Spectrometry. J Proteome Res. 2018 1 5;17(1):635–46. [PubMed: 29161047]

25. Angel PM, Baldwin HS, Gottlieb Sen D, Su YR, Mayer JE, Bichell D, et al. Advances in MALDI imaging mass spectrometry of proteins in cardiac tissue, including the heart valve. Biochim Biophys Acta Proteins Proteom. 2017 7 1;1865(7):927–35. [PubMed: 28341601]

26. American Diabetes Association. 4. Comprehensive Medical Evaluation and Assessment of Comorbidities: Standards of Medical Care in Diabetes—2019. Diabetes Care. 2019 1;42(Suppl 1):S34–45. doi: 10.2337/dc19-s004 [PubMed: 30559230]

27. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, et al. The evolution of Protégé: an environment for knowledge-based systems development. Int J Hum Comput Stud. 2003 1 1;58(1):89–123.

28. Lamy J-B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artif Intell Med. 2017 7 1;80:11–28. [PubMed: 28818520]

29. Paulheim H Knowledge graph refinement: A survey of approaches and evaluation methods. Semant Web. 2017;8(3):489–508.

30. Balcan M-F, Beygelzimer A, Langford J. Agnostic active learning. J Comput System Sci. 2009 1 1;75(1):78–89.

31. Bodenreider O The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 1 1;32(Suppl 1):D267–70. [PubMed: 14681409]

32. Liu Y, Ge T, Mathews KS, Ji H, McGuinness DL. Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion. arXiv:1804.04225 [Preprint]. 2018 4 11 Available from: http://arxiv.org/abs/1804.04225. Accessed 2019 Oct 15.

33. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell. 1997 4 1;91(2):183–203.

34. Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. Detecting gene relations from Medline abstracts. Pac Symp Biocomput. 2001;6:483–95.

35. Pratt W, Yetisgen-Yildiz M. LitLinker: Capturing Connections Across the Biomedical Literature. In: K-CAP ‘03: Proceedings of the 2nd International Conference on Knowledge Capture; 2003 Oct 23–25; Sanibel Island, FL, USA. New York (NY, US): Association for Computing Machinery; 2003 p. 105–12.

36. Pedro SDS, Appel AP, Hruschka ER. Autonomously Reviewing and Validating the Knowledge Base of a Never-ending Learning System. In: WWW ‘13 Companion: Proceedings of the 22nd International Conference on World Wide Web; 2013 May 13–17; Rio de Janeiro, Brazil. New York (NY, US): Association for Computing Machinery; 2013 p. 1195–204.

37. Dessimoz C, Škunca N, editors. The Gene Ontology Handbook New York (NY, US): Springer; 2017 (Methods in Molecular Biology Vol 1446).

38. Imran M, Castillo C, Lucas J, Meier P, Vieweg S. AIDR: Artificial Intelligence for Disaster Response. In: WWW ‘14 Companion: Proceedings of the 23rd International Conference on World Wide Web; 2014 Apr 7–11; Seoul, Korea. New York (NY, US): Association for Computing Machinery; 2014 p. 159–62.

39. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics. 2005 5 24;6(Suppl 1):S3.

40. Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics. 2008 4 11;9(Suppl 3):S3.

41. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform. 2005 12;6(4):357–69. [PubMed: 16420734]

42. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus. 2016 9 20;5(1):1608. [PubMed: 27652181]

43. Song M, Kim SY. Detecting the knowledge structure of bioinformatics by mining full-text collections. Scientometrics. 2013;96:183–201. doi: 10.1007/s11192-012-0900-9

44. Zhao J, Feng Q, Wu P, Warner JL, Denny JC, Wei W-Q. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (LPA). PLoS One. 2019;14(2):e0212112. doi: 10.1371/journal.pone.0212112

45. Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artif Intell Rev. 2017;47:1–66. doi: 10.1007/s10462-016-9475-9

46. Parveen D, Ramsl H-M, Strube M. Topical Coherence for Graph-based Extractive Summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17–21; Lisbon, Portugal. Red Hook (NY, US): Association for Computational Linguistics; 2015. doi: 10.18653/v1/d15-1226

47. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. J Biomed Inform. 2011 8;44(4):688–99. [PubMed: 21440086]

48. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. Nat Methods. 2015 11;12(11):1033–8. [PubMed: 26389570]

49. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018 5;6(5):361–9. [PubMed: 29503172]

50. Naulaerts S, Moens S, Engelen K, Vanden Berghe W, Goethals B, Laukens K, et al. Practical Approaches for Mining Frequent Patterns in Molecular Datasets. Bioinform Biol Insights. 2016 5 2;10:37–47. doi: 10.4137/bbi.s38419 [PubMed: 27168722]

51. Alagukumar S, Lawrance R. A Selective Analysis of Microarray Data Using Association Rule Mining. Procedia Comput Sci. 2015;47:3–12. doi: 10.1016/j.procs.2015.03.177

52. Huang L-T. An integrated method for cancer classification and rule extraction from microarray data. J Biomed Sci. 2009 2 24;16:25. [PubMed: 19272192]

53. Hou W-J, Chen H-Y. Rule extraction in gene-disease relationship discovery. Gene. 2013 4 10;518(1):132–8. [PubMed: 23235120]

54. Chen Z, Li J, Wei L. A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artif Intell Med. 2007 10;41(2):161–75. [PubMed: 17851055]

55. Stevens R Ontology-based knowledge representation for bioinformatics. Brief Bioinform. 2000;1:398–414. doi: 10.1093/bib/1.4.398 [PubMed: 11465057]

56. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics. 2009 2 5;10(Suppl 2):S1.

57. Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. Cold Spring Harb Mol Case Stud. 2015 10;1(1):a000372.

58. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006 Summer;10(2):185–98. [PubMed: 16901225]

59. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. Biomed Res Int. 2017 2 12;2017:2858423.

60. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics. 2015 5 14;16:157. [PubMed: 25971816]

61. Jiang Z, Li L, Huang D, Jin L. Training word embeddings for deep learning in biomedical text mining tasks. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2015 Nov 9–12; Washington, DC, USA. Piscataway (NJ, US): IEEE; 2015. doi: 10.1109/bibm.2015.7359756

62. Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics. 2018 12 1;34(23):4138. [PubMed: 29933431]

63. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform. 2018 11;87:12–20. [PubMed: 30217670]

64. Abdulla AAA, Lin H, Xu B, Banbhrani SK. Improving biomedical information retrieval by linear combinations of different query expansion techniques. BMC Bioinformatics. 2016 7 25;17(Suppl 7):238. [PubMed: 27455377]

65. Munir K, Sheraz Anjum M. The use of ontologies for effective knowledge modelling and information retrieval. Appl Comput Inform. 2018;14:116–26. doi: 10.1016/j.aci.2017.07.003

66. Saeedi A, Peukert E, Rahm E. Using Link Features for Entity Clustering in Knowledge Graphs. In: Gangemi A, Navigli R, Vidal ME, Hitzler P, Troncy R, Hollink L, et al., editors. The Semantic Web: 15th International Conference, ESWC 2018; 2018 Jun 3–7; Heraklion, Greece. Cham

(Switzerland): Springer: 2018 p. 576–92. (Lecture Notes in Computer Science; vol. 10843) doi: 10.1007/978-3-319-93417-4_37

67. Sevon P, Eronen L, Hintsanen P, Kulovesi K, Toivonen H. Link Discovery in Graphs Derived from Biological Databases. In: Leser U, Naumann F, Eckman B, editors. Data Integration in the Life Sciences. DILS 2006 Berlin, Heidelberg (Germany): Springer; 2006 p. 35–49. (Lecture Notes in Computer Science; vol. 4075). doi: 10.1007/11799511_5

68. Frey LJ, Bernstam EV, Denny JC. Precision medicine informatics. J Am Med Inform Assoc. 2016 7;23(4):668–70. [PubMed: 27274018]

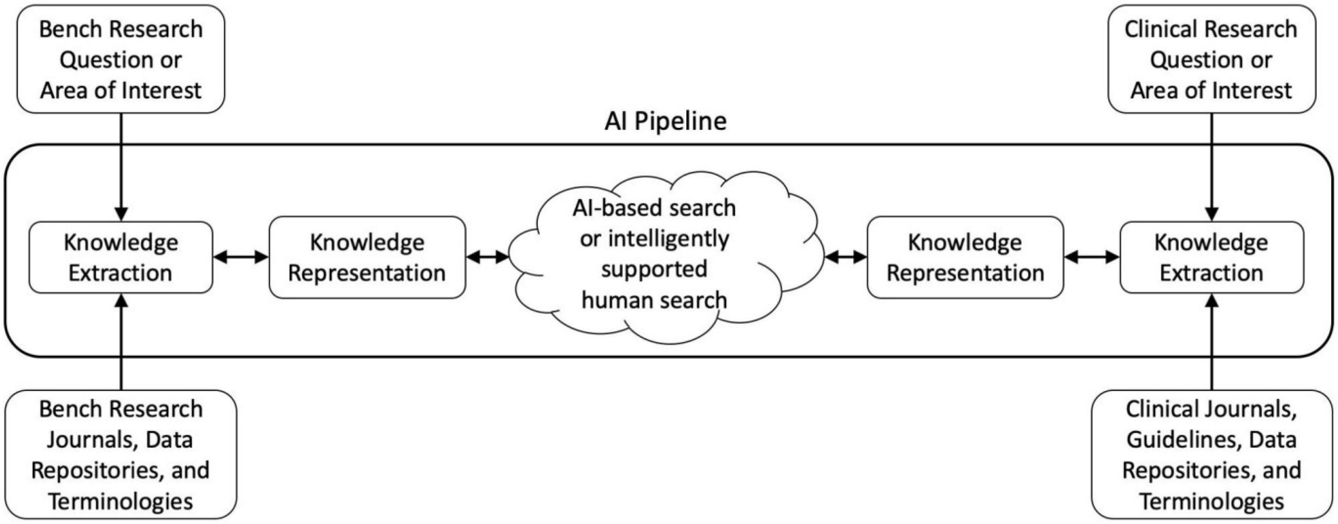
Bench Research Question or Area of Interest
Clinical Research Question or Area of Interest
AI Pipeline
Knowledge Extraction
Knowledge Representation
AI-based search or intelligently supported human search
Knowledge Representation
Knowledge Extraction
Bench Research Journals, Data Repositories, and Terminologies
Clinical Journals, Guidelines, Data Repositories, and Terminologies

Figure 1.

Intelligent Precision Medicine Pipeline overview.

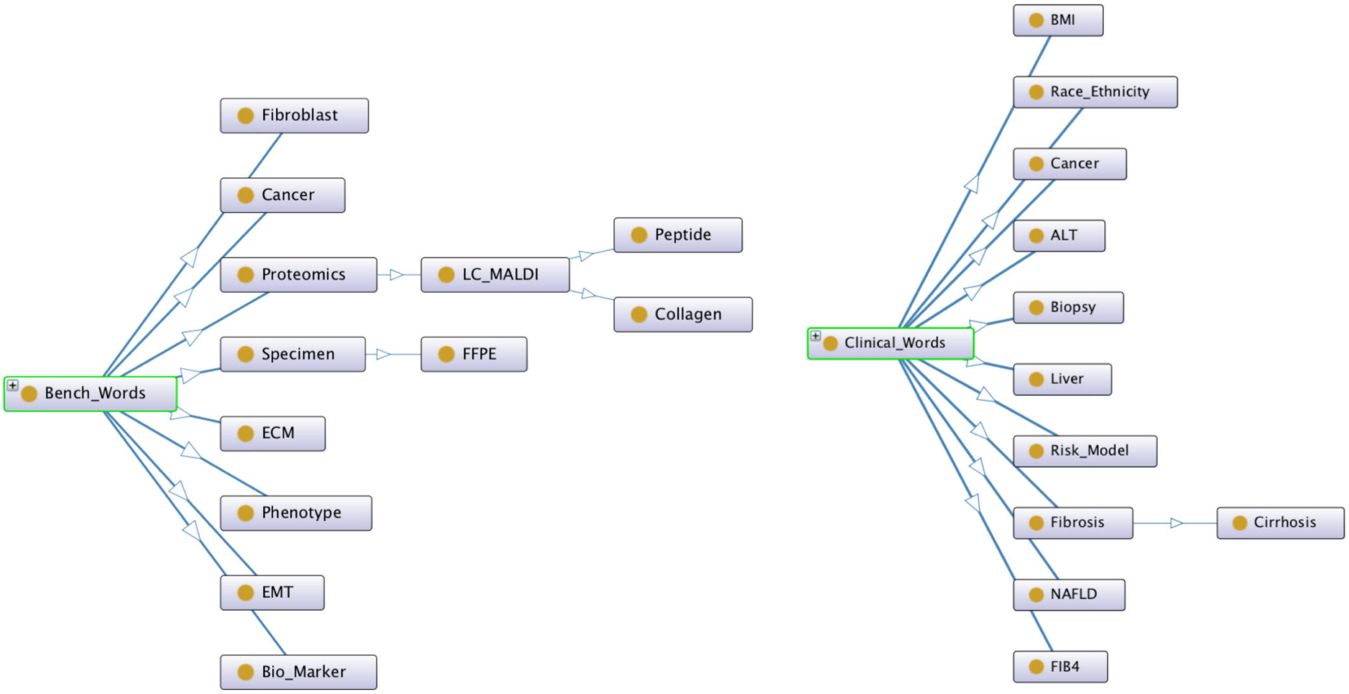

Figure 2.
Example of using NAFLD research keywords provided by bench (Left) and clinical (right) researchers to generate knowledge representations.

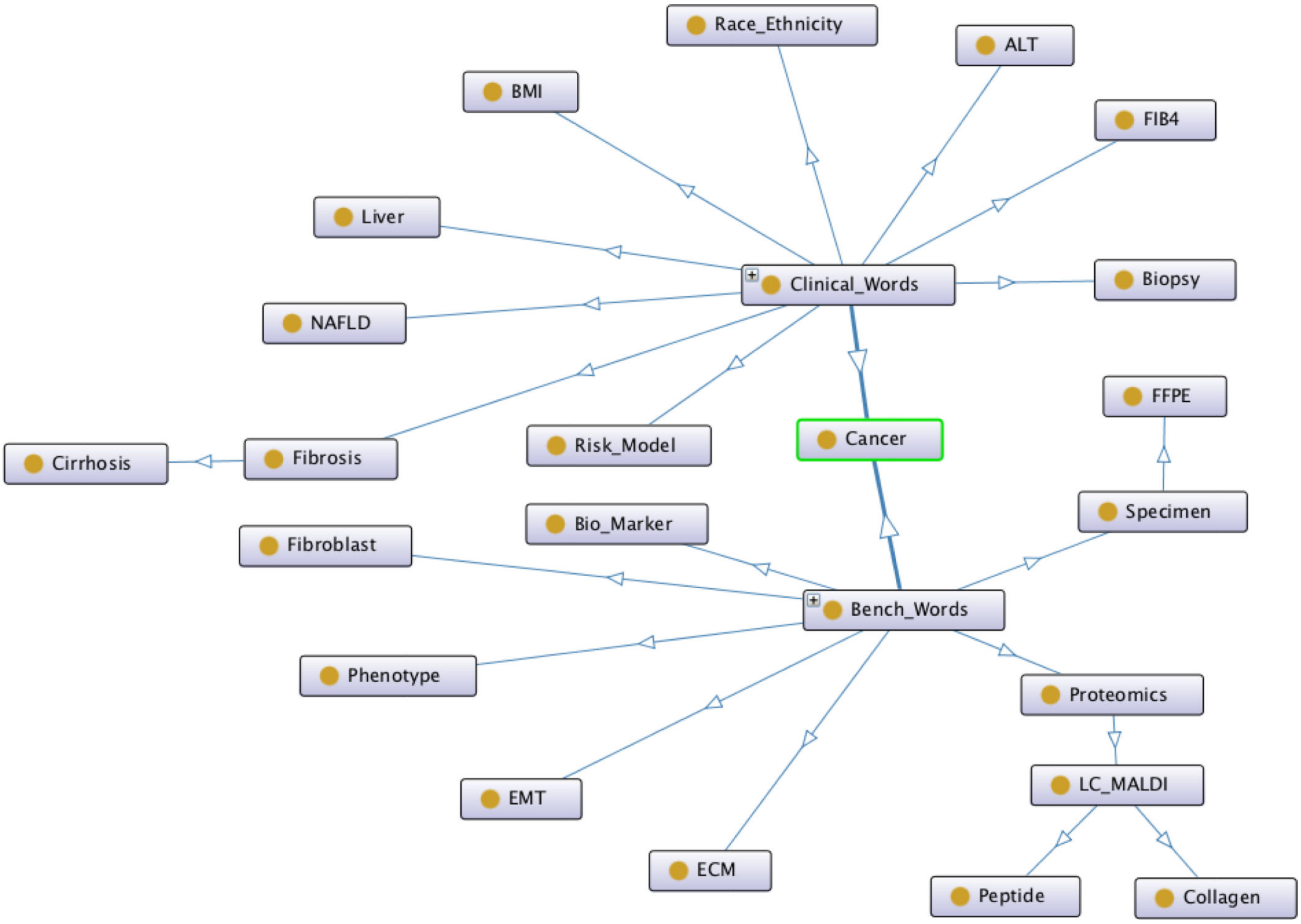

Figure 3.

Example collaborative knowledge representation that combines the key words for both clinical and bench researchers with subset relations between the keywords.

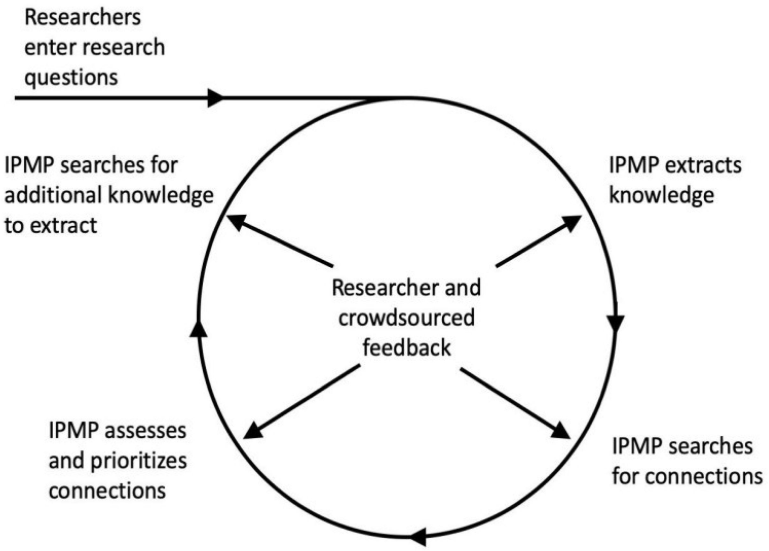
Researchers enter research questions
IPMP extracts knowledge
IPMP searches for connections
IPMP assesses and prioritizes connections
IPMP searches for additional knowledge to extract
Researcher and crowdsourced feedback

Figure 4.

Visualization of IPMP activities. Note the continuous nature of the search with multiple opportunities for interactions with the researchers and (if desired) a larger research community through crowdsourcing.