**Specific Aims**

Alzheimer's disease (AD) is the most common form of dementia. cby Alzheimer's, making it the fifth leading cause of death in adults older than 65. The cost of AD in 2020 alone surpassed the $305 billion mark, and it is estimated that by 2050 the number of Alzheimer's patients will have doubled to reach an expected cost of more than $1 trillion [1]. Due to its prevalence, AD is a major economic burden on society and a primary reason scientists constantly search for ways to fight this disease. One of the most popular research resources for AD is PubMed [2], a free online database of more than 190,000 citations from MEDLINE, life science journals, and online books. However, while PubMed allows users to search by author name or publication year to retrieve citations and full-text articles, it is still limited in compiling and associating information about AD. For example, PubMed does not allow users to associate and visualize specific proteins that are linked to a particular AD experiment (e.g., amyloid precursor protein (APP), presenilin-1/2 or tau) or by other related terms such as "amyloid plaques" or "neurofibrillary tangles." This is a significant limitation, as it constrains the ability of scientists to use PubMed as a research resource to uncover promising new associations between variables in their hypothesis discovery process. We want to highlight that AI in the healthcare market is projected to reach $194.4 billion by 2030; only the natural language processing (NLP) tools to aid drug discovery are projected to reach a staggering US$ 11.9 billion by the end of 2030 [3].

Knowledge graphs are a promising framework for understanding biomedical literature [4]. They can capture the relationships between entities and concepts in large databases, which enables users to explore this information in novel ways. Knowledge graphs allow users to visualize the complex relationships between entities in a network. In particular, they can be used to discover new associations between variables that were previously unknown or difficult using traditional text-based methods. However, a key challenge to making knowledge graphs an effective tool for biomedical research is transforming the unstructured and often ambiguous textual data from PubMed into a structured format that can represent relationships between entities. This process involves abstracting, disambiguation, and linking information to become more easily accessible and usable by researchers. At the same time, to truly make such a technological tool applicable to Behavioral and Social Science Research, a user-centered development approach is needed to understand better how users interact with the data and how the visualizations influence their perceptions of information.

AI POW LLC is a Texas-based company that integrates automation and interpretability technologies for Machine Learning (ML) to help researchers and stakeholders better understand the data they generate. The company's products include a suite of tools that can be used to assist with data cleaning, collating, and visualization. The company recently launched an alpha version of DiscoveryPath.ai, a platform designed to enable users to understand and explore the data they generate through visualization and natural language generation. As such, DiscoveryPath.ai is an online-based prototyping platform that helps us perform three distinct functions (1) automatically collects, cleans, and organizes data from PubMed, (2) transforms textual data into a structured format to uncover relationships in AD-focused literature, and (3) it offers a graphic way for the user to search for data and explore its relationships.

The user-centered nature of DiscoveryPath.ai, which aims to provide researchers with an overview of the relevant literature, allows for associations between variables to be identified that might otherwise remain hidden. In this Phase I work, the project team will be conducting user testing to test the efficiency of its discovery engine over other available systems. The project team proposes a platform for user testing, which involves having researchers use the DiscoveryPath.ai platform and documenting their experiences with it. The team found that users could navigate the system easily in the preliminary work leading to this proposal. However, there are some areas for improvement, such as the high complexity of the original knowledge graph. To solve this issue, our proposed tool now presents subgraphs in the form of Paths to present a simplified and clear view of the knowledge graph. This will allow users to navigate through the different subgraphs easily and, thus, make the search process more efficient and effective. In Phase I, the AI POW LLC team and UTHealth will determine the feasibility and user acceptability of the proposed DiscoveryPath.ai platform to curate and associate literature on AD to accelerate scientific discovery.

**Aim 1:** Engineer a user-centered fast-prototyping platform for data correlation from PubMed Library **Aim 2.** knowledge graph generation to conduct user testing on different user interfaces to interact with DiscoveryPath.ai and gauge its efficiency and user interest. **Aim 3**: Test DiscoveryPath.ai with ten researchers working on AD from McGovern Medical School at UTHealth for six months, iterating monthly improvements in the platform, and evaluate the usability outcomes among researchers and its ability to effectively visualize AD literature for hypothesis discovery.

**Impact:** The successful development of a DiscoveryPath.ai platform has the potential to accelerate scientific discovery in this area. By providing a way to easily explore the existing body of research, scientists can more quickly identify gaps in knowledge and areas that require further study. This could lead to a better understanding of the disease and how to treat it, ultimately benefiting patients.

**Research Strategy**

**(A) Significance**

Alzheimer's disease (AD) is the most common form of dementia, affecting about 5 million Americans. By 2020, AD will cost more than $305 billion a year—and this number is expected to surge above $1 trillion by 2050 [1]. Due to its prevalence, AD is a major economic burden on society. However, despite a large amount of research funding and the enormous effort put into finding cures for AD, no effective ways to visualize AD literature have been developed. For example, PubMed allows users to search by author name or publication year to retrieve citations and full-text articles—but it cannot identify associations between specific AD data. This limitation makes the PubMed database less useful for identifying previously unknown relationships between different variables in scientific research. In contrast, knowledge graphs are a promising framework for a deeper understanding of biomedical literature [4]. They enable users to explore this information novelly by capturing the relationships between entities and concepts—e.g., proteins or drugs—in large databases. Knowledge graphs allow users to visualize complex relationships between entities in a network and discover new associations that were previously unknown. However, a significant challenge to making knowledge graphs a useful tool for biomedical research is transforming PubMed's unstructured and ambiguous textual information into data about entities that relationships can represent.

AI POW LLC, based in Texas, is a company that integrates automation and interpretability technologies for Machine Learning (ML) to help researchers better understand the data they generate. Recently, the company launched an alpha version of DiscoveryPath.ai, a platform designed to enable users to understand and explore data generated through visualization. Because of this, DiscoveryPath.ai is an online-based prototyping platform that combines an intelligent search engine that uses natural language processing to analyze sentences and identify important concepts related to AD and a visualization tool that shows how each concept relates to one another in a network graph format. To fully understand how researchers will interact with DiscoveryPath.ai, the platform was designed to be tested by having users complete tasks using it and then documenting their experience.

**(B) Innovation**:

DiscoveryPath.ai, a proposed cloud-based prototyping platform that combines automated processing of literature available in PubMed, which involves using natural language processing, information extraction-correlation techniques, ML to generate insights from the literature, and a visualization module to display these insights in a user-friendly manner. In addition, the platform supports active learning by presenting users with an interactive visualization tool that displays the latest research findings and allows them to curate this information into a visual representation of their hypothesis discovery process. The AI POW team has already tested our initial prototype with scientists working on AD at McGovern Medical School at UTHealth in a joint interest to help us establish the feasibility of our proposed ML-based tool.

While there are a variety of tools for researchers to search for articles and data relevant to their research questions and determine how they should structure their studies, these tools have some limitations; they often provide a limited view of the literature, require users to have broad and deep technical skills to understand the relationship of different literature in AD, and are not optimized as research visualization tools. For example, Table 1 shows that the **Euretos** [5] tool offers in silico discovery and validation of targets, biomarkers, and molecular mechanisms. It helps scientists evaluate these mechanisms but doesn't provide a relationship across different documents; **EvidScience** [6] uses an automated text analysis process to extract data from medical articles. However, the company does not seem to aggregate or visualize this information in any way for end users; **Araicom** [7] allows users to visualize variables and generate hypotheses, yet the way the information is presented overwhelms users at a glance. It also seems that updates have ceased since 2014. **As a result, the current tools to explore scientific literature and the data generated from experiments are simply not enough to efficiently guide AD scientists in their hypothesis discovery process in the highly complex and dynamic nature of research.**

| | Automatic text analysis | Cross document analysis | Visualization | User-friendly interface | Graph database management |
|---|---|---|---|---|---|
| Euretos [5] | ✘ | ✘ | ✓ | ✓ | ✘ |
| EvidScience [6] | ✓ | ✘ | ✘ | ✘ | ✘ |
| Araicom [7] | ✘ | ✓ | ✓ | ✘ | ✘ |
| DiscoveryPath's opportunity to innovate | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 1.** Comparison of available tools in the market for literature visualization.

**(C) Approach**

**Preliminary studies**

For our preliminary study, we built a proof-of-concept system using readily available open-source tools to obtain, organize and visualize articles from PubMed. The resulting alpha prototype, DiscoveryPath v0.1, can be accessed at www.DiscoveryPath.ai. Below we elaborate on the construction of our proof-of-concept system, its current limitations, and the feedback from researchers working on AD at UT Health.

*API usage to download PubMed papers:* Several E-utilities APIs are provided to download the documents from the NIH National Library of Medicine database. In our research, we acquired the documents only corresponding to AD. *Three APIs are adopted in this preliminary work: ESearch, ELink, and EFetch [8].* The whole crawling process is composed of three parts using three APIs respectively. First, ESearch collects related document IDs with the keyword "Alzheimer's disease." Considering the retrieval of highly relevant documents, the collected document IDs returned from ESearch are not guaranteed to be highly related. Thus, in the second stage, the retrieved document IDs are further utilized as the query to ELink. ELink can acquire relevant documents based on the given document IDs. In this case, the collected papers are ready to be analyzed. Finally, EFetch is here to gain the abstract of each selected document ID. This study constructs the biomedical-related knowledge graph based on the downloaded abstract.

Natural language preprocessing for abstracts: Preprocessing of raw textual data is necessary to analyze the natural language data. The raw textual data is usually unconstructed and noisy when extracted from the natural language dataset. In this preliminary work, biomedical knowledge graphs are built upon the abstracts of documents related to AD. The first step is to lowercase the text, eliminating the confusion caused by lower and uppercase letters. Second, removing stop words and special symbols reduces the impact of extracting textual features from noise phrases and commonly-used terms. Tokenization in the third step decomposes the sentences into several unit tokens for further data processing. The token can be not only a single vocabulary but a phrase. After receiving the preprocessed unit tokens, part-of-speech tagging is the fourth step to derive the part-of-speech labels for each unit token. The tagging step is essential in our research since the part-of-speech labels particularly determine the entities and the relations in the knowledge graphs. Next, phrase chunking integrates the multiple terms or phrases usages in different documents, which can bridge the relation between two related papers even if the phrases appear in various formats. The final step is coreference resolution, which replaces coreference items with their original phrases. The coreference resolution mitigates the entity noises and prevents mislinks while constructing the knowledge graphs. *The open-source package NeuralCoref [9] was adopted for our preliminary coreference step to conduct correlation resolution.* In Figure 1, we depict the process of organizing and correlating textual data in the abstracts of [10, 11, 12] to find in-paper relationships to the term *Mild Cognitive Impairment (MCI)* and the corresponding correlations among the three papers (cross-paper correlation).

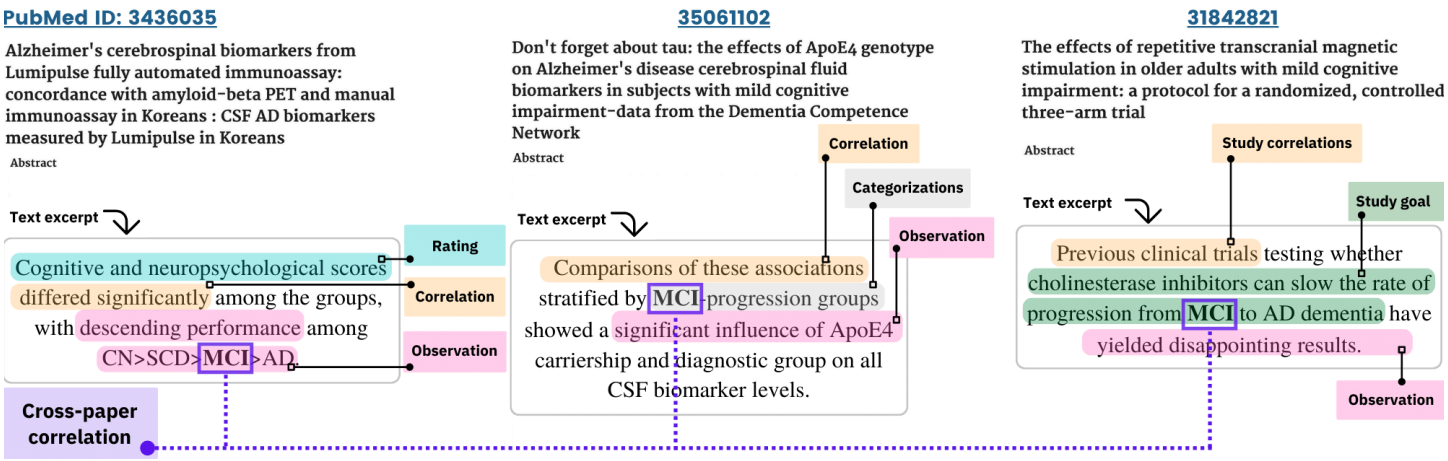


**Figure 1. Example of natural language preprocessing of searching the 'MCI' term in the abstracts of PubMed papers [10, 11, 12]. The sentences highlighted represent the relationships within the manuscript and to other manuscripts. This is the foundation for the Knowledge Graph of our proof-of-concept tool DiscoveryPath v0.1.**

Knowledge graph construction**:** We extracted the relation from the raw sentences using the pre-trained part-of-speech tagging model from NeuralCoref. The Knowledge graphs obtain "entities" and "relations" to structurally illustrate the

knowledge extracted from preprocessed textual data. Typically, a knowledge graph comprises several triplets extracted from the sentences. The triplets are in the format of head entity, relation, and tail entities. In this preliminary work, our team built a knowledge graph by condensing the knowledge of AD into triplets. The pipeline of constructing a knowledge graph is proposed in two steps: recognizing biomedical-related entities and creating meaningful and informative triplets. *In the first stage, the (NER) techniques from BERN [13] are adopted to extract the potential entities in the sentences*. BERN is a neural biomedical entity recognition tool trained on articles from PubMed, which can yield precise biomedical-related terms from the given input sentences. In summary, our first step provides the NER entities related to biomedical domains. Our second step links the NER entities with the extracted relations from each sentence of preprocessed textual data.

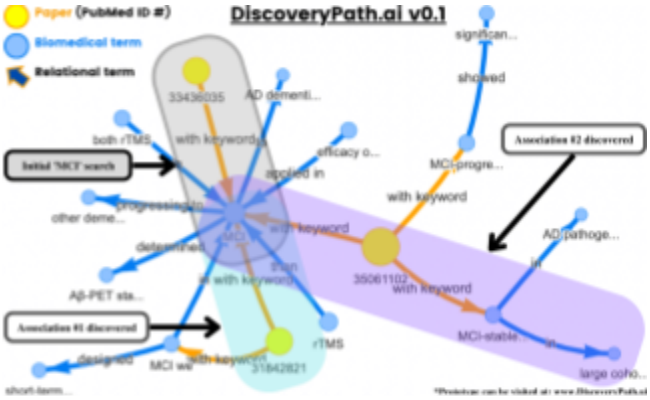


**Figure 2.** DiscoveryPath v0.1, visual representation of term correlation of AD literature for 'MCI' query.

Figure 2 shows the visualization of the KG constructed from the information processed for abstracts represented in Figure 1. We tested its usability with scientists at UT Health, who expressed enthusiasm about this tool's potential; however, our early users also pointed out various limitations of DiscoveryPath.ai v0.1. For example, the input query still *requires a simpler format* tailored to users that are non-programmers; second, while our team has improved the output visualization, there's still *more user feedback needed to improve our graphic user interface*; and third, because our preliminary work only included the abstracts of 1,000 AD works it made the search very limited, for this, *we propose to include all the AD literature available in PubMed* in our the Phase I work described below.

**Phase I Work Plan**:

**Aim 1: Engineer a user-centered fast-prototyping platform for data collection and correlation from PubMed Library.** Figure 3 depicts the flowchart of the proposed KD construction process. After collecting the articles from PubMed [1] in our preliminary work with downloading API [8], filtering out research works that are irrelevant to AD, and transforming the raw textual data of their abstracts via natural language preprocessing tools, another critical question we need to answer (The path from "Processed AD abstracts" to "Extracted AD entities" to "AD KG") is: how to organize and store the preprocessed textual data in a structured format, from which the non-expert can easily search and acquire informative information (e.g., latest biomedical solutions or scientific papers)?

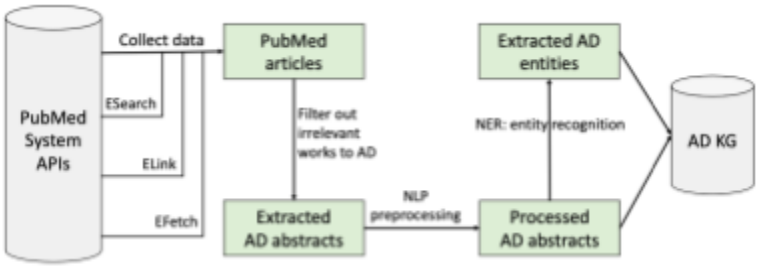


**Figure 3.** Overview of the proposed KG construction approach for Alzheimer's disease (AD).

Recently, knowledge graphs (KG)[2, 14, 15, 16], as a special graph-structured knowledge base, have been widely adopted to enhance the retrieval quality of real-world systems, ranging from social networks [17, 18], recommendation systems[19,20], question answering[21, 23], to education system [20, 21]. KG has been proven effective in reasoning answers located multiple hops away from an input query [24, 25]. Inspired by the remarkable achievements of KG in these open-world retrieval systems, this research aims to build a biomedical knowledge graph to store the preprocessed biomedical knowledge to facilitate biomedical information retrieval effectively. Following standard KG definition[26], our biomedical knowledge graph is a collection of triplets, where each triplet is a triple-tuple describing the interlink between the head entity and tail entity via a specific relation type, i.e., <head, relation, tail>. Specifically, the head/tail entities refer to words in the vocabulary or the IDs of research papers.

There are two major challenges when building an effective biomedical KG for AD. First, the KG is tailored specifically for this disease, so we should only focus on words relevant to AD. Second, given the plain words extracted from the abstracts from our preliminary work, discriminating between related words and entity words is a challenge. To tackle this challenge, we suggest building the KG in two stages.

We aim to emphasize AD-related keywords in the first stage and propose a biomedical-related entity recognition approach (The “Extracted AD entities” in Figure 3). The high-level idea is to filter out irrelevant entities to the target, i.e., AD. To be specific, we regard the task as a named entity recognition[27] (NER) problem and adopt the popular NER techniques used in BERN to extract the potential entities in the sentences. Specifically, BERN is a neural biomedical entity recognition tool trained on the articles from PubMed; it is publicly accessible and can yield precise biomedical-related terms from the given input sentences. Using this tool, the entities (i.e., words) related to AD could be easily identified.

In the second stage, we aim to extract a list of triples from the recognized entities in the first stage (The “AD KG” box in Figure 3). In practice, triples are created from the sentences in the head, relation, and tail format using Algorithm 2 in [28]. It contains three major steps. First, head and tail entities are extracted with their relations from the sentences, creating a list of triples. Second, a graph is created from those triples to reveal the relations among named entities in separate sentences. Based on the relations of prepositions such as in, on, and at, more triples are created to provide more links between named entities in the graph. Finally, the triples created by these two steps are joined to make the full list of triples for the given text. To clean the entities, we removed some tokens, including articles (e.g., a, an, the), possessive pronouns (e.g., its, their), and demonstrative pronouns (e.g., that, these) from the head and tails of each triple. As a result, the related phrases are selected by prepositions, postpositions, and verb phrases in the sentences.

**Aim 1 Milestone, Potential Pitfalls, and Alternative Strategies:** At the end of this project, our team will have built an effective knowledge graph database from scientific papers. This database represents the relationships between research concepts and highly-relevant scientific publications in a triple format, from which the non-expert can effectively acquire state-of-the-art research outcomes. A potential problem may arise if the constructed knowledge graph is noisy and incomplete due to the inevitable NER error and limited references related to AD. To tackle this pitfall, the team will develop novel machine-learning techniques to update the biomedical knowledge graph using either the structural information among entities or human experts in the loop to create a more accurate knowledge graph.

**Aim 2. Knowledge graph generation to conduct user testing on different user interfaces to interact with DiscoveryPath.ai and gauge its efficiency and user interest.** Despite the informative biomedical knowledge graph, accessing structural information in the knowledge graph is still a problem for non-computer scientists. Thus, providing a user-friendly and easy-to-manipulate interface system is necessary for users to further explore the information in our developed knowledge graph. Our demonstration system shows clear visualization of the knowledge graph and provides several auxiliaries functionalities for Scientists on their personalized retrieval usage. Besides, our developed system can input and delete information from the existing knowledge graphs, allowing domain experts to manually adjust the developed knowledge graph with their domain knowledge. To make functionalities more efficient and effective, we utilize the knowledge graphs to establish our graph database using Neo4j [29]. The graph database has numerous advantages: fast retrieval progress with simple query commands and a flexible database structure for rapid adjustment [30, 31]. By taking advantage of the graph database, our developed system can provide comprehensive functions for biomedical experts to explore new concepts and hypotheses. Neovis.js [32] is then utilized to construct the API for connecting the graph database to the front-end demonstration. As for the front-end website design, we adopted HTML5 to develop the frame of the web page and embedded a visualization block to show the retrieved results from the graph database.

This research showcases the developed knowledge graph by providing a user interface that allows users to retrieve, adjust, and highlight from the given keyword queries. For the first function, the demonstration system provides users to retrieve the related knowledge based on the given keywords. The retrieval results are shown in subgraphs, enabling users to explore new hypotheses and concepts. Considering the hypotheses and discovery usually occurring with multiple keywords, the system allows users to provide numerous keywords as search queries. The search results reveal whether there exists an unknown relationship between the keywords that are supposed to be not correlated. The second function is adjusting, which enables domain experts to filter unrelated information or append related knowledge. The noise information may happen during the automated knowledge graph construction. One of the reasons is the occurrence of homonyms or misunderstandings in collected documents, which is hard to detect if there is no prior domain knowledge. The same phenomenon happens when the latest information is required to add to the knowledge graph. The adjusting function shows its flexibility in fulfilling these requirements. The third function is highlighting. After the users acquire the related subgraph with their on-demand query, our developed system supports them in highlighting the specific path in the

retrieved subgraph. This function guides the users to concentrate on the detailed information for further concept and hypothesis exploration. Our research created an AI-based system recommending relevant scientific material to scientists studying AD. The system plans to use a dataset of more than 1 million papers on AD collected from PubMed Central, an open-access archive provided by the National Institutes of Health.

From the observations of current users of biomedical databases, we realize that the needs for biomedical professionals are not only to search the related information under the given keywords but also to provide further information exploration, which AI can assist. Knowledge graphs are a powerful tool for data exploration, helping scientists perform more efficiently [33]. We discover that scientists are eager to explore the existing data in a new way, meaning they're not just looking at it from one angle but using multiple angles and methods to look at the same thing. Obtaining a personalized knowledge graph system is essential to researchers, which keeps their collected information matching their research interests and saves the surveying time on filtering the non-related information. However, current object-relational databases [34, 35, 36] in biomedical are not eligible for users to easily update the latest information on themselves. Our DiscoveryPath system can create a personalized knowledge platform for individual users to acquire their own research needs. By providing a customized and flexible knowledge profile system for each researcher, our developed DiscoverPath system can accelerate the process of conducting deeper and broader research. In addition, scientists can manually add and delete the information shown inside the personalized DiscoverPath system. The accessibility of the latest information is also essential to researchers. It is challenging to discover new concepts without specific keywords. Researchers are required to find the latest citation or use the new keywords beforehand to acquire the latest information. Our system brings forth potential paths between prior arts and the latest academic documents, which create the environment for researchers to explore unaware knowledge.

The DiscoveryPath platform helps scientists find patterns in their data and see things differently than others initially considered. Compared with other biomedical platforms [5, 6, 7], the DiscoveryPath platform obtains more customized and powerful functionalities and shows great potential to explore multiple topics in the biomedical domain. The expected users of the developed system are biomedical experts and researchers, who will use it to explore new knowledge and for biomedical newcomers to initiate with more keywords and related expertise to acquire. In this research, the DiscoveryPath system is built upon the knowledge graph based on AD. The links provided in the knowledge graph indicate the relation beyond different biomedical-related entities and documentation. In this case, our system connects biomedical-related entities from different research work, contributing to efficient data exploration for discovering new concepts. After the first retrieved results, users can preserve the data based on their needs and keep it as their personalized database. Since the new relations between prior arts and the latest work are automatically generated, the new information can be broadcast to researchers in real time. This makes the DiscoveryPath system flexible and powerful for knowledge exploration.

**Aim 2 Milestone, Potential Pitfalls, and Alternative Strategies:** Our team will build an interactive user interface that demonstrates the advantage of the DiscoveryPath platform by retrieving, adjusting, and highlighting functionalities. A potential problem may occur if users ask for more functionalities supported in our platform. Our team will solve this by constructing a well-designed API structure that interacts with the front-end user interface, which can speed up the upgrading progress once there is a requirement to add a new procedure.

**Aim 3: Test of DiscoveryPath.ai with researchers working on AD to guide product development.** This Aim will require six months to test and evaluate the usability of DiscoveryPath.ai; at the end of each month, our team will iterate improvements and evaluate the outcomes to effectively visualize AD literature for hypothesis discovery using the lean methodology pictured in Fig. 4. Upon UT Health approval, we plan to reach out to ten UT Health (informatic) scientists currently working on AD research, each participant will be incentivized to join the study with a gift card. This aim will require (1) study recruitment, (2) a pilot study to **measure** the usability of DiscoveryPath.ai with five UT Health scientists, (3) an analysis of usability data will help us **learn** from the previous platform iteration and to inform design changes, and (4) final **build** of the platform will be tested testing it with a larger cohort of 10-20 UT Health scientists.

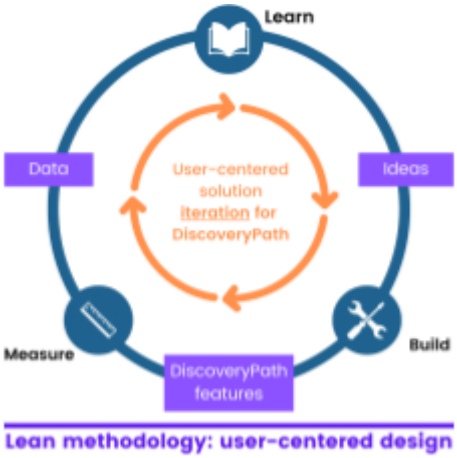


**Figure 4.** Aim 3 activities will set the stage to create a high-fidelity minimum viable product using an iterative lean approach based on a build-measure-learn iterative work.

1. Our team will recruit a small group of beta testers to help us test the platform's usability and provide feedback on its design and functionality. We plan to recruit five scientists working on AD research at UT Health in Houston, including senior researchers with PhDs, junior researchers with Master's degrees, and graduate students enrolled in an AD-tailored program. We will use this feedback to improve the design and functionality of DiscoveryPath.ai, for example, by adding a tutorial that teaches scientists how to use the platform's key features.
2. In our pilot usability study of DiscoveryPath.ai, we will evaluate the platform's performance utilizing a task analysis approach. Using a set of tasks developed by our partners at UT Health, we will evaluate how well scientists use DiscoveryPath.ai to search for and find research articles relevant to their needs. We will also evaluate how quickly they can find these articles and whether or not any barriers are preventing them from doing so.
3. Analysis of usability data to inform design changes. Following our usability study, we will provide feedback to our data science team on applying ML algorithms to biomedical literature to improve DiscoveryPath.ai's performance. This will include an analysis of the performance of several algorithms and their ability to predict the relevance of publications, as well as a discussion of different types of visualization elements to better serve our users.
4. Our team has planned a final evaluation of the platform by testing it with a larger cohort of 10-15 UT Health scientists. We will evaluate the performance of DiscoveryPath.ai over 2-3 weeks and provide feedback to our data science team on how they can improve the platform based on the user's experiences with its features. In this final evaluation (for Phase I), we will compare how well scientists use the new features on DiscoveryPath.ai to search for and find research articles relevant to their needs. We will measure again and compare to our previous results in *step 2* regarding the time and effort it takes to researchers to find relevant information.

**<u>Aim 3.</u> Milestone, Potential Pitfalls, and Alternative Strategies**: At the end of this Aim, our team will have demonstrated the feasibility and usability of DiscoveryPath. This will be demonstrated through data collected and analyzed and feedback from actual end users participating in the study. A potential problem may arise if there is insufficient access to articles or the search results are irrelevant. To avoid this pitfall, the team will expand the number of articles analyzed from an initial pool of 10,000 articles from PubMed to create a more comprehensive search.

**Summary**: At the end of Phase 1, our team will have implemented DiscoveryPath.ai 1.0, as illustrated in Figure 5, this working prototype will have demonstrated its usability through the data collected and analyzed. Phase 2 efforts will focus on expanding indexed libraries; based on DiscoveryPath.ai 1.0 user feedback, we will enhance our product's user experience and functionality for DiscoveryPath.ai 2.0; scale up efforts to recruit a larger pool of participants. At the end of Phase 2, DiscoveryPath.ai will have shown that it can outperform human teams in identifying and extracting key information from relevant research, enabling researchers and clinicians to research smarter and faster. After validating our results, our team will have developed a fully functioning, scalable product that can be used by a broader audience.

**Commercialization**: AI in the healthcare market was valued at $8.23 billion in 2020 and is projected to reach $194.4 billion by 2030, growing at a CAGR of 38.1% from 2021 to 2030. In particular, NLP tools to aid drug discovery are projected to reach US$ 11.9 billion by the end of 2030, and AI-powered search engines can for this market reach 1.1 Billion [3]. A tool like DiscoveryPath aims to streamline doctors' research process, increase our understanding of AD, and create true economic value for the researchers and hospitals that adopt it. For this, the AI POW team has partnered with UTHealth to develop the DiscoveryPath framework proposed for phase 1. As a small business, this partnership will be key to positively impacting more hospitals, researchers, patients, and their families.


**Figure 5.** Wireframe of DiscoveryPath v1.0.

**Reference**


[1] Wong, Winston. "Economic burden of Alzheimer disease and managed care considerations." The American journal of managed care 26.8 Suppl (2020): S177-S183.

[2] Doms, Andreas, and Michael Schroeder. "GoPubMed: exploring PubMed with the gene ontology." Nucleic acids research 33.suppl_2 (2005): W783-W786.

[3] Bio.IT World Press. AI in Drug Discovery Market Size US$ 11.9 Billion by 2030. Retrieved 2022.

[4] Li, Michelle M., Kexin Huang, and Marinka Zitnik. "Graph Representation Learning in Biomedicine." arXiv preprint arXiv:2104.04883 (2021).

[5] Euretos AI Platform. 2022 (Accessed Nov. 2, 2022). https://www.euretos.com.

[6] EvidScience. 2020 (Accessed Nov. 2, 2022). https://www.evidscience.com.

[7] Araicom Bioinformatics. 2014 (Accessed Nov. 2, 2022). http://araicom.com.

[8] Kans, Jonathan. "Entrez direct: E-utilities on the UNIX command line." Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US), 2022.

[9] M. Honnibal, I. Montani, S Van. Landeghem, and A. Boyd. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303.

[10] Moon, Sohee, et al. "Alzheimer's cerebrospinal biomarkers from Lumipulse fully automated immunoassay: concordance with amyloid-beta PET and manual immunoassay in Koreans." Alzheimer's research & therapy 13.1 (2021): 1-12.

[11] Benson, Gloria S., et al. "Don't forget about tau: the effects of ApoE4 genotype on Alzheimer's disease cerebrospinal fluid biomarkers in subjects with mild cognitive impairment—data from the Dementia Competence Network." Journal of Neural Transmission (2022): 1-10.

[12] Taylor, Joy L., et al. "The effects of repetitive transcranial magnetic stimulation in older adults with mild cognitive impairment: a protocol for a randomized, controlled three-arm trial." BMC neurology 19.1 (2019): 1-15.

[13] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

[14] Noy, Natasha, et al. "Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done." Queue 17.2 (2019): 48-75.

[15] Kertkeidkachorn, Natthawut, and Ryutaro Ichise. "T2kg: An end-to-end system for creating knowledge graph from unstructured text." Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[16] Chen, Xiaojun, Shengbin Jia, and Yang Xiang. "A review: Knowledge reasoning over knowledge graph." Expert Systems with Applications 141 (2020): 112948.

[17] He, Qi, Jaewon Yang, and Baoxu Shi. "Constructing knowledge graph for social networks in a deep and holistic way." Companion Proceedings of the Web Conference 2020. 2020.

[18] Qian, Jianwei, et al. "Social network de-anonymization and privacy inference with knowledge graph model." IEEE Transactions on Dependable and Secure Computing 16.4 (2017): 679-692.

[19] Guo, Qingyu, et al. "A survey on knowledge graph-based recommender systems." IEEE Transactions on Knowledge and Data Engineering (2020).

[20] Bao, Junwei, et al. "Constraint-based question answering with knowledge graph." Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. 2016.

[21] Huang, Xiao, et al. "Knowledge graph embedding based question answering." Proceedings of the twelfth ACM international conference on web search and data mining. 2019.

[22] Chen, Penghe, et al. "Knowedu: A system to construct knowledge graph for education." Ieee Access 6 (2018): 31553-31563.

[23] Chen, Penghe, et al. "An automatic knowledge graph construction system for K-12 education." Proceedings of the fifth annual ACM conference on learning at scale. 2018.

[24] Liao, Jinzhi, et al. "To hop or not, that is the question: Towards effective multi-hop reasoning over knowledge graphs." World Wide Web 24.5 (2021): 1837-1856.

[25] Saxena, Apoorv, Aditay Tripathi, and Partha Talukdar. "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.

[26] D'souza, Jennifer, and Nandana Mihindukulasooriya. "The State of the Art on Knowledge Graph Construction from Text: Named Entity Recognition and Relation Extraction Perspectives." The Knowledge Graph Conference 2022. 2022.

[27] Nothman, Joel, et al. "Learning multilingual named entity recognition from Wikipedia." Artificial Intelligence 194 (2013): 151-175.
[28] Stewart, Michael, Majigsuren Enkhsaikhan, and Wei Liu. "Icdm 2019 knowledge graph contest: Team uwa." 2019 IEEE international conference on data mining (ICDM). IEEE, 2019.
[29] The Neo Database. Daniela Florescu, Oracle, in the “ACM Queue”, Vol. 3, No.8, 2005.
[30] Angles, Renzo, and Claudio Gutierrez. "Survey of graph database models." ACM Computing Surveys (CSUR) 40.1 (2008): 1-39.
[31] Zheng, Dongyang, et al. "Scholar-Course Knowledge Graph Construction Based on Graph Database Storage." International Symposium on Emerging Technologies for Education. Springer, Cham, 2021.
[32] Neovis.js Toolkit. Sep. 2022 (Accessed Nov. 2, 2022). https://github.com/neo4j-contrib/neovis.js/.
[33] Qiu, Yuchen, et al. "Tax-KG: Taxation Big Data Visualization System for Knowledge Graph." 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP). IEEE, 2020.
[34] Stinson, Barry. PostgreSQL essential reference. Sams Publishing, 2001.
[35] Hinz, S., P. DuBois, and J. Stephens. "MySQL 5.7 Reference Manual." (2005).
[36] Microsoft.com. Editions and supported features of SQL Server 2019. Retrieved 2022.

**DiscoveryPath: A user-centered knowledge graph to curate and associate literature on Alzheimer's disease to accelerate scientific discovery.**

## Due date: November 4, 2022, 5 PM ET

**NIH request:**

3. **NIH/NIA 009 - AI/ML Tool for Visualizing Behavioral and Social Science Research:** Develop an AI-based tool for literature visualization and hypotheses discovery, which might be marketed ultimately to various institutions that consume scientific research, with an emphasis on AD.

### BSR – AI/ML Tool for Visualizing Behavioral and Social Science Research

#### Scope of Work

- To develop an AI-based tool for BSR-specific literature visualization and hypothesis discovery, that can be marketed to behavioral science investigators and research institutions that consume scientific research with a particular emphasis on AD/ADRD.

#### Background

- Behavioral and social science research represents an important segment of the research funded by the NIH and is key to developing effective new approaches for supporting individuals with cognitive impairment.
- A major barrier lies in the challenge of compiling, collating and comprehending the literature.
- No tool exists that allows users of BSR research to quickly derive a causal overview of the relevant literature, so that they can uncover promising new associations between variables.

#### Phase I

- Establish feasibility of an AI/ML based tool for BSR specific literature visualization and hypotheses discovery.
- Conduct user testing to prove efficiency over "standard" search.
- Conduct user feedback survey to gauge interest.

#### Phase II

- Develop user management system for enhanced literature navigability and result accuracy.
- Verify of the tool's efficiency, output comprehensibility, and efficacy for hypothesis discovery.
- Broaden the type of science indexed – with an emphasis on AD – in conjunction with ontology development for more comprehensive literature exploration.

**Question 2:** **What are the current tools used by behavioral and social scientists (Alzheimer's researchers) to understand the AD portfolio?**

*Answer:* *See chart below.*



| TOOL | DESCRIPTION | DOMAIN INDEXED | EXTRACTION | | VISUALIZATION | |
|---|---|---|---|---|---|---|
| | | | VAR | C. REL | VAR | C. REL |
| **Semantic Scholar** (owned by AI2) | Free, non-commercial tool; extracts concepts from research papers; no visualization | General | ✓ | - | - | - |
| **BASE** | Free, non-commercial tool; provides semantic matching, based on EuroVoc thesaurus; no visualization | General | - | - | - | - |
| **WolframAlpha** | Computational knowledge engine; visualizes facts across a variety of domains | Comput. Sciences | - | - | - | - |
| **Meta** (owned by FB) | Tracking research connections and trends; not launched yet | General | ✓ | - | - | - |
| **Resolute Innovation** | Literature review tool; suggests relevant concepts and terms; visualizes semantic relationships between concepts | General | ✓ | - | ✓ | - |
| **Dimensions** (owned by Digital Science) | Linked-research citation database; visualizes connections between publications, grants, patents, and research developments & trends | General | ✓ | - | ✓ | - |
| **Iris AI** | Extracts topics from research abstracts; visualizes research by related topics; they have plans for extracting hypotheses | General | ✓ | - | ✓ | - |
| **Metabus/INN** | Meta-analysis tool; uses manually curated results (no AI yet); extracts variables and correlation strengths between variables; visualizes concept taxonomy; cluttered visual interface | Beh. & Social Sciences | ✓ | - | ✓ | - |
| **Euretos** | In silico discovery & validation of targets and biomarkers; helps scientists evaluate molecular mechanisms; no visualization | Molecular Biology | ✓ | ✓ | - | - |
| **OccamzRazor** | Extracts disease info from medical research related to Parkinson's; provides knowledge graph visualization for Parkinson's disease | Parkinson's Research | ✓ | - | ✓ | - |
| **Robot Reviewer** | Non-commercial, open-source (potential resource for us); automatically extracts *PICO* study characteristics from papers; results are provided in PDF format; relatively low accuracy | Biomed. Sciences | ✓ | ✓ | - | - |
| **EvidScience** | Extracts intervention and outcome variables from medical papers; no plans for data aggregation or visualization discernable | Biomed. Sciences | ✓ | ✓ | - | - |
| **Araicom** | Visualizes variables & generates potential hypotheses; business appears to have stagnated | Biomed. Sciences | ✓ | ✓ | ✓ | - |
| **Quertle** | Displays connections between variables as correlation matrix (no directionality, aggregation, or drill-down options); founded in 2008, but appears somewhat limited in market reach | Biomed. Sciences | ✓ | ✓ | ✓ | - |
| **Researchably** | Literature review tool, with PV-specific features; extracts variables (and potentially relationships) and categorizes scientific papers; no plans for data aggregation or visualization discernable | Biomed. Sciences | ✓ | ✓ | - | - |