Article published online: 2025-04-08

Knowledge Representation and Management in the Age of Long Covid and Large Language Models: a 2022-2023 Survey

Jonathan P. Bona, Ph.D.
Department of Biomedical Informatics, University of Arkansas for Medical Sciences

Summary

Objectives: To select, present, and summarize cutting edge work in the field of Knowledge Representation and Management (KRM) published in 2022 and 2023.

Methods: A comprehensive set of KRM-relevant articles published in 2022 and 2023 was retrieved by querying PubMed. Topic modeling with Latent Dirichlet Allocation was used to further refine this query and suggest areas of focus. Selected articles were chosen based on a review of their title and abstract.

Results: An initial set of 8,706 publications were retrieved from PubMed. From these, fifteen papers were ultimately selected matching one of two main themes: KRM for long COVID, and KRM approaches used in combination with generative large language models.

Conclusions: This survey shows the ongoing development and versatility of KRM approaches, both to improve our understanding of a global health crisis and to augment and evaluate cutting edge technologies from other areas of artificial intelligence.

Keywords

Knowledge Representation and Management; Biomedical Ontologies; Artificial Intelligence; Post-Acute Covid-19 Syndrome; Natural Language Processing

Yearb Med Inform 2024:
http://dx.doi.org/10.1055/s-0044-1800747

1. Introduction

The goal of this survey is to present and summarize cutting edge work in the field of Knowledge Representation and Management (KRM) published in 2022 and 2023. Efforts to understand and combat the ongoing COVID-19 pandemic remain prominent topics in medical informatics. The 2021 Yearbook's Knowledge Representation and Management survey paper highlighted the role of clinical knowledge in the COVID-19 pandemic in moving towards a global learning health system [1]. The 2022 survey paper focused on KRM work seeking to address health inequities, including those exposed and exacerbated by the COVID-19 pandemic [2].

Now in the fourth year of the pandemic, KRM approaches continue to be adapted and applied to address many areas of this multifaceted crisis. One example is the Coronavirus Infectious Disease Ontology (CIDO) effort [3], which provides a semantically rich framework for representing and managing information about coronavirus diseases. The application of KRM approaches to COVID-19 issues includes efforts aimed at the growing health crisis posed by long COVID. One special area of focus for this survey is emerging knowledge representation work that seeks to address the growing health crisis posed by long COVID ("post-COVID syndrome", "Postacute sequelae of SARS-CoV-2 infection (PASC)"). Characterized by the long-term persistence of symptoms following SARS-CoV-2 infection [4], and occurring following at least 10% of COVID-19 cases, long COVID is complex, associated with over 200 distinct symptoms [5] and serious complications [6]. A January 2023 estimate placed the number of persons afflicted with long COVID worldwide at over 65 million. Long COVID negatively impacts health and quality of life for those afflicted by it [7] and is creating significant economic burdens and stresses on healthcare systems [8,9]. In the face of this complex and baffling health crisis, there is a critical need for research to collect, organize, integrate, and interpret large and diverse sets of relevant data [10-12]. Knowledge representation approaches are well-poised to help address this need, and this review examines several such efforts now under way.

Another area of special focus in this survey is the growing field of combined use of knowledge representation techniques with generative AI and Large Language Models.

The last two years have seen a revolutionary development in Artificial Intelligence (AI) and Natural Language Processing (NLP): the emergence of generative Large Language Models (LLMs) [13] with the ability to use and produce natural language at a level of competence far exceeding any previous efforts. OpenAI's Generative Pre-trained Transformer (GPT) [14] (and its public-facing chat interface ChatGPT) have attracted significant public interest, as well as interest in its capabilities from experts in fields that have already been leveraging AI technologies, including in healthcare and medicine [15–17]. LLMs, and GPT/ChatGPT in particular, have received attention for their potential to assist in medical

diagnosis [18,19]. To become such competent users of language, GPT and similar generative LLMs were trained on massive amounts of text harvested from the Internet. This has allowed for the creation of open domain conversational agents like ChatGPT, with an impressive ability to discuss nearly any topic in natural language. Sometimes dismissed as "stochastic parrots" [20] or, at the other extreme, considered as a harbinger of Artificial General Intelligence (AGI) [21], these models do show some ability to engage in tasks requiring reasoning. However, their limited ability to understand [22], propensity to "hallucinate" (make things up) [23], and limited ability to reason [24] pose challenges that may best be addressed by solutions that combine LLMs with other AI techniques best-suited for such tasks.

These challenges (understanding, veracity, and reasoning) are all areas that are well suited to solutions from the areas of KRM. The use of these approaches in combination with LLMs is a rapidly emerging field, and this survey highlights recent projects in this area.

Table 1. PubMed queries and result counts

Query		#Results
Q1	(„knowledge representation" OR „description logic" OR „semantic web" OR „ontology" OR „ontologies")	8,706
Q2	Q1 AND NOT („gene ontology") NOT „review"[sb] NOT „systematic review"[sb]	1,264

2. Methods

A comprehensive set of KRM-relevant articles published in 2022 and 2023 was retrieved by querying PubMed with a large disjunctive query including terms "ontology", "knowledge representation", "semantic web" and "description logic", filtered to results from 2022 and 2023 (see query Q1 in Table 1). The resulting corpus of 8,706 titles and abstracts was judged too large to be evaluated manually.

The corpus was analyzed using topic modeling (Latent Dirichlet Allocation (LDA) [25], implemented in Python using scikit-learn) to identify significant topics common across large subsets of these abstracts. LDA does not label or categorize the topics it identifies, but presents each topic as a distribution over a set of words (Figure 1). Inspecting the top words for a topic is one way of coming to understand what the topic is about. This topic analysis of the 8700 or so abstracts retrieved by our query identified a particularly relevant topic for this survey, which may best be characterized as "biomedical KRM" (topic #3). Table 2 shows the ten most important terms for this topic, along with two others: discovered by this analysis identifiable as "covid-19" (topic #9) and "cancer" (topic #2), among others (genetic epidemiology (topic #6), genomics (topic #1), non-coding RNA research (topic #7), pharmacology and drug discovery (topic #4)).

A manual inspection of a subset of abstracts within these topics revealed that the widespread use of the Gene Ontology (GO) to annotate scientific literature was a contributing factor to the large amount of search results over this timespan. A second query (Q2) was developed to exclude most mentions of the GO, as well as excluding systematic reviews. The resulting set of abstracts was manually reviewed to identify abstracts matching two main topics of interest: COVID-19, and specifically long COVID, and generative NLP / LLMs.

Table 2. A sample of recognizable abstract topics and their top terms.

Topic#	2	3	9
Description	Cancer	Biomedical KRM	Covid-19
Top 10 Terms	cancer	data	covid-19
	immune	knowledge	ad
	expression	information	sars-cov-2
	genes	health	infection
	analysis	research	severe
	patients	semantic	viral
	prognostic	ontologies	virus
	cell	model	disease
	gene	use	acute
	tumor	approach	infected

3. Results

3.1. Long COVID

Long COVID negatively impacts health and quality of life for the tens of millions of people worldwide who are or have been afflicted by it [7] and is creating significant economic burdens and stresses on healthcare systems [8,9]. There is a critical need for research to collect, organize, integrate, and interpret large and diverse sets of relevant data [10-12].ledge representation and

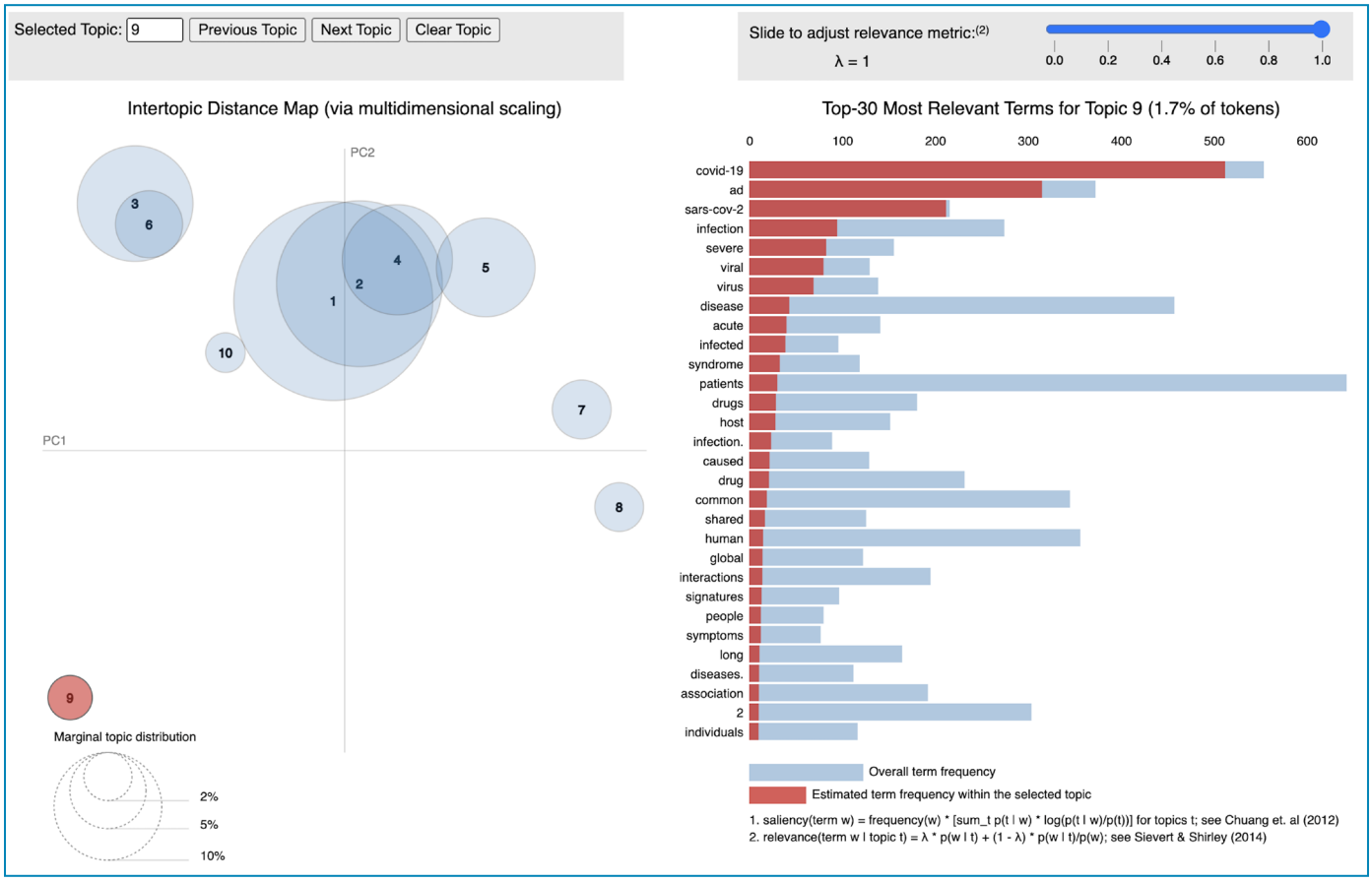

Figure 1. PyLDAVis topic model visualization focused on topic #9 (COVID-19), for which the 30 most relevant terms are highlighted at the right. Numbered circles on the left show the distance between topics when reduced to a 2D space. The marginal topic distribution shows how relatively popular a topic is within the entire corpus.

knowledge management approaches are well-poised to help address this need. Below we highlight a few such reports seeking to clarify our understanding of long COVID through the application of KRM techniques. There is a rich body of KRM work targeting COVID-19 more generally, some of which has been reported in past years' surveys, so we emphasize here recent additions to this literature that are focused on long COVID. Identifiable sub-themes in this section include:

General knowledge representation and knowledge management approaches to handling the massive amount of complex data relevant to developing an understanding of long COVID;

Development and use of ontologies/terminologies to capture, organize, and understand information about the diverse clinical presentations of long COVID;

Use of the GO specifically within analyses that investigate similarities between long COVID and conditions such as myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS).

Providing an overview of health informatics approaches to managing long COVID, Ambalavanan et al. [26] highlight the large-scale collection of electronic health record data for the purpose of understanding the COVID-19 pandemic and managing long COVID. Among these efforts in the United States is the National Covid Cohort Collaborative (N3C) [27] data enclave, which has assembled de-identified medical records including 3 billion observations for over 22 million from across 83 sites [28]. Ambalavanan et al. review long COVID identification and characterization techniques, endorsing an approach to interpreting health data in pursuit of an understanding of long COVID that includes knowledge representation techniques in combination with other AI techniques in the areas of machine learning and deep learning. Useful for standardizing and managing complex data and knowledge and data are phenotypic ontologies, like the HPO which provides "a hierarchical classification of standardized human pathological features that are used to define phenotypes" [29].

In an ontology-driven effort to organize emerging information about clinical manifestations of long COVID, Deer et al. [30] conduct an analysis of almost sixty long COVID studies, identifying 287 distinct clinical manifestations ("signs, symptoms, and laboratory as well as imaging abnormalities") of long COVID, and mapping these symptoms to terms in the Human Phenotype Ontology. They also create new terms (HPO term synonyms and plain-language definitions) aimed at making this information more accessible to laypersons, and developing

common data elements to support interoperability across long COVID studies.

A project with a similar goal, to understand the symptomology of long COVID as expressed in clinical notes, Wang et al. [31] report on the development of a comprehensive long COVID symptom lexicon based on their analysis of over 300,000 notes from long COVID patients. The approach uses the MTERMS [32] NLP tool to identify relevant symptoms mentioned in these notes, augmented by chart review and other evaluation. The result is PASCLex, a publicly-available lexicon of 355 long COVID symptom terms, as well as detailed information about the frequency with which these symptoms are mentioned in notes for long COVID patients.

In a long COVID related project with a small KRM component, Lee et al. [33] seek to identify distinguishing characteristics in the immunological profiles of long COVID and severe COVID patients. This project uses the GO in gene set enrichment ultimately finding differences in DNA methylation between those who developed long COVID and those who did not.

Lv et al. [34] also conduct GO and pathway enrichment analysis, aimed at identifying common genes among long COVID patients and patients with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). Symptoms experienced by sufferers of ME/CFS and long COVID overlap significantly, and there are a number of similarities in the underlying biology [35]. This study conducted GO analysis of genes associated with long COVID and ME/CFS, and identified nine common genes, also predicting/proposing candidate drugs that might be used as interventions to treat long COVID based on their interactions with these genes. In a systematic review and analysis seeking to understand the commonalities between long COVID and ME/CFS, Tziastuodi et al. [36] also examine the genetics of COVID-19 and ME/CFS, combining findings from 71 COVID-19 studies and 26 ME/CFS studies, conducting GO analysis, finding evidence of gene overlap, and suggesting possible interventions to be explored.

In another long COVID project that uses electronic medical record (EMR) data and ontologies, Reese et al. [37] perform a semantic clustering of nearly 6500 long COVID patients based on multicenter EMR data collected within the National Covid Cohort Collaborative (N3C). N3C data uses the OMOP [38] common data model. In order to use the HPO with this set of patient data, this project used the OMOP2OBO [39] algorithm to map N3C OMOP codes to HPO terms. They deployed the Phenomizer algorithm to compute a similarity matrix of long COVID patients based on their phenotypic features expressed in HPO, and organized these patients into clusters using k-means clustering. This analysis yielded six long COVID subtype clusters, which the authors labeled as following based on their characteristic features: multisystem+lab, pulmonary, neuropsychiatric, cardiovascular, pain/fatigue, and multisystem-pain. The approach was developed using a set of data from one of six data partner sites, and then applied to the remaining data to demonstrate generalizability.

3.2. Knowledge Representation and Large Language Models

The emergence of generative LLMs [13] is a major scientific development that has attracted significant public interest. LLMs like GPT/ChatGPT and its competitors are also seeing widespread interest in their use in healthcare and medicine [15–17]. Possessing an unprecedented ability to use and generate natural language text, these tools are nonetheless limited in their ability to perform intelligent tasks involving understanding and reasoning, and are known to "hallucinate" – making up answers rather than trying to figure out what is known to be true. Knowledge representation and knowledge management approaches can help to address these challenges. The use of KRM approaches in combination with LLMs is a rapidly emerging field, and this survey highlights recent projects in this area.

Among projects presented in this survey that combine KRM with LLMs in 2022-2023, a few sub-themes are identifiable, including:

LLM-based natural language question & answer interfaces used to facilitate access to ontology-aligned biomedical datasets;

Use of ontologies as sources of ground truth/knowledge or evaluation of LLM-based solutions;

Use of LLMs to augment GO term enrichment and gene set analyses.

In a topical and recent survey and analysis, Denecke et al. [40] queried experts in healthcare NLP to study how transformer-based models – LLMs like GPT are the most well-known examples – may shape the practice of healthcare. The contribution of transformer models to healthcare knowledge management was the most immediately relevant among the eight categories of use cases identified in this qualitative analysis. Respondents were also enthusiastic about potential applications in the areas of documentation and clinical coding, and decision support, both areas where integration with existing KRM approaches will likely be fruitful.

The GO is one of the largest and most widely-used biomedical ontologies in the world [41]. As discussed earlier, GO is so heavily used to annotate genes and gene products in the scientific literature and is also so extensively used for term enrichment in bioinformatics, that explicitly excluding "gene ontology" in literature queries may be necessary to focus those queries on knowledge representation work. As of January 2024, GO has over 40,000 active terms [42], and in 2023 alone it was referenced in over 4,500 publications indexed by PubMed.

To assist researchers with protein function prediction work using GO, Giri et al. [43] have developed the GO2Sum tool, which uses an LLM (Text-to-Text Transfer Transformer (T5) [44]) fine-tuned on GO terms and UniProt descriptions. GO2Sum generates descriptions of protein functions by aggregating and summarizing descriptions of individual genes associated with GO terms. GO2Sum was evaluated by comparing its output to existing functional descriptions in the UniProt database.

Similarly, Hu et al. [45] seek to facilitate gene set analyses by deploying LLMs as "functional genomics assistants" capable of generating names and summaries for

gene sets of interest. This project created a pipeline that uses the standard GPT-4 API, providing an initial prompt that instructs GPT to act as an "efficient and insightful assistant to a molecular biologist", then given a set of genes and asked to generate biologically descriptive names and analysis text. Unlike the GO2Sum project described above, this use of GPT-4 relies on knowledge latent in the model, and does not use GO terms as inputs. Rather, the GPT pipeline is evaluated by comparing its outputs to knowledge stored in the GO. Note however, that GPT's large training corpus of texts of all sorts harvested from the web, likely includes GO annotations [46]. The authors of this study conclude that GPT-4 has potential for this use, though in gene set analyses it produces "highly similar names" to the GO 50% of the time, and produces "unverifiable statements" in its analyses 22% of the time. It is worth noting that for projects like this and the preceding, the availability of the GO as a high-quality curated knowledge source appears to be invaluable as a source of truth and comparison.

In a recent PubMed-indexed preprint on related work, with conclusions more critical of LLM-based methods, Joachimiak et al. [46] report on their development of the Structured Prompt Interpolation of Natural Language Descriptions of Controlled Terms for Ontology Reporting (SPINDOCTOR) method, which performs gene set function summarization using GPT based on different sources of information, including ontologies and latent knowledge baked into the language model. The report concludes that results produced by this LLM-based approach are "typically plausible, relevant, and largely free of hallucination", but they lack precision and the ability to reliably quantify the relevance of terms and perform worse on less well-known genes. These results highlight the continued need for curated knowledge bases to keep LLM-based projects grounded in reality.

Munarko et al. [47] report on their development of a tool for using natural language queries to conduct exploratory searches on repositories of data that have been semantically annotated with ontologies, the Biosimulation Model Search Engine (BMSE) [47]. While they claim generalizability of the approach to other domains for data annnotated using RDF and ontologies, their specific use case is a repository of biosimulation models encoded in CellML and using annotation from ontologies such as the Ontology of Physics for Biology (OPB), the Foundational Model of Anatomy (FMA), and Chemical Entities of Biological Interest (ChEBI). This work builds on the CASBERT mechanism [48] developed by the same group, which uses the Bidirectional Encoder Representations from Transformers (BERT) [49] LLM to support text-based searches over ontology-aligned data, as a more accessible alternative to SPARQL queries. Previous work in this area includes SPBERT [50], which used multilayer bidirectional transformers like BERT for SPARQL query generation and result verbalization, training their models on SPARQL query logs to create a tool that translates natural language queries into SPARQL, and to generate natural language answers based on the query results.

LLMs, and GPT/ChatGPT in particular, have received attention for their potential to assist in medical diagnosis [18,19]. Reese et al. [51] join this conversation with an ontology-based assessment of GPT-4's performance at differential diagnostic reasoning. The group performed two assessments: one based on de-identified clinical notes, and one based on clinical features extracted from those notes as HPO terms. The second evaluation approach is motivated by noting that, in practice, clinical notes ordinarily cannot be transmitted to ChatGPT because they will contain personally identifiable information. When working with narrative-based text, the performance of GPT-4 at this task was "good" at 38.7%, and similar to the performance reported in Kanjee et al. [18]. However, on the more practical task of producing diagnoses based on extracted clinical features, GPT performed much worse. The paper concludes with the suggestion to combine use of LLMs with knowledge representation approaches for clinical diagnosis solutions.

Authoring summary notes is a complex and time-consuming process for clinicians. To assist with this, Searle et al. [52] have developed an approach to generating inpatient Brief Hospital Course (BHC) summaries, pulling together information from a diverse set of source notes written during a patient's hospitalization. The basic approach uses Bidirectional and Auto-Regressive Transformers (BART) [53]. This is extended by augmenting the model with "guided summarization" by using SNOMED-CT terms for problems and interventions extracted from the notes via more traditional concept extraction approaches. The resulting ontology-guided summarization model is more focused on clinically relevant information and outperforms baseline models at the brief hospital course summarization task with real-world data.

According to Wang et al. [54], because the use of ontologies and terminologies for rare diseases in clinical data is limited, researchers often need to perform manual phenotyping or apply traditional NLP methods to identify patients with these diseases. To help address this gap, their project fine-tuned LLaMA2 LLMs using training sentences constructed from HPO and OMIM terms. Evaluation of a concept identification task compared standard ChatGPT (3.5) to the fine-tuned LLaMA2 models. ChatGPT performed poorly at this task, hallucinating some terms and concept identifiers that do not exist in the target ontologies.

4. Discussion

This survey focuses on 2022 and 2023 papers reporting KRM work in two areas of special relevance currently, namely 1) KRM techniques applied to the emerging long COVID crisis (seven papers), and 2) KRM techniques applied in combination with generative LLMs within the area of medical informatics (eight papers), along with supporting background and other relevant work. The work summarized in this survey shows the impressive versatility of KRM approaches both to improve our understanding of a global health crises and to evaluate and augment cutting edge technologies from other areas of AI.

Work to understand the etiology of long COVID is complicated by the complexity of the disease and its diverse manifestations.

The use of ontologies and other KRM approaches in this emerging area provides solutions for organizing, managing, understanding, and using complex information about long COVID, often in combination with complementary approaches from other areas (NLP, semantic clustering, genomics). The recent development and rapid proliferation of generative AI and LLMs are impacting all areas where processing or generation of natural language text is relevant, including widespread interest and applications in many areas of biomedicine. A major limitation in the design of these models is their tendency to confidently produce results without consideration for their accuracy or correctness. Integration of ontologies, knowledge bases, and other KRM approaches with LLMs looks promising as a solution for this problem, providing verifiable curated sources of truth that can improve trust and reliability of systems using these models for biomedical and healthcare applications. LLMs can also help with KRM approaches, by providing a natural interface for humans to interact with complicated tools and resources without requiring deep knowledge of ontologies or related technologies.

References

1. Verspoor K. The evolution of clinical knowledge during COVID-19: towards a global learning health system. Yearb Med Inform. 2021;30[01]:176–84. DOI: 10.1055/s-0041-1726503
2. Hastings J. Achieving Inclusivity by Design: Social and Contextual Information in Medical Knowledge. Yearb Med Inform. 2022 Aug;31(1):228–35. DOI: 10.1055/s-0042-1742509
3. He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Sci Data. 2020 Jun 12;7(1):181. DOI: 10.1038/s41597-020-0523-6
4. Raveendran AV, Jayadevan R, Sashidharan S. Long COVID: An overview. Diabetes Metab Syndr Clin Res Rev. 2021 May 1;15(3):869–75. DOI: 10.1016/j.dsx.2021.04.007
5. Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol. 2023 Mar 1;21(3):133–46. DOI: 10.1038/s41579-022-00846-2
6. Di Toro A, Bozzani A, Tavazzi G, Urtis M, Giuliani L, Pizzoccheri R, et al. Long COVID: long-term effects? Eur Heart J Suppl. 2021 Oct 1;23(Supplement_E):E1–5. DOI: 10.1093/eurheartj/suab080
7. O' Mahony L, Buwalda T, Blair M, Forde B, Lunjani N, Ambikan A, et al. Impact of Long COVID on health and quality of life. HRB Open Res. 2022;5:31. DOI: 10.12688/hrbopenres.13516.1
8. Faghy MA, Owen R, Thomas C, Yates J, Ferraro FV, Skipper L, et al. Is long COVID the next global health crisis? J Glob Health. 2022 Oct 26;12:03067. DOI: 10.7189/jogh.12.03067
9. Mirin AA. A preliminary estimate of the economic impact of long COVID in the United States. Fatigue Biomed Health Behav. 2022 Oct 2;10(4):190–9. DOI: 10.1080/21641846.2022.2124064
10. Rischard F, Altman N, Szmuszkovicz J, Sciurba F, Berman-Rosenzweig E, Lee S, et al. Long-Term Effects of COVID-19 on the Cardiopulmonary System in Adults and Children: Current Status and Questions to be Resolved by the National Institutes of Health Researching COVID to Enhance Recovery Initiative. Chest. 2024 Apr;165(4):978–89. DOI: 10.1016/j.chest.2023.12.030
11. Bonilla H, Peluso MJ, Rodgers K, Aberg JA, Patterson TF, Tamburro R, et al. Therapeutic trials for long COVID-19: A call to action from the interventions taskforce of the RECOVER initiative. Front Immunol. 2023 Mar 9;14:1129459. DOI: 10.3389/fimmu.2023.1129459
12. Astin R, Banerjee A, Baker MR, Dani M, Ford E, Hull JH, et al. Long COVID: mechanisms, risk factors and recovery. Exp Physiol. 2023;108(1):12–27. DOI: 10.1113/EP090802
13. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Comput Surv. 2023;56(2):1–40. DOI: 10.1145/3605943
14. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. ArXiv; 2023. DOI: 10.48550/arXiv.2303.08774
15. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023 Mar 30;388(13):1233–9. DOI: 10.1056/NEJMsr2214184
16. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. DOI: 10.3389/frai.2023.1169595
17. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4: a new era of artificial intelligence in medicine. Ir J Med Sci. 2023 Dec;192(6):3197–200. DOI: 10.1007/s11845-023-03377-8
18. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023 Jul 3;330(1):78–80. DOI: 10.1001/jama.2023.8288
19. Ito N, Kadomatsu S, Fujisawa M, Fukaguchi K, Ishizawa R, Kanda N, et al. The Accuracy and Potential Racial and Ethnic Biases of GPT-4 in the Diagnosis and Triage of Health Conditions: Evaluation Study. JMIR Med Educ. 2023 Nov 2;9:e47532. DOI: 10.2196/47532
20. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big?In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. p. 610–23. DOI: 10.1145/3442188.3445922
21. Zhang C, Zhang C, Li C, Qiao Y, Zheng S, Dam SK, et al. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. ArXiv; 2023. DOI: 10.48550/arXiv.2304.06488
22. Titus LM. Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. Cogn Syst Res. 2024 Jan 1;83:101174. DOI: 10.1016/j.cogsys.2023.101174
23. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. DOI: 10.1145/3571730
24. Arkoudas K. ChatGPT is no stochastic parrot. But it also claims that 1 is greater than 1. Philos Technol. 2023;36(3):54. DOI: 10.1007/s13347-023-00619-6
25. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3[Jan]:993–1022. [cited 2024 Jul 1]. Available from: https://dl.acm.org/doi/pdf/10.5555/944919.944937
26. Ambalavanan R, Snead RS, Marczika J, Kozinsky K, Aman E. Advancing the Management of Long COVID by Integrating into Health Informatics Domain: Current and Future Perspectives. Int J Environ Res Public Health. 2023 Sep 26;20(19). DOI: 10.3390/ijerph20196836
27. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021 Mar 1;28(3):427–43. DOI: 10.1093/jamia/ocaa196
28. N3C Dashboard - Home. [cited 2024 Feb 1]. Available from: https://covid.cd2h.org/dashboard/
29. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res. 2024 Jan 5;52(D1):D1333–46. DOI: 10.1093/nar/gkad1005
30. Deer RR, Rock MA, Vasilevsky N, Carmody L, Rando H, Anzalone AJ, et al. Characterizing Long COVID: Deep Phenotype of a Complex Condition. EBioMedicine. 2021 Dec;74:103722. DOI: 10.1016/j.ebiom.2021.103722
31. Wang L, Foer D, MacPhaul E, Lo YC, Bates DW, Zhou L. PASCLex: A comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. J Biomed Inform. 2022 Jan;125:103951. DOI: 10.1016/j.jbi.2021.103951
32. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to Process Medication Information in Outpatient Clinical Notes. AMIA Annu Symp Proc. 2011;2011:1639–48.
33. Lee Y, Riskedal E, Kalleberg KT, Istre M, Lind A, Lund-Johansen F, et al. EWAS of post-COVID-19 patients shows methylation differences

in the immune-response associated gene, IFI44L, three months after COVID-19 infection. Sci Rep. 2022 Jul 7;12(1):11478. DOI: 10.1038/s41598-022-15467-1
34. Lv Y, Zhang T, Cai J, Huang C, Zhan S, Liu J. Bioinformatics and systems biology approach to identify the pathogenetic link of Long COVID and Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. Front Immunol. 2022;13:952987. DOI: 10.3389/fimmu.2022.952987
35. Komaroff AL, Lipkin WI. ME/CFS and Long COVID share similar symptoms and biological abnormalities: road map to the literature. Front Med (Lausanne). 2023 Jun 2;10:1187163. DOI: 10.3389/fmed.2023.1187163
36. Tziastoudi M, Cholevas C, Stefanidis I, Theoharides TC. Genetics of COVID-19 and myalgic encephalomyelitis/chronic fatigue syndrome: a systematic review. Ann Clin Transl Neurol. 2022 Nov;9(11):1838–57. DOI: 10.1002/acn3.51631
37. Reese JT, Blau H, Casiraghi E, Bergquist T, Loomba JJ, Callahan TJ, et al. Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. EBioMedicine. 2023 Jan;87:104413. DOI: 10.1016/j.ebiom.2022.104413
38. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.
39. Callahan TJ, Wyrwa JM, Vasilevsky NA, Robinson PN, Haendel MA, Hunter LE, et al. OMOP2OBO: Semantic Integration of Standardized Clinical Terminologies to Power Translational Digital Medicine Across Health Systems. 2020. [cited 2024 Jul 1]. Available from: https://www.ohdsi.org/wp-content/uploads/2020/10/Tiffany-Callahan-Callahan_OMOP2OBO_2020_OHDSI_Symposium_Callahan_Poster.pdf
40. Denecke K, May R, Rivera Romero O. How Can Transformer Models Shape Future Healthcare: A Qualitative Study. Stud Health Technol Inform. 2023 Oct 20;309:43–7. DOI: 10.3233/SHTI230736
41. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May 4;224(1):iyad031. DOI: 10.1093/genetics/iyad031
42. Gene Ontology Resource. Gene Ontology Resource. [cited 2024 Jan 30]. Available from: http://geneontology.org/stats.html
43. Giri SJ, Ibtehaz N, Kihara D. GO2Sum: Generating Human Readable Functional Summary of Proteins from GO Terms. bioRxiv : the preprint server for biology. United States; 2023. p. 2023.11.10.566665. DOI: 10.1101/2023.11.10.566665
44. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J Mach Learn Res. 2020;21[140]:1–67. [cited 2024 Jul 1]. Available from: https://dl.acm.org/doi/pdf/10.5555/3455716.3455856
45. Hu M, Alkhairy S, Lee I, Pillich RT, Bachelder R, Ideker T, et al. Evaluation of large language models for discovery of gene set function. Res Sq [Preprint]. United States; 2023;rs.3.rs-3270331. DOI: 10.21203/rs.3.rs-3270331/v1
46. Joachimiak MP, Caufield JH, Harris NL, Kim H, Mungall CJ. Gene Set Summarization using Large Language Models. arXiv; 2023. DOI: 10.48550/arXiv.2305.13338
47. Munarko Y, Rampadarath A, Nickerson D. Building a search tool for compositely annotated entities using Transformer-based approach: Case study in Biosimulation Model Search Engine (BMSE). F1000Research. 2023 Feb 10;12:162. DOI: 10.12688/f1000research.128982.1
48. Munarko Y, Rampadarath A, Nickerson DP. CASBERT: BERT-based retrieval for compositely annotated biosimulation model entities. Front Bioinforma. 2023;3:1107467. DOI: 10.3389/fbinf.2023.1107467
49. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2018. DOI: 10.48550/arXiv.1810.04805
50. Tran H, Phan L, Anibal J, Nguyen BT, Nguyen TS. SPBERT: an Efficient Pre-training BERT on SPARQL Queries for Question Answering over Knowledge Graphs. In: Proceedings of the International Conference on Neural Information Processing. 2021. p. 512–23. DOI: 10.1007/978-3-030-92185-9_42
51. Reese JT, Danis D, Caulfied JH, Casiraghi E, Valentini G, Mungall CJ, et al. On the limitations of large language models in clinical diagnosis. medRxiv [Preprint]. 2024 Feb 26:2023.07.13.23292613. DOI: 10.1101/2023.07.13.23292613
52. Searle T, Ibrahim Z, Teo J, Dobson RJB. Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. J Biomed Inform. 2023 May;141:104358. DOI: 10.1016/j.jbi.2023.104358
53. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv; 2019. DOI: 10.48550/arXiv.1910.13461
54. Wang A, Liu C, Yang J, Weng C. Fine-tuning Large Language Models for Rare Disease Concept Normalization. J Am Med Inform Assoc. 2024 Jun 3:ocae133. DOI: 10.1093/jamia/ocae133

Copyright