Abstract

Background: The Unified Medical Language System (UMLS) has been a critical tool in biomedical and health informatics, and the year 2021 marks its 30th anniversary. The UMLS brings together many broadly used vocabularies and standards in the biomedical field to facilitate interoperability among different computer systems and applications.

Objective: Despite its longevity, there is no comprehensive publication analysis of the use of the UMLS. Thus, this review and analysis is conducted to provide an overview of the UMLS and its use in English-language peer-reviewed publications, with the objective of providing a comprehensive understanding of how the UMLS has been used in English-language peer-reviewed publications over the last 30 years.

Methods: PubMed, ACM Digital Library, and the Nursing & Allied Health Database were used to search for studies. The primary search strategy was as follows: UMLS was used as a Medical Subject Headings term or a keyword or appeared in the title or abstract. Only English-language publications were considered. The publications were screened first, then coded and categorized iteratively, following the grounded theory. The review process followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Results: A total of 943 publications were included in the final analysis. Moreover, 32 publications were categorized into 2 categories; hence the total number of publications before duplicates are removed is 975. After analysis and categorization of the publications, UMLS was found to be used in the following emerging themes or areas (the number of publications and their respective percentages are given in parentheses): natural language processing (230/975, 23.6%), information retrieval (125/975, 12.8%), terminology study (90/975, 9.2%), ontology and modeling (80/975, 8.2%), medical subdomains (76/975, 7.8%), other language studies (53/975, 5.4%), artificial intelligence tools and applications (46/975, 4.7%), patient care (35/975, 3.6%), data mining and knowledge discovery (25/975, 2.6%), medical education (20/975, 2.1%), degree-related theses (13/975, 1.3%), digital library (5/975, 0.5%), and the UMLS itself (150/975, 15.4%), as well as the UMLS for other purposes (27/975, 2.8%).

Conclusions: The UMLS has been used successfully in patient care, medical education, digital libraries, and software development, as originally planned, as well as in degree-related theses, the building of artificial intelligence tools, data mining and knowledge discovery, foundational work in methodology, and middle layers that may lead to advanced products. Natural language processing, the UMLS itself, and information retrieval are the 3 most common themes that emerged among the included publications. The results, although largely related to academia, demonstrate that UMLS achieves its intended uses successfully, in addition to achieving uses broadly beyond its original intentions.

(JMIR Med Inform 2021;9(8):e20675) doi: 10.2196/20675

KEYWORDS

Unified Medical Language System; systematic literature analysis; biomedical informatics; health informatics
Introduction

Background
The Unified Medical Language System (UMLS) [1] is a critical resource in biomedical and health informatics. It was created and released by the National Library of Medicine, an institute of the National Institutes of Health (NIH). The first edition of UMLS Knowledge Sources was distributed in 1991 [1], although its conceptualization can be traced to 1986 [2]. Currently, there are three UMLS Knowledge Sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. The Metathesaurus contains approximately 4.4 million concepts and 16 million unique concept names, which are from 218 source vocabularies in 25 languages worldwide (2021AA release). The Semantic Network provides consistent categorization for all concepts included in UMLS [3]. The SPECIALIST Lexicon and Lexical Tools provide large syntactic lexicon tools that have been used broadly in the biomedical and health fields to normalize strings and lexical variants.

UMLS brings together many broadly used vocabularies and standards in the biomedical field to facilitate interoperability and semantic understanding among different computer systems and software applications [4,5]. UMLS has been maintained and further developed by the National Library of Medicine over the past 30 years. In the initial publication, UMLS was intended to be used in four main areas: patient care, medical education, library service, and product development [1]. A comprehensive evaluation of the UMLS would be a large project; however, a close examination of the literature in the form of peer-reviewed publications can provide a perspective on how the UMLS is used in academia, which is the rationale for this literature review.

Objective
The year 2021 is the 30th anniversary of UMLS. Despite its longevity, there is no comprehensive publication analysis of UMLS. To call attention to the importance of UMLS and highlight its critical role in advancing biomedical informatics, health informatics, medicine, and health care, this systematic analysis was conducted to demonstrate how UMLS has been used, based on peer-reviewed publications in English over the past 30 years, which is the objective of this literature review.

Methods

Literature Search Sources and Strategies

Overview
PubMed, ACM Digital Library, and the Nursing & Allied Health Database were used for the search. The primary strategy was to search literature that either used UMLS as a MeSH (Medical Subject Headings) term or a keyword or had UMLS or unified medical language system in the title or abstract.

Search Strategy in PubMed on April 28, 2020
unified medical language system [MeSH term]

Search Strategy in ACM Digital Library on April 28, 2020
Searches were conducted within the ACM Guide to Computing Literature:

The following journals were excluded because they are indexed in PubMed: Journal of Biomedical Informatics, Artificial Intelligence in Medicine, and Bioinformatics.

Search Strategy in the Nursing & Allied Health Database on April 28, 2020
Searches were conducted within peer-reviewed publications:
mesh (unified medical language system) OR ti(umls) OR ti(unified medical language system) OR ab(umls) OR ab (unified medical language system)

Literature Examination

Literature Examination Process
The literature examination process followed the grounded theory. The steps for the content analysis were as follows: all duplicate publications were removed before the literature examination. The exclusion criteria included the following: UMLS not mentioned in the abstract, abstract unavailable, or non-English publications.

The first step of the content analysis was to go over and then code (or index) each title and to record the repeated themes or topics. The second step was to go over each abstract one by one to code (or index) each abstract again, record the repeated themes or topics, and exclude the irrelevant publications. The third step was to organize the themes and group them according to their similarities. Subsequently, each publication was classified into the corresponding theme, and additional themes were created during the process.

The classification step was conducted iteratively. The first round began with obvious and repeated themes. Each publication was examined and, as appropriate, categorized by theme. I began with the relatively obvious themes, each of which had relatively fewer publications. The initial group of themes included artificial intelligence (AI) tools and applications, other language UMLS studies, medical education, patient care, medical subdomains, digital library, and degree-related theses. The publications were then classified, one by one, for the following themes: UMLS itself, information retrieval, terminology study, natural language processing (NLP), ontology and modeling, data mining, and knowledge discovery. The publications that fell outside of these themes during the coding (or indexing) process were classified last. The classification process stopped when all publications were classified into themes without the need for additional consideration. The themes were adjusted whenever needed during the iterative classification processes. The publications were then analyzed, categorized, synthesized iteratively, counted, and recorded into each category.
A word cloud picture (Multimedia Appendix 1) based on the titles included in this comprehensive literature analysis was generated by removing all commonly used words. The Pro Word Cloud function within Microsoft Word (Microsoft Corporation) was used to generate the word cloud picture.

**Literature Classification Principles**

The following principles were followed during classification: the primary principle is that when a publication is analyzed, the objectives of the publication, not the methods implemented, are the prioritized reasons for the categorization. The secondary principle is to maximize the possibility that a publication will stand out among the publications in each category; that is, if a publication has an approximately equal possibility to be classified into 2 categories, the one with fewer publications wins. The third principle is to give publications on applications and patient care a higher priority than methodology development or foundational studies, in general. The fourth principle is to classify a publication into the most specific category whenever possible. The rationale for following these principles is based on the literature review. Instead of providing a comprehensive evaluation of all aspects of the UMLS, I attempted to determine how the UMLS is used in the real world. I focused on its application as a critical factor. As the UMLS is found in medicine, patient care is a higher priority.

In addition, I used this opportunity to recognize my peers’ contributions by maximizing the possibility of their publications standing out because only a small fraction of the work can be awarded a prize. These principles helped me to classify all the publications in a more consistent, clear, reproducible, and objective manner.

**Literature Review Guideline**

The systematic literature analysis protocol has not been registered. The data items used in this literature review including title, author, publication year, journal or conference proceeding, abstract, MeSH terms or keywords, PubMed ID (if available), full-text for some publications, and what was UMLS used for. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [6] were followed and most of the checklist items were included. The PRISMA checklist is provided in Multimedia Appendix 2.

**Results**

**Overview**

The search strategies yielded 1061 records in PubMed, 322 in the ACM Digital Library, and 60 in the Nursing & Allied Health Database. After removing the duplicates, records without abstracts, non-English records, and abstracts that did not mention UMLS, 943 records were retained for the final analysis. Figure 1 [6] shows detailed records of the literature search, screening, and analysis.

![Flowchart of the literature search, screening, analysis, and its records](image)

Multimedia Appendix 3 shows the yearly number of the included UMLS publications over the last 30 years. Table 1 presents the themes that emerged and the corresponding number of publications for each category. This table provides an overview of the results of the systematic analysis. Multimedia Appendix 4 presents the major themes, topics, and corresponding publication counts.
<table>
<thead>
<tr>
<th>Themes and subtopics</th>
<th>Publication counts (n=975), n (%)</th>
<th>After removing the duplicates (n=943), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Artificial intelligence tools and applications</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic annotation or interpretation</td>
<td>7 (15.2)</td>
<td>7 (15.2)</td>
</tr>
<tr>
<td>Automatic coding</td>
<td>7 (15.2)</td>
<td>7 (15.2)</td>
</tr>
<tr>
<td>Automatic summarization</td>
<td>15 (32.6)</td>
<td>15 (32.6)</td>
</tr>
<tr>
<td>Question-answering systems</td>
<td>10 (21.7)</td>
<td>10 (21.7)</td>
</tr>
<tr>
<td>Other intelligent tools</td>
<td>7 (15.2)</td>
<td>7 (15.2)</td>
</tr>
<tr>
<td>Data mining and knowledge discovery</td>
<td>25 (2.6)</td>
<td>25 (2.7)</td>
</tr>
<tr>
<td>Degree-related theses</td>
<td>13 (1.3)</td>
<td>8 (0.8)</td>
</tr>
<tr>
<td>Digital library</td>
<td>5 (0.5)</td>
<td>4 (0.4)</td>
</tr>
<tr>
<td><strong>Information retrieval</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image retrieval</td>
<td>20 (16)</td>
<td>20 (16.8)</td>
</tr>
<tr>
<td>Indexing</td>
<td>33 (26.4)</td>
<td>30 (25.2)</td>
</tr>
<tr>
<td>Information retrieval</td>
<td>34 (27.2)</td>
<td>32 (26.9)</td>
</tr>
<tr>
<td>Information retrieval system and search engine</td>
<td>8 (6.4)</td>
<td>8 (6.7)</td>
</tr>
<tr>
<td>Performance</td>
<td>13 (10.4)</td>
<td>12 (10.1)</td>
</tr>
<tr>
<td>Query</td>
<td>17 (13.6)</td>
<td>17 (14.3)</td>
</tr>
<tr>
<td>Medical education</td>
<td>20 (2.1)</td>
<td>19 (2)</td>
</tr>
<tr>
<td>Medical subdomains (34 subdomains)</td>
<td>76 (7.8)</td>
<td>76 (8.1)</td>
</tr>
<tr>
<td><strong>NLP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbreviation</td>
<td>11 (4.8)</td>
<td>11 (4.8)</td>
</tr>
<tr>
<td>Feature identification or extraction or phenotyping</td>
<td>4 (1.7)</td>
<td>4 (1.7)</td>
</tr>
<tr>
<td>Lexicon and/or inventory</td>
<td>7 (3)</td>
<td>7 (3)</td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept recognition or extraction</td>
<td>42 (25.5)</td>
<td>42 (25.5)</td>
</tr>
<tr>
<td>Name entity recognition or extraction</td>
<td>18 (10.9)</td>
<td>18 (10.9)</td>
</tr>
<tr>
<td>Natural language, vocabulary, question generation</td>
<td>3 (1.8)</td>
<td>3 (1.8)</td>
</tr>
<tr>
<td>Natural language understanding</td>
<td>3 (1.8)</td>
<td>3 (1.8)</td>
</tr>
<tr>
<td>Relationship recognition or extraction</td>
<td>45 (27.3)</td>
<td>45 (27.3)</td>
</tr>
<tr>
<td>Semantic similarity, relatedness, or distance</td>
<td>20 (12.1)</td>
<td>20 (12.1)</td>
</tr>
<tr>
<td>Word sense disambiguation</td>
<td>34 (20.6)</td>
<td>34 (20.6)</td>
</tr>
<tr>
<td><strong>Syntax</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td>18 (7.8)</td>
<td>18 (7.8)</td>
</tr>
<tr>
<td>Tagging</td>
<td>5 (27.8)</td>
<td>5 (27.8)</td>
</tr>
<tr>
<td>Terminology extraction</td>
<td>5 (27.8)</td>
<td>5 (27.8)</td>
</tr>
<tr>
<td>Text classification</td>
<td>8 (44.4)</td>
<td>8 (44.4)</td>
</tr>
<tr>
<td>Other NLP-related publications</td>
<td>15 (6.5)</td>
<td>15 (6.5)</td>
</tr>
<tr>
<td><strong>Ontology and modeling</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification or taxonomy</td>
<td>21 (26.3)</td>
<td>21 (26.6)</td>
</tr>
<tr>
<td>Modeling</td>
<td>18 (22.5)</td>
<td>17 (21.5)</td>
</tr>
<tr>
<td>Ontology</td>
<td>29 (36.3)</td>
<td>29 (36.7)</td>
</tr>
<tr>
<td>Representation</td>
<td>12 (15)</td>
<td>12 (15.2)</td>
</tr>
</tbody>
</table>
Themes, Subtopics, and Publications Under Each Category

After the included publications were examined carefully, the following themes emerged during analysis and synthesis.

**UMLS Is Used in AI Tools and Applications**

The UMLS has been used in developing AI tools and applications since 1994 [7] (publication; the actual work started many years ago). The AI tools include question-answering systems, automatic summarization, automatic coding, automatic annotation, and plagiarism detection. Question-answering systems focus on the medical domain. Some question-answering systems focus specifically on answering consumers’ questions. Automatic summarization focuses mainly on summarizing medical literature, textbooks, and patient records. This category also includes methodology exploration. Multimedia Appendix 5 includes the 46 UMLS publications in this category.

I recognize that there is an overlap between AI tools and NLP. The criterion used concerned whether a publication focused on the final products. If so, it was classified into the AI tools and applications category; if a publication focused on the middle-layer methodology to enhance performance, it was classified into the NLP category.

Automatic translation can also be categorized into this theme; however, the publications were categorized on automatic translation into the other language UMLS studies category, using a more detailed description. Similarly, intelligent tutoring systems were classified into medical education instead of AI tools and applications. These categories should be cross-referenced accordingly.

**UMLS-Based Data Mining and Knowledge Discovery**

UMLS is used broadly as a critical tool in data mining and knowledge discovery in the biomedical field. However, there are large overlaps between this category and the subcategory under NLP, namely, relationship extraction. The following categorization criteria were implemented: if a publication could be dissected into a relationship (eg, drug-drug interaction, condition-treatment, and association rule mining) extraction, identification, or discovery, the publication was classified under the relationship extraction subcategory of NLP; otherwise, the publication was included in the data mining and knowledge discovery category. Multimedia Appendix 6 lists all 25 included UMLS publications related to data mining, knowledge discovery, data analysis, and text analysis.

**UMLS in Degree-Related Theses**

Notably, there are 13 doctoral theses [8-20] included from the ACM Digital Library that used the UMLS as a key component in conducting the research. I believe that it is very likely that there is greater use of the UMLS in doctoral or master theses that might not be captured through the title, abstract, or keywords. My own doctoral thesis used UMLS as a critical foundational tool to build a knowledge base; however, UMLS was not listed as a keyword.
UMLS for Digital Libraries

A digital library is another initial goal of UMLS. In this systematic literature analysis, 5 publications related to the UMLS and a digital library were identified. Of these, one publication used machine learning to process information extracted from a digital library, in which UMLS served as an information source [21]. In terms of a digital library, UMLS is also used for navigation purposes [22], for the semantic query [23], to improve the functions of the digital library [24], and to extract knowledge from a digital library [25]. There could be additional publications on the topic that do not necessarily use digital library as the key term.

UMLS in Information Retrieval

Since its inception, UMLS has been used to achieve and improve information retrieval. A total of 125 publications were identified in this theme, which is the third most active theme in this review. The subtopics of this emerging theme include image retrieval (eg, radiological images, pathological images, microscopic images, computed tomography scans, and electrocardiograms), indexing, information retrieval (including information needs), information retrieval systems, and search engines (eg, PubMed, MEDLINE, electronic health record systems, books, databases of texts, images, and sounds), performance or correct measures (including ranking), and query (from generic queries, query formulation, query expansion, and more accurate queries to evaluations). The information sources for retrieval purposes included documents, information within documents, metadata, scientific literature, and patient records. Multimedia Appendix 7 lists all 125 UMLS publications related to information retrieval.

UMLS in Medical Education

UMLS was planned for use in medical education [1,26-29]. Most of the publications in this category included curriculum mapping [30], continuing education [31,32], problem-based learning [33], tutoring systems [33-41], and educational resource development [31,32,42-44].

UMLS in Different Medical Subdomains

As the most comprehensive collection of medical terminologies, UMLS has been used in 34 medical subdomains in a variety of ways. The subdomains in which UMLS has been used include Alzheimer disease [45,46], anatomical structure [47-64], appendectomy [65], asthma [66,67], blood transfusion [68,69], breast biopsy [65], breast cancer [70,71], cardiovascular diseases [72-74], colorectal cancer [75,76], depression [77,78], dilated cardiomyopathies [79], epidemiology [80,81], falling injury risk assessment [82], HIV [83], hypertension [84-86], Kawasaki disease [87], liver cancer [88], liver diseases [89,90], lupus [91], neuropsychiatric disorders [92-94], occupational medicine [95,96], oncology [97,98], Parkinson disease [99], pneumonia [100], physical therapy [101], primary care [102-104], prostate cancer [105,106], rare diseases [107-111], respiratory tract infection [112], stroke thrombolysis [113], surveillance [114-116], traditional Chinese medicine [117,118], urology [119,120], and Zika virus [121]. There are significantly more publications about anatomy than about any other medical subdomain.

UMLS in NLP

UMLS is used as a critical component in NLP, the most active theme in the review, with 230 publications identified. The specific use of UMLS in this category includes abbreviation-related studies, feature identification, lexicon and inventory, semantic-related studies, syntax-related studies, text classification, and other NLP-related UMLS publications.

Semantic-related publications (165/230, 71.7%) included concept recognition and extraction, named entity recognition, natural language, vocabulary, question generation, natural language understanding, relationship recognition and extraction, semantic similarity or relevance or distance, and word sense disambiguation. Named entity recognition also included negation recognition. For concept recognition or extraction, the following groups were included: adverse drug event identification, contextual property identification, disorder recognition, and identification of treatment information. Relationship recognition and extraction included association recognition, medication-indication relationships, drug-drug interaction, and disease-manifestation relationships.

Syntax-related publications (18/230, 7.8%) included part-of-speech tagging, parsing, and terminology extraction. Other NLP-related publications (47/230, 20.4%) included rule-based NLP, statistical NLP, corpus development, morphological similarity, word embedding, and stemming.

The source document types used in NLP are very rich and include discharge summaries, problem lists, clinical trial eligibility criteria, clinical trial protocols, clinical narrative notes, patient records, radiology reports, neuroradiology reports, pathology reports, histology reports, emergency department reports, surgical operative reports, medical progress notes, literature, social media, emails, and forum posts. Multimedia Appendix 8 presents a list of all 230 publications classified into the NLP category.

UMLS-Based Ontology and Modeling-Related Publications

UMLS is also a common tool used in ontology, classification, taxonomy, modeling, knowledge representation, and their associated studies. Although UMLS and terminology study are 2 existing categories, there are still some publications that cannot be categorized into either of these categories. If a publication can be included in a more specific subcategory, for example, a model of an information retrieval system, then it will be classified into the information retrieval system and search engine subcategory instead of the modeling subcategory. In this category, the publications were classified into corresponding subcategories only if the publication could not be included in any other category. Multimedia Appendix 9 presents a list of all 80 publications in this category.

UMLS English-Language Publications About Non-English Languages

There are efforts related to using UMLS in languages other than English, as well as multilingual studies. In this category, 10 additional languages and 53 publications were identified. Some publications are related to automatic translation, whereas others
are related to the coverage of an additional language of medical terms in addition to English. Languages other than English that relate to multilingual or cross-language uses of UMLS include Bulgarian [122], Dutch [123], French [63,124-145], German [76,146-149], Italian [150], Japanese [151-153], Korean [154-158], Portuguese [159], Spanish [160-162], and Swedish [146,163]. A total of 12 publications included more than two languages [146,159,163-172]. Clearly, there are more French-related UMLS publications than any other non-English language.

**UMLS in Patient Care**

One of the original goals of UMLS is to facilitate patient care directly or indirectly. As planned, UMLS has been used in patient care in many different ways, including the prediction of bariatric surgery outcomes, ensuring patient safety, development of a fall injury risk assessment instrument, patient outcome measurement, functional status measurement, clinical care quality assurance, computerized physician order entry, and clinical decision support systems. Multimedia Appendix 10 presents a list of all 35 publications in this category.

**UMLS for Terminology Studies**

As a critical tool, UMLS is used to conduct terminology studies. A total of 90 publications were classified into this category. The scope of the work includes a comparison of terminologies, construction of terminology, harmonization, interoperability, quality assurance, and other UMLS publications of terminology. UMLS is critical for achieving and advancing interoperability. The publications about the UMLS itself were classified into the UMLS category instead of under terminology studies. The roles of the UMLS in terminology studies include data sharing, aggregating data, harmonizing (including mapping among different terminologies), and vocabulary foundation. The publications on lexical mapping were classified into NLP. Multimedia Appendix 11 presents a list of all 90 UMLS publications on terminology studies.

**Studies About the UMLS Itself**

A total of 150 publications about the UMLS itself, which is the second most active theme after NLP, were identified. The scope of the publications ranged from auditing and enhancement of UMLS to the development of its own components, including MetaMap, SPECIALIST Lexicon and Lexical Tools, and Semantic Network, as well as its application tools MetaMap, MMTx, and SemRep. Furthermore, many efforts were related to increasing the coverage of UMLS in different subdomains, for example, in nursing, radiology, genetic disease, anatomy, and herbal supplements. In this category, the subtopics included applications or tools for UMLS, auditing of UMLS, components of UMLS, and coverage of UMLS. All studies in this category used UMLS as the study object. For example, auditing of UMLS includes publications on auditing-related studies that focus on the auditing of UMLS. If UMLS was used for other auditing purposes in a publication, then the publication was classified into UMLS in the other purposes category.

This category of publications also included modeling in UMLS. Other modeling-related publications that used UMLS were classified into the ontology and modeling categories. The publications that used UMLS to achieve different objectives (eg, identification of associations in texts) were classified into other categories based on their corresponding objectives. Multimedia Appendix 12 presents a list of all 150 UMLS publications on studies of the UMLS itself.

**UMLS for Other Purposes**

This category is used mainly for publications that use UMLS to achieve other purposes that cannot be covered by the themes noted above. In this category, auditing (not for UMLS auditing), consumer health, integrated system or data, and other research uses (including profile construction, management use, and deidentification) were included. Multimedia Appendix 13 presents a list of all 27 publications in this category.

**Discussion**

**Summary and Interpretation of the Results**

The results of the literature analysis showed the broad scope of the impact of UMLS in the academic world in the form of peer-reviewed journal publications, peer-reviewed conference publications, book chapters, and degree theses. What has been captured here, however, is only a small fraction of the real impact of UMLS. This literature analysis does not capture the following possible uses or impacts if no paper was published or if UMLS was not included in the title, abstract, or keywords: use of UMLS in the health information technology industry, health care delivery, software development, and any patent-related output.

The results show that UMLS has been broadly used, from basic science to applied projects in biomedical and health informatics. From the perspective of the number of publications, NLP, UMLS itself, and information retrieval were the 3 themes with the most publications. Anatomy is the medical subdomain with the most publications. French is the most active language, with a higher number of UMLS English-language publications of non-English languages. The large number of publications shows that certain themes are very active, although this literature analysis does not examine the overlap in different themes among different research projects. In addition, the number of publications should be used in a relative sense and with caution because a special issue of a journal or focused workshops or contests can skew the number of publications significantly.

In the Results section, the themes that repeatedly emerged during the literature analysis and synthesis have been listed. However, this is only an observation and a recording. From a purely ontological perspective, the same publications can be classified into different categories, depending on the axis. For example, a publication that focuses on automatic translation can be included in AI tools or applications; it can also be included in the multilingual category. Ideally, it will be useful to cross-reference each publication, which can then be classified into different categories. However, because of the large number of publications included in this literature analysis, such publications have been listed in only one category mostly (only 32/943, 3.4% of the publications was categorized into 2 categories; Table 1) instead of all possible categories. It is recognized that what was provided in this review is a snapshot
of the publications at the gross anatomy level, not a panoramic view of the publications with every single detail at the molecular level. This literature analysis serves as an archive of English-language UMLS peer-reviewed publications. The themes and subtopics and the publications under each theme or subtopic show only one perspective, not the only perspective, on the publications and their organizations. It is recognized that the search strategies can find only those publications for which UMLS plays a critical role. Some additional publications may use UMLS in their work; however, if UMLS was not listed in the title, MeSH terms, or abstract, then these publications will not be found through the search strategies. Therefore, the real impact of UMLS, even as academic output, is far larger than this review can represent.

**Comparison With Existing UMLS-Use Publications**

No systematic review or comprehensive literature analysis of UMLS was found during the literature search; however, there are publications on the use of UMLS through an analysis of UMLS annual reports [2] and the collection of surveys of UMLS users [173]. Nevertheless, the content of this literature analysis is complementary to these 2 studies [2,173]. The study by Fung et al [2] reported the geographical distribution of the users, the organizations of the UMLS license holders, types of information processed by UMLS, and areas of use of UMLS as well as users’ support, communications, and feedback. The study [2] drew conclusions from 1427 UMLS annual reports for the year 2004. Chen et al [173] reported the results of a 26-item survey sent to those on a UMLS mailing list (>600 subscribers). The research team analyzed the responses from 70 respondents, provided detailed categories of the users’ employment and areas of use, and concluded that the top uses of UMLS were to access the source terminologies through UMLS and to achieve mapping among these terminologies. In addition, terminology research, information retrieval, terminology translation, UMLS research, and NLP, as well as UMLS auditing, were identified as the categories for the use of UMLS and as future priorities [173]. By comparison, this literature analysis paints a more comprehensive picture of publications in the last 30 years with regard to UMLS, by UMLS, and with UMLS. In analog language, this literature analysis is still at the level of gross anatomy; however, this review does provide more comprehensive categories, more detailed classifications, and clusters of publications on the topic. This literature analysis also lists degree-related doctoral theses in which the UMLS plays a critical role.

**About UMLS**

The original intended uses of UMLS involved four main areas: patient care, medical education, library service, and product development [1]. Comparing the results of this literature analysis with the originally intended uses, it is concluded that, although the literature analysis reflects an output largely within academic settings, the original intended uses have been achieved successfully. There are multiple themes and subtopics that can be matched to each of the 4 areas. For example, the patient care and medical subdomains can be placed in the patient care category. It was, however, recognized that such a literature analysis is not the best way to capture all the uses of UMLS in the real world, especially with regard to product development. Nevertheless, it is acknowledged that many electronic health records, AI, and NLP applications in the health field commonly use UMLS [5].

UMLS has been a cornerstone of academic activities in biomedical informatics, health informatics, and health information technology as a way to facilitate interoperability in broad medical and health fields. This literature analysis demonstrates only a small fraction of the true impact of UMLS. UMLS can be used as a terminology hub that hosts the most commonly used biomedical and health terminologies worldwide by using a universal concept unique identifier. A terminology hub is different from terminology in the same way that SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) and UMLS are different but, at the same time, have some similarities. The 2 resources overlap but have mainly complementary purposes in the biomedical and health fields. SNOMED-CT is the most comprehensive medical terminology in the world, and UMLS includes SNOMED-CT and many additional terminologies. A common use of the UMLS is to provide machine-processable codes and meanings, which is similar to the use of SNOMED-CT; UMLS also provides mapping among different source terminologies. UMLS is critical for processing historical data and heterogeneous data sources, which will be a reality in health care in the near future. Therefore, to achieve seamless and effortless interoperability with a finer level of granularity in health care delivery sufficient to completely solve the puzzle described in e-patient Dave case study [174], at least at the front end, we need both SNOMED-CT and UMLS as well as many other resources.

However, UMLS is beyond a terminology hub. The intended uses of UMLS are mainly through software programs or systems. Many listed applications of UMLS include linking terms and codes in practice, pharmacy, and laboratory; facilitating mapping among different terminologies by providing terminology services; and serving as a lexical tool for NLP and AI, among others. Many additional UMLS applications have never been captured in the form of peer-reviewed publications. For example, my colleagues and I use UMLS as a teaching tool to introduce the concept of using controlled vocabularies to code medical records for health science major undergraduates.

**Future Work**

This literature analysis provides a descriptive observation of English-language peer-reviewed publications on UMLS over the last 30 years. It is an overview of the publications in terms of scope, as well as major themes and subtopics. More detailed content and literature analysis can be conducted for each theme. In this study, most of the publications were examined through an analysis of titles and abstracts, with some full-text publications when necessary. A more detailed full-text publication analysis may provide a more in-depth understanding of this topic.

Another possible direction is to examine the overlap among different themes and subtopics. For example, future research could analyze the overlaps by classifying a publication into as many categories as possible. If a publication has only 1 position within 1 theme or one subtopic, a theme graph can be generated.
with all themes and subtopics (a graphical representation of Table 1) and all publications within each theme and each subtopic. Each publication would then have multiple positions in the theme graph. A visualization to consider the aggregated overlap (the same publication with multiple positions among multiple subtopics) among themes and subtopics can show or even inspire possible research collaboration opportunities among themes and subtopics.

Conclusions
This comprehensive literature analysis provides an overview with systematic evidence of the UMLS English-language peer-reviewed publications in the last 30 years. The analysis provides a descriptive observation of the themes and their subtopics of the publications and provides a detailed list of the publications in each category. UMLS has been used and published successfully in patient care, medical education, digital libraries, and software development in biomedicine, as well as in degree-related theses, building AI tools, data mining and knowledge discovery, and many more foundational works in methodology and middle layers that may lead to advanced products. The results, although largely in academia, demonstrate that UMLS achieves its intended uses successfully and has been used successfully and broadly beyond its original intentions. NLP, UMLS itself, and information retrieval are the three themes with the most publications. Anatomy is the most active medical subdomain. French is the most active language among the UMLS English-language publications of non-English languages. Nevertheless, this systematic literature analysis only captures publications in the English language; therefore, it should not be treated as a comprehensive impact description of UMLS, which should include English-language peer-reviewed publications and much more (eg, other language publications, patents, software, apps, care quality, and patient safety).

Acknowledgments
This study was partially supported by the National Library of Medicine of the NIH under award number R15LM012941 and partially supported by the National Institute of General Medical Sciences of the NIH under award numbers P20 GM121342 and R01GM138589. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Word cloud for all titles included in this systematic literature analysis (publication counts).
[PNG File, 335 KB-Multimedia Appendix 1]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for Scoping Reviews checklist.
[PDF File (Adobe PDF File), 543 KB-Multimedia Appendix 2]

Multimedia Appendix 3
The yearly number of publications included in the literature analysis from 1990 to 2020.
[PNG File, 32 KB-Multimedia Appendix 3]

Multimedia Appendix 4
Themes and major topics of applications of the Unified Medical Language System.
[PNG File, 58 KB-Multimedia Appendix 4]

Multimedia Appendix 5
Unified Medical Language System is used for building artificial intelligence applications and tools.
[PDF File (Adobe PDF File), 246 KB-Multimedia Appendix 5]

Multimedia Appendix 6
Publications on data mining, knowledge discovery, and text analysis using the Unified Medical Language System.
[PDF File (Adobe PDF File), 208 KB-Multimedia Appendix 6]

Multimedia Appendix 7
Unified Medical Language System publications related to information retrieval.
Multimedia Appendix 8
Unified Medical Language System in publications related to natural language processing.

Multimedia Appendix 9
Unified Medical Language System publications in ontology and modeling.

Multimedia Appendix 10
Publications that used the Unified Medical Language System in patient care.

Multimedia Appendix 11
Unified Medical Language System publications about terminology studies.

Multimedia Appendix 12
Publications about studies of the Unified Medical Language System itself.

Multimedia Appendix 13
Unified Medical Language System publications for purposes other than the themes noted above.

References

2. Fung KW, Hole WT, Srinivasan S. Who is using the UMLS and how - insights from the UMLS user annual reports. AMIA Annu Symp Proc 2006:274-278 [FREE Full text] [Medline: 17238346]
11. McInnes BT. Supervised and Knowledge-Based Methods for Disambiguating Terms in Biomedical Text Using the UMLS and Metamap. Minneapolis, Minnesota: University of Minnesota; 2009.


64. Lowe HJ, Huang Y, Regula DP. Using a statistical natural language Parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports. AMIA Annu Symp Proc 2009 Nov 14;2009:386-390. [Medline: 20351885]


78. Kossman S, Jones J, Brennan PF. Tailoring online information retrieval to user’s needs based on a logical semantic approach to natural language processing and UMLS mapping. AMIA Annu Symp Proc 2007 Oct 11:1015. [Medline: 18694113]


102. Westberg EE, Miller RA. The basis for using the internet to support the information needs of primary care. J Am Med Inform Assoc 1999;6(6):6-25 [FREE Full text] [Medline: 9925225]


Abbreviations

AI: artificial intelligence
MeSH: Medical Subject Headings
NIH: National Institutes of Health
NLP: natural language processing
SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms
UMLS: Unified Medical Language System

Please cite as:

Jing X
The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis
JMIR Med Inform 2021;9(8):e20675
URL: https://medinform.jmir.org/2021/8/e20675
doi: 10.2196/20675
PMID: 34236337

Edited by C Loui; submitted 25.05.20; peer-reviewed by A Wang, J Varghese, M Dugas, M Torii; comments to author 28.10.20; revised version received 25.11.20; accepted 02.07.21; published 27.08.21

https://medinform.jmir.org/2021/8/e20675 JMIR Med Inform 2021 | vol. 9 | iss. 8 | e20675 | p. 17
(page number not for citation purposes)