http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

基于文献知识发现的传染病继发疾病挖掘分析研究

李晓瑛 门佩璇 李永洁 黄雅兰 孙轶楠 唐小利
中国医学科学院医学信息研究所

摘　要：目的/意义 传染病的早期预警受到国内外相关机构的有效监管，然其继发风险尚未引起广泛关注；通过挖掘公开发表的文献大数据揭示传染病相关的继发知识，助推继发感染监测平台研建，提升我国感染性疾病预警防控水平。方法/过程 首次利用基于文献的知识发现方法，以病毒性肝炎、HIV 感染和结核感染三种常见的传染病为例，从大规模生物医学文献数据中挖掘分析其继发疾病。结果/结论 从超过 3680 万篇 PubMed 文献中发掘出三种传染病的继发疾病，包括感染性疾病、非感染性疾病以及肿瘤。结论：研究结果不仅验证了该方法对系统全面地揭示传染病相关的继发知识提供一种新思路，而且为其继发疾病的预警监测和尽早干预提供更加有力的文献证据。

关键词：知识发现；传染病；继发感染

分类号：G254

Research on the Secondary Incidence Analysis of Infectious Diseases Based on Literature Knowledge Discovery

LI XiaoYing, MEN Peixuan, LI Yongjie, HUANG Yalan, SUN Yinan, TANG XiaoLi

Institute of Medical Information, Chinese Academy of Medical Sciences

Abstract： Purpose/Significance The early warning of infectious diseases has been effectively regulated by international and domestic specialized organizations, but there is limited significant attention paid to its secondary risk. This paper aims to reveal secondary knowledge related to infectious diseases by mining public literature data, and promote the research and construction of secondary infection monitoring platform, which will improve the prevention and control level of infectious diseases in China. Method/Process The literature based discovery method is firstly adopted to mine and analyze the secondary diseases from large-scale biomedical literature data, taking viral hepatitis, HIV infection and tuberculosis infection as the examples. Result/Conclusion Three kinds of secondary diseases including infectious diseases, non-infectious diseases and even tumors, were found from over PubMed citations. Conclusion：The research results not only validate that this method provides a new approach for systematically and comprehensively reveal secondary knowledge related to infectious diseases, but also provide more powerful literature evidences for effective monitoring and early intervention of secondary diseases.

Keywords： Knowledge Discovery； Infectious Diseases； Secondary Infection

CLC Code：G254

1.引言

[基金项目]国家重点研发计划项目“重大传染病重症患者继发感染多维度监测与预警系统”（2023YFC23088）；
[通讯作者]唐小利，tangxl@imicams.ac.cn。

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

由各种病原体引起的能够在人与人之间相互传播的传染病，给人民生命健康和经济社会发展造成巨大影响。据世界卫生组织统计，2022 年全球乙肝新发病例 110.5 万[1]，慢性乙肝和慢性丙肝导致的死亡人数达 49.6 万和 17.7 万；结核与 HIV 感染的病死率更高，2022 年全球因结核病死亡的人数（不含艾滋病）高达 111.2 万，全世界有 49.7 万死于 HIV 感染（含艾滋病）。

长期以来，国内外疾病预防控制相关机构高度重视新发、原发传染病监测预警工作，在新发、原发传染病防控方面做出了重要贡献。世界卫生组织建立了 GISRS 全球流感监测和响应系统[2]，通过与各国家、区域的哨点机构合作，实现了全球季节性和大流行性传染病的监测防范、积极应对和全球警报。美国国防部开发的 GEIS 全球新兴感染监测系统是一个基于大数据的全球公共卫生、传染病研究和疾病监测网络[3]，已覆盖全球 72 个国家，提供了全球传染病威胁信息。我国已建成全球规模最大的国家级传染病网络直报系统，并打通了直报系统与医院信息化系统数据交互壁垒[4]，平均报告时间从 5 天缩短到了 4 小时，提升了防控科学性和精准性。然而，关于传染病继发感染等疾病的监测预警仍有待加强。

已感染传染病的患者在治疗过程中一旦出现其他疾病（包括微生物感染），将极大增加其康复难度和病死率[5]，而科学准确的早期预警成为降低传染病继发风险和改善预后的有效手段。全面了解传染病可能继发的疾病，及早预测传染病继发风险、更准确地识别继发高风险患者，是降低病死率的关键突破口和重大科学问题之一。

基于文献的知识发现（Literature-Based Discovery, LBD）是一种著名的情报学方法[6]，利用自然语言处理和机器学习等技术，从公开发表的科技文献中挖掘领域知识及知识之间的隐性关联，预期发现潜在的“新”知识[7]。该方法最早于 1986 年由美国芝加哥大学 Swanson 等人提出[8]，先后从文献中发现了鱼油与雷诺氏病、镁与偏头痛、生长调节素 C 与精氨酸等生物医学知识并相继得到证实，从而在国内外得到广泛关注。现今，以 LBD 方法为代表的数据驱动的领域知识发现，已成为信息资源管理、医学信息学等交叉学科的重要研究内容。本研究首次将情报学中经典的 LBD 方法应用于传染病继发预警监测领域，从大规模生物医学文献数据中系统全面地挖掘分析传染病继发性相关知识，为继发疾病的预警监测和早期干预提供更加有力的文献证据。

2. 国内外相关研究综述

2.1 基于文献的知识发现研究综述

通过文献调研可知，国内外关于 LBD 方法的相关研究大致经历了三个阶段。第一个阶段聚焦关联文献的知识发现，主要利用基于单词或短语的词频统计法，从多篇相关文献发现潜在的关联信息，由此诞生了 ABC 知识发现模型。第二个

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

阶段融入了语义过滤和专家智慧，采用语义聚类强化潜在知识筛选，并引入文献作者对结果进行审查。第三个阶段整合自然语言处理和人工智能技术，从大规模的科技文献中自动高效地完成领域知识发现，从而迈向数智驱动的知识发现新阶段。期间，产生了基于长短时记忆网络的药物-疾病知识发现模型[9]、双向长短时记忆网络的药物-药物知识发现模型[10]、基于深度学习的药物-疾病-基因深层次关联发现模型[11]，以及基于词表与规则的 Semantic Knowledge Representation (SemRep)语义知识抽取工具等成果[12,13]。这些研究推动了 LBD 方法在生物医学、智慧农业等领域的广泛应用[6]，提升了科技文献的学科知识服务能力。

2.2 传染病继发性相关研究综述

目前，国内外关于传染病继发性的相关研究主要集中在病原学分析、临床表现鉴别、有效治疗等方面。Nagao、Tabriz、Mitrokhin 等研究呈现了肺结核患者继发感染的病原学分析[14,15]和耐药性监测[16]结果，林明贵[17]、郭新美[18]、谭守勇[19]等对肺结核继发肺部真菌感染的临床特点和诊治进展等进行了阐释，Ivushkina 等探讨了部分机会性菌群在肺结核患者继发感染中发挥的作用，另有大量的病例报告报道了肺结核继发脓气胸[20]、气管食管瘘[21]、肺动脉高压[22]等病例。乙型肝炎、HIV 感染等其它传染病的继发性研究现状大体类似，尚缺少对于传染病继发疾病的系统性研究。

3. 资料与方法

3.1 研究资料

本研究所挖掘分析的文献数据来自美国国立医学图书馆研发的 PubMed 生物医学文献数据库，共有自建库至 2024 年 2 月 8 日收录的 36,840,441 篇文献（因其为题目数据库仅有文章摘要，不含全文）。此外，将文献中揭示的传染病能够继发的疾病划分为感染性和非感染性（包括肿瘤），划分的依据为国际上具有高影响力的《医学主题词表》（Medical Subject Headings, MeSH）。该词表从 16 个维度组织医学知识，包括疾病、解剖结构、有机体、化学物质和药物、诊疗技术及设备等（图 1 左）。其中，疾病又包括感染、肿瘤、消化系统疾病、呼吸道疾病、心血管疾病、免疫系统疾病等 23 个分支（图 1 右）。本文参考 MeSH 词表中的疾病组织方式，将感染（MeSH 编号为 C01）及其下位疾病归为感染性疾病，而将其它 22 个分支的疾病划分为非感染性疾病。

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

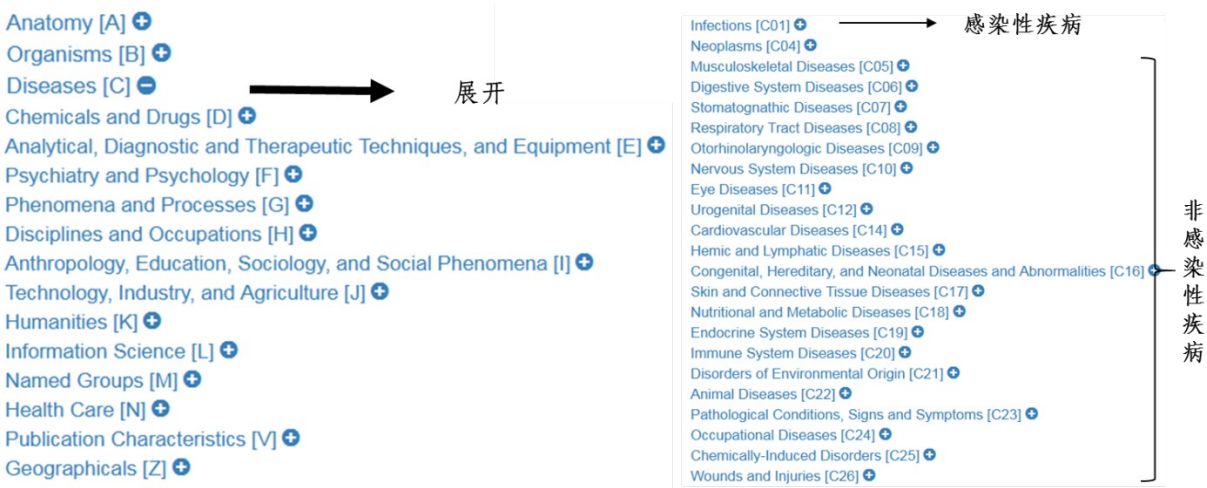

图 1.MeSH 词表结构及疾病组织方式

3.2 研究方法

依据基于文献的知识发现方法相关原理，利用 SemRep 工具完成传染病继发性相关的知识发现。该工具是由美国国立医学图书馆研发的一个知识发现工具，基于一体化医学语言系统（Unified Medical Language System, UMLS）、自然语言处理的句法分析和语义规则，自动发现生物医学文本中的语义知识（共 58 种）[23]。UMLS 词表在 SemRep 知识发现中发挥了重要的作用，包括同义异形词归并、语义类型组织等。如下表 1 所示，UMLS 词表辅助 SemRep 工具将“hepatitis B viral (HBV) infection”、“hepatitis B virus (HBV) infection”统一归并为“Hepatitis B”，实现同义词的概念汇聚；并且为“Hepatitis B”、“Chronic hepatitis”分配语义类型“Disease or Syndrome”，有助于本研究自动筛选出传染病与其继发疾病之间的因果关联知识。此外，SemRep 工具通过自然语言处理的句法分析，制定一系列语义规则，完成语义谓词的识别与归一；例如，从表 1 所列的三条文本句中抽取出继发性知识特征词“results in”、“caused”、“cause”，并进一步整合为“CAUSES”。最终，SemRep 工具将所发现的生物医学知识表示为机器可读取、可理解、可推理的结构化三元组 Subject – Predicate - Object(S-P-O),S、P、O 分别称为主语、谓语、宾语。

表 1.SemRep 工具从生物医学文献发现传染病继发性知识示例

PMID	生物医学文本句	继发性知识
33969140	**Chronic hepatitis B** is a necro-inflammatory of the liver parenchyma **caused by hepatitis B virus (HBV) infection.**	**Hepatitis B – CAUSES – Chronic Hepatitis B**
33685080	…… providing a new strategy for the treatment of **chronic hepatitis B caused by HBV infection**.	
37670813	**HBV infection**, **leading to chronic hepatitis B** (CHB).	

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

如图 2 所示，本研究基于文献知识发现的传染病继发性挖掘分析共包括文献数据采集、S-P-O 语义三元组抽取、传染病继发性知识筛选、专业人员审核、继发性知识 S-P-O 去重，以及感染性疾病与非感染性疾病区分共六个关键步骤。

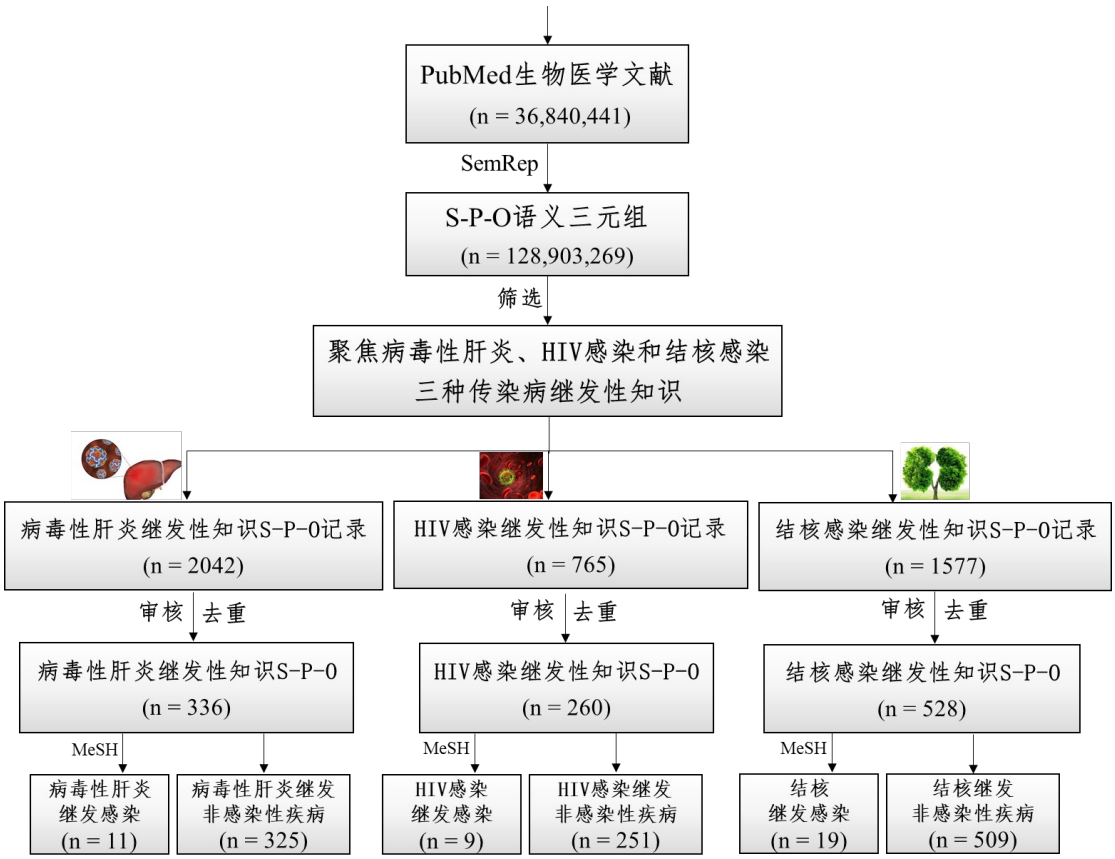

图 2.研究数据处理流

①生物医学文献大数据采集。本文一共收集了 PubMed 数据库自建库至 2024 年 2 月 8 日以来，所收录的 36,840,441 篇生物医学文献摘要数据，全部保存为 TXT 文本数据格式用于自动化文献知识挖掘处理。

②面向生物医学文献的 S-P-O 语义三元组抽取。将采集的所有生物医学文献数据作为输入，提交 SemRep 工具进行 S-P-O 语义三元组抽取，最终获得 128,903,269 条 S-P-O 记录，并带有来源文献及相应的文本句信息。

③聚焦三种传染病完成继发性知识筛选。本研究聚焦三种常见且继发疾病发生率较高的传染病开展挖掘分析，即病毒性肝炎、HIV 感染和结核感染。具体而言，首先在 S-P-O 语义三元组的 S 中筛选这三种传染病的名称；其中，病毒性肝炎具体细分为五种常见的类型，即甲型肝炎“Hepatitis A”、乙型肝炎“Hepatitis B”、丙型肝炎“Hepatitis C”、丁型肝炎“Hepatitis D”及戊型肝炎“Hepatitis E”，通过在 S 中选择“Hepatitis %”完成；HIV 感染和结核

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

感染分别在 S 中筛选“HIV Infections”、“Tuberculosis”来完成；其次，通过限定谓语 P 为“CAUSES”，筛选继发性知识；最后，利用宾语 O 的语义类型“Disease or Syndrome”、“Neoplasm”，筛选出继发性疾病。至此，筛选出病毒性肝炎、HIV 感染和结核感染三种传染病的继发性知识，分别为 S-P-O 语义三元组 2042 条、765 条、1577 条。

④专业人员结合相应的文献文本句开展继发性知识审核。筛选得到的三种传染病继发性知识 S-P-O 三元组，全部经两名专业人员背对背地结合关联文献的摘要文本句审核确认，依据为摘要文本句是否表达了 S-P-O 三元组所代表的传染病继发性知识。例如，表 2 所示的前两行数据，S-P-O 三元组为原句内涵的结构化表示，审核结果为正确；第三行记录，因文本句揭示的医学知识“Acute delta hepatitis”为“hepatitis B”合并感染“hepatitis delta”的结果，而 S-P-O 三元组“Hepatitis B-CAUSES-Acute delta hepatitis”遗漏了部分病因，未能完全表达 PMID 为 26253093 一文的主旨内容，故人工审核为错误，不再纳入后续处理及分析讨论。

表 2.传染病继发性知识审核结果示例

PMID	生物医学文本句	继发性知识	审核结果
28856052	Hepatic cirrhosis is a major cause of morbidity and mortality worldwide due to hepatitis C.	Hepatitis C-CAUSES-Liver Cirrhosis	正确
30536422	TTP secondary to human immunodeficiency virus (HIV) infection is unique to sub-Saharan Africa.	HIV Infections-CAUSES-Thrombotic Thrombocytopenic Purpura	正确
26253093	Acute delta hepatitis, as a result of coinfection with hepatitis B and hepatitis delta, is rare.	Hepatitis B-CAUSES-Acute delta hepatitis	错误

⑤继发性知识去重处理。此处的去重指忽略来源文献及相应的文本句信息，（参见表 1 的第 1 列和第 2 列），按照 S-P-O（表 1 的第 3 列）去除重复项后保留单一项。此外，本研究逐一计算每条 S-P-O 关联的来源文献 PMID 个数，作为继发性知识的出现频次。该出现频次具有两方面的意义；其一，出现频次越大，表明该知识的准确性越高；其二，在大规模知识图谱自动构建时，依据知识出现频次，可在无专业人员审核的情况下，批量筛选出具有高准确性的传染病继发性知识，自动过滤掉低频次、可能有误的语义三元组 S-P-O。去重后的三种传染病继发性知识 S-P-O 三元组分别为 336 条、260 条、528 条。

⑥将传染病继发疾病区分为感染性疾病与非感染性疾病。依据 MeSH 词表对感染性疾病（即感染，编号 C01）的组织结果，将属于感染及其下位概念的疾病划分为感染性疾病，其它疾病归为非感染性疾病（包括肿瘤）。处理后的三种传染病继发感染 S-P-O 三元组分别为 11 条、9 条、19 条，继发非感染性疾病 S-P-O 分别为 325 条、251 条、509 条，纳入后续的结果分析讨论。

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

4. 结果与讨论

4.1 基于文献的病毒性肝炎继发性知识发现结果

病毒性肝炎是全球面临的严峻公共卫生问题之一，在我国法定传染病中归为乙类。本研究按照 PubMed 文献编号（即 PMID）去重后的肝炎继发性知识 S-P-O 共有 336 条，其中病毒性肝炎继发感染相关的 S-P-O 为 11 条。图 3 为文献知识发现结果的可视化，五种常见的病毒性肝炎（包括甲型肝炎“Hepatitis A”、乙型肝炎“Hepatitis B”、丙型肝炎“Hepatitis C”、丁型肝炎“Hepatitis D”及戊型肝炎“Hepatitis E”），能够引发慢性乙肝“Chronic Hepatitis B”、慢性丙肝“Chronic Hepatitis C”及其它类型的病毒性肝炎“Viral hepatitis”，详见图 3 上半部分（线条粗细代表关联的文献量，线条越粗反映揭示该继发性知识的文献越多）。此外，病毒性肝炎可能继发其它感染，特别是慢性感染“Chronic Infection”，以及持续性感染“Persistent infection”。我国患者感染最多的亚型为乙型肝炎，严重者可发展至重型乙肝，患者病死率较高；此时，由于肝细胞大量坏死出现肝功能障碍，肝脏免疫功能减退、解毒功能下降，易出现各种继发感染。据统计，临床重型乙型肝炎患者继发真菌感染的病死率约为 80%以上[24]，感染部位多为肺部、肠道、腹腔、泌尿系统和中枢神经系统，其中肺部真菌感染发生率最高。

如图 3 下所示，病毒性肝炎与多种非感染性疾病的发生存在因果关联。这些疾病多发生在肝脏，具体为肝硬化“Liver Cirrhosis”、肝纤维化“Liver Fibrosis”、脂肪肝“Fatty Liver”、失代偿肝硬化“Decompensated cirrhosis”、肝功能衰竭“Liver Failure”、急性肝功能衰竭“Acute Hepatic Failure”、慢性肝病“Chronic liver disease”、终末期肝病“End Stage Liver Disease”，以及肝癌“Liver carcinoma”（来自 PMID 为 25206272 一文的摘要句“Hepatitis B virus (HBV) infection is a major global health challenge leading to serious disorders such as cirrhosis and hepatocellular carcinoma”）；人体感染乙肝病毒后，病毒的复制会导致长期的肝细胞受损和炎症，引发肝脏纤维化和肝硬化的形成，为肝癌这种恶性肿瘤的发生创造了病理基础。此外，肝炎还可能继发胆汁淤积“Cholestasis”、肾病（包括膜增生性肾小球肾炎“Membranoproliferative Glomerulonephritis”、膜性肾小球肾炎“Membranous Glomerulonephritides”、肾小球肾炎“Glomerulonephritis”）、皮肤病（迟发性皮肤卟啉病“Porphyria Cutanea Tarda”、皮疹“Exanthema”、贾-克综合征“Gianotti-Crosti Syndrome”）、血管疾病（具体为冷球蛋白血症“Cryoglobulinemia”、结节性多动脉炎“Polyarteritis Nodosa”、血管

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

炎“Vasculitis”、B 淋巴细胞病“B lymphocyte disorder”），以及代谢疾病（如非胰岛素依赖型糖尿病“Non-Insulin-Dependent Diabetes Mellitus Diabetes”、混合性冷沉球蛋白血症“Mixed cryoglobulinemia”）。其中，乙型肝炎患者继发膜性肾小球肾炎等肾病，与乙肝病毒感染后通过免疫反应形成的免疫复合物在肾小球膜上沉积有关，这些沉积物会导致肾小球膜受损，临床上需加强早期预警监测，尽可能避免乙肝患者继发肾病的发生。可见，大量的文献知识挖掘分析结果证实，感染肝炎病毒后患者不仅容易继发慢性肝炎和真菌感染，而且累及患者肾脏、血管、皮肤等其它部位，严重者恶化为肝癌。

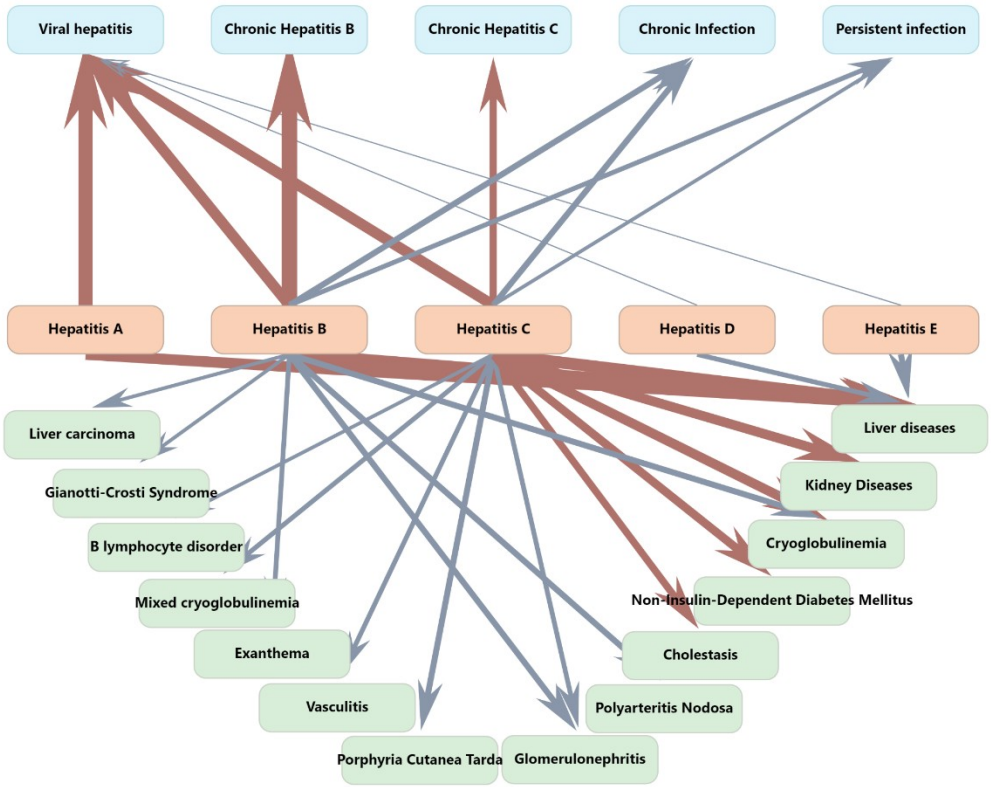

图 3.基于文献的病毒性肝炎继发性知识发现结果可视化

4.2 基于文献的 HIV 感染继发性知识发现结果

HIV 感染为一种血源性传染病，其病原体是一种逆转录 RNA 病毒，主要攻击 CD4+T 细胞；人体感染后引起机体免疫功能缺陷，造成免疫系统损害，出现一系列继发疾病。在本研究中，经去重后的 HIV 感染继发性知识 S-P-O 共有 260 条，其中继发感染相关的 S-P-O 为 9 条。图 4 左半部分展示了 HIV 感染可能继发的其它感染性疾病。其中，获得性免疫缺陷综合征“Acquired Immunodeficiency Syndrome”的继发感染性最高，相关文献支持量达到 135 篇。由于 HIV 病毒的攻击，感染者的免疫功能受到破坏，一般经历急性感染期、无症状感染期和艾滋病期，最后可继发结核和多种机会性感染“Opportunistic Infections”，甚至死亡； HIV 感染常因其继发的感染性疾病为首发症状，包括艾滋病痴呆复合征“AIDS Dementia Complex”、艾滋

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

病相关性肾病“AIDS-Associated Nephropathy”、HIV脑病“HIV Encephalopathy”、HIV消耗综合征“HIV Wasting Syndrome”、“HIV肠病”HIV Enteropathy，以及病毒血症“Viremia”（来自PMID为26714714一文的摘要句“HIV-1 infection of hu-mice results in chronic viremia and CD4+ T cell loss, ……”）。

如图4右半部分所示，文献大数据揭示了HIV感染能够继发多种非感染性疾病。具体而言，包括免疫系统疾病（如淋巴细胞减少症“Lymphopenia”）、营养不良（如脂质营养不良“Lipodystrophy”）、代谢疾病（特别是骨质减少“Osteopenia”、骨质疏松“Osteoporosis”、血脂异常“Dyslipidemias”）、血液病（血小板减少“Thrombocytopenia”、血栓性血小板减少性紫癜“Thrombotic Thrombocytopenic Purpura”、高丙种球蛋白血症“Hypergammaglobulinemia”、贫血“Anemia”）、神经系统疾病（脑病“Encephalopathies”、脑炎“Encephalitis”、感觉性神经元神经病“Sensory neuropathy”、多发性神经病“Polyneuropathy”）、心血管疾病（扩张型心肌病“Dilated Cardiomyopathy”、动脉粥样硬化“Atherosclerosis”）、肝疾病、肾疾病（如慢性肾功能衰竭“Chronic Kidney Failure”）、肺疾病（如肺动脉高压“Pulmonary arterial hypertension”）、皮肤病（如银屑病“Psoriasis”）及视网膜疾病“Retinal Diseases”等。此外，从文献PMID:26836785的摘要“In patients living with HIV infection with hepatitis C (HCV) is common. HIV/HCV co-infection results in more rapid liver fibrosis progression”发现，HIV感染合并丙型肝炎存在继发肝硬化的趋势；尽管丙型肝炎。

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

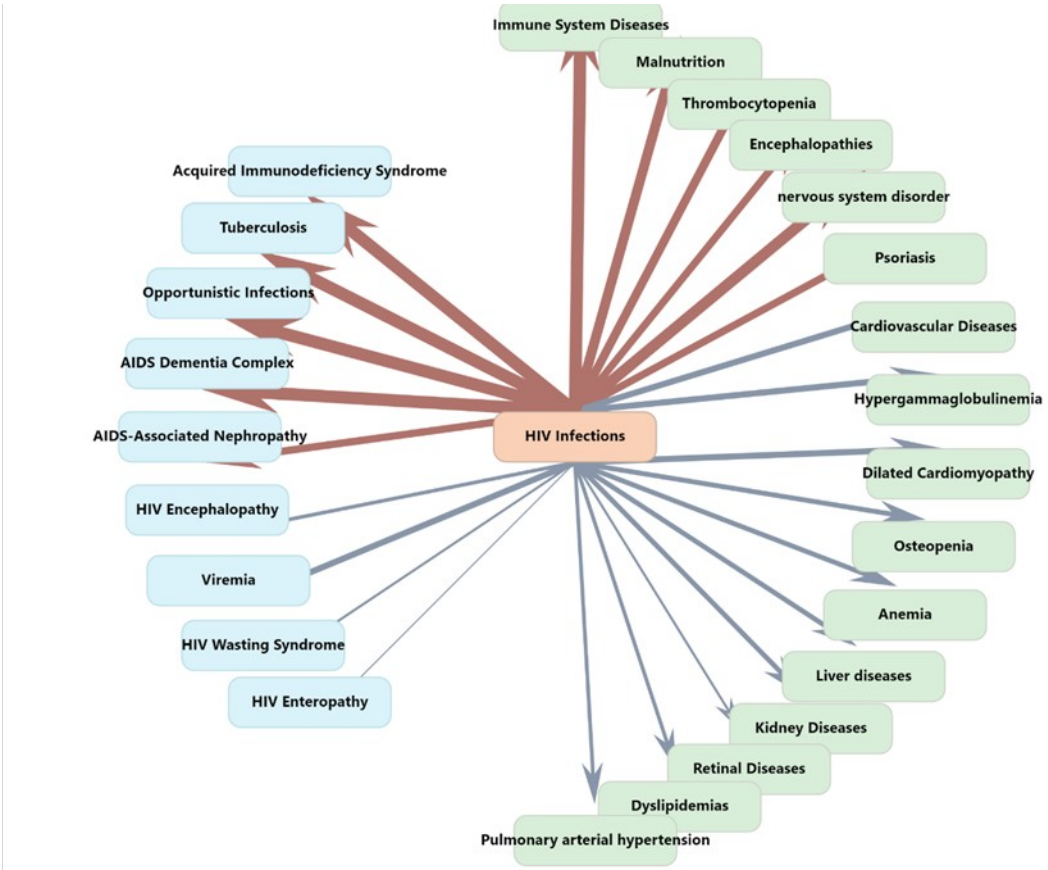

图 4.基于文献的 HIV 感染继发性知识发现结果可视化

值得提出的是，HIV 感染继发恶性肿瘤的临床证据较为缺乏，因果关联较为薄弱。然而，HIV 病毒破坏机体免疫系统，导致免疫功能丧失，从而增加恶性肿瘤发生风险。虽然本研究未从大规模生物医学文献中自动挖掘出 HIV 感染继发恶性肿瘤的相关知识及直接证据，但这与设定 SemRep 工具输出结果 S-P-O 中谓词 P 为“CAUSES”有关。进一步扩大谓词范围包涵“COEXISTS_WITH”、“PREDISPOSES”之后，发现 HIV 感染与淋巴瘤的发生高度关联；PMID 为 21526314 一文的摘要结论句“In this series of patients infected with HIV, a predominance of aggressive histological subtypes of lymphomas and low complete remission rates, were observed”，即证实了这一发现。目前，世界卫生组织已将“免疫缺陷相关的淋巴增殖性疾病”，增为淋巴瘤分类目录的一种类型，表明 HIV 感染在诱发恶性肿瘤中起到不容忽视的作用。

4.3 基于文献的结核感染继发性知识发现结果

结核为一种细菌感染后的慢性消耗性疾病，一般病程较长，患者免疫功能长期处于受损状态，严重者由于肺部等组织病变广泛，存在坏死干酪组织和空洞，局部抵抗力降低。加之免疫抑制剂或广谱抗生素长期使用等因素影响，结核感染造成患者呼吸道组织损伤和防御功能降低，导致免疫功能紊乱，对病原

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

菌的易感性增高，进而继发多种细菌、真菌感染；尤其是重症肺结核患者的肺部损害严重，易继发肺部混合感染。本研究去重处理后的结核感染继发性知识 S-P-O 共有 528 条，其中继发感染相关的 S-P-O 为 19 条。图 5 左揭示了结核可能继发的其它感染性疾病，根据关联的文献量，依次为肺炎“Pneumonia”、肺结核“Pulmonary Tuberculosis”、脊椎关节盘炎“Spondylodiscitides”、骨髓炎“Osteomyelitis”、胸膜炎“Pleurisy”、脓毒症“Sepsis”、真菌性动脉瘤“Mycotic Aneurysm”、结核瘤“Tuberculoma”、结核疹“Tuberculid”、抗药性结核“Drug-Resistant Tuberculosis”、粟粒性结核“Miliary Tuberculosis”、结核性腹膜炎“Tuberculous Peritonitis”、抗多种药物性结核“Multidrug-Resistant Tuberculosis”、脊柱结核“Spinal Tuberculosis”、关节结核“Joint Tuberculosis”、瘰疬性皮肤结核“Scrofuloderma”、皮肤结核“Cutaneous Tuberculosis”、淋巴结结核“Lymph Node Tuberculosis”和泌尿生殖系统结核“Urogenital Tuberculosis”。

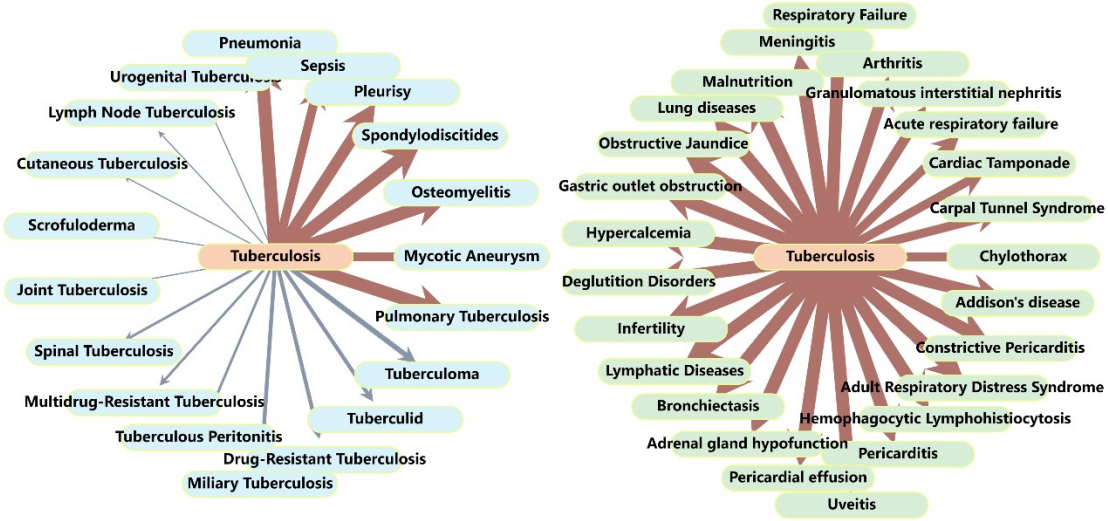

图 5.基于文献的结核继发性知识发现结果可视化

本研究通过大规模生物医学文献的挖掘分析，发现结核感染能够进一步继发多种非感染性疾病（图 5 右）。这些疾病发生在人体的呼吸系统（如支气管扩张症“Bronchiectasis”、成人呼吸窘迫综合征“Adult Respiratory Distress Syndrome”、肺疾病“Lung diseases”、呼吸衰竭“Respiratory Failure”、急性呼吸衰竭“Acute respiratory failure”、乳糜胸“Chylothorax”等）、内分泌系统（具体为艾迪生病“Addison's disease”、肾上腺机能减退“Adrenal gland hypofunction”）、心血管系统（包括心包炎“Pericarditis”、缩窄性心包炎“Constrictive Pericarditis

bmr.202408.00057V1

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

”、心包积液“Pericardial effusion”、心包病变“Pericardial disease”、心脏压塞“Cardiac Tamponade”）、淋巴系统（具体为噬血细胞性淋巴组织细胞增多症“Hemophagocytic Lymphohistiocytosis”）、眼部（葡萄膜炎“Uveitis”）、消化系统（胃出口梗阻“Gastric Outlet Obstruction”、吞咽障碍“Deglutition Disorders”）、神经系统（脑膜炎“Meningitis”、腕管综合征“Carpal Tunnel Syndrome”），还包括关节炎“Arthritis”、不育“Infertility”、营养不良“Malnutrition”、 高钙血症“Hypercalcemia”、阻塞性黄疸“Obstructive Jaundice”、肉芽肿性间质性肾炎“Granulomatous interstitial nephritis”等。可见，通过基于文献的知识发现方法，能够系统全面地揭示结核感染的继发疾病，包括肺炎等感染性疾病，以及内分泌、心血管、淋巴、呼吸、消化等系统部位的非感染性疾病，表明本文研究结果有助于尽早预测其继发风险、减少疾病负担、降低病死率。

5.结论

新发、突发传染病威胁全人类健康，有效监测预警已为公共卫生筑牢安全防线。然而，传染病的继发性不容忽视，其早期预测和风险评估是减少疾病负担、降低病死率的关键突破口。本研究首次引入情报学中经典的基于文献的知识发现方法，聚焦病毒性肝炎、HIV 感染和结核感染三种常见传染病，从超过 3680 万篇的生物医学文献大数据中挖掘分析出相关的继发疾病，包括感染性疾病、非感染性疾病以及肿瘤。研究结果为上述三种传染病继发疾病的预警监测和尽早干预提供了更加有力的文献证据，且证实了基于文献的知识发现方法能够作为系统全面揭示传染病继发性的一种新思路。不足之处体现在缺少对 SemRep 工具的更新完善，后续工作体现在两个方面；其一，应用先进的人工智能技术（包括大语言模型），从算法模型等方面研究针对文献大数据的知识发现方法，拓展 LBD 方法学及挖掘工具的创新研究；其二，融合深度学习、自主学习，研建高效精准的传染病继发感染多维度监测平台，提升我国感染性疾病预警防控水平。

bmr.202408.00057V1

参考文献

[1]World Health Organization. Indicators[EB/OL].[2024-05-03]. https://www.who.int/data/gho/data/indicators.
[2]World Health Organization. Global Influenza Surveillance and Response System (GISRS)[EB/OL]. [2024-05-05].https://www.who.int/initiatives/global-influenza-surveillance-and-response-system.
[3]Johns M, Burke R, Vest K, et al. A growing global network's role in outbreak response: AFHSC-GEIS 2008-2009[J].BMC Public Health,2011,11(S2):S3.
[4]丁超.国家传染病网络直报系统优化策略初探——以应对新冠肺炎疫情为例[J].医学与法学,2020,12(05):14-18.

http://www.biomedrxiv.org.cn/article/doi/bmr.202408.00057
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

[5]徐凌燕,杜诗霖,徐宏斌,等.慢性重型肝炎并发侵袭性肺部真菌感染患者的临床 CT 影像学特征分析[J].临床研究,2017,25(3):14-15.
[6]田瑞强,姚长青,潘云涛.关联文献的知识发现与创新研究进展[J].情报理论与实践,2013,36(08):117-123.DOI:10.16353/j.cnki.1000-7490.2013.08.013.
[7]代冰,胡正银.基于文献的知识发现新近研究综述[J].数据分析与知识发现,2021,5(04):1-12.
[8]Swanson D. Fish Oil,Raynaud's Syndrome,and Undiscovered Public Knowledge[J]. Perspectives in Biology and Medicine,1986,30(1):17-18.
[9]Liu T,Pan X,Wang X, et al. GrEDeL:A Knowledge Graph Embedding Based Method for Drug Discovery From Biomedical Literatures[J]. Journal of Artificial Intelligence for Medical Sciences,2021,1(03):30-42.DOI: 10.2991/jaims.d.201208.001.
[10]Xu B，Shi X，Zhao Z，et al. Leveraging Biomedical Resources in Bi-LSTM for Drug-Drug Interaction Extraction[J].IEEE Access,2018,06:33432-33439.
[11]刘景陶,柳耀花.计算机分子模拟技术及人工智能在药物研发中的应用[J].科技创新与应用,2018(02):46-47.
[12]Kilicoglu H, Rosemblat G, Fiszman M, et al. Broad-coverage biomedical relation extraction with SemRep[J]. BMC Bioinformatics, 2020,21:1-28.
[13]Rindflesch T, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text[J]. Journal of Biomedical Informatics, 2003,36(6):462-477.
[14]Nagao G, Okuzumi S, Kakimoto T, et al. Pulmonary Mycobacterium kyorinense infection secondary to cavitary pulmonary tuberculosis[J]. IDCASES, 2022,28(e1476).
[15]Tabriz N, Lavrinenko A, Kozhamuratov M, et al. The etiological structure and antibiotic sensitivity of pathogens from patients with multidrug-resistant tuberculosis with secondary infections[J]. INTERNATIONAL JOURNAL OF INFECTIOUS DISEASES, 2018,73S:342-343.
[16]Mitrokhin S, Ivushkina L, Mironov A. Rational antimicrobial pharmacotherapy of secondary infections in patients with pulmonary tuberculosis[J]. ANTIBIOT KHIMIOTER, 2005,50(7):32-36.
[17]林明贵,金关甫,王巍,等.肺结核病继发真菌感染 69 例临床分析[J].中华医院感染学杂志,2003(06):49-50.
[18]郭新美,荣学东,马秀丽.74 例肺结核继发肺部真菌感染情况分析[J].中国防痨杂志,2005(06):380-382.
[19]谭守勇,邝浩斌.肺结核继发肺真菌感染的诊治进展[J].中国防痨杂志,2010,32(10):673-676.
[20]Bajpai J, Tewari J, Roy S, et al. Pyopneumothorax Secondary to Pulmonary Tuberculosis Superadded by Congenital Factor XIII Deficiency: A Case Report[J]. CUREUS JOURNAL OF MEDICAL SCIENCE, 2023,15(10): e47350.
[21]Güçsav M, Gayaf M, Aksel N, et al. Tracheoesophageal fistula secondary to tuberculosis: A case report[J]. RESPIRATORY CASE REPORTS. 2019,8(2):49-53.
[22]Raine R. Screening for pulmonary hypertension secondary to pulmonary tuberculosis[J]. AFR J THORAC CRIT CARE MED. 2020.26(4).
[23]李晓瑛,李军莲,李丹亚.一体化医学语言系统及其在知识发现中的应用研究[J].数字图书馆论坛,2019(09):24-29.
[24]张建中,姚玲娣,陈国春.重型肝炎并发医院内真菌感染 68 例临床分析[J].实用临床医药杂志,2011,15(21):157-158.

bmr.202408.00057V1