基于实体共现与引用的潜在共病关系发现*

关陟昊 单治易 林紫洛 杨雪梅 唐小利

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 利用 SPO 语义挖掘和实体识别技术提取文献中具有共病关系的疾病实体，在此基础上构建共病网络，运用链路预测方法发现潜在共病组合。糖尿病领域实证结果表明模型具有良好的有效性和准确性，能够对共病发病机制、疾病预防、临床诊疗等方面起到辅助作用。

〔关键词〕 共病；链路预测；SPO 语义挖掘

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.04.005

Discovery of Potential Comorbidity Relationship Based on Co－occurrence and Citation of Entities *GUAN Zhihao, SHAN Zhiyi, LIN Ziluo, YANG Xuemei, TANG Xiaoli, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China*

〔**Abstract**〕 SPO semantic mining and entity recognition technology are used to extract the disease entities with comorbidity relationship from literature. On this basis, the comorbidity network is constructed, and the link prediction method is used to predict the potential comorbidity combination. The empirical results in the field of diabetes show that the model has good effectiveness and accuracy, and can provide references for the pathogenesis, disease prevention, clinical diagnosis and treatment of comorbidity.

〔**Keywords**〕 comorbidity; link prediction; SPO semantic mining

1 引言

我国是世界上老年人口最多的国家，老年与共病密切相关，60 岁以上居民中有 75.8% 被 1 种以上慢性病困扰[1]。共病与日益增加的不良健康结果相关，如死亡率高、残疾、生活质量差、住院以及医疗资源和支出增加[2]。疾病防治重在预防，我国大力推进的健康中国战略核心在于“治未病”这一预防理念。如果能够发现疾病共患的关联规律、预测潜在的共病关系，对临床诊疗方案有效制定和国家医疗资源合理配置具有参考意义。

2 研究现状

2.1 概述

共病这一概念最早由美国 Feinstein A R 提出，英文表达形式为“comorbidity”，是指患有所研究的某种索引疾病的患者同时还伴发其他疾病[3]。本研究中所指共病为多种疾病同时发生在同一机体内的现象，包括并发症、合并症和继发症等。目前国内外关于共病的医学研究主要分为两个方向，分别是

〔修回日期〕 2022－03－16

〔作者简介〕 关陟昊，硕士研究生；通讯作者：唐小利，研究馆员，硕士生导师。

〔基金项目〕 中国医学科学院医学与健康科技创新工程 2021 年重大协同创新项目“生物医学文献信息保障与集成服务平台”（项目编号：2021－I2M－1－033）。

共病模式研究和共病预测研究。医学领域的“疾病关联”多指疾病与病因的关联，包括：宿主病因，即基因、蛋白、通路等组学角度的病因[4-5]；环境病因，即社会、物理、化学等流行病学角度病因[6]。因此相比于“疾病关联”，“共病”一词更适合描述疾病之间的关联关系。

2.2 共病模式研究

共病模式研究目的是了解调查人群的共病患病现状，挖掘常见高发共病组合或共现关系较强的疾病诊断集群[7]。共病模式研究较为成熟，但多基于共现和统计分析思想，提取、描述能力较强，预测能力较弱，研究重点在于挖掘常见疾病之间关联关系、发现高频疾病组合，以达到疾病预警、共病防治的目的。

2.3 共病预测研究

2.3.1 研究策略 随着自然语言处理和网络分析技术发展，共病预测正在成为共病研究中重要研究方向。目前国内外已有大量关于共病预测的相关研究成果，研究策略主要包括以下 3 个方面。一是从生物信息学角度：基于高通量基因组学、蛋白组学数据，利用生物信息学方法，从基因表达角度计量疾病间关联关系，进而预测可能共现的疾病[8]。二是从临床医学角度：基于电子病历数据，提取疾病共现关系，根据疾病在真实世界中的共现频次和关联网络特点预测未出现的并发疾病[9]。三是从情报学角度：基于临床病例构建共病网络，适用于挖掘发病率较高的常见病共病关系[10-11]，但对于发病率非常低的罕见病，可能不会在所研究的临床病例样本中出现，也可能被多次误诊[12]。解决上述问题的方法之一是使用严谨准确的科学文献数据，生物医学文献包含科研人员对疾病的明确表述。

2.3.2 基于知识网络的相关研究 大量的文献集聚使研究内容彼此之间的关系呈现为一种高度复杂性的网络，研究人员可以通过知识网络对相关隐性知识进行挖掘[13]。Xu R、Li L 和 Wang Q[14]将两个疾病概念在同一个句子中的共现视为具有共患风险的疾病对，通过提取疾病概念对建立疾病风险网络，该数据集随后被一些学者[15-16]用于共病网络研究，这说明基于语义模型提取共病关系是可行的。但是从文本挖掘角度来说，共现关系并不能完整体现概念间基于文献建立的关联，因为概念除了在同一篇文章中共同存在，还会通过文献间引用建立关联，被称为实体计量学。Song M、Kang K 和 An J Y[17]对比基于共现和基于引用构建的实体关联网络，提出基于引用关系构造的网络能够发现更为多样但链接关系较弱的关联，而利用基于共现关系构造的网络可以得到更高准确率。由此可知在实体关联网络的构造过程中，基于引用提取的关系偏重于“全”，基于共现提取的关系偏重于“准”，将二者融合起来可能会达到“全”和“准”的平衡。国内外已有基于单一关系（共现或是引用）进行潜在关系发现的研究成果，并没有将二者结合的先例。

2.3.3 链路预测 其作为分析复杂网络的有效手段，是指如何通过已知网络节点以及网络结构等信息，预测网络中尚未产生连边的两个节点之间产生连接的可能性，在共病预测领域已有广泛应用，但都局限于从共现层面提取共病关系，忽略了实体间通过引用行为建立的关联。

2.3.4 创新研究路径 为解决以上问题，本研究探讨将共现与引用关系相结合的潜在共病关系发现方法。以糖尿病领域为例，通过时间切片方法说明所提方法的优越性，并对该领域的共病组合进行预测，提出未来可能的共病组合，结合相关文献分析疾病间有可能发生关联的途径。

3 方法论

3.1 概述

本研究选用文献数据作为研究对象，基于语义模型和实体计量学提取共病关系构建共病网络，利用链路预测算法计算网络结构特征指标，选取预测效果最好的指标进行共病关系的预测。本研究设计 4 个步骤：数据收集、共病关系提取、共病网络构建以及共病关系预测，见图 1。

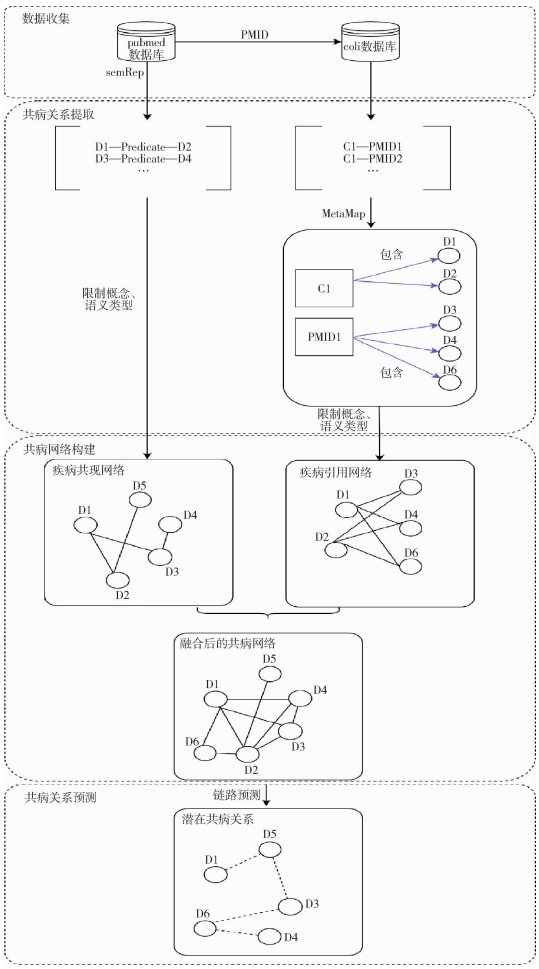

图 1 潜在共病发现模型

3.2 数据收集

PubMed 数据库是美国国立医学图书馆（National Library of Medicine，NLM）开发的免费文献检索系统，提供生物医学文摘信息及相关数据链接。本研究旨在发现可以为临床诊疗与疾病预防提供参考的共病组合，因此选取数据库中时效性较强的文献类型。Colil 数据库是日本学者基于 PubMed Central Open Access Subset（PMC－OAS）全文本构建的生物医学领域引用语句数据库[18]，本研究选取 Colil 数据库获取文献对应的施引语句。

3.3 共病关系提取

3.3.1 主谓宾（Subject－Predicates－Object，SPO）结构提取 使用 SemRep 工具提取文献摘要中的共病对，SemRep 是 NLM 基于统一医学语言系统（Unified Medical Language System，UMLS）开发的从生物医学文本中抽取语义三元组的工具，这个三元组被称为语义谓词。语义谓词由主语、宾语和它们之间的关系组成，形成 SPO 三元组。利用 SemRep 工具从下载的 MEDLINE 摘要数据中抽取语义三元组，通过限制实体类型为“dsyn”（疾病或综合征）；限制语义类型为“COMPLICATES”（并发）、“ASSOCIATED_ WITH”（与…相关联）、“CAUSES”（引起）、“AFFECTS”（影响）、“PREDISPOSES”（诱发）、“MANIFESTATION_ OF”（现象表达）、“PRECEDES”（先于…发生）、“COEXISTS_ WITH”（与…同时发生）可以筛选出具有共病关系的疾病对[14]。

3.3.2 引用语句实体提取 MetaMap 是 NLM 开发的医学实体抽取工具，可以将生物医学文本与 UMLS 叙词表中的概念匹配起来。使用 MetaMap 工具识别施引语句中的医学实体，通过限制实体类型为疾病或综合征（disease or syndrome）可以筛选出施引语句中所包含的疾病实体。例如 PMID 为 33450530 的文献的施引语句中包含的疾病实体为“Diabetes Mellitus”，假设该篇文献摘要中包含的疾病实体为“Ketoacidosis”和“Asthma”，那么基于引用关系建立的共病对为“Diabetes Mellitus－Ketoacidosis”和“Diabetes Mellitus－Asthma”。

3.4 共病网络构建

对抽取出的共病关系进行数据清洗，首先排除 Disease、Syndrome、Disorder 等无意义的泛指概念[19]。同一种疾病可能有不同表达方式，例如妊娠性糖尿病可能被表达为 gestational diabetes 或 diabetes during pregnant。因此要对提取出的疾病概念做消歧处理。具体而言是将实体列表导入德温特数据分析平台（Derwent Data Analyzer，DDA）通过人工建立叙词表的方式完成清洗工作。对基于共现的共病关系和基于引用的共病关系做取并集处理，得到完整共病网络。

3.5 共病关系预测

3.5.1 预测方法 采取链路预测方法进行潜在疾

病关系发现。其基本假设是如果两个节点相似性越大，那么它们之间存在连边的可能性也就越大[20]。分别是基于局部信息、基于路径和基于随机游走的相似性指标。本研究分别从这 3 类指标中选取最有代表性的 14 种进行预测，根据预测结果的准确性选取预测性能最好的指标。表 1 中变量的定义如下：$\varphi(x)$ 表示节点 x 的邻居节点集合，$\varphi(y)$ 表示节点 y 的邻居节点集合，k_x 表示节点 x 的度，k_y 表示节点 y 的度，A 表示网络的邻接矩阵，$(A^n)_{xy}$ 表示节点 x 和节点 y 之间长度为 n 的路径数目，α 和 β 均为可调参数，l^+_{xy} 表示该网络拉普拉斯矩阵的伪逆矩阵中相应位置的元素。

3.5.2 模型评价指标 AUC 是常用的准确性评估指标，表示预测的正例排 S 在负例前面的概率[21]，选取 AUC 作为模型评价的指标。

表 1 链路预测指标及计算公式

相似性指标	计算公式
CN（Common Neighbors）	$S^{CN}_{xy} = \lvert \varphi(x) \cap \varphi(y) \rvert$
Salton	$S^{Salton}_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{\sqrt{k_x * k_y}}$
JC（Jaccard’s Coefficient）	$S^{Jaccard}_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{\lvert \varphi(x) \cup \varphi(y) \rvert}$
Sorenson	$S^{sorensen}_{xy} = \frac{2 \lvert \varphi(x) \cap \varphi(y) \rvert}{k_x + k_y}$
HPI（Hub Promoted Index）	$S^{HPI}_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{min\{k_x, k_y\}}$
HDI（Hub Depressed Index）	$S^{HDI}_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{max\{k_x, k_y\}}$
LHN－I	$S^{LHN1}_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{k_x * k_y}$
PA（Preferential Attatchment）	$S^{PA}_{xy} = k_x * k_y$
AA（Adamic－Adar）	$S^{AA}_{xy} = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{1}{log k_z}$
RA（Resource Allocation）	$S^{RA}_{xy} = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{1}{k_z}$
LP（Local Path）	$S^{LP}_{xy} = (A^2)_{xy} + \alpha (A^3)_{xy}$
Katz	$S^{Katz}_{xy} = \sum_{l=1}^{\infty} \beta \cdot \lvert pathS^l_{xy} \rvert = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots$
ACT（Average Commute Time）	$S^{ACT}_{xy} = \frac{1}{l^+_{xx} + l^+_{yy} - l^+_{xy}}$
Cos	$S^{cos+}_{xy} = cos(x,y)^+ = \frac{v_x^T v_y}{\lvert v_x \rvert \cdot \lvert v_y \rvert} = \frac{l^+_{xy}}{\sqrt{l^+_{xx} ? l^+_{yy}}}$

3.5.3 预测方法有效性验证 由于共病网络形成是具有时序性的，因此预测方法的有效性可通过时间切片方法进行验证，即将第 1 年至第 $n-1$ 年数据作为训练集，第 n 年的数据作为测试集。为比较基于共现关系的方法、基于引用关系的方法和本研究所提出的方法在预测新共病关系方面的性能差异，分别对这 3 种方法对应的共病网络进行链路预测并用 AUC 评估模型的预测性能。

4 糖尿病领域实证

4.1 数据收集

从两个维度收集数据，一是获取 2016－2020 年 PubMed 数据库中糖尿病相关文献，二是通过 Colil 数据库获取这些文献的引用语句。在 PubMed 中检索糖尿病相关文献，时间限定为 2016－2020 年，共收集到 213 199 篇文献和对应的 1 024 427 条引用语句。

4.2 共病网络对比分析

基于引用关系提取的唯一疾病实体数量大约是基于共现关系提取的唯一疾病实体数量的 5 倍，二者交集占前者的 4%、占后者的 23%。在共病对数量方面，两种方法提取出的重复疾病对数量为 40 对，占基于共现方法提取数量的 3%，占基于引用方法提取数量的 2‰，可以看出仅基于共现或基于引文不能获取完整的共病网络，这说明将二者结合是有意义的，见表 2、图 2。

表 2 基于共现、基于引用和融合后网络的疾病和关系数量

项目	共现网络	引用网络	融合网络
疾病数	545	3 070	3 508
关系数	1 277	26 532	27 769

图 2 2016－2020 年共病网络

4.3 指标分析

各项指标均大于 0.5，说明在糖尿病的共病网络中边不是随机产生的，可以利用链路预测算法对未来共病网络进行预测。整合后的网络在各项预测指标上总体优于仅基于共现和仅基于引用构建的网络，说明整合后的网络能够很好地描述糖尿病领域的共病现象，将二者结合是有意义的。其中基于随机游走的 Cos 指标预测效果最好，见表 3。因此利用基于随机游走的 Cos 指标对全部数据进行预测，列出了相似度最高的前 10 条边，即最有可能产生连边的疾病对，见表 4。

表 3 链路预测各指标的 AUC 值

指标	融合网络	共现网络	引用网络
CN	0.857 672	0.691 01	0.850 784
Salton	0.775 248	0.670 645	0.751 757
JC	0.769 166	0.658 218	0.745 178
Sorenson	0.769 05	0.658 18	0.745 64
HPI	0.735 145	0.518 687	0.716 329
HDI	0.768 319	0.657 911	0.742 625
LHN – I	0.642 015	0.508 548	0.602 458
PA	0.829 384	0.597 092	0.832 093
AA	0.863 168	0.692 336	0.857 414
RA	0.865 625	0.691 839	0.861 611
LP	0.815 557	0.621 378	0.819 213
Katz	0.816 472	0.618 473	0.819 192
ACT	0.812 197	0.654 637	0.808 871
Cos	0.882 513	0.679 768	0.870 099

表 4 相似度最高的前 10 个疾病对

疾病 1	疾病 2
mobius syndrome	chronic granulomatous disease
edema disease	navajo neurohepatopathy
lipoidosis	class Ⅲ malocclusion
infection by trypanosoma vivax	thoracic outlet syndrome
lupus renal disease	hypotestosteronism
bilateral pneumonia	empyema
familial generalized lipodystrophy	internal carotid artery diseases
inflammatory disease of the central nervous system	severe combined immunodeficiency due to adenosine deaminase deficiency
diabetic_ gastroparesis	juvenile hemochromatosis
bilateral pneumonia	head and neck disorder

4.4 预测出的共病关系的科学性验证

通过查找表中所列疾病的相关文献进行分析，发现疾病对之间的发病机制存在关联。针对部分疾病组合进行解读和说明。mobius syndrome – chronic granulomatous disease：Mobius 综合征是一种罕见的出生缺陷[22]，其致病基因之一与 B 细胞的存活有关[23]。慢性肉芽肿是一类基因突变引起的免疫缺陷病[24]。这两种疾病均在患者幼年起病，影响免疫系统正常功能。edema disease – navajo neurohepatopathy：纳瓦霍神经肝病多发于严重金属污染地区[25]，而体内累积过多重金属会对神经、血液、消化等系统造成损害，水肿可能这些基础疾病的结果。这两种疾病的发病可能都与患者居住环境有关。lipoidosis – class Ⅲ malocclusion：类脂蛋白沉积症是指透明蛋白样物质沉积在皮肤、黏膜及内脏而引起的疾病，牙齿发育异常是常见的并发症[26]。三类牙错合是颌骨大小与牙齿大小不成比例的临床表征之一。二者均在幼年发病并进行性发展，到患者成年时期自然静止，且都与口腔黏膜异常有关。lupus renal disease – hypotestosteronism：狼疮性肾病患者体内的促炎细胞因子升高会影响脂类代谢，这是低胆固醇血症的病因之一[27]。狼疮性肾病和低胆固醇血症均与细菌、病毒感染以及免疫系统的异常炎性反应有关。综上疾病间可能通过症状、生活环境、发病时期等途径产生关联。疾病之间的关联并非偶然，患者当前所患疾病可能是另一种疾病的危险因素，发现共病的共同机制对疾病的早期干预和防控措施制定具有一定意义。

5 结语

本研究利用实体提取技术和复杂网络分析方法，从生物医学文献中提取疾病实体并根据语义和引用关系构建共病对，融合实体共现与引用关系，构建共病网络，运用链路预测方法对潜在疾病组合进行预测，为疾病的病因、病理、治疗等方面研究提供新的参考方向。研究不足之处在于：受链路预测算法限制，只能预测网络中已有节点间的新链

接，不能预测网络中尚未出现的节点间的链接；受科研条件和专业知识的限制，仅能通过已发表的文献解释潜在疾病组合间产生关联的可能途径，未能通过一定实验手段进行验证。

参考文献

1 王丽敏，陈志华，张梅，等．中国老年人群慢性病患病状况和疾病负担研究［J］．中华流行病学杂志，2019（3）：277－283.

2 张丽，李耘，钱玉英，等．老年共病的现状及研究进展［J］．中华老年多器官疾病杂志，2021，20（1）：67－71.

3 Feinstein A R. Pre－Therapeutic Classification of Co－Morbidity in Chronic Disease［J］. Journal of Chronic Diseases，1970，23（7）：455.

4 浦建宇，陈蕾，邵楷．基于 Katz 增强归纳型矩阵补全的基因－疾病关联关系预测［J］．计算机科学与探索，2019，13（7）：1154－1164.

5 庞永佳，翟桂影，李瑞，等．脂滴包被蛋白的结构、功能及其与脂类疾病关联的研究进展［J］．中国畜牧杂志，2021，6（57）：1－18.

6 吕阳，王志盟，陈滨．室内空气环境与高龄者心脑血管疾病关联性研究进展［J］．建筑科学，2018，34（2）：124－130.

7 刘莉，姚京京，李俊，等．基于共词分析和可视化的高血压疾病关联性挖掘［J］．中国医学物理学杂志，2019，36（5）：614－620.

8 冯程程．基于临床共病现象分析潜在的分子机理［D］．上海：华东师范大学，2016.

9 云科，石锋，穆润清，等．基于病案首页的心血管病种间共病关系分析［J］．中国病案，2019，20（10）：66－69.

10 Xu Z，Zhang J，Zhang Q，et al. Explainable Learning for Disease Risk Prediction Based on Comorbidity Networks［J］. 2019 IEEE International Conference on Systems，Man and Cybernetics（SMC），2019.

11 Xu Z，Zhang J，Zhang Q，et al. A Comorbidity Knowledge－Aware Model for Disease Prognostic Prediction［J］. IEEE Trans Cybern，2021，PP.

12 Zurynski Y，Gonzalez A，Deverell M，et al. Rare Disease：a National Survey of Paediatricians´Experiences and Needs［J］. Bmj Paediatrics Open，2017（1）：e0001721.

13 李东巧，陈芳，韩涛，等．基于二模复杂网络的隐性知识发现方法研究——以潜在药物靶点挖掘为例［J］．图书情报工作，2020，64（21）：120－129.

14 Xu R，Li L，Wang Q. dRiskKB：a Large－scale Disease－Disease Risk Relationship Knowledge Base Constructed from Biomedical Text［J］. BMC Bioinformatics，2014，15（105）.

15 Lee D，Shin H. Disease Causality Extraction Based on Lexical Semantics and Document－clause Frequency from Biomedical Literature［J］. BMC Medical Informatics and Decision Making，2017，171（53）.

16 Jhee J H，Bang S，Lee D，et al. Comorbidity Scoring with Causal Disease Networks［J］. IEEE－ACM Transactions on Computational Biology and Bioinformatics，2019，16（5）：1627－1634.

17 Song M，Kang K，An J Y. Investigating Drug－Disease Interactions in Drug－Symptom－Disease Triples via Citation Relations［J］. Journal of The Association for Information Science and Technology，2018，69（11）：1355－1368.

18 Fujiwara T，Yamamoto Y. Colil：A Database and Search Service for Citation Contexts in the Life Sciences Domain［J］. Journal of Biomedical Semantics，2015，6（38）.

19 Xu R，Li L，Wang Q. Towards Building a Disease－Phenotype Knowledge Base：Extracting Disease－Manifestation Relationship from Literature［J］. Bioinformatics，2013，29（17）：2186－2194.

20 吕琳媛．复杂网络链路预测［J］．电子科技大学学报，2010，39（5）：651－661.

21 Hanley J A，Mcneil B J. The Meaning and Use of The Area Under a Receiver Operating Characteristic（ROC）Curve［J］. Radiology，1982，143（1）：29－36.

22 Bianchi B，Zito F，Perlangeli G，et al. Long－term Results of Facial Animation Surgery in Patients with Moebius Syndrome［J］. Journal of Cranio－Maxillofacial Surgery，2020，48（12）：1132－1137.

23 Fagerberg L，Hallstrom B M，Oksvold P，et al. Analysis of the Human Tissue－specific Expression by Genome－wide Integration of Transcriptomics and Antibody－based Proteomics［J］. Molecular & Cellular Proteomics，2014，13（2）：397－406.

24 唐湘凤，卢伟，井远方，等．单倍体造血干细胞联合第三方脐血移植治疗慢性肉芽肿临床研究［J］．中国当代儿科杂志，2019.

25 Bitting C P，Hanson J A. Navajo Neurohepatopathy：A Case Report and Literature Review Emphasizing Clinicopathologic Diagnosis［J］. ACTA Gastro－Enterologica Belgica，2016，79（4）：463－469.

26 Hamada T. Lipoid Proteinosis［J］. Clinical and Experimental Dermatology，2002，27（8）：624－629.

27 宋俊贤，任景怡，陈红．原发性与继发性低胆固醇血症［J］．北京大学学报（医学版），2010，42（5）：612－615.