• 医学信息研究 •

构建患者相似性分析计算体系*

李昊旻

（浙江大学医学院附属儿童医院 杭州 310052）

〔摘要〕 详细阐述临床类比推理和患者相似性分析、特定临床语义空间中概念可计算性扩展、患者层面多概念空间融合、患者相似组的建立和利用等方法，分析该研究领域面临的机遇与挑战，为国内开展相关研究提供参考。

〔关键词〕 患者相似性；概念可计算；个性化诊疗；类比推理

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.09.004

Construction of the Computing System of Patient Similarity Analysis *LI Haomin, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310052, China*

〔**Abstract**〕 The methods of clinical analogical reasoning and patient similarity analysis, extension of concept computability in specific clinical semantic space, fusion of patient－level multi－concept space, establishment and utilization of patient similarity group are expounded in detail, and the opportunities and challenges in this research field are analyzed, so as to provide references for domestic related research.

〔**Keywords**〕 patient similarity; computability of concept; personalized diagnosis and treatment; analogical reasoning

1 引言

1.1 研究背景

临床技术的不断进步和医学知识的爆炸式增长推动临床实践中解决临床问题能力的提升，但同时带来更为复杂的临床决策环境。单纯依靠个人学习能力和知识技能较难应对日益复杂的临床决策需求，这一现状严重制约医疗服务质量提升[1]。医疗大数据的积累为医学大数据分析和人工智能技术应用提供新的基础[2]，促使医疗服务从基于熟练技艺转向数据驱动的科学发展[3]。

1.2 患者相似性分析概述

目前医疗人工智能范式快速发展，包括符号主义、贝叶斯主义、联结主义、类比主义等多种范式，应用场景逐渐多样化。其中患者相似性分析是基于大量已知案例通过衡量患者之间的距离建立患者相似组，并通过相似组特征获取传统仅能通过医学实践才能获得的临床经验知识，以此定量化评估患者状态、推荐治疗方案和预测患者预后[4－8]。具体来说患者相似性分析是指在特定医疗环境下，选取临床概念（如诊断、症状、检查检验、家族史、既往史、暴露环境、药物、手术、基因等）作为患者的特征项，定量化分析即计算复杂概念语义空间中概念间的距离，通过某种模型融合多维度特征，

〔收稿日期〕 2021－10－21

〔作者简介〕 李昊旻，博士，高级工程师，发表论文 90 余篇。

〔基金项目〕 国家自然科学基金“基于患者相似性分析的普适性临床决策支持方法研究”（项目编号：81871456）；国家重点研发计划“精准医学知识库管理与共享平台开发”（项目编号：2016YFC0901905）。

从而度量患者间的距离，筛选出与索引患者相似的患者相似组并以此模拟临床类比推理的思维模式，同时可通过患者相似组的其他多维特征开展各类评估、推荐和预测。其相较针对特定目标的基于机器学习的人工智能模型具有更好的普适性、临床可解释性等优势。

1.3 患者相似性分析步骤

共有 3 个核心步骤[9]：首先计算复杂概念语义空间中概念间的距离；其次利用多维临床概念度量进一步评估患者间的距离；最后建立合适的患者相似组。上述过程依赖于一套临床概念相似性和患者相似性的可计算体系。本文将介绍本课题组近年来进行的临床概念的可计算范围扩展情况以及利用这些可计算性服务于构建患者相似性分析计算体系的方法、路径，同时探讨当前患者相似性分析技术面临的机遇和挑战。

2 临床类比推理和患者相似性分析

2.1 临床推理的原理

医学分析哲学专家 Sadegh – Zadeh 在其专著中细致分析了临床推理的原理：临床推理的对象是患者“p”，医生面对患者时，患者提供一个非空的数据 $D_1 = \{\delta_1, \cdots, \delta_m\}$ 其中 $m \geqslant 1$，每个 δ_i 代表一个关于患者问题、主诉、症状等的声明。通常认为临床推理是临床医生寻求一个诊断能够解释为什么 D_1 可以发生的过程，这也是传统的基于知识工程的临床决策支持解决方案的理论基础。然而 Sadegh – Zadeh 认为此观点是对于临床实践本质和意图的误解。临床实践以 D_1 作为一个临床问题，临床推理是解决这个问题的过程，解决方案瞄准的不是诊断而是采取什么措施。在寻找和优化治疗措施时往往需要从患者身上获取更多信息，其中包括诊断。因此临床实践可以看作是一个在临床医生控制下通过问答、生成信息实现路径寻找以处置好临床问题的过程。

2.2 患者相似性分析方法

基于 Sadegh – Zadeh 的这一理论可以将临床决策过程抽象为函数 F：

$$F(D_i) = A_i$$

这个推理函数 F 可在面对一个临床问题 D_i 时输出下一步干预的措施 A_i，并基于此干预下的新的临床问题 D_{i+1} 可以迭代输出进一步的措施 A_{i+1}。这一过程不等同于直接寻求诊断。在临床决策过程中最典型的一类知识类型被称为命题式知识，这类知识简单描述就是个体心智状态“knowing that something is the case”（知道这属于什么类型）即类比推理。具有丰富经验的医生可以快速地将一名患者 p 归入到某一个案例模式 P_i（$P_i \in P \{P_1, P_2, \cdots, P_n\}$），而案例模式 P 本质上是患者群体的一个聚类，针对每个患者聚类 P_i 临床医生具有确定下一步采取何种干预的知识技能。大多数误诊和不当处置缘于这个匹配过程不准确或者相应知识技能不完善。因此寻找一个具有普适性的函数能够把患者 p 映射到特定模式 P_i 即可实现对于临床思维过程的计算机化的模拟。患者相似性分析正是基于这样的理论基础，从最初的基于少量典型案例的推理逐步发展为面向海量数据的患者相似性分析。

2.3 基于患者相似性分析的人工智能研究现状

近年来基于患者相似性分析的人工智能研究成为热门研究领域，涉及精神和行为异常、传染病、癌症等[9]。其中所使用数据类型、技术手段各不相同，预测效果也不一致，甚至部分研究对于同一方法的表现优劣存在矛盾性结论。患者相似性分析效果优劣的关键在于构建的相似性分析计算体系是否能够真实评估患者临床意义的相似性。本文将对这些关键问题和挑战进行阐述。

3 特定临床语义空间中概念可计算性扩展

3.1 概述

临床信息中包含了不同语义空间的概念，如诊断、药物、表型、检查检验以及遗传分子信息等，患者相似性分析首先需要建立各特定语义空间相似性计算方法。但是很多临床概念（如诊断、药物、表型等）通常是以文字符号表征的抽象概念而不具有定量细化的可计算性。早期相似性分析计算体系中往往简化计算通过某个特征是否存在来构建二值化的特征空间。这种方式忽略了概念在语义层面的

相似性，往往并不能很好地反映临床意义上的概念距离，而扩展不同语义空间概念的可计算性是相似性分析的重要研究内容。

3.2 基于具有层次语义关系的标准概念编码体系的相似性计算方法

通过分层的方式逐步细化概念是组织领域知识的通常做法。在临床领域同样存在较多此类具有层级结构的语义空间，最典型的是服务于诊断的疾病与有关健康问题的国际疾病分类（International Classification of Diseases，ICD），目前广泛使用的 ICD－10 版本中，疾病和健康问题被分为 22 章、262 节、2 051 个类目、9 505 个亚目以及 22 908 个具体概念编码。显然同一个类目下的疾病比不同类目下的疾病更相似，因此借由此类具有良好空间层次定义的概念编码可以更精细评估概念之间的距离。在这一体系中，评估概念相似性的最优方法是基于信息量（Information Content，IC）的距离计算。目前有多种 IC 以及概念距离计算方法[10]。基于此开展领域性的本体建设可以服务扩展领域概念的可计算性。

3.3 利用临床数据关联构建具有层次结构的可计算概念语义空间

由于体系性的标准术语体系或者概念本体建设依赖大量专家资源建立和维护，并不能在所有临床概念空间均建立或应用这些层次体系，对于这类概念往往需要通过其他方式完成可计算语义空间的扩展。以临床药物为例，虽然化学药物体系中构建了类似层次结构的概念体系解剖学治疗学及化学分类系统（Anatomical Therapeutic Chemical，ATC），但是在国内实际临床环境中该概念体系并不覆盖临床大量使用的复合药物、生物制剂、中成药以及中草药等。针对这类缺乏统一层次化概念语义空间的情况，需要探索利用大数据资源中的关联信息构建全新、广覆盖的可计算语义空间。本课题组针对临床药物的层次分类语义空间构建问题，利用临床用药记录和患者诊断信息的关联信息，采用统计检验获取药物和诊断的显著关联关系，通过诊断空间的特征向量构建临床药物的可计算方法[11]。通过验证，利用临床数据构建的语义空间和传统专家定义的 ATC 具有很好的相关性，同时覆盖更多的临床常用药物，为开展临床药物处方的相似性分析提供了计算基础。这种临床药物距离评估方法除服务于定量评估药物距离外，还可以通过非监督聚类形成药物分类，服务于特定群体用药评估[12]。

3.4 概念标准值参考体系

3.4.1 *体系构建方法* 许多概念描述是数值型，然而计算临床概念间的距离不能忽视实际的临床意义。更特殊的情况是由于年龄、性别甚至人种差异，不同数值在不同群体中具有不同的临床意义。因此对于此类存在人群分布差异的概念需要构建标志值参考体系，然后将原始数值转换为 Z 值（标准分数）。Z 值代表原始分数中减去群体的平均值，再依照群体的标准差分割成不同差距。对于分布不对称或者单边异常的临床概念，通常需要结合临床意义矫正 Z 值，对于正常范围的数值，定义为 0，低于下限或高于上限则处理为该值与下/上限的差值与群体标准差的比值。

3.4.2 *基于研究人群数据构建特定标准值体系* 由于临床实践中还有大量临床数值型概念缺乏公开广泛接受的标准值体系，在实际应用中可以基于研究人群的数据构建特定标准值体系，本研究组曾就中国儿童人群中心脏的超声心动图常规测量数值和髋关节发育不良评估的测量值构建并评估相关标准值参考体系[13－14]，这也从侧面说明基于临床大数据可以有效地构建标准参考体系，并服务于相关概念的相似性分析计算。

3.5 集合或者时序上的概念集相似性计算

3.5.1 *多值概念集相似性计算* 许多临床概念空间可以给一名患者赋值一组概念值，如一名患者可以诊断多个疾病同时使用多组药物，因此同一概念空间中还存在不同大小概念集上如何计算相似性的问题。由于涉及不同长度的集合概念之间的匹配和距离计算，不同匹配策略会带来不同效果，在实际测试[10]中发现最小加权二分匹配（Minimum Weighted Bipartite Matching，MWBM）的算法对于不同长度概念集的匹配效果更佳。

3.5.2 *时序分析方法* 除这类多值概念集情况外，还有一些概念是由时间序列数据组成的，如术中监护的血压数据，这些序列数据的长度通常偏差

更大，从几十到上百，而且具有明确时间特性。传统的相似性计算仅通过统计特性，如均值、方差、斜率变化等反映动态数据特征，但是在这一过程中丢失了序列本身较多变化特征，因此需要引入更多时序分析方法。如可利用 soft - DTW 计算序列血压数据之间的相似性作为人工智能模型的输入来获得更多动态数据相似性[15]，同时一些针对时序数据的聚类方法如 kml 等也可以方便应用于此类数据的聚类分析，并基于聚类信息提供动态相似性。

4 患者层面多概念空间的融合

4.1 概述

医疗大数据背景下的所有医疗数据，如诊断、症状、检查检验、家族史、既往史、暴露环境、药物、手术等，可以作为相似性计算的输入。如何融合不同概念空间到统一的体系中获得最终患者层面的相似性是最核心的挑战。

4.2 传统方法

4.2.1 *方法 1* 传统方法中，多通过简单的映射不同概念空间将患者描述为一个多维空间中的特征向量，然后利用数学方法定量地度量多维概念语义空间中特征向量之间的距离，基于排序或聚类分析筛选出患者相似组。这种方法的局限在于为所有特征都构建独立维度，容易导致维度灾难，同时所有特征都享有统一权重可能带来大量无效特征稀释空间有效特征分布的问题，最终影响患者相似性分析效果。

4.2.2 *方法 2* 针对特定临床场景和临床问题，利用专家知识挑选特征和构建特征权重可以解决一部分问题，基于领域知识数字模型的患者相似性分析通常可以取得更好效果。但是这样的融合模式丧失了患者相似性分析技术路线的普适性，必须依赖专家资源，同时在复杂临床场景下构建此类可计算领域模型的可行性较差。因此需要探索一种能够从临床数据中自学习的融合机制。

4.3 创新方法

4.3.1 *步骤* 本研究团队受心理学领域关于类比推理的结构映射理论（structure - mapping theory）启发，将计算机化的类比推理分为两步：第 1 步是计算属性相似度，在此过程中仅就特定概念空间中对应项的属性之间进行比较和计算相似距离，通常是逻辑和计算清晰的过程。第 2 步是计算关系相似度，通常是高级神经活动和专业知识发挥作用的过程，在计算上引入机器学习模型，通过大量案例学习训练完成不同概念空间属性距离的融合[16]，这类似于人类医生的经验训练过程，不同在于机器训练过程可以更快速地完成并获得人类医生通常需要数年训练才能取得的经验。

4.3.2 *存在的问题* 目前此框架的主要问题在于学习目标的特异性有可能会减弱患者相似性分析的通用性，需要进一步探索人类经验学习机制。

5 患者相似组建立和利用方法

5.1 概述

类似于基因组、蛋白组用来描述某个层次上的全部信息，患者相似组[8]用来描述一个大规模患者群体中具有相似特征的患者群体。该相似组中蕴含了临床实践的各种知识，为计算机获取医学知识提供基础。患者相似组本质上代表的是一个群体特性，这个群体特性是否具有针对特定个体、特定任务的特异性的表征能力，是最终决定相关智能任务效果的关键。

5.2 建立方法

5.2.1 *方法 1* 通常在获取患者层面的定量相似距离评估结果后，可以直接通过筛选距离最近的 N 个患者构建患者相似组，但是对于 N 如何定义缺少理论的支持，同时在不同的空间分布下 N 所代表的距离关系也会有很大的变异。另外一个策略是通过一个距离域值来过滤患者获得一个患者相似组，但同时面临阈值过高相似组的构成太少不具备群体特性，或者阈值过低相似组构成不够单一的问题。在实际操作中通常采用两种策略补充的方式，在适度放松 N 和阈值的情况下，通过满足两个条件来构建患者相似组。

5.2.2 *方法 2* 通过非监督的聚类方法来自动完成群体的分组，根据群体的分布特征完成相关聚类分组，通过某些分组之间距离的评估来评价当前

分组的优劣，一些具有层次聚类的方法还可以进一步丰富构建患者相似组的颗粒度，相似组内部可进一步划分为多个不同的子群体，称为子相似组。

5.2.3 相似组质量控制 无论何种策略构建的相似组，在群体数量不足或者目标患者异质性很高的背景下，很难构建真正意义的相似组，因此基于相似组获取的知识、给出的建议有可能是错误的或者存在偏差的，因此在利用患者相似组开展各类智能任务之前需要对于相似组进行质量控制，一方面可以通过患者相似组中的各类属性的统计分布来检验这个群体中相关特征是否具有很好的一致性，例如要预测的指标为住院时间，那么在这个相似群体中住院时间是否比较集中在一个特定取值范围，和非相似组或者全部群体相比是否有更小的分布方差，在均值分布上是否具有统计意义的偏差等。

5.3 利用方法

对于通用场景下的患者相似组，可以探索一些可视化的方式综合展现个体、相似组和群体的关系[17]，从而更好地理解 3 者之间的关系并基于相似组信息进行临床决策，或者扩展人工智能模型的可解释性。

6 挑战与机遇

6.1 概述

患者相似性分析提供一种通用的计算机辅助临床决策支持的理论框架，在医疗大数据不断积累的背景下其潜力将会逐步被认识、发现和利用。患者相似性分析也是今后医疗大数据产业的一项关键基础技术。目前在开展的一个针对罕见病诊断的项目中[18]，以表型相似性分析为基础，借助可视化方法，试图为临床罕见病患者特别是新生儿提供一种快速的鉴别诊断方法，弥补相关分子诊断周期过长的问题，从而为需要快速诊断和处置的危重新生儿提供决策支持。同时针对先天性心脏病领域，正在探索基于领域知识的相似性分析。未来患者相似性分析利用领域相关研究尚待进一步深入开展。

6.2 数字孪生在医疗健康领域的应用

6.2.1 概述 数字孪生（Digital Twins）是一个工业领域的概念，通常用来评估复杂系统，如航空发动机。其核心是为一个真实的实体构建一个可计算的数字孪生模型，可以满足一些具有不可重复和侵害性的测试需求。近年来有学者将此概念应用于医疗健康领域，希望构建数字孪生患者以提高诊断和治疗能力[19]。

6.2.2 面临的挑战 数字孪生和患者相似性在理论本质上具有同源一致性，因此患者相似性分析可以用来生成数字孪生模型。但是其中最大的挑战是构建动态模型，患者是一个生物动态系统，其生命体征是随时间变化的，干预效果也是动态波动的，目前大多数研究仅利用静态时间点的各种数据或者单一维度下的时序数据进行相似性分析，还不能完整地反映患者动态的相似性。

6.2.3 应对措施 针对这一问题，有研究者将时间信息纳入到患者相似性分析中以寻求突破。动态数据的相似性搜索要求子序列匹配、趋势分析，虽然在统计学和信号处理中，对时间序列分析已有大量研究，但是对于一个高纬度模型来说，动态所带来的往往是灾难性的计算需求。如何在一个时间多分辨率的情况下开展高纬度模型的相似性分析依然是挑战。

6.3 在患者相似性分析实践过程中需处理好精准医学相关要求

在患者相似性分析实践过程中需要处理好精准医学强调的个性化与患者相似性分析的群体特征之间的对立统一，以及基于专家知识和大数据的对立统一。精准医学强调患者的个性化，认为需要针对性地给予个性化治疗，但是患者相似性假设患者在一个特定相似组中具有共性，能够根据共性特征来开展诊疗，从字面上理解两者是对立的，但是本质上患者相似性分析也是在多样的群体中寻求具有个性化特征的群体，当群体足够大时个性化就变成一个小群体的个性化；同时引入领域专家知识能够提高患者相似性分析的准确度，但是大数据中同样蕴含很多未知或者没有系统总结过的新知识，相似性分析可以为知识发现提供支持。

6.4 开展个体和群体多维临床特征可视化研究的意义

对于医疗问题，模型的性能和可解释性同等重

要。尽管应用深度学习模型在特定影像处理领域取得很多成果，但是在通用临床领域如何解释其输出结果以及逻辑还缺乏成熟的机制。患者相似性分析相比黑盒的预测模型具有更好的可解释性，但在复杂多维环境中，这种相似性表现得较抽象，通常需要借助数据可视化工具将聚类、分布、排列、比较、关联等信息以可视化的方式呈现给医生[17]，直接提升对信息认知的效率，引导医生从可视化的结果中分析和推理出有效信息。利用可视化的患者相似性分析其实是在综合人脑对于数据模式的认知以及电脑对于数据计算的高效处理，通过一种互动模式构建人机交互的知识转化框架，因此开展个体和群体多维临床特征的可视化研究对于推动患者相似性分析具有重要意义。

7 结语

本文从患者相似性的理论基础以及构建患者相似性分析计算体系中的若干核心问题出发，结合项目团队近年来的工作实践，系统介绍在临床概念层面构建可计算体系、融合多维特征、构建患者相似组以及评估患者相似组等技术的路径和方法，分析该领域需要重点突破的难点所在。患者相似性分析是医学人工智能综合展现的一个核心领域，该技术的突破能够破解很多长期困扰医疗体系的问题，推动医学人工智能发展到新的层次。

参考文献

1 James J. A New, Evidence – based Estimate of Patient Harms Associated with Hospital Care [J]. Journal of Patient Safety, 2013, 9 (3): 122 – 128.

2 Belle A, Thiagarajan R, Soroushmehr S M R, et al. Big Data Analytics in Healthcare [J]. BioMed Research International, 2015 (2015): 370194.

3 Snyderman R. Personalized Health Care: From Theory to Practice [J]. Biotechnology Journal, 2012, 7 (8): 973 – 979.

4 Hoogendoorn M, El Hassouni A, Mok K, et al. Prediction Using Patient Comparison vs. Modeling: A Case Study for Mortality Prediction [J]. 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016 (2016): 2464 – 2467.

5 Sharafoddini A, Dubin J A, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review [J]. JMIR Medical Informatics, 2017, 5 (1): e7.

6 Zhang P, Wang F, Hu J Y, et al. Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics [J]. AMIA Joint Summits on Translational Science Proceeding, 2014 (2014): 132 – 136.

7 Ng K, Sun J M, Hu J Y, et al. Personalized Predictive Modeling and Risk Factor Identification Using Patient Similarity [J]. AMIA Joint Summits on Translational Science Proceeding, 2015 (2015): 132 – 136.

8 Brown S A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine [J]. Frontiers in Physiology, 2016 (7): 561.

9 贾峥，宗瑞杰，段会龙，等，基于电子病历的患者相似性分析综述 [J]. 中国生物医学工程学报，2018，37 (3)：353 – 366.

10 Jia Z, Lu X, Duan H, et al. Using the Distance between Sets of Hierarchical Taxonomic Clinical Concepts to Measure Patient Similarity [J]. BMC Medical Informatics and Decision Making, 2019, 19 (1): 91.

11 Zeng X, Jia Z, He Z, et al. Measure Clinical Drug – drug Similarity Using Electronic Medical Records [J]. International Journal of Medical Informatics, 2019 (124): 97 – 103.

12 Yu G, Zeng X, Ni S, et al. A Computational Method to Quantitatively Measure Pediatric Drug Safety Using Electronic Medical Records [J]. BMC Medical Research Methodology, 2020, 20 (1): 9.

13 李昊旻，俞劲，王雨虹，等. 基于大数据的儿科超声心动图标准参考体系建设 [J]. 中华超声影像学杂志，2019，28 (3)：1 – 7.

14 Li H, Shu L, Yu J, et al. Using Z – score to Optimize Population – specific DDH Screening: A Retrospective Study in Hangzhou, China [J]. BMC Musculoskeletal Disorders, 2021, 22 (1): 1 – 9.

15 Zeng X, Hu Y, Shu L, et al. Explainable Machine – learning Predictions for Complications after Pediatric Congenital Heart Surgery [J]. Scientific Reports, 2021 (11): 17244.

16 Jia Z, Zeng X, Duan H, et al. A Patient – similarity – based Model for Diagnostic Prediction [J]. International Journal of Medical Informatics, 2020 (135): 104073.

17 Cheng F, Liu D, Du F, et al. VBridge: Connecting the Dots Between Features, Explanations, and Data for Healthcare Models [J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28 (1): 378 – 388.

18 Yang J, Dong C, Duan H, et al. RDmap: A Map for Exploring Rare Diseases [J]. Orphanet Journal of Rare Diseases, 2021 (16): 101.

19 Bjornsson B, Borrebaeck C, Elander N, et al. Digitial Twins to Personalize Medicine [J]. Genome Medicine, 2020 (12): 4.