科学通报 2023 年 第 68 卷 第 19 期: 2537 ~ 2543
观 点
《中国科学》杂志社
SCIENCE CHINA PRESS

医学科学研究的牵引力和驱动力建模：健康需求与未知探索

杜建[1,4*], 赵文静[1], 郭欣[3,4], 董尔丹[2,3,4,5,6,7*]

1. 北京大学健康医疗大数据国家研究院, 北京 100191;
2. 北京大学第三医院心内科, 血管医学研究所, 北京 100191;
3. 北京大学心血管研究所, 北京 100191;
4. 分子心血管学教育部重点实验室, 北京 100191;
5. 康复大学青岛医院(青岛市市立医院)心脏中心, 青岛 266075;
6. 中国医学科学院医学科研管理/代谢性心血管疾病基础与临床研究创新单元, 北京 100191;
7. 细胞生态海河实验室, 北京 100191
* 联系人, E-mail: dujian@bjmu.edu.cn; donged@bjmu.edu.cn

当前，医学已成为全球科技创新最活跃的领域之一. 我国高度重视面向人民生命健康的科技创新. 医学科技创新在保障人民健康、促进经济增长、维护健康安全中发挥着越来越重要的作用. 从国际组织的行动上看，世界卫生组织、世界知识产权组织和世界贸易组织联合建立了关于公共卫生、知识产权和经济贸易的三边合作机制[1]. 这3个组织分别代表了健康促进、科技创新和经济增长3个方面. 关注医学科技创新的健康效应、知识产权(核心技术)和经济影响等重大问题，凸显了医学科技创新的多维度属性. 我国对科技发展的指导思想从2016年“三个面向”到2020年“四个面向”的跨越, 标志着“面向人民生命健康”已经上升到与“面向世界科技前沿”“面向经济主战场”“面向国家重大需求”同样的高度.

医学科学研究是指为增加或证实有关身心疾病和残疾及其原因、诊断、预防、治疗或控制的信息而进行的研究和实验，其目标是通过增进对疾病的了解，开发新的药物或治疗手段以改善健康. 医学科学研究包括一系列广泛的研究: 基础研究、转化研究、临床研究(https://www.nature.com/subjects/medical-research). 其中, 基础研究提供的是治病救人的“理论可能性”; 转化研究是将“理论可能性”转化为“实践可能性”; 临床研究则是通过开展以人为对象的试验, 证实“医疗确定性”. 医学科技进步依靠科技创新和科技管理“两翼齐飞”, 科技创新又受“自下而上”的自由探索和“自上而下”的需求导向“双力驱动”. 学者曾在*Nature Reviews Drug Discovery*发表文章，探讨新药研发的牵引力和驱动力，认为新药研发方向的布局决策过程中，以疾病负担表征的健康需求为牵引力，以科学突破可能转化出来的新治疗机会为驱动力[2]. 此外，医学科学政策界的学者利用疾病负担指标与科学基金、学术文献和临床试验等数据，定量测度健康需求与科技创新的匹配性与关联性[3–6]. 其总体结论认为: 当前的基金资助、科学论文、临床试验等科学活动并未完全与疾病负担相匹配, 存在着“欠资助-过资助”“欠研究-过研究”的疾病领域, 但尚未对这种不平衡性的原因和机理进行深入分析.

杜建 北京大学健康医疗大数据国家研究院助理教授，博士生导师. 主要研究方向为医学科技大数据分析.

董尔丹 中国工程院院士, 北京大学博雅讲席教授, 北京大学心血管研究所所长, 血管稳态与重构全国重点实验室主任. 任中国医学科学院学部委员、教育部科学技术委员会科技伦理专门委员会主任委员等. 长期从事心血管基础研究和医学科技管理工作.

本文试图探索“健康需求与科研活动间不平衡性的原因和机理”这一问题的思路与路径. 我们认为, 以疾病负担为代表的健康需求仅是医学科学研究的牵引力. 这种牵引力属于

© 2023《中国科学》杂志社

宏观范畴, 表现为从健康需求层面“牵引”着科研方向的宏观布局. 例如, 我国心血管疾病死亡人数为300万/年, 在各类疾病中居全因死亡率首位, 超过第二位(肿瘤)和第三位(呼吸系统疾病)之和, 且远高于全球平均水平[7,8]. 科学家在微观层面的科研选题决策过程中, 主要的驱动力和出发点为尚未解决的科学问题, 是具体的问题. 而且是, 科学问题的解决过程可能会引发对更多科学问题的探索(然后再引发更多问题), 类似一种科学发现的引擎, 这正是通向新的前沿科学和原始创新的源泉. 因此, 科学和知识的增长始于科学未知问题, 科学研究尤其是基础研究的本质就是通过提出和解决科学问题, 实现知识体系的扩展, 引领和支撑科学技术的发展[9]. 但目前尚未有一个科学未知问题的数据库, 以方便科学家和资助者查询和理解当前悬而未决的科学问题, 以及这些科学问题之间的关系. 美国国立卫生研究院也正在资助科学问题挖掘的项目(https://reporter.nih.gov/search/vWGjv1mWGEGuhqdcP-moNqA/project-details/10069773). 受以上综合启发, 本文从健康需求和科学未知两个角度, 对医学科学研究的牵引力和驱动力进行综合讨论, 并提出建模方法, 为当前国际上科学政策领域关于科学研究和社会需求的不平衡性问题提供新的见解.

本文提出: 可从牵引力和驱动力两个维度, 更好地理解医学科技创新的过程, 宏观优化医学科学研究的结构和布局; 同时辅以有效的医学科技管理, 为医学科技创新活动提供保护力. 具体来看, 第一, 以未满足的健康需求作为牵引力. 例如, 针对高负担疾病, 降低患病率、死亡率、伤残调整生命年(disability adjusted life years, DALYs); 针对未满足的临床需求, 尤其是现有医疗技术疗效的不确定性和安全性风险, 研发更适宜的诊断、药物、疫苗等医疗技术. 第二, 以悬而未决的科学问题作为驱动力, 包括解决已有的科学问题和探索新的科学未知. 第三, 以医学科技管理作为医学科技创新的保护力, 贯穿科技创新活动的全过程. 其核心在于资源配置(如针对高负担疾病和科学难题开展规划布局)和风险管控(如以人为本的伦理治理)(图1).

1 科技管理: 医学科技创新的保护力

医学科技管理有两大核心任务. 一是通过宏观管理活动提高有限科技资源的配置效率, 即配置资源的问题. 如通过科技规划活动遴选优先发展领域, 通过组织揭榜挂帅促进关键核心技术研发. 二是通过微观管理活动管控医学研究和开发可能带来的对人、社会和环境的风险, 即规制风险的问题[10]. 伦理审查、科研诚信管理、人类遗传资源管理等属于这类范畴. 医学科技创新的最终目的是服务于人的生命和健康, 因此有关伦理问题较之其他领域更为突出, 需要坚持以人为本、保护受试者的总体原则. 规制创新可能带来的伦理风险成为医学科技管理区别于其他领域科技管理的重要内容.

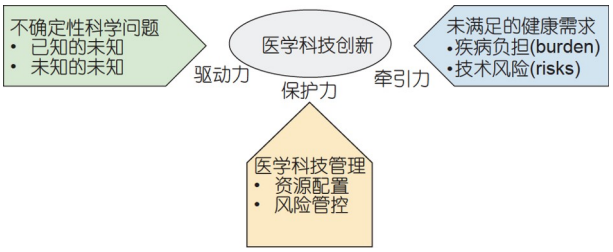

图 1 (网络版彩色)医学科技创新的驱动力、牵引力和保护力
Figure 1 (Color online) The protective, pulling and pushing powers in medical research

医学科技管理的对象和内容多涉及复杂性问题. 首先, 随着学科交叉和深度融合, 医学领域的新兴技术发展迅猛, 高等院校、科研机构都越来越多地成为医疗机构以外“涉及人”的科学研究的主体, 而且这些活动因为所涉及技术的不确定性, 可能具有比经典的药物临床试验更高的伦理风险. 如何既鼓励原始创新, 又能及时管控风险, 比以往任何时候都更加重要. 其次, 医学强调整体性和一体化, 涵盖了基础研究、转化研究、临床试验和实践应用的整体链条. 不同的研究类型需要自适应且适时演化的管理模式和评价机制, 不可一刀切. 同时, 具备“全链条”的医学科技创新活动并不完全遵循商品经济学角度的创新规律. 例如, 新药和疫苗的研发管理需同时兼顾其纯商品属性和公共服务的特殊商品属性[11]. 总体来看, 医学科技管理具有极高的复杂性和特殊性, 发展医学科技管理学, 有助于提高医学科技管理的效率和水平, 支撑和保护我国医学科技创新的高质量发展.

2 健康需求: 医学科技创新的牵引力

美国罗斯福总统的科技顾问范内瓦·布什和拉什·D. 霍尔特[12]撰写的《科学: 无尽的前沿》实际上开启了政府组织和规划科学发展的行动. 随着科学研究组织化程度的不断提高, 优先面向社会需求和现实复杂难题布局科研方向已成为基本共识. 在疾病层面, 量化资源集中程度的一种方法是将资助金额与疾病负担(如残疾调整生命年)进行比较. 1999年, 卫生研究全球论坛发布报告, 提出了著名的“10/90 gap”问题, 即占全球负担90%的疾病, 只获得了全部疾病研究经费的10%, 引起了对健康需求与科技创新间不平衡性及其供需机制的关注[13]. 实际上, 1998年, 美国智库机构——美国医学研究所(Institute of Medicine, IOM)已应国会要求, 对美国国立卫生研究院(National Institutes of Health, NIH)优先资助领域的设置标准进行评估, 并建议在资金分配决策过程中将定量表征各类疾病致死致残的健康需求指标纳入考量[14]. 2022年发表的数据分析结果表明, NIH的基金资助布局和疾病负担结构有明显的相关性和匹配度. 然而, 也存在一些异常值, 某些疾病领域的资金与其负担相比相对充足(如艾滋病毒/艾滋病、糖尿病和乳腺癌), 而其他领域的资金相对不足(如霍奇金病、慢性阻塞性肺疾病和肝癌)[15]. 目前, NIH共资助309个研

究领域/疾病领域. 从2016年开始, 在疾病领域资助金额的网站上均附有美国人群的死亡率和患病率等疾病负担数据(https://report.nih.gov/funding/categorical-spending#/). 这些疾病负担的数据和资助预算表一起被上报国会, 供决策者参考, 以明确不同疾病间资助强度与疾病的公共卫生需求之间的关系. 尽管在实际决策过程中, 很难确定资金在不同疾病之间的最佳分配比例, 但疾病负担仍是一个重要的参考指标. 这一过程需要注意两点: 一是, 认识到并纠正以前的不平衡(如某些疾病在历史上被过度资助); 二是, 认识到疾病发展成熟的领域(如新发现增加了取得突破的可能性)[16].

科技创新与健康需求的供需机制问题逐渐受到学界和国际组织的关注. 2013年, 国际著名权威医学期刊*Lancet*报道了一篇针对健康领域的科研及其需求之间不平衡性的综述性文章[17]. 该文章表明, 当代多数医学科学研究的动机与其改善人类健康的原始目的之间存在巨大差距. 同时, 该文章进一步从研究构架的角度提出, 可利用科学基金、临床试验与学术文献等科技大数据, 分别从投入、过程与产出3个维度量化科研活动, 并开展医学科学研究结构与供需平衡机制的研究. 2014年, *Science*发表研究称, 建立全球健康研发观测站(Global Health R&D Observatory)有助于将有限的医疗科研资源与公共卫生优先事项相匹配, 优化医疗领域的科研资源配置与优先领域设置[18]. 在上述背景下, 世界卫生组织成立了全球卫生研发观测站以收集更精准的科研数据信息, 包含科研投入、文献产出、临床试验、专利数据等, 为开展医学科学研究结构与供需平衡机制的研究提供了数据基础设施.

3 科学未知: 医学科技创新的驱动力

科学研究始于尚未解决的科学问题, 通过知识积累不断取得科学进步. 美国哥伦比亚大学Firestein[19]在其著作《未知: 如何驱动科学发展》一书中指出, “科学往往就像在黑暗的房间里寻找一只黑猫, 而房间里可能根本就没有猫. 这个过程比想象的更偶然, 在线索或假设出来之后有很多绊脚石, 也有很多摸索、试错的过程. 但正是这种未知和不确定性, 这种对棘手问题或离群数据的困惑, 是科学的真正驱动力”. 由此可见, 尚未解决的科学问题是医学科技创新的驱动力. 科学中的未知驱动着科学的发展, 是科学进步的内在引擎. 问题诱导好奇, 好奇驱动探索. 真正的科学进步始于不完备、悬而未决、处于争议矛盾的科学问题[20].

受著作《未知: 如何驱动科学发展》[19]的启发, 笔者从主观感知程度和客观知识状态两个维度, 将知识分为4类: 已认识到的已知知识、尚未认识到的已知知识、已认识到的未知知识、尚未认识到的未知知识(图2). *Science*前主编丹尼尔·科什兰(Daniel E. Koshland)提出了科学研究Cha-Cha-Cha分类方法: 攻克型(charge)、机遇型(chance)和挑战型(challenge). 攻克型解决了已经存在但别人无法解决的问题, 挑战型为现有科学理论无法解释的事实提供新的概念、理论, 机遇型是研究人员偶然发现的现象并证明了其重要意义[21]. 实际上, 攻克型科学研究侧重解决已认识到的未知知识, 机遇型和挑战型科学研究则针对尚未认识到的未知知识.

(1) 已认识到的已知知识. 实际上就是人类的先验知识体系, 如维基百科. 在生物医学领域, 有很多的公共知识库, 例如GenBank、DrugBank等就属于这一类, 汇集了关于基因、药物的基础知识. 已知的已知还可从另外一个角度来理解, 即与新的科学问题探索相比, 针对已知知识的重复研究在一定程度上构成了冗余研究. 例如, 尽管已有权威指南推荐他汀类药物作为冠心病的一线用药, 属于已知的已知, 但仍有大量研究他汀类药物对冠心病疗效的临床试验. 这样的冗余研究, 对临床诊疗新知识贡献相对较少[22].

(2) 已认识到的未知知识. 实际上绝大部分的科学研究都是解决已知的未知, 试图填补已知的空白. 对“*Science*年度十

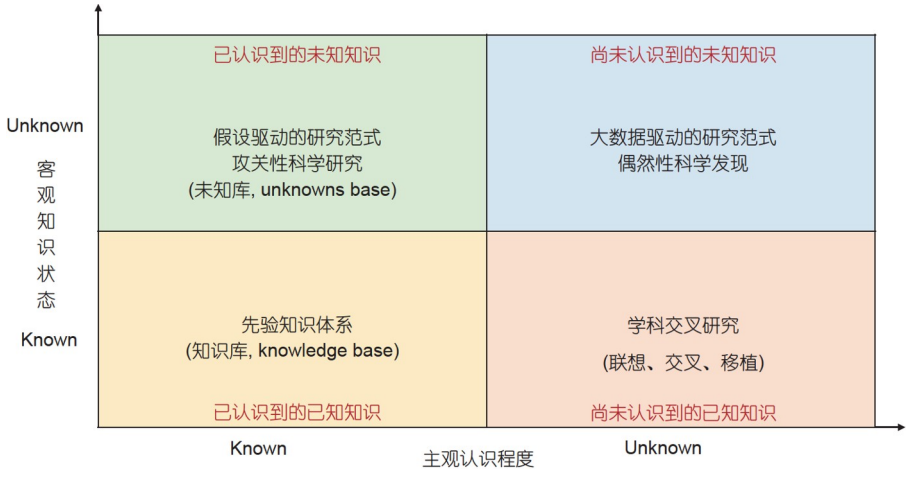

图 2 (网络版彩色)从主观认识程度和客观知识状态对知识进行分类
Figure 2 (Color online) A taxonomy of knowledge by matrixing known and unknown

大科学突破”(共335篇) 的分析发现, 只有很少部分(2.1%)的研究属于意外发现的新物质、新材料、新现象等全新的科学问题; 大部分研究(76.4%)属于已有的研究问题驱动, 专注于解决已存在而尚未攻破的难题, 且未发生明显的范式转移[23]. 例如, 即使癌症疗法取得了革命性发展, 但关于癌症生物学领域仍有3个已知的未知尚待解决: 关于耐药性、肿瘤扩散和正常组织作用的基本问题仍未得到解答[24]. *Science*公布的“125个最具挑战性的科学问题”也属于已知的未知. 其中有些议题经久不衰, 一直是悬而未决的科学问题. 在现有范式框架内, 提出一个新假设, 能证实它, 只是渐进性研究; 无法证实它(即试验得不到期待的结果), 甚至能证伪它(得到阴性结果), 才更重要, 可能孕育着变革性研究. 在随机和偶然情况下, 实验结果出乎预料, 这是导致认识“未知的未知”的源泉[25].

(3) 尚未认识到的未知知识. 通常, 通过意外、惊讶、偶然的科学发现来了解未知的未知. 笔者认为, 已知的未知和未知的未知对应着医学领域的两大研究范式: 假设驱动和数据驱动. 假设驱动对应着已感知的未知. 我们感知到有一个假设, 然后去做实验, 去验证, 把知识变成已知的已知. 而大数据驱动的研究, 恰好对应着未知的未知. 我们也不知道有哪些假设存在, 先通过大数据挖掘(或探索性数据分析), 形成待检验的假设集合, 然后逐一去验证. 大数据驱动的医学研究, 似乎是探索和发掘未知的未知类知识的最佳途径, 也是促进“从0到1”科技创新的重要手段.

(4) 尚未认识到的已知知识. 跨学科研究属于这一类, 通过联想、交叉, 把其他学科已知的知识进行消化和移植, 用于解决本学科的问题, 相当于把知识进行了盘活. 另外, 还有“知识链接”的思路, 例如非相关文献的知识发现. 通过寻找非相关文献之间的“桥接”, 为发掘“未知的已知”提供路径. 例如, 如果在一组文献集中包含“A影响B”, 而在另一组文献集中包含“B影响C”, 那么就有理由假设“A影响C”. 连接A与C的B, 被称为“隐特征”[26]. 发现未知的已知需要计算科学家和实验科学家的交叉与合作. 知识发现是数据驱动的, 是计算科学家的主要工作; 科学发现是实验驱动的, 是实验科学家的主要工作(https://blog.sciencenet.cn/blog-3422014-1303670.html).

综上, 4种知识状态各自具有典型的特征. 已知的已知主要表现为先验知识体系和常用的公共知识库. 已知的未知是科学研究最常见的出发点和驱动力, 适用于假设驱动的研究范式和攻关型的科学研究. 而未知的未知, 适于大数据驱动的研究范式和偶然性的科学发现, 这是探索人类认知领域的新手段. 而已知的未知, 则主要依靠学科交叉和知识链接来实现.

4 医学科技创新的牵引力和驱动力建模思路与方法

在从理论框架层面构建牵引力和驱动力模型的基础上, 本文进一步提出, 可以基于大规模数据对医学科学研究的牵引力和驱动力建模. 进而从解决科学问题和降低疾病负担两个角度, 支持疾病层面研究资助的结构优化管理与决策. 具体而言, 医学科技创新牵引力和驱动力模型中主要涉及两个对象, 即健康需求(牵引力)和科学问题(驱动力). 需通过将两类对象的定量数据映射至统一的疾病名称, 以实现融合分析. 最后, 通过识别重大需求与知识缺口的疾病类型, 聚焦不同疾病医学科研的重难点, 为医疗健康领域研究重点设置与基金资源优化配置提供证据基础.

(1) 牵引力(健康需求)方面. 疾病负担(burden of disease)是全球使用最为广泛和普遍的人口健康状况定量测度方法. 疾病负担的常用测量指标包括单一型指标(如发病率、患病率、死亡人数等)和复合型指标伤残调整生命年(disability adjusted life years, DALYs). 其中, DALYs作为以时间为单位的生命数量和生命质量的综合度量, 合理性与简便性使其成为世界卫生组织及各国卫生统计广泛利用表征健康需求的主要指标, 也是该模型中牵引力的定量表征. 数据获取层面, 世界卫生组织的全球健康监测项目(Global Health Estimates, GHE)定期发布全球各国的疾病负担监测结果(https://www.who.int/data/global-health-estimates), 可作为该模型中牵引力方面的核心数据.

(2) 驱动力(科学问题)方面. 科学文献、科学基金项目申请书、临床试验报告中蕴含着丰富的有关悬而未决科学问题(或医疗问题)的陈述, 这些陈述对应着已知的未知. 这3类数据也常纳入数据集成平台, 用于测度科学活动和疾病负担的不平衡性问题. 我们提出, 用大数据分析加计算语言学的方法, 识别“尚未验证、悬而未决、处于争议矛盾中”的研究主题, 能够反映面向科学前沿、聚焦重大需求和挑战的科学问题. 从这些科学问题的句子中提取疾病的概念, 可间接反映存在着未解决的科学问题涉及的疾病, 同时也表征了疾病的复杂性. 在识别方法层面, 可以利用表征“机制不清、原因不明、结果有争议、发现不一致、报告相互矛盾等”的词语作为触发词(如unknown、unclear、uncertain、controversial、incomplete、conflicting、inconsistent、unexpected、confusing等), 在科学文本中进行语句查询、提取和人工判断. 构建医学领域悬而未决的科学问题清单和未知库, 并实现自动更新. 进而, 通过与疾病负担数据进行关联分析, 可从疾病层面分析悬而未决的科学问题与疾病负担之间的关系, 为优化疾病层面的资助管理与决策提供科学证据.

(3) 疾病映射层面, 将上述不同类型的科研数据与疾病负担的疾病分类体系相互映射, 是实现健康需求和科学问题关联分析的必要步骤. 科学基金、临床试验以及科学文献数据自身的字段信息各异, 数据结构也均不相同. 如何基于各类数据的已有字段信息辨别其所关注的疾病, 并且将科研数据中所提取的“疾病信息”与疾病负担的疾病分类体系相互映射, 是该模型构建的核心研究内容. 本文认为, 可以通过利用自然语言处理工具(如Medical Text Indexer)提取各类科研文

本中所涉及的医学主题词(Medical Subject Headings, MeSH), 以表征该文本所涉及的主要内容. MeSH是美国国立医学图书馆编制的兼具权威性、规范性与动态性的主题词表, 可以作为科学活动数据与疾病负担数据两个体系之间映射的桥梁. 进而, 通过构建“MeSH-疾病类型”的映射词表, 实现多类型科研数据与疾病类型的映射.

整体而言, 多类型的科研资助、临床研究活动和产出报告数据、疾病负担指标等, 以及大数据分析挖掘技术, 为医学研究的牵引力和驱动力建模提供了数据与方法基础.

5 结论

本文提出医学科技创新的牵引力、驱动力和保护力模型. 以未满足的健康需求作为牵引力、以悬而未决的科学问题作为驱动力, 以更好地理解医学科技创新的过程. 同时, 以科技管理为抓手, 从资源配置与风险管控两个角度为医学科技创新的良性发展提供保护力. 最后, 尝试在大数据驱动的视角下对医学科技创新的牵引力和驱动力进行建模, 从解决科学问题和降低疾病负担两个角度, 支持疾病层面的医学研究结构和布局优化, 为医学科技管理(保护力)的有效运行提供决策基础. 我们将在下一步工作中, 通过收集医学科技创新全流程的科研活动数据(科学投入、试验、产出等)和疾病负担指标数据, 在构建多源数据融合模型与疾病分类术语映射体系的基础上, 对所提出的理论模型进行实证研究, 以进一步优化模型以及通过实证研究促进医学科技管理与决策实践.

致谢 感谢国家重点研发计划(2022YFF0712000)、国家自然科学基金(72074006, 82202820, L2124021)、中国工程院战略研究与咨询项目(2022-GJ-02)、中国医学科学院医学与健康科技创新工程项目(2021-I2M-5-003)和细胞生态海河实验室创新基金(HH22KYZX0047)资助.

推荐阅读文献

1 WIPO. WHO, WIPO, WTO Trilateral Cooperation on Public Health, IP and Trade. 2019, https://www.wipo.int/policy/en/global_health/trilateral_cooperation.html
2 Agarwal P, Searls D B. Can literature analysis identify innovation drivers in drug discovery? Nat Rev Drug Discov, 2009, 8: 865–878
3 Evans J A, Shim J M, Ioannidis J P. Attention to local health burden and the global disparity of health research. PLoS One, 2014, 9: e90147
4 Confraria H, Wang L. Medical research versus disease burden in Africa. Res Policy, 2020, 49: 1–18
5 Zhao W, Wang L, Zhang L. How does academia respond to the burden of infectious and parasitic disease? Health Res Policy Syst, 2022, 20: 89
6 Atal I, Trinquart L, Ravaud P, et al. A mapping of 115000 randomized trials revealed a mismatch between research effort and health needs in non-high-income regions. J Clin Epidemiol, 2018, 98: 123–132
7 The Writing Committee of the Report on Cardiovascular Health and Diseases in China. Report on cardiovascular health and diseases in China 2021: An updated summary (in Chinese). Chin Circul J, 2022, 36: 553–578 [中国心血管健康与疾病报告编写组. 中国心血管健康与疾病报告2021概要. 中国循环杂志, 2022, 36: 553–578]
8 Vos T, Lim S S, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet, 2020, 396: 1204–1222
9 The Writing Committee. Cases of Condensing Scientific Questions (in Chinese). Beijing: Science Press, 2023 [本书编写组. 凝练科学问题案例. 北京: 科学出版社, 2023]
10 Dong E D. Introduction to Medical Science and Technology Management (in Chinese). Beijing: Peking University Medical Press, 2022 [董尔丹. 医学科技管理学概论. 北京: 北京大学医学出版社, 2022]
11 Wu L S. Consideration on constructing theoretical and methodological system of medical science and technology management (in Chinese). Chin J Med Sci Res Manag, 2009, 22: 14–15 [吴乐山. 构建医学科技管理学理论方法体系的思考. 中华医学科研管理杂志, 2009, 22: 14–15]
12 Bush V, Holt R D. Science, the Endless Frontier. Princeton: Princeton University Press, 2020 [范内瓦·布什, 拉什·D. 霍尔特. 科学: 无尽的前沿. 北京: 中信出版社, 2021]
13 Global Forum for Health Research, World Health Organization. The 10/90 Report on Health Research. In: Global Forum for Health Research. 1999
14 Institute of Medicine (US) Committee on the NIH Research Priority-Setting Process. Scientific Opportunities and Public Needs: Improving Priority Setting and Public Input at the National Institutes of Health. Washington DC: National Academies Press (US), 1998
15 Nimgaonkar A, Mughal A Y, Heimer H, et al. Exploring static and dynamic relationships between burden of disease and research funding in the United States. Health Res Policy Sys, 2022, 20: 60
16 Khurana M P, Madsen E B. Perspectives on the narrowing and clustering of research trajectories: An epistemic threat to medical progress? Sci Public Policy, 2022, scac074
17 Røttingen J A, Regmi S, Eide M, et al. Mapping of available health research and development data: What's there, what's missing, and what role is

there for a global observatory? Lancet, 2013, 382: 1286–1307
18 Terry R F, Salm Jr J F, Nannei C, et al. Creating a global observatory for health R&D. Science, 2014, 345: 1302–1304
19 Firestein S. Ignorance: How it Drives Science. New York: Oxford University Press, 2012
20 Sarewitz D. The voice of science: Let's agree to disagree. Nature, 2011, 478: 7
21 Koshland Jr D E. The Cha-Cha-Cha Theory of scientific discovery. Science, 2007, 317: 761–762
22 Jia Y, Wen J, Qureshi R, et al. Effect of redundant clinical trials from mainland of China evaluating statins in patients with coronary artery disease: Cross sectional study. BMJ, 2021, 372: n48
23 Wuestman M, Hoekman J, Frenken K. A typology of scientific breakthroughs. Quant Sci Stud, 2020, 1: 1203–1222
24 Bourzac K. Biology: Three known unknowns. Nature, 2014, 509: S69–S71
25 David C L. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. J Exp Bot, 2009, 60: 712–714
26 Smalheiser N R. Rediscovering Don Swanson: The past, present and future of literature-based discovery. J Data Inf Sci, 2017, 2: 43–64

Summary for “医学科学研究的牵引力和驱动力建模: 健康需求与未知探索”

Modeling the pulling and pushing power in medical research: Unmet health needs and unsettled scientific questions

Jian Du[1,4*], Wenjing Zhao[1], Xin Guo[3,4] & Erdan Dong[2,3,4,5,6,7*]

[1] *National Institute of Health Data Science, Peking University, Beijing 100191, China;*
[2] *Institute of Vascular Medicine, Department of Cardiology, Peking University Third Hospital, Beijing 100191, China;*
[3] *Institute of Cardiovascular Sciences, Peking University, Beijing 100191, China;*
[4] *Key Laboratory of Molecular Cardiovascular Science, Ministry of Education, Beijing 100191, China;*
[5] *Heart Center, Qingdao Municipal Hospital, University of Health and Rehabilitation Sciences, Qingdao 266075, China;*
[6] *Research Unit of Medical Science Research Management/Basic and Clinical Research of Metabolic Cardiovascular Diseases, Chinese Academy of Medical Sciences, Beijing 100191, China;*
[7] *Haihe Laboratory of Cell Ecosystem, Beijing 100191, China*
* Corresponding authors, E-mail: dujian@bjmu.edu.cn; donged@bjmu.edu.cn

At present, medical research has become one of the most active fields of science and technology innovation in the world, and our country attaches great importance to the scientific and technological innovation oriented to health. The advancement of medical science and technology is propelled by two factors: Scientific and technological innovation and management. In the meanwhile, two forces—Free exploration and demand-oriented innovation—Drive scientific and technical advancement. This article proposed a “three-force” model of medical science and technology innovation, namely pulling, pushing and protecting forces. The protecting force of science and technology management places a strong emphasis on resource allocation (e.g., strategic planning for diseases with high burden and scientific conundrum) and risk management (e.g., ethics governance). The pushing and pulling forces could contribute to a better understanding of innovation process. Unmet health needs, the pulling force, focus more on the health-research priorities for lowering burden of diseases. In particular, unmet health need is also the fundamental motivation for the development of diagnostics, medications, vaccinations, and other health-related treatments. Exploration of the unknown, the pushing power, addresses the existing scientific problems and explores the scientific unknowns. This paper divided knowledge states into four categories, which are known knowns, known unknowns, unknown knowns and unknown unknowns. Through the interpretation of four knowledge states, this paper provides a thorough discussion on how scientific unknowns drive scientific and technological innovation.

The approach of big-data analysis provides the methodological basis for further structural optimization of the research landscape based on the pulling and pushing power model. Hence, on the basis of constructing the “three-force” model from the theoretical conceptual level, this paper further proposes that the pulling and pushing force model can be built in a quantitative approach with big data analysis. To be specific, three research objects, namely, health needs, scientific problems and science and technology innovation, could be quantified using various indicators and methods. In terms of measuring health needs, morbidity, mortality, and disability adjusted life years (DALYs) are the widely used and universal quantitative indicators, which can be applied in the “three-force” model. As for the identification of scientific unknowns, we argue that using big data analysis and computational linguistics to identify research topics that are “controversial, conflicting, unverified, unsolved” can reflect scientific problems that are oriented to the scientific frontier and focus on major needs and challenges. For example, terms like “unknown, suspect, unclear, unusual, controversial, consensus, incomplete, conflicting, contrary, debatable, etc”. can be employed as trigger words, and sentences containing such triggers words can be extracted to represent medical claim with uncertainty. In terms of disease mapping, natural language processing tools (such as Medical Text Indexer) can be used to extract Medical Subject Headings (MeSH) involved in various scientific research texts to represent the main content. With the construction of “MeSH-Diseases” concordance table, the mapping of multi-types of scientific research data and disease can be realized. Such research could assist scientific policy in the identification of major scientific issues, the allocation of scientific resources, and the encouragement of coordinated, multidisciplinary research.

medical research, disease burden, scientific problems, ethical governance, big data analysis

doi: 10.1360/TB-2022-1310