

循证医学方法学在人工智能时代背景下面临的挑战

江华 杨浩 彭谨 陈伟 Charles Damien Lu 曾俊

传统循证医学作为临床医学研究的标准性方法,为解决若干重要临床争论提供了很好的思路。然而,新近发表的研究表明:以 Meta 分析为代表的循证医学方法,其理论支柱之一的“异质性检验”存在方法学缺陷,而且受到样本数影响、抽样误差导致假异质性的挑战,导致同类患者群采用同样干预却出现自相矛盾的结果,使得临床医师无所适从。我们从经典循证医学的缺陷着手,将临床与基础研究的数据相结合,提出新型的计算机技术模型,将机制性解释和临床宏观表型吻合,使个体化精细医疗成为可能。

1 Meta 分析是循证医学方法的支柱之一

过去 20 年中,统治整个临床研究领域的方法学范式是循证医学^[1-4]。对于临床医生而言,循证医学主要以 3 种方法为基准:随机对照试验(RCT)、Meta 分析、循证指南。在权威的循证医学国际组织 Cochrane 协作网的定义中^[5],基于 Meta 分析的系统评价,是所谓“最高级别”的临床证据,被广泛接受为是制定指南以及指导医生诊疗决策最重要的科学依据^[6-9]。

2 Meta 分析的缘起

既然可以做临床试验,为什么还需要 Meta 分析? ① 大多数 RCT 的样本量不够大,把多个研究者的 RCT 通过 Meta 分析合并在一起,可以大大降低达到科学客观结论的成本;② 即使单个 RCT 可以做到大样本量,但受到研究者和资助者主观、客观因素的影响较多,设计可能不够合理,得到的结论往往是良莠不齐。Meta 分析能够梳理这些影响因素,从而为有争论的问题理清思路。如果没有 Meta 分析初期的成功,“循证医学”的概念将可能很难获得广泛认可,并很难成为统治临床医学实践的方法论。

3 Meta 分析在方法学上的先天不足

Meta 分析具有某种先天缺陷,并从其诞生之初就一直如影随形,即所谓“异质性”(heterogeneity)。换言之,即使临床试验是针对同一类疾病、同种治疗手段而开展的,由于人和人之间的差异性,临床试验的设计以及试验环境的差异性,并不可能找到绝对相同的两个研究。显然,根据 Meta 分析的理论假说和哲学基础,只有当在不同时间、不同地点或不同研究者所进行的试验具有相当的一致性时,合并才是合理的。为达到这一目的,就必须从大量研究中找出具有相似性的研究,并确认其本质上的相似性,进而与其他本质上完全不同的研究区分开来。所谓异质性就是指研究之间存在的、具有本质性的差异。

4 经典异质性检验:理论与统计学上的缺陷

Meta 分析的先驱者们深知解决异质性的重要性。首先对研究设计相似性进行定性评估,后来又发展出基于统计的定量化评估方法,即目前流行的以 Q 和 I^2 为代表的所谓“异质性检验”。然而,我们最近发表的研究从数学上证明,这些经典的、已经沿用 10 余年的“异质性检验”在方法学上也有缺陷^[10]。Meta 分析最初出现时,人们已经发现不同的临床试验在数据采集、样本的具体情况方面可能具有的差异属性非常多,要证明能够将来自不同研究的数据合并在一起分析,在数学上是很难得出合理性证明的(legitimate)。如何定义异质性并加以定量评价,是整个循证医学发展最初 20 年中在方法学方面最重要的问题之一。2003 年 Higgins 等^[11]提出 I^2 统计量后,似乎得到了一个可接受的解决方案。 I^2 实际是基于之前提出的 Q 检验的一

DOI: 10.3760/cma.j.issn.2095-4352.2015.09.002

基金项目:四川省科技计划项目(2012SZ0181, 2012SZ0219, 2014FZ0125, 2015SZ0110);四川省成都市科技计划项目(11PPYB099SF-289, 12PPYB181SF-002)

作者单位:610072 四川成都,计算数学与生物统计学教研组,创伤代谢组多学科实验室,四川省医学科学院·四川省人民医院〔江华、杨浩、彭谨、陈伟、Charles Damien Lu(美国)、曾俊〕;610072 四川成都,电子科技大学临床医学院(江华、曾俊);100730 北京,中国医学科学院北京协和医院肠外科(江华、陈伟);300071 天津,南开大学数学科学学院计算数学系(杨浩)

通讯作者:陈伟, Email: chenw@pumch.cn

种改进。 Q 的定义为^[12]: $Q = \sum_k \hat{\omega}_k (\mu_k - \hat{\mu}_{\hat{\omega}})^2$, 其中: $\hat{\omega}_k = n_k / \sigma_k^2$ 。公式中 n_k 为第 k 个研究的样本数, σ 为研究样本方差。Higgins 等^[11]发现, Q 值虽然可以用于度量 k 个研究的异质性, 但当 k 数量逐渐增大时, 其检验过程由于没有包含研究的个数 k (也就是自由度), 所以其值是不稳定的。随着 k 数量的逐渐增加, Q 值会持续增加, 从而造成检验结果的假阳性。为了解决这一问题, Higgins 等进一步提出了 I^2 , 其计算方法为: $I^2 = (Q - df) / Q \times 100\%$ 。

从上述公式中可以发现, I^2 的计算方法消除了自由度引起的超额变异, 使 I^2 比 Q 更为合理。Higgins 和 Julian^[13]有关 I^2 的论文在 *BMJ* 杂志上发表后, I^2 迅速被业界接受为异质性度量的标准, 并被写入了包括 Cochrane 系统评价手册在内的几乎所有循证医学教科书中^[14], 是如今几乎每一篇 Meta 分析都会用到的方法。

但是, 我们的研究通过数值仿真证明: 当样本数逐渐增大时, I^2 随着研究数的增加, 其上升趋势基本上是单调不降的^[10]。这意味着只要研究样本量足够大, 哪怕是本来没有异质性的、来自同一样本的抽样也会出现假异质性。进一步证明了传统异质性检验量 Q 同样依赖于样本量的大小。

异质性检验本质上就是为保障 Meta 分析的可靠性, 使其能够将来自多个 RCT 的数据进行合并, 扩大样本量, 从而实现检验假设所必需的效应量。但我们的研究证明, 随着研究数量的增加, 合并 RCT 并使得样本量增加的 Meta 分析, 其异质性检验的结果并不可靠^[10]。

5 对基于 Meta 分析循证医学的反思

循证医学的创始人之一 David Sackett 曾经在其名著 *Evidence-Based Medicine: How to Practice and Teach EBM* 中指出, 慎重、准确和明智地应用当前所能获得的最好的研究依据, 同时结合医生的个人专业技能和多年临床经验, 并考虑患者的价值和愿望, 制定出的治疗措施称为循证医学^[15]。然而, 在循证医学实践发展的过程中, 随着对证据分析体系的制定以及对高等级证据的不断强调, 使得临床研究者和广大医务人员在实践中逐渐把最佳证据理解为: 大样本的 RCT 以及基于这类研究的 Meta 分析。但随着时间推移, 越来越多的 RCT 和 Meta 分析逐渐显示出各种自相矛盾性, 反而使得临床医师无所适从。例如临床营养学界最近对 2 项发表于知名杂志的、结果完全矛盾的危重病患者最佳能量和宏量营养素摄入的 RCT 应当如何取舍和解读, 就陷入到巨大争论中^[16-17]。

很多时候 RCT 结果相互矛盾, 存在着学术因素之外的影响。但即使数据本身没有问题的临床试验, 其假设检验以及对数据的认识仍存在着深层次的矛盾和困境。例如: 从影响临床结局的原因来看, 实际上影响最终考察目标的因素实在太多, 单纯的随机分组并不能在数学上保证把每个样本的各个参量都平衡地分配到各个研究组当中。根本上, 个体差异反映的是从基因组到宏观表型的差异。随着对基因组学认识的深化, 研究者已经发现, 影响特定临床表型的基因数量可以有几十个到数百个之多。进一步考虑在转录水平、表达水平的影响因素, 那么可能影响到临床结局的分子因素将以数量级增加。

即使假设这些影响因素在人群中的分布都是随机的, 即正态分布, 但存在于个体间数以万千计的影响因素实际上已经构成了维度高达上万个的超高维空间。面对上述本质上分布于超高维空间中的个体差异, 即使有数以千计的研究对象, 也不可能完全做到真正随机地对个体差异进行分配。

综上, 我们可以看出 RCT 的假设基础: 随机化能够保证的仅仅是每个参加实验的对象均有“同等机会”被分配到试验组和对照组中, 但是不能保证每个影响实验结局的因素都“均匀”地被分配到两组中。

6 引入新方法, 建立新一代循证医学的研究模型, 重新认识“循证临床实践”

归根结底, 无论是 RCT、队列研究、亦或是病例对照研究, 其方法学的哲学本体其实并没有太大差别: 即观察和收集数据。根据数据, 对宏观表型背后的机制作出推断。随着新兴的计算技术、新一代统计方法和物理学方法的迅猛发展, 以及快速将其方法论渗透进入临床研究领域, 面对复杂的疾病生物现象, 应大胆承认: 第一代循证医学及其背后的、基于古典统计学思想的研究方法, 到了应该被质疑并更替的时刻。

在人工智能和大数据时代, 首先需要肯定, 所有临床数据不论是来自前瞻性试验还是源于实践中收集的, 在经过适当的统一化和清理之后都具有同等的应用地位。最佳的新一代循证医学研究方法应建立在广泛性的原始数据开源基础上。使用数据挖掘和机器学习等方法对数据进行深入挖掘, 将可以得到很多新知识。

与此同时, 我们也深知大范围的数据开源或许还需等待一个较长的认同和接受的过程。在这一逐步转变的过程中, 对于已经经过研究者整理的数据, 依然有进行“二次研究”的需求和必要性。在这一过渡时期,

各学科应该加强对数据报告规范性的要求,同时引入并开发一些适合于这些数据的模型,将会对证据的整合和产生起到很好的作用。

在我们最近进行的一项对危重病患者最佳能量摄入区间的研究中发现,与经典的、基于“专家经验”的 Meta 分析比较,基于无偏倚聚类的人工智能方法对各研究间的相似性和异质性的判断更为合理^[18]。这意味着我们将不再依赖于几个统计模型,强行要求数据必须“适合”这些模型,而是反过来,我们将谨慎细致地评估数据类型,选择最适合于数据的模型^[19]。换言之,新一代循证医学将是“数据驱动,模型适应”,而非“模型驱动,数据适应”。

针对有关统计学数据报告的规范性和研究重现性问题,统计学家谢益辉教授发布了基于 R 语言的一个工具包 (package)^[20],并倡导利用该工具包来撰写动态的统计报告,目的是使读者能够重复研究中的统计计算过程,并得到其报告的统计结果。通过这样的报告范式,可以更好地增加研究结果的可信性,方便发现并避免一些统计方法的使用错误,并有利于后来的研究者对已发表研究的数据进行深入利用和挖掘。

在此基础之上,正如转化医学和整合医学研究哲学所提倡的,我们必须将临床数据与基础研究的数据相结合,运用适当的手段,开发合适的计算技术,使机制性解释和临床宏观表型之间能够形成结合。临床研究的前提是接受疾病的动态发展和充满复杂性的现实,起点是认真细致地描述从分子机制到临床过程的各种尺度上的复杂性^[21]。不再寄希望于仅仅使用“病死率”“并发症发生率”“住院时间”“敏感度”“特异度”等统一指标衡量临床干预或诊断措施的成败。针对复杂疾病,一旦机制和机制性过程能够与临床表型相结合,我们就能发展出疾病特异和患者个体化相结合的、真正有用的临床评估体系。实现上述目标,必须要有临床工作者和基础医学以及数学、物理学专业人员紧密结合的多学科研究团队来共同完成。

需要强调指出,由循证医学先驱们所开创的临床研究哲学是以证据为基础,结合患者需求以及考虑卫生经济的可持续发展,依然是有效的。作为新一代的医生和科学家,我们应当勇敢地直面挑战,努力学习,毫不犹豫地跨过学科之间的界限与鸿沟,从而发展出全新一代的循证医学研究范式。从 1991 年“Evidence-based Medicine”这一概念的提出至今,已经 25 年过去了。25 年前,没有人能够预见到如今计算技术和互联网技术快速发展的情形,我们相信,临床医学的一场彻底革命,也将以我们难以预计的速度到来。解决世界性医疗资源紧缺难题的钥匙,或许也正蕴含其中。

参考文献

- [1] Straus SE, Glasziou, Richardson WS, et al. Evidence-base Medicine [M]. 4th Ed. London: Churchill Livingstone, Elsevier, 2011.
- [2] 李幼平, 刘鸣. 循证医学——21 世纪的临床医学[J]. 实用医学杂志, 2000, 16 (7): 517-520.
- [3] 罗侃, 罗畅, 刘雄. 从循证医学与中医辨证论治谈中西医结合的前景[J]. 中国中西医结合急救杂志, 2002, 9 (6): 311-313.
- [4] 李银平. 循证医学与中西医结合医学[J]. 中国中西医结合急救杂志, 2001, 8 (5): 259-261.
- [5] Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations [J]. BMJ, 2008, 336 (7650): 924-926.
- [6] 蒋朱明, 江华, 詹文华, 等. 制定肠外肠内营养指南和规范的“指南”: 方法学、推荐意见分级与通过程序[J]. 中华临床营养杂志, 2006, 14 (5): 283-288.
- [7] 中华医学会重症医学分会. 急性肺损伤/急性呼吸窘迫综合征诊断和治疗指南 (2006) [J]. 中华危重病急救医学, 2006, 18 (12): 706-710.
- [8] 高戈, 冯喆, 常志刚, 等. 2012 国际严重脓毒症及脓毒性休克诊疗指南[J]. 中华危重病急救医学, 2013, 25 (8): 501-505.
- [9] 陈炜. 急性心力衰竭的诊断和治疗指南解读: 2012 年新版 ESC 急性心力衰竭诊断和治疗[J]. 中华危重病急救医学, 2013, 25 (11): 698-702.
- [10] Li SJ, Jiang H, Yang H, et al. The dilemma of heterogeneity tests in meta-analysis: a challenge from a simulation study [J]. PLoS One, 2015, 10 (5): e0127538.
- [11] Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses [J]. BMJ, 2003, 327 (7414): 557-560.
- [12] Borenstein M. Fixed-Effect versus Random-Effects Models [M]//Borenstein M, Hedges L, Higgins J, et al. Introduction to Meta-Analysis. U.S.: John Wiley & Sons, Ltd, 2009: 79-94.
- [13] Higgins, Julian PT. Cochrane Handbook for Systematic Reviews of Interventions [M]. Chichester: Wiley-Blackwell, 2008.
- [14] 刘鸣. 系统评价、Meta 分析设计与实施方法[M]. 北京: 人民卫生出版社, 2011.
- [15] Sackett D L. Evidence-Based Medicine [M]. London: John Wiley & Sons, Ltd, 2000.
- [16] Heidegger CP, Berger MM, Graf S, et al. Optimisation of energy provision with supplemental parenteral nutrition in critically ill patients: a randomised controlled clinical trial [J]. Lancet, 2013, 381 (9864): 385-393.
- [17] Casaer MP, Wilmer A, Hermans G, et al. Role of disease and macronutrient dose in the randomized controlled EPaNIC trial: a post hoc analysis [J]. Am J Respir Crit Care Med, 2013, 187 (3): 247-255.
- [18] 江华. 能量摄入模式与危重症患者死亡风险的相关性. 中华医学会第九届全国肠外与肠内营养学学术会议[C]. 北京: 中华医学会, 2015.
- [19] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 北京: 中国人民大学出版社, 2012.
- [20] Xie YH. Knitr: A general-purpose package for dynamic report generation in R [EB/OL]. [2015-06-30].
- [21] 江华, 杨浩, 曾俊, 等. 分析肿瘤学[J]. 肿瘤代谢与营养电子杂志, 2015, 2 (2): 26-30.

(收稿日期: 2015-07-30) (本文编辑: 李银平)