诊断研究≠诊断试验，做好诊断研究不容易

作者：李楠 赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/439.html

说到与疾病诊断相关的研究，可能大家最熟悉不过的就是诊断试验了。没错，诊断试验在针对研究中的确是非常重要的一种设计形式。在针对研究的四个阶段中，前三个阶段的主要形式也是诊断试验。即通过探讨不同构成特征待诊断人群中，某一分类指标的灵敏度、特异度，或是某一连续指标的ROC曲线特征，从而探讨该指标的诊断效能。

问题来了

小明是个影像科医生，他发现最近有一种新的MRI成像方法“ MRI-X ”，虽然扫描时间长很多(总共大概比常规MRI序列多出30分钟)，但是图像比常规MRI更清晰，看清某特定病灶B的可能性更高，但是与诊断病灶B的标准方法MRI造影还是差了一些。小明想，这个方法可能具有一些优势，那我们是不是应该做个针对性研究呢?如何设计才能得到更好的结果呢?

这个问题看起来好像已经很明确了，但是如果小明已经开始思考上面两个问题，并且构思如何实施，可能还真为时尚早。因为最核心的问题好像并没有解决：这个研究的选题到底是什么?

诊断研究的临床问题

在设计与病因、治疗、预后相关的研究时，我们可能会花费不少时间在琢磨选题上面。但是当到了针对研究的时候，很多研究者往往抱着这样的认识：“ 问题已经细化到这个程度了，不就是看这个方法诊断的对不对么，可直接开展了 ”。其实不然，就拿小明的例子来说，看起来是个诊断问题，但是作为辅助科室医生的小明，有没有想过“ MRI-X ”对应的临床需求是什么呢?

在针对研究中，诊断方法的诊断效能与否本身并不是研究要解决的最终目标。一个方法诊断效能再高，在临床中找不到合理的应用场景，都不能认为这一诊断试验是有价值的。比如对于小明的方法来说，如果病灶B是决定患者手术与否的关键征象，常规MRI具有足够的灵敏度，但是特异度不足，意味着能够满足筛查的需求;同时MRI造影能够确诊是否有病灶B。此时如果再提出特异度由于常规MRI(代价是时间更长)，但又远不及金标准MRI造影的方法，似乎就没什么必要了。因此，很可能最终研究的结果和小明预期的一样，但是这一的结果一出现就变成了某些院士口中的“ 垃圾论文 ”。问题出在哪儿呢?关键的问题就是，作为影像学诊断专家的小明并没有跳出自己的影像科，没有站到临床需求的全局上看待这一指标到底可以解决什么临床问题。

因此，对于一个诊断试验的选题，最核心的问题之一是：这一诊断方法是否有其恰当的临床应用场景?

如果答案是没有：那么完全可以放弃这一研究;

如果答案是不清楚：弄清楚可能的应用场景，就是完善选题最好的出发点;

如果答案是清楚：那么只要后续的的设计细节一直围绕这一场景，那针对性研究本身的价值就能够得到保障了。

如何考虑针对研究的人群代表性

在开展针对性研究的时候，从方法上看，要么选取诊断明确的患者或非患者(如针对研究第一阶段)，要么选取不同患病可能性的患者(如针对研究的第二阶段)，要么选取真实的待诊断人群(针对研究的第三阶段)。看起来文字似乎都不难理解，但是操作起来就不会出问题么?

其实，确保合理的人群代表性至关重要，这涉及到针对研究的结果到底能够外推到什么人群中去。以看似已经具备了较好代表性的第三阶段研究来说。我们定义的“ 真实待诊断人群 ”是否合理至关重要。还拿小明的例子来说事儿，小明认为“ 临床中所有开具检查，需要进行B病灶MRI常规筛查的患者 ”我们的目标人群。真的是这样么?当问好出现的时候，相比您已经认识到问题的复杂性了。没错，对于MRI-X这个“ 不上不下 ”的诊断方法来说，似乎临床上用它来完全替代常规MRI的可能性并不大。毕竟每例患者都额外增加30分钟的检查时长，这对很多机构来说并不现实。类似这样的例子还有很多，比如这种“ 不上不下 ”的检查还是有创的，或者价格奇高……这些都限制了该方法作为临床常规应用的方法。因此从上面提到的应用场景角度考虑，目前常规检查方法的待诊断人群似乎并不一定是将来该方法的真实应用人群。

对于这类诊断方法，他们可能的应用人群显然要比常规方法应用人群更偏向“ 患者 ”，否则不会愿意负担额外多出来的这部分费用、时间或是创伤。因此我们考虑将来结果外推性的时候，也应该预先估计到这一点，从设计层面上予以考虑。即便是以所有常规方法检查人群作为研究对象，我们至少也应该考虑到高风险人群(潜在的真实应用人群)这个亚组的结果如何，以及与常规方法的优势大小如何。

我们提出的针对研究真的有必要开展么?

当然上面提到的都是假定真的要开展这一研究。其实在针对研究的顶层设计时，还有一个至关重要的问题没有回答：我们选定的这个研究方向到底有没有必要走下去。

如果我们面临小明的情景，很可能最终会选择放弃这一研究。因为从成本上看，该研究的成本显然更高，从收益上评价，对于患者MRI-X并不能作为最终诊断，患者少不了还是要做个MRI造影。因此即便是做下去，得到结果，很可能也没什么临床价值、无法在临床推广该应用。当然，回到第一个观点就是，该诊断方法的优势点运气不好，几乎找不到临床应用的场景。

那么小明真的走投无路了么?那倒也是不见得，有时候我们偶然也会发现一些不大的临床应用场景，可以把这种方法完美的嵌入进去。比如，对于某些高危人群来说(比如年龄大于65岁，或是有XX病史)，B病灶被小明提出的MRI-X检出的可能性极高，特异度几乎达到了和MRI造影相当的水平。此时我们完全可以用MRI-X代替“ MRI筛查+MRI造影 ”的组合，直接作为最终诊断。

看上去这种规律的发现需要运气，实际上不然，这里面涉及到了MRI-X这一指标在不同人群中诊断效力不同的现象。这一现象的本质有可能是背后的某种交互作用。因此从，当我们看到过类似

的文献报道，或是自己前面第一、二阶段的研究中有类似的现象，完全可以针对这一特殊人群设计诊断试验，探讨在较小人群上，诊断方法的效能如何。虽然看上去外推的人群小了，但是至少临床应用的场景更明确了，不管是不是苍蝇肉，但至少算是找到肉了吧。

同样，很多研究者在一、二阶段甚至第三阶段的诊断研究中，会探讨某些患者/疾病特征对诊断效能的影响，但背后的初衷却没说清楚。其实后续发现苍蝇肉的可能就是这类分析的价值。更好的策略是，在一二阶段诊断研究的效能影响因素分析中，预先想好临床中的应用场景(结合现有诊断方法的不足)，结合未来的应用提出诊断效能影响因素的分析点。

只有第四阶段结果才能决定诊断方法的去留

在诊断研究中，绝大多数待评价的方法，其研究都止步于前三个阶段。原因很简单，经过前三个阶段的研究，关于该方法诊断效能的评价已经完成了，完全可以用于制定临床常规、临床指南以及选择诊断方法的需求了。因此，很多研究者并不重视诊断研究的第四阶段。而第四阶段的目的是什么呢?就是探讨某一诊断方法，在临床中应用后是否能进一步改善临床结局，带来实际收益。

让我们举一个例子。某肿瘤M诊断较容易，但具体分型却不容易明确。如果经过穿刺后基因检测，能够很好的明确肿瘤分型。但是问题来了，这一组织基因检测应该在临床中常规应用么?

我们再给几个不同的背景：

1、如果肿瘤M只有一种可选治疗方案，无论分型如何都用同一种方法治疗。此时该诊断是否有存在的必要?

2、如果针对不同分型，有不同的治疗方案。但是当基因诊断的方法出现后，我们发现根据该基因调整治疗方案后，并没有改变患者的总体预后。此时该诊断是否有存在的必要?

答案当然都是“ 没有 ”。从第2个背景不难看出，如果有一种方法能够对患者作出某个状态的判断，但是不管知道该状态与否、按该状态调整治疗策略与否，都不会带来结局的差异，那么对该状态的诊断就没有实际临床价值。这样的诊断也就是所谓的“ 过度诊断 ”。而第四阶段的诊断研究，恰恰就是从临床中找出这类过度诊断的手段。而第四阶段诊断研究往往也不是通过诊断试验的方法完成的，而是可以通过RCT的方法进行评价，探讨诊断与否的患者是否有临床结局的差异或是卫生经济学差异。