PROBAST：一个用于评价预测模型研究偏倚风险和适用性的工具：说明和详述文件

Karel G.M. Moons, PhD*; Robert F. Wolff, MD*; Richard D. Riley, PhD; Penny F. Whiting, PhD; Marie Westwood, PhD; Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Jos Kleijnen, MD, PhD; and Sue Mallett, DPhil

在卫生保健领域，预测模型是在一个个体上，使用多个预测因子来推估其现有（诊断模型）或将来发生（预后模型）某健康状况或疾病的可能性。

近些年，预测模型相关文献已经变得越发普遍。针对同一个临床结局或目标人群，往往会存在有多个预测模型。卫生保健人员、指南制定者及政策制定者，经常无法确定可以使用或推荐哪个预测模型，要在哪些人或者医疗场所中使用或推荐这些模型。因此，该类研究的系统评价越发变得亟需且必要，该类系统评价的制作也越来越多。

预测模型系统评价的一个关键要素，是评价偏倚风险以及模型在预期目标人群和临床场所使用时的适用性。为了能在这个过程中帮助到系统评价员，本文作者开发了 PROBAST（预测模型偏倚风险评估工具），可用于评价以建立、验证或更新（如，拓展）诊断或预后预测模型为目的的临床研究。

PROBAST 是由该领域的一个专家组，经共识形成的标准流程而制定的。该工具涵盖了涉及四个领域的 20 个信号问题（即研究对象、预测因子、临床结局和数据分析）。此说明和详述文本将解释纳入每个领域和信号问题的原因，并指导研究人员、系统评价员、读者及指南制定者该如何使用它们来评估偏倚风险和适用性。所有这些均基于来自不同研究主题的案例给出了讲解。最新版本的 PROBAST 清单、随附文件和填写示例都可以从 www.probast.org 网站下载。

Ann Intern Med. 2019; 170: W1-W33. Doi:10.7326/M18-1377

作者单位见文末

* Moons 和 Wolff 博士对该文做出同等贡献。

在卫生保健中，预测模型是为了预测一个个体是否存在一个特定的临床结局，如疾病（诊断模型），或是否将来会发生该临床结局（预后模型）（1-6）。诊断模型可用于引导患者做进一步检查、用于启动治疗方案或使患者知晓其现况。预后模型可用于支持预防性生活方式改变、治疗性干预或监测策略等有关的决策制定，或在随机试验设计和数据分析中用于划分风险区组（7,8）。预测模型的潜在用户包括卫生保健执业人员、政策制定者、指南制定者、病人和普通公众。

在医学文献中，有成千上百的研究都致力于建立和验证预测模型。针对同样的目标人群和临床结局，通常也可以见到多个预测模型。例如，有超过 60 个模型可用于预测乳腺癌预后（9）；在妇产科领域，就有超过 250 个模型（10）；有近 800 个模型可用于预测心血管疾病患者的临床结局（11）。预测模型如此井喷的现象，只会随着个体化或精准医疗的发展而进一步加剧。

系统评价被认为是处理治疗类随机对照试验以及诊断准确性试验的最可靠证据形式（12）。在现今这个个体化和精准医疗时代，我们对于预测模型系统评价的研究兴趣正在极速提升。对此，Cochrane 预后研究方法学组的成立，即为支持预后研究（包括预后模型研究）系统评价的制作（13，14）。支持预测模型系统评价制作的指导性文献，已被发布（见表 1），包括文献检索策略的制定（15，41-43），系统评价研究问题的制定（16，17），数据提取（16）和 meta 分析（17，22-25，40，44，45）。

在任何类型的系统评价中，偏倚风险评价都是一个基本步骤。研究设计、实施和数据分析上的不足，会导致研究估计值存在一定偏倚风险，意即：研究结果有出现偏差或失真的风险。在解读一项系统评价的结果时，相比于偏倚风险高或风险不清楚的原始研究，读者可以得出更稳健的结论如果该系统评价是基于低风险的原始研究（46）。另外，检索收集与系统评价所关注的临床场所及目标人群最贴切的原始研究，也是极其重要的，这是考量原始研究之于系统评价研究问题的适用程度。因此，我们开发了 PROBAST（预测模型研究偏倚风险评价工具），以解决适于评价预测模型原始研究偏倚风险和适用性之专用评价工具匮乏的问题。

PROBAST 由 4 个领域，共 20 个有助于评价风险偏倚的信号问题构成（39）。其结构和评价体系类似于评价随机对照试验的修订版 Cochrane 偏倚风险评价工具（ROB 2.0）、评价

诊断准确性研究的 QUADAS-2（诊断准确性研究质量评价工具）和评价系统评价偏倚风险的 ROBIS（37，47，48）。尽管 PROBAST 是专为预测模型研究系统评价而设计，它同样可以用作对预测模型原始研究进行批判性评估的一个常规工具。

表一、制作预测模型研究系统评价的指导性文件

用途	**指导文件**
原始研究的报告	诊断和预后预测模型研究的透明报告规范（TRIPOD）（7，8）
定义系统评价研究问题及制定符合标准*	定义系统评价研究问题和设计预后研究系统评价的指导文件（CHARMS）（16，17），另见表四中的诊断准确性研究系统评价计划书之指导文件（18，19）
文献检索*	预测研究的检索滤器（15，41-43） https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/filters-to-identify-studies-about-prognosis 诊断准确性研究检索（20）
文献筛选和数据提取*	预测模型研究数据提取和批判性评价的指导文件及清单（CHARMS）（16） 诊断准确性研究指导文件（19，21）
纳入研究的偏倚风险和适用性评价*	PROBAST（39）
数据分析和 meta 分析*	预测模型研究的 meta 分析（17，22-25，40，44，45） 诊断准确性研究 meta 分析（26-33）

结果解释和形成结论*	PROBAST（39）和预测模型研究系统评价结果解释的指导性文件（17，22-24，40） 诊断准确性研究系统评价结果解释的指导性文件（19）
系统评价的报告	系统评价和 meta 分析透明报告标准（PRISMA 和 PRISMA-DTA）（34-36）
系统评价的偏倚风险评价	ROBIS（37）

CHARMS = 预测模型研究系统评价批判性评价和数据提取清单; DTA = 诊断准确性；PRISMA = 系统综述和 Meta 分析报告条目清单；PROBAST = 预测模型偏倚风险评估工具; ROBIS = 系统评价风险偏倚评价工具；TRIPOD = 用于个体预后或诊断的多因素预测模型透明报告。

*制作步骤与 Cochrane 系统评价的基本步骤相符

在此，我们会讲解 PROBAST 各领域及信号问题的合理性，如何评价它们，如何总结各领域的偏倚风险和总体偏倚风险，以及如何总结原始研究之于系统评价研究问题的适用性。在我们的网站（www.probast.org）上，有五个已评价好的、来自不同领域的范例可帮助理解该评价过程。因为该领域的研究较活跃，在必要时，我们会更新该工具、范例及其附属指导文件。从上述网站上，读者应该可以随时下载到最新版的 PROBAST。

PROBAST 关注的问题

PROBAST 是设计用于评价以建立、验证或更新（如扩展）诊断或预后多变量预测模型为目的的原始研究（文本框一和二）。多变量预测模型指，由旨在评估一个个体现有或将来发生某特定临床结局可能性或风险的两个或多个预测因子（如年龄、性别、体征、临床症状、疾病分期或生物标至物）而构成的任意组合或方程式（1，4，6-8，49，50）。预测模型的其他名称还包括风险预测模型、推测模型、预测指数或工具和风险评分（1，3-8，49-51）。

文本框一. PROBAST 考虑的诊断和预后模型研究类型*

预测模型建立（无外部验证）：这类研究致力于用一个特定的建模数据集来建立预后或诊断预测模型。其目的是识别和在研临床结局相关的重要预测因子，用某种形式的多因素分析给每个预测因子分配权重（例如，回归系数），建立一个可用于个体化预测的预测模型，并量化该模型在建模数据中所表现出的预测性能。有时候，模型建立研究可能也关注将新预测因子添加到此前已知的预测因子。过度拟合在任何预测模型研究中可能都会出现，尤

其会发生在数据集较小的时候。因此，建模研究应该包含某种形式的重抽样或“内部验证”（内部验证：是一份数据被同时用于模型建立和模型内部验证），例如使用自助抽样法或交叉验证。这些方法可以量化所建模型上出现的预测性能乐观偏差（偏倚）。

预测模型建立和外部验证：这类研究具有和上一类研究相同的目的，但是会用建模数据之外的（外部）数据（例如，来自于不同研究对象）来量化模型的预测性能。这些数据可以是由相同调查人员收集、通常会用相同的预测因子和临床结局定义以及测量方法，但却是从稍晚的时间段上抽样而来（时序验证）；也可以是由其他调查人员在另一个医院或国家、有时数据的收集会使用不同的定义和测量方法（地理验证）；还可以出自于相似的研究对象，但却有意选择了另一种临床场合（例如，一个模型在二级卫生保健系统中被建立，但却在初级卫生保健系统的同类研究对象中被验证）；甚至也可以用不同类型的研究对象（比如，一个模型在成年人中建立，但在儿童中验证）。随机将一个数据集拆分为建模数据和验证数据的做法，常会被错误地看作是一种外部验证形式，但实际上这是一种低效能的内部验证方法；因为用这种方式构建的两个数据集之间的差异是随机的，且建模的样本量也会被降低。当一个正在进行外部验证的模型表现出的预测性能较差时，可以按验证数据集对模型做调整（或更新现有模型，例如通过再校准基线风险或危险，或调整模型中预测因子的权重），甚至可以添加新的预测因子来扩展现有模型。这两种情况下，在现有模型的外部验证之后，事实上已经建立了一个新的预测模型。

预测模型外部验证：这类研究的目的是使用不同于建模样本的外部数据（即，来自不同研究对象的数据），来评价现有模型的预测性能。

*改编自 TRIPOD（个体诊断和预后的预测模型研究之透明报告规范）和 CHARMS（预测模型研究数据提取和批判性评价的指导文件及清单）（8，16）

文本框二. 诊断和预后预测模型研究间的差异

诊断预测模型是为了评估一个个体**现在**有否出现一个目标症状、或经参考标准（在PROBAST 中被称为临床结局）所确诊疾病的可能性。在诊断预测模型研究中，预测是针对于一个已经出现的临床结局，因此首选研究设计是横断面研究。然而,有时候为确定目标症状或疾病是否在做出预测时即已出现，随访也会被用作参照标准的一部分。

预后预测模型评估一个个体**在将来一段时间内**（数分钟至数小时，数天，数周，数月或者数年）会不会经历一个具体的临床事件或结局；该相关关系都是纵向性的。

尽管所预测临床结局具有不同的时机，诊断和预后预测模型仍有许多共同点，包括以下几个：

- 临床结局类型往往是二分类(目标症状是否出现;或临床事件在将来是否会发生)。
- 主要兴趣点都是，以告知个体和指导决策制定为目的，以多个预测因子为基础，来评估一个临床结局现存或将来出现的可能性。
- 在建立或验证多因素预测模型的过程中，会遇到相同的挑战。它们都可以用相同指标去评价模型预测性能，尽管诊断模型常会从对预测性能的评价外延到对临床上相关阈值的关注。

诊断和预后模型研究间还存在多个术语上的差异，包括以下几点：

诊断预测模型研究	预后预测模型研究
预测因子	
诊断试验或检验指标	预后因子或预后指示因子
临床结局	
用于评价或验证目标症状的参照标准	临床事件(一个事件将来是否会发生),结局

	事件的测量
结局评价结果缺失	
部分验证，失访	失访和删失

PROBAST = 预测模型偏倚风险评价工具

诊断和预后模型

诊断预测模型用于评估现在出现某特定临床结局（即目标健康状态）的可能性。诊断预测模型研究通常招募具有疑似（但暂不确定）目标健康状态的个体。

预后预测模型用于评估将来发生某临床结局或事件（如死亡、复发、疾病并发症或治疗反应）的可能性。预测的时间跨度少则数小时（如在术前预测术后恶心呕吐会否发生），多则数年（如预测冠状动脉血管事件发生的终生风险）。尽管许多预后模型研究会招募具有明确诊断的患者，这并不意味着必须将诊断已知作为评价预后的起始节点，比如有的模型即是为了预测健康孕妇将来是否会发生糖尿病（52），或者预测健康人群将来是否会发生骨质疏松性骨折（53）。因此，和 TRIPOD 声明一样，PROBAST 将那些旨在预测有某临床事件潜在风险的人群将来出现该结局可能性的模型统称为预后模型。

对于预测因子和临床结局，诊断和预后模型研究常常会使用不同的术语（见文本框二）。肿瘤文献通常将预后和指示性模型区别对待，这种指示性模型用于鉴别那些存在差异性治疗疗效的患者（54）。这些指示性模型并不在该文的研究范围内。

预测因子、临床结局和建模方法分类

PROBAST 可用于评价任意一种以个体化风险预测为目的的诊断或预后预测模型，不管该模型中纳入了哪些预测因子，想要预测什么临床结局，或者采用怎样的模型建立、验证或更新（如扩展）方法。

预测因子的类别范围可以从人群流行病特征、病史和体格检查结果，一直延伸到影像学检查结果、电生理、血、尿或机体组织检测、和疾病分期/特征，再到组学[译者注：如基因

组学]和其他新兴生物测量结果。预测因子也可被称为协变量、风险因子、预后因子、决定因素、检测指标或独立变量（4，6-8，49，50，55-57）。

PROBAST 会将候选预测因子和最终模型中纳入的预测因子区别对待（57）。候选预测因子是那些被认为有可能能够预测现有临床结局（诊断）或临床事件将来发生与否（预后）的全部变量，也即在一项研究中所评估的全部变量，不管它们是否最终被纳入到多因素模型中去。

PROBAST 主要适用于二分类及事件发生时间临床结局的预测模型，因为该类型的临床结局在医学中最常见。然而，该工具也可用于评价非二分类临床结局相关的预测模型，例如连续性评分（如疼痛评分或胆固醇水平）或分类临床结局（如 Glasgow 昏迷指数）。在 PROBAST 中，除了一些与连续型临床结局不相关的信号问题，如变量平均事件发生数和某些模型性能测量指标（如 c 统计量），几乎所有 PROBAST 信号问题都同等地适用于连续型临床结局和分类型结局。

预测模型通常是用回归建模技术，例如逻辑回归或生存模型。而预测模型的建立或验证也可以使用非回归技术，如神经网络、随机森林或支持向量机。随着常规医疗数据（及大数据）应用的兴起，更多的建模技术也变得流行起来，其中包括机器学习和人工智能模型。使用回归的研究和应用其他建模技术的研究之间，其主要差异点在于数据分析方法。由非回归建模技术开发的模型，常常会在数据量较小时出现较高的过度拟合风险；并且缺乏透明度也会影响到模型的适用性和可用性。在讨论修改 PROBAST 信号问题的相关部分，我们将会提供相应建议，以指导该如何调适 PROBAST 来处理其他类型的临床结局和建模技术。

系统评价研究问题分类

PROBAST 可用于各种研究问题的预后模型系统评价。对有些系统评价研究问题而言，所有预测模型研究（包括预测模型建立和验证研究）都能适用；然而，有一些问题是只和模型验证研究相关。文本框三提供了一些可能适于使用 PROBAST 的、预后和诊断预测模型相关的系统评价研究问题案例。CHARMS（预测模型研究系统评价批判性评价和数据提取清单）及表格二提供了更详尽的解释，指导该如何构建一个清晰明确的预测模型研究系统评价研究问题（16，17）。

文本框三. PROBAST 适用的系统评价研究问题范例

一个明确的目标人群:

为预测一般人群中 2 型糖尿病新发风险而建立或验证的所有模型的系统评价（58）。

为用于已确诊患有急性中风的患者而建立或验证的所有预后模型的系统评价（59）。

一个明确的临床结局:

为诊断——不管其患者类型——有没有深静脉血栓而建立或验证的所有诊断模型的系统评价（60）。

为预测任意患者日常活动丧失而建立或验证的所有预后模型的系统评价（61）。

一个特定临床领域:

在生殖医学领域建立或验证的所有预后模型的系统评价（62）。

在创伤性脑损伤急诊领域建立或验证的所有预后模型的系统评价（63）。

一个明确的预测模型:

EuroSCORE（一个用于预测心脏术后手术死亡风险的模型）在其所有外部验证研究中所报告的预测性能系统评价（64）。

比较那些可预测全人群之中年个体发生心血管疾病风险的各种预测模型，在所有验证研究中的预测性能的系统评价（65）。

一个明确的预测因子:

当被添加进 Framingham 风险模型时，C 反应蛋白所增添的额外预测价值的 meta 分析（66）。 颈动脉成像对一个现有心血管风险预测模型，所增添的额外预测价值的 meta 分析（67）。

预测模型系统评价可以解决的研究问题有许多。以上是 PROBAST 适用的不同类型系统评价范例。EuroSCORE = 欧洲心脏手术风险评估系统；PROBAST = 预测模型偏倚风险评估工具

表格二. PICOTS

条目	备注
1. 研究对象	定义会用到待评估预测模型的目标人群
2. 试验指标	定义待评估的预测模型
3. 对照组	若适用，定义是否评价其他预测模型，并与参照模型做比较
4. 结局	定义由待评价预测模型所预测的临床结局
5. 时间要素	定义在什么时候或时间节点（如，病情检查时）可以用到待评价预测模型，以及在哪个期间的临床结局才被预测（后者适用于预后模型的情况）
6. 研究场景	定义待评估预测模型的既定临床研究场景和模型的用途

PICOTS = 研究对象、试验指标、对照组、临床结局、时间要素和临床背景

* 构建系统评价研究问题的关键信息已经在此前的指南（16，17）中做过说明。PICOTS 是对治疗类干预研究系统评价中常用的 PICO 系统的一个改编，其中加入了时间要素（使用预测模型的时机和预测所跨时间段）和临床背景（17）。

预测模型研究的类型

PROBAST 可评价那些讨论对个体做诊断和预后预测（即个体化预测）之多因素模型的原始研究（见文本框一）。这包括（1）建立新的预测模型；（2）建立和验证同一个预测模型；（3）验证已经存在的预测模型；（4）建立新预测模型，且和已经存在的预测模型做比较；（5）更新（如调整模型中各因素系数）或扩展（如加进去新的预测因子）已有的预测模型；和（6）结合多个上述目的的研究。

PROBAST 不适用于评价预测因子识别的研究，这类研究的目的通常是鉴别与临床结局相关的预测因子，而不是建立用于个性化预测的模型（16，68，69）。QUIPS 工具已被开发用于此类研究的评价（70）。

PROBAST 也不适用于以量性评估使用和未使用某预测模型（或替代模型）对研究对象健康结果影响的疗效比较研究。这种比较模型效益的研究能使用随机或非随机临床研究设计（8，55，71-74），因此可选用适用于随机试验或非随机临床研究的偏倚风险评估工具（47，75）。

另一个偏倚风险评估工具（QUADAS-2）已经被开发用于评价诊断试验准确度研究（48）。然而，需要注意的是，有些诊断试验准确性研究包括一个诊断预测模型而不是诊断试验。在这种情况下，可酌情考虑用 PROBAST。

偏倚风险和适用性

偏倚风险 （ROB）

偏倚通常被定义为在研究中出现的系统错误，导致研究结果失真或缺陷，并妨碍研究的内部有效性。对于预测模型的建立与验证，虽然只有有限的实证证据显示最重要的偏倚来源，已知可导致一个研究存在偏倚风险的因素确实存在。我们认为，如果研究设计、实施或分析上的不足，导致一个模型的预测性能评估出现系统性偏差时，偏倚风险即会发生。模型预测性能通常可用校准度和区分度的测量指标进行评价，有时（尤其是在诊断模型研究）也可用分类指标（见文本框四）（8）。试想该如何设计、实施和分析一个假设的、方法上严谨的预

测模型研究，可有助于理解模型预测性能评估上的偏倚。

文本框四. 预测模型性能指标

*校准度*反映模型预测和所观测到的临床结局之间的一致性。校准度常首选以图示法进行呈现：实际观测风险画在 y 轴，和 x 轴上的预测风险做对比。该图通常将预测风险值范围的每十分之一做为一个单位制作，并且可以沿着整个预测风险值范围添加一条平滑线（LOWESS，局部加权回归散点平滑法），从而增强显示此图。这既适用于经逻辑回归而建立的预测模型（59，76，77），也可用于生存回归建立的模型（78，79）。校准图可呈现整个预测风险值范围上的任何校准错误的方向和大小，这可以和校准曲线斜率以及截距一并呈现（79，80）。校准度通常通过计算 Hosmer-Lemeshow 拟合优度检验来做评价；然而，该检验方法对于校准度较差模型的适用性有限，且对分组数目和样本量比较敏感：此检验通常对于小样本量数据不显示统计显著性；但对大样本数据几乎总是有显著性。仅报告 Hosmer-Lemeshow 拟合优度检验，但缺少校准图或表格去对比结局事件预测和观察值频率的研究，是没有提供任何和预测风险精确度有关的信息的（见信号问题 4.7）。

*区分度*是指一个预测模型在个体之间区分现在有或无（诊断）、或将来发生或不发生（预后）临床结局事件的能力。对于逻辑和生存回归，最常用和最被广泛引用的区分度指标是一致性指数（c 统计量），也等同于逻辑回归模型的受试者工作特征曲线下的面积。

*校准度*和*区分度*指标应该考虑到待评价的临床结局类型。对于生存模型，研究者应该恰当地考量事件发生时间数据和删失数据（50，81，82），比如使用 Harrell 的 c 统计量或 D 统计量。

其他许多模型预测性能指标也是可用的，包括呈现模型分类能力（如，敏感度和特异度）以及再分类参数（如，净再分类指数）的指标（80）。这些指标只有在模型预测风险值范

围中引入一个（或多个）阈值后，才可被评估。分类指标常常被用于诊断试验准确性研究，较少用在预测模型研究中。为评估模型分类指标而对预测风险值范围做分类化处理，会导致信息的丢失，因为模型的预测风险值的整个范围没有被完整地使用。使用阈值可以让区分度报告潜在临床相关的临界值，而不是横跨所有的（尽管可能缺乏临床重要性）潜在临界值。然而，引入风险阈值意味着所选阈值适用于临床实践，但通常并不是这样，因为这些阈值多是基于数据诱导而得出的，所以会得出有偏倚的分类参数（83）。相反，作者应该基于预先拟定（风险）阈值的总体原则来评价这些指标（见信号问题 4.2），以避免在阈值上做多重检验和因数据驱使而导致阈值的潜在的选择性报告。

还有许多其他模型预测性能指标，包括净收益指标和决策曲线分析，但是这些不常见于预测模型研究（84）。这些指标中，有许多是将风险阈值和假阳性以及假阴性结果整合联系起来。

当在建模数据中评估时，所有上述模型性能指标，大多都会因过度拟合或选用较好的阈值而呈现出乐观偏差；因此，应该用自助重抽样法或交叉验证法对它们做评估（见信号问题 4.8）。

适用性

当研究对象、预测因子或临床结局与系统评价研究问题中所要求的这些要素不同时，就会有这样一种担心[译文称之为“*关切*”]：原始研究是否适用于系统评价研究问题。例如，当预测模型研究中的研究对象与系统评价研究问题中所定义的目标人群，分别来自不同的医疗环境时，该关切即可出现（见表二）。在二级医疗场所建立起来的一个预测模型，在初级医疗场所中可能会有不同的区分度和校准度，因为医院里的患者通常比初级医疗场所的病情更重（71，85）。

即使原始研究的研究对象、预测因子和临床结局与系统评价研究问题能直接匹配，这些研究的适用性关切仍可能存在。更不要说，系统评价的纳入标准通常要比系统评价研究问题的确切关注点更加宽泛。

在此，不可将偏倚及适用性关切与某特定模型在各验证研究之间的预测性能异质性相混淆，该异质性可能来自于，比如，病例混合、或差异较大的疾病严重程度（17，40，44）。使用 meta 分析方法研究异质性时，可以使用相应的预测区间报告特定模型在验证研究中的预测性能差异（17，40）。

例如，在一个包含了所有验证研究的某特定预测模型系统评价和 meta 分析中，对偏倚风险和适用性的评价，还可以参考对验证研究之间模型预测性能异质性的评估。可想而知，在其他研究中得到验证的某模型，其预测性能，可能会因（比如）研究对象特征、医疗环境、地理位置或时间的差异而不同。这并不意味着在原始验证研究中存在着偏倚风险，也不表示存在适用性问题；这仅仅反映了一个特定模型其预测性能在各研究之间所存在的预期差异。研究间异质性的潜在来源，可以使用 meta 分析进行探索，或者依据研究间存在着差异的临床特征来分组呈现（17，40，44）。

实施 PROBAST 评价价

PROBAST 评价共分四步进行（见表格三）。PROBAST 评价应该要对符合系统评价研究问题的每一个模型去做。我们将通过多个案例来阐述与偏倚风险和适用性相关的关键问题（见表格四）（85）。这些示例涉及到诊断和预后模型；考虑到了不同的医疗领域、研究设计以及预测因子和临床结局类型；并包括有模型建立和验证研究。这些案例的评价可以在 www.probast.org 网站上找到。

表格三. PROBAST 的四个步骤

步骤	任务	何时完成
1	明确系统评价的研究问题	对每个系统评价各评价一次*[，均进行一次步骤 1 相应的任务]*

2	区分预测模型评估的类型	对每个待评价研究中的每个预测模型以及每个相关临床结局各评价一次*[，均进行一次步骤 2 相应的区分]*
3	评价偏倚风险和模型适用性	对研究中的每个预测模型的每次建立与验证各评价一次*[，均进行一次步骤 3 相应的评价]*
4	总体评价	对研究中每个预测模型的每次建立与验证各评价一次*[，均进行一次步骤 4 相应的评价]*

PROBAST = 预测模型风险偏倚评价工具

表格四. 示例论文

作者，年代（参考文献）	研究主题	预测模型研究的类型		数据来源	研究对象	预测因子类型	结局指标	研究总样本量（发生临床事件的研究对象数），n	模型预测性能	
		建立/验证	诊断/预后						区分度	校准度
Aslibekyan, 2011 (86)	心肌梗塞（MI）	建立和验证	预后	非巢式病例对照研究，来自哥斯达黎加中央山谷的人群（1994-2004）	首次非致命急性心肌梗塞病案和没有非致命心肌梗塞的对照组患者	病史采集，体格检查	首次非致命性心肌梗塞	4547（1984）	有	无
Han, 2014 (87)	严重创伤性脑损伤（TBI）	验证	预后	队列研究，新加坡一家医院（2006 年 2 月-2009 年 12 月）	诊断有严重创伤性脑损伤的患者（GCS ≤ 8）	病史采集，体格检查，实验室指标	三个临床结局：14 天和六个月时的死亡事件，	300（143 例 14 天死亡事件；162 例六个月死亡事件；213 例六	有	有

						和 CT	以及六个月时的不良事件	个月不良事件）		
Oudega, 2005 (85)	深静脉血栓（DVT）	验证	诊断	前瞻性横断面研究，荷兰的 110 家初级卫生保健诊所（2002 年 1 月-2003 年 3 月）	有深静脉血栓症状或体征的患者	病史采集，体格检查	深静脉血栓	1295（289）	无	无
Perel, 2012 (88)	创伤性出血	建立和验证	预后	建模：随机对照试验，40 个国家的 274 家医院（日期未报告）	建模：创伤和大出血患者或 8 小时内有大出血风险的患者	病史采集，创伤类型，生理学检查	28 天内的死亡事件	建模：20127（3076）	有	有
				验证：登记数据，英格兰和威尔士的 60%创伤医院（2000-2008）	验证：创伤患者和估计失血量≥20%的患者			验证：14220（1765）	有	有

Rietveld, 2004 (89)	感染性结膜炎	建立	诊断	队列研究，荷兰的 25 家初级卫生保健诊所（1999 年 9 月-2002 年 12 月）	有感染性结膜炎体征的患者（眼睛发红和黏浓性渗出液或眼睑粘连）	病史采集，体格检查	阳性细菌培养	184（57）	有	有

CT = 计算机断层扫描； GCS = Glasgow 昏迷量表

第一步：明确系统评价的研究问题

首先，系统评价员需可以参考以下几个要素来明确他们的系统评价问题：预测模型的预期用途、模型中纳入的预测因子、目标人群以及待预测的临床结局。结构化地呈现这些要素有助于评价适用性。正如表格二中所总结的，已有的指导性文件（CHARMS 清单）可以帮助系统评价员定义一个清晰且明确的系统评价研究问题（16）。

每个系统评价都需要完成一次第一步的评价。表格五提供了一个参考示例。

表格五. 步骤一在 Perel 示例论文中的使用示范*

标准	明确系统评价的研究问题
模型的使用意图	预后；在医院急诊室就诊时[译者注：使用]
研究对象，包括选择标准和临床背景	8 个小时或以内有大出血风险的创伤患者，在医院急诊室就诊
预测因子（预测模型中使用到的），包括预测因子类型（如，病史、临床检查、生化指标、影像检查）、测量时间、具体的测量问题（如，对特定医疗设备的需求和禁忌）	病人人口学特征、生理学变量、创伤特征、创伤时长，全部都在医院急诊室就诊时测量 影像学检查结果在就诊 4 小时内可获得
待预测的临床结局	创伤 28 天内的死亡

*参考文献 90

第二步：区分预测模型评估的类型

在第二步，为了和 PROBAST 中的相关信号问题相衔接，要识别出预测模型评价的研究类型。当单个研究同时报告了某特定模型的建立以及验证（见文本框一），或验证和模型调整（或扩展）时，需要对它们分别做 PROBAST 评价。模型扩展（即现有模型中加入新的预

测因子）应该被视为一个新的模型建立去评价。

对系统评价中评估的每个预测模型，都要做一次第二步操作；表格六提供了示例可供参考。

表格六. 步骤二在 Perel 示例论文中的使用示范*

预测研究类型	要填的 PROBAST 方格	恰当处打勾	预测模型研究类型定义
仅以模型建立为目的	建立		预测模型建模研究（无外部验证）。这类研究可以包含模型的内部验证，例如使用自助抽样法（bootstrapping）和交叉验证技术（cross-validation techniques）
模型建立和验证	建立和验证	√	在同一篇文献中，既有预测模型建立，又包含在其他研究对象对此模型所做的外部验证
仅以模型验证为目的	验证		基于其他研究对象对已知（先前建立的）预测模型进行的外部验证

PROBAST = 预测模型偏倚风险评价工具

*参考文献 90

第三步：评价偏倚风险和适用性

评价偏倚风险

基于四个领域和相关信号问题，PROBAST 提供了一个结构式方法来找出潜在的偏倚风险。信号问题需要基于事实去做判断，在偏倚风险评价中，每个问题都可以被回复成“是（Y）”、“可能是（PY）”、“否（N）”、“可能否（PN）”或“缺少信息（NI）”。在 PROBAST 中，所有信号问题均经过合理地遣词造句，“是”即表示低偏倚风险，“否”表示高偏倚风险。评级“可能是（PY）”和“可能否（PN）”，在既有信息不足以使我们很确信地回答“是”或“否”时可以使用。和其他风险评价工具一样，“是”实际上与“可能是（PY）”（同样地“否”与“可能否（PN）”）有相似的含义和影响，但两者却可以将已知的及可能存在的情况区分开（37，47，75）。评价员应该仅在无法找到任何信息来回答信号问题时，才可使用“缺少信息（NI）”。

信号问题的答案可以帮助系统评价员判断每个领域的总体偏倚风险。如果一个领域中的所有信号问题得到的全是“是”或“可能是”的评价结果，该领域可被判定为低偏倚风险。如果一个或多个问题，得到“否”或“可能否”的评价，则意味着存在出现偏倚的可能性，而“缺少信息（NI）”即指相关信息不够充分，这不意味着一定存在偏倚。例如，在一个预后研究中，如果预测因子在临床事件出现并得以测量之前即已确定，然而研究报告中却没有说明是否掩盖临床结局而盲测预测因子，此时该问题（信号问题 2.2）实际应该为“缺少信息（NI）”。然而，评价员可能仍然会将总体偏倚风险判定为低，因为由推测可知，预测因子是在临床结局判定之前很长一段时间测量的。所以，当判断某特定领域的偏倚风险时，系统评价员需要基于自己的判断来确定，经信号问题识别出的问题是否可能会给模型建立或验证带入偏倚。

评价有关适用性的问题

使用表格五（系统评价问题）和表格七至表格九中的信息，可从三个领域评价原始研究之于系统评价研究问题的适用性。数据分析领域仅涉及和数据或如何进行数据分析有关的局限性，这些并不牵涉到系统评价研究问题，因此，该领域无需做适用性评价。适用性关切可以被判定为“低”、“高”或“不清楚”。“不清楚”应该仅在所报告的信息不充分时使用。

表格七. 领域一研究对象—偏倚风险和适用性评定指导说明

偏倚风险评价

背景

预测模型的总体目的，是想给出在新的个体上也足够正确的绝对风险值预测。有些数据源或研究设计并不适用于获得绝对风险值。如果一个研究不恰当地纳入、或从入组样本中排除一些研究对象的话，也会出现问题。

1.1 是否使用了恰当的数据来源，如队列、RCT 或巢式病例对照研究数据？

是（Y）/可能是（PY）：如果使用了队列研究设计（包括 RCT 或恰当的登记数据）或巢式病例对照或病例对列设计（并在数据分析中恰当地调整了基线风险/危险）。

否（N）/可能否（PN）：如果使用了一个非巢式病例对照设计

缺少信息（NI）：如果研究对象的抽样方法不够明确

1.2 研究对象的纳入和排除标准是否恰当？

是（Y）/可能是（PY）：如果研究对象的纳入和排除标准比较恰当，由此研究对象足以代表未能被选入的符合标准的研究对象。

否（N）/可能否（PN）：如果纳入的研究对象本就已经被认定发生有某临床结局事件，因此不再是带有疑似疾病的研究对象（诊断研究），或具有该事件发生风险的研究对象（预后研究）。

或者，如果从样本中排除掉可能足以改变预测模型在预期目标人群中使用时的预测性能的特定研究对象亚组[译者注：即排除标准不恰当，预期使用该模型的目标人群

已被排除，以至于实际所评估到的模型预测性能与预期理应表现出的性能存在差异]

缺少信息（NI）：如果没有信息说明是否发生不当的纳入或排除时

由研究对象或数据来源所引入的偏倚风险

低偏倚风险：如果所有信号问题答案都是“是（Y）”或“可能是（PY）”，该偏倚风险可以被认为较低。如果一个或超过一个答案是“否（N）”或“可能否（PN）”的话，该偏倚风险仍可被判断为较低，只要能够给出为什么判断为低风险的具体原因。

高偏倚风险：如果任一信号问题的答案是“否（N）”或“可能否（PN）”，即会有出现偏倚风险的可能性；上述低偏倚风险的情况除外。

偏倚风险不清楚：如果有些信号问题的有关信息缺失，同时没有任何一个信号问题的答案足以使该领域被认定为高偏倚风险。

适用性评价

背景

所纳入的研究对象、所用的筛选标准以及原始预测模型研究所处的临床场合，应该和系统评价研究问题相接近。

研究对象和场所与系统评价研究不匹配所导致的关切

适用性关切低：所纳入的研究对象和临床场景与系统评价研究问题匹配。

适用性关切高：所纳入的研究对象和临床场景与系统评价研究问题完全不同。

适用性关切不清楚：如果没有报告研究对象和临床场景的相关信息。

RCT = 随机对照试验

表格八. 领域二预测因子—偏倚风险和适用性评定指导说明

偏倚风险评价

背景

当预测因子定义和测量存在缺陷时，模型性能上的偏倚即可能会发生。预测因子是指正在对其与临床结局之间的相关性做评估的变量。例如，当没有对所有研究对象采用足够相似的方式测量预测因子，或对临床结局的了解影响到预测因子测量的时候，偏倚即会发生。

2.1 是否对所有研究对象均通过相似的方法来定义和评测预测因子？

是（Y）/可能是（PY）：如果预测因子的定义和测量方法在所有研究对象均足够相似

否（N）/可能否（PN）：如果不同的定义被用于同一个预测因子，或者如果需要主观解释的预测因子是由临床经验各不相同的测量人员来测定。

缺少信息（NI）：如果缺少关于预测因子是如何定义或测量的信息。

2.2 预测因子的评测是否是在不知晓临床结局数据的前提下做出的？

是（Y）/可能是（PY）：如果已声明结局指标信息不会被用于预测因子测量，或在评价预测因子时，结局指标的相关信息很明确是无法获知的。

否（N）/可能否（PN）：如果相当明确，结局指标信息被用于测量预测因子。

缺少信息（NI）：缺少信息说明，是否在测量预测因子时，测量员并不知晓结局指标信息。

2.3 在想要使用预测模型的时间节点上，模型中所有预测因子的信息是否都可以获取？

是（Y）/可能是（PY）：在想用模型做预测的时候，所有纳入的预测因子都是可以测量到的。

否（N）/可能否（PN）：在想用模型做预测的时候，预测因子信息无法被采集到。

缺少信息（NI）：没有信息说明，是否在使用模型的时候能够测量到预测因子。

由预测因子所引入的偏倚风险

低偏倚风险：如果所有信号问题答案都是“是（Y）”或“可能是（PY）”，该偏倚风险可以被认为较低。如果一个或超过一个答案是“否（N）”或“可能否（PN）”的话，该偏倚风险仍可能会被判断为较低，但应该提供为什么该偏倚风险被判断为低的具体原因，例如，用了客观性预测因子并不需要做主观解释。

高偏倚风险：如果任一信号问题的答案是“否（N）”或“可能否（PN）”，即会有出现偏倚风险的可能性。

偏倚风险不清楚：如果有些信号问题的有关信息缺失，同时没有任何一个信号问题的答案足以使该领域被认定为高偏倚风险。

适用性评价

背景

原始预测模型研究中的预测因子定义、测量和使用时机，应该适用于系统评价临床问题，例如，预测因子应该经能用在系统评价所探讨的日常临床实践中的操作/方法去做测量。

因预测模型其预测因子的定义、评价或时机与系统评价研究问题不匹配所导致的关切

适用性关切低：预测因子的定义、测量和时机与系统评价研究问题足够匹配。

适用性关切高：预测因子的定义、测量和时机与系统评价研究问题不同。

适用性关切不清楚：没有报告预测因子相关的信息。

表格九. 领域三临床结局—偏倚风险和适用性评定指导说明

偏倚风险评价

背景

如果用于判定临床结局的方法，无法正确地将有或无临床结局事件的个体区分开的话，模型预测性能上的偏倚即可能会出现。结局指标判定方法上出现的偏倚，可以来自于：对不良测量方法、检验或标准的使用，这会导致临床结局测定上出现无法接受之高的误差；当研究对象之间所用的测量方法不一致；或对预测因子信息的知晓，影响到对结局指标的判定。不恰当的结局判定时间也会导致偏倚出现。

3.1 临床结局是否被恰当地判定？

是（Y）/可能是（PY）：如果已经使用了在指南或此前文献中被认为较优或者可接受的结局指标判定方法。

注意：该问题事关结局指标判定方法中的测量误差水平（见适用性关切中有关结局指标的*定义*是否足够合适的问题）。

否（N）/可能否（PN）：如果很明显使用的是一个存在缺陷的方法，其在判定研究对象的临床结局状态上，会导致不可接受的错误。

缺少信息（NI）：缺少有关结局指标是如何判定的信息。

3.2 是否使用了预先设定的或标准的临床结局定义？

是（Y）/可能是（PY）：如果临床结局判定方法是足够客观的，或如果使用了标准的结局指标定义，或如果使用了预先确定的方法对临床结局做归类。

否（N）/可能否（PN）：如果临床结局的定义不够标准或没有预先明确。

缺少信息（NI）：缺少信息说明是否临床结局定义是预先明确的或足够标准的。

3.3 预测因子是否从临床结局定义中被排除掉？

是（Y）/可能是（PY）：如果没有预测因子被用于定义临床结局。

否（N）/可能否（PN）：如果有一个或多个预测因子构成临床结局定义的一部分。

缺少信息（NI）：缺少信息说明是否从临床结局定义中将预测因子信息排除出去。

3.4 对所有研究对象来说，临床结局是否均经相似的方法而被定义和判定？

是（Y）/可能是（PY）：如果所有研究对象都采用了相似的方法来定义及确定临床结局。

否（N）/可能否（PN）：如果对有些研究对象，很显然用到了不同的方法去定义和判定临床结局。

缺少信息（NI）：缺少信息说明，是否对所有研究对象，都用了相似的方法去定义和判定临床结局。

3.5 结局指标是否是在不知晓预测因子信息的情况下而被判定？

是（Y）/可能是（PY）：如果在判定临床结局状况时，预测因子信息是无法获知的；或明确说明，临床结局状态是在不知晓预测因子信息的情况下才被判定的。

否（N）/可能否（PN）：如果很显然在确定临床结局时参考了预测因子信息。

缺少信息（NI）：缺少信息说明，结局指标是否是在不知晓预测因子信息的情况下被确定的。

3.6 预测因子测量和临床结局确定之间的时间间隔是否恰当？

是（Y）/可能是（PY）：如果预测因子测量和临床结局测定之间的时间间隔比较恰当，使得足够可以观测到研究中所定义的临床结局以及满足样本代表性的结局事件数。

或者，如果完全无需要求时间间隔相关的信息，去让满足代表性的临床结局事件发生；或如果预测因子测量和临床结局判定，是取自于时间间隔足够合理的信息源。

否（N）/可能否（PN）：如果预测因子测量和临床结局测定之间的时间间隔，过短或过长，使得无法记录到研究中所定义的临床结局和样本量足够的结局数量。

缺少信息（NI）：如果没有提供关于预测因子测量和临床结局测定之间时间间隔的信息。

由临床结局所引入的偏倚风险

低偏倚风险：如果所有信号问题答案都是“是（Y）”或“可能是（PY）”，该偏倚风险

可以被认为较低。如果一个或更多答案是“否（N）”或“可能否（PN）”的话，该偏倚风险仍旧可以被判断为较低，只要能够提供做出该判断的合理且具体的原因。例如，在判定临床结局的时，即已知晓预测因子相关的信息，但是临床结局的判定无需评价员给出很多主观解释（如，全因死亡）。

高偏倚风险：如果任一信号问题的答案是“否（N）”或“可能否（PN）”，即会有出现偏倚风险的可能性。

偏倚风险不清楚：如果有些信号问题的有关信息缺失，同时没有任何一个信号问题的答案足以使该领域被认定为高偏倚风险。

适用性评价

背景

原始研究中的临床结局定义，应该和系统评价研究问题中的结局指标定义相切合。

临床结局、其定义、测量时机或确定方法与系统评价研究问题不匹配所导致的关切

适用性关切低：临床结局定义、测量时机或判定方法，可以很好地定义系统评价研究问题想研究的临床结局

适用性关切高：临床结局定义、测量时机或判定方法，在定义另一个、不是系统评价想要研究的临床结局

适用性关切不清楚：临床结局定义、测量时机或判定方法相关的信息不够明确

如果系统评价研究问题及原始研究匹适度良好，适用性有关的关切极可能较低。一个系统评价可能想解决一个具有针对性的研究问题，但纳入研究标准却相对宽泛。

判断依据和评级理由

为提升评价过程的透明度，PROBAST 的每个领域都包括有两个文本框。第一个文本框可以让系统评价员记录判断依据，即用于回答偏倚风险信号问题或辅助评价该领域适用性的相关信息。文本既可是概括小结性的，也可是从评价文献中直接复制粘贴的。第二个文本框是“评级理由”，这里可以让系统评价员记录为什么将预测模型判定为具有高、低或不清楚偏倚风险，或高、低或不清楚适用性关切。例如，如果一个领域被判定为高偏倚风险，系统评价员能够汇总导致该评级的原始研究特征。或者，当一个或多个信号问题得到“否”、“可能否”或“缺少信息（NI）”的答案，但一个领域仍被判断为低偏倚风险时，此文本框可以用来解释为什么经信号问题找出的问题不太可能在研究中引入偏倚。

更多指导性文件及案例详见各领域的相应部分，以及表格七至十。最新版本的指导性文件可以从 www.probast.org 网站下载。

表格十. 领域四数据和分析—偏倚风险和适用性评定指导说明

偏倚风险评价 背景 统计分析是预测模型建立和验证中的重要部分。使用不恰当的统计分析方法，会增加模型性能指标上出现偏倚的可能性。模型建立研究包括有多个步骤，其中错误的方法可能会扭曲研究结果。我们建议，在做数据分析领域的评价时，系统评价员可以寻求专业的统计学意见。

4.1 有特定临床结局事件的研究对象例数是否足够合理？

是（Y）/可能是（PY）：对于模型建立研究来说，如果发生临床结局事件的研究对象例数，和候选预测因子参数数目的比值 ≥ 20 （即 $EPV \geq 20$）*。

对于模型验证研究来说，如果发生临床结局事件的研究对象例数 ≥ 100.

否（N）/可能否（PN）：对于模型建立研究，如果发生临床结局事件的研究对象例数，和候选预测因子参数数目的比值 < 10 （即 $EPV < 10$）*。

对于模型验证研究，如果发生临床结局事件的研究对象例数 < 100.

缺少信息（NI）：对于模型建立研究，缺少发生临床结局事件的研究对象例数，或候选预测因子参数数目的信息，以至于无法计算 EPV。

对于模型验证研究，缺少发生临床结局事件的研究对象例数信息。

4.2 连续型预测因子和分类型预测因子是否被恰当地处理？

是（Y）/可能是（PY）：在纳入到模型中的时候，如果一个连续型预测因子没有被转换为两个或更多类（即，二分类化，或分类化）。

或，如果经（比如）分式多项式或限制性立方样条技术，检验连续型预测因子的非线性特征。

或如果用预先确定的分类组别划分预测因子。

对于模型验证研究而言，如果连续型预测因子经用和模型建立研究相同的定义或转换而被纳入到模型分析，且分类变量是用相同的阈值结点被划分。

否（N）/可能否（PN）：如果分类化的预测因子定义没有被提前确定。

对于模型建立研究而言，在纳入进模型的时候，如果一个连续型预测因子被转换为两个或更多分类。

如果与模型建立研究相比，模型验证研究是用不同的定义或转换将连续型预测因子纳入到模型分析；或者分类变量是用不同的阈值结点被划分。

缺少信息（NI）：缺少信息说明，有无检测连续型预测因子的非线性特征，并且没有说明预测因子的分类组别是如何被划分的。

对于模型验证研究而言，没有信息说明，其是否和建模研究一样使用相同的定义或转换以及相同的阈值结点。

4.3 是否所有参与研究的研究对象均被纳入到数据分析？

是（Y）/可能是（PY）：如果所有参与研究的研究对象都被纳入到数据分析中。

否（N）/可能否（PN）：如果有些或一个亚组的研究对象，被从数据分析中不合宜地删除掉。

缺少信息（NI）：缺少信息说明，是否所有入组的研究对象都被纳入到数据分析中。

4.4 出现缺失数据的研究对象是否被恰当地处理？

是（Y）/可能是（PY）：如果预测因子或临床结局没有出现缺失值，并且该研究明确地说明没有研究对象因为缺失值而被删除。

或者，如果缺失值通过多重插补法进行处理。

否（N）/可能否（PN）：如果有缺失值的研究对象被从数据分析中删除。

或者，如果处理缺失值的方法明显是错误的（例如，缺失指示因子法，或不恰当地使用最后观察值推估法）。

或者，如果该研究没有明确提及如何处理缺失值。

缺少信息（NI）：如果用于判断缺失值处理方法是否合理的信息不充分。

4.5 是否避免依据单因素分析来筛选预测因子？†

是（Y）/可能是（PY）：如果预测因子在被纳入到多因素模型中之前，没有基于单因素分析而被预先筛选。

否（N）/可能否（PN）：如果在被纳入到多因素模型中之前，预测因子先基于单因素分析被筛选过。

缺少信息（NI）：如果缺少信息说明，是否避免使用单因素分析的筛选过程。

4.6 是否恰当地考量处理了数据中出现的复杂问题（如，删失数据、竞争风险数据、对照组研究对象的抽样）？

是（Y）/可能是（PY）：如果数据中存在的复杂问题被合理地处理了。

或者，如果很显然任何潜在的数据复杂问题，都可以很恰当地被认为并不重要。

否（N）/可能否（PN）：如果数据中出现的足以影响模型性能的复杂问题被忽略掉。

缺少信息（NI）：缺少信息说明，是否数据中出现复杂问题，或如果有，有无被合理处理。

4.7 是否恰当地评估了相应的模型性能测量指标？

是（Y）/可能是（PY）：如果校准度和区分度均被恰当评价（包括那些评价生存结局预测模型而制定的指标）。

否（N）/可能否（PN）：如果校准度和区分度均未被恰当评价。

或者，如果仅用拟合优度检验（如 Hosmer-Lemeshow 检验）来评价校准度。

或者，如果对预测生存结局的模型而言，没有采用可考量数据删失的模型性能指标。

或者，如果分类指标（像敏感度、特异度或预测价值），依据从已有数据推导而出的预测风险阈值而被呈现出来。

缺少信息（NI）：没有报告校准度或区分度。

或者，缺少信息说明，是否使用了恰当的生存结局相应的性能指标（例如，参考相关文献，或具体提及的方法，例如使用 Kaplan-Meier 估计量）。

或者，没有提供涉及到可用于估算分类指标之阈值的信息。

4.8 模型过度拟合和模型性能上的乐观偏差是否被恰当地考量处理？†

是（Y）/可能是（PY）：如果内部验证技术（如，自助重抽样法和交叉验证，此过程需要涵盖到全部的模型建立流程），已经被用于处理模型拟合中出现的乐观偏差，并且此后也对模型性能估值做出了调整。

否（N）/可能否（PN）：如果不曾进行内部验证。

或者，如果内部验证仅仅是一个在研究对象上做的简单随机样本拆分。

或者，如果自助重抽样或交叉验证，并没有囊括所有的每一步模型建立流程（包括变量筛选）。

缺少信息（NI）：缺少信息说明，是否已经使用内部验证技术（包括所有的模型建立流程）。

4.9 最终模型中的预测因子及其权重分配是否和多因素分析中的结果相符？†

是（Y）/可能是（PY）：如果最终模型中的预测因子和回归系数与多因素分析所呈现的结果相匹配。

否（N）/可能否（PN）：如果最终模型中的预测因子和回归系数，与多因素分析所报告的结果不匹配。

缺少信息（NI）：如果并不清楚，最终模型中的预测因子和回归系数是否与多因素分析所呈现的结果匹配。

由数据分析所引入的偏倚风险

低偏倚风险：如果所有信号问题答案都是“是（Y）”或“可能是（PY）”，该偏倚风险可以被认为较低。如果一个或超过一个答案是“否（N）”或“可能否（PN）”的话，该偏倚风险仍旧可以被判断为较低，只要能够提供该判断的合理且具体的原因。

高偏倚风险：如果任一信号问题的答案是“否（N）”或“可能否（PN）”，即会有出现偏倚风险的可能性。

偏倚风险不清楚：如果有些信号问题中缺少与数据分析有关的信息，同时没有任何一个信号问题的答案足以使该领域被认定为高偏倚风险。

EPV = 变量平均事件发生数

* 对于 EPVs 在 10 至 20 之间的情况，该条目可以被定级为“可能是（PY）”或“可能否（PN）”，这取决于临床结局事件的发生频率、预测模型总体性能和模型预测因子的数据分布。更多信息，见参考问题文献 145 至 147.

† 仅适用于模型建立

领域一：研究对象

该领域涵盖了与所用数据源和如何选取入组的研究对象相关的潜在偏倚来源和适用性有关的问题。在判断依据文本框，系统评价员应该描述数据来源--比如，队列研究、随机对照研究或常规数据登记，以及原始研究研究对象的选取标准。

偏倚风险。两个信号问题可帮助评价该领域的偏倚风险（见表格七）。

1.1 是否使用了恰当的数据来源，如队列、RCT 或巢式病例对照研究数据？

许多数据来源和研究设计都可以用在预测模型研究中。

预后模型研究。当采用前瞻性纵向队列设计时，预后模型研究具有低偏倚风险，在这样的研究设计中，一般会提早定义、并在既定随访时间内执行统一的：研究对象纳入排除标准、预测因子评价及结局指标测量（1）。采用预先确定的、统一的研究方法能确保研究对象的数据可被系统且真实地被采集。

如果研究对象数据是来自于已存的数据源，例如已有的队列研究或常规医疗登记数据，模型建立和验证研究即存在更高的偏倚风险，这是因为通常这些数据并不是为了建立、验证或更新预测模型而被采集的，且通常没有预定的研究方案。在常规医疗登记库中，涉及到纳入排除标准的数据通常未经统一测量和记录（44，90）。例如，关于英国临床实践研究数据链，Herrett 和同事（90）说到“初级卫生保健的数据质量差异性较高，因为这些数据是全科医生在常规问诊过程中录入的，而不是以研究为目的进行收集的。因此，研究者在开展一项研究之前，一定要全面地检查数据质量。”

来自随机对照试验的单个或多个试验组的数据也可用于预后模型建立、验证或更新。然而，随机分配的治疗措施可能需要被看成独立预测因子，并纳入到模型分析中，以权衡治疗效应，这是因为有效的治疗措施本身就是临床结局的预测因子（91，92）。此外，随机对照试验往往具有较严格的纳入排除标准，这通常导致预测因子的分布比较窄（案例混合更小）。相比于用预测因子分布略宽的数据源来建立或验证模型，使用预测因子分布较窄的数据所建立或验证的模型倾向于表现出更低的区分能力（93-96）。这是因为在后者，一个模型所预测的可能性范围更小，也因此区分能力更差。

在病例队列（case-cohort）或巢式病例对照研究中，发生临床结局（即病例组）和未发生该临床结局的研究对象（即非病例组或对照组）都是从已存的明确描述的队列、或者样本量大小已知的常规医疗登记中抽样而来。只要研究者在数据分析中恰当地对原始队列或登记库中临床结局出现的频次进行过调整，仍可认为这类研究仅具有较低的偏倚风险（见信号问题 4.6）（57，97-100）。如果未做调整，对于预测模型研究来说，病例队列和巢式病例对照研究有较高的偏倚风险。例如，在逻辑预测模型中，经逆采样分数法（inverse sampling fraction）对对照组和病例组的样本进行权重再分配，可有助于正确评估基线风险，这可以让研究者能够获得矫正后的绝对预测概率模型校准指标（97-100）。如若病例对照研究中的病例组和对照组之研究对象并不是从预先确定且明确定义的队列或登记数据库抽样而来，该类研究具有较高的偏倚风险，这是因为相比于来源人群，病例组和对照组中所选的研究对象之定义和例数均不清楚。这样的话，并无法正确地矫正基线风险或危险值以及临床结局发生风险绝对值（57）。

诊断模型研究。诊断模型可预测在测量指标试验或预测因子的时间点上，是否已出现了某特定临床结局（或目标疾病）（见文本框二）。据此，对诊断模型研究来说，具有最低偏倚

风险的设计是横断面研究；该设计是选取一组（队列）疑似具有目标疾病的某些临床症状或体征的个体作为研究对象。之后，测量所有研究对象的预测因子（指标试验），并参照金标准测量临床结局（目标疾病的有或无）（101-104）。如若无法在同一个时间节点上确定所有研究对象的疾病状态（例如，许多可能带有恶性肿块的研究对象，其影像图像上并没有病灶，这可以做病理穿刺来确认），具有横断面设计的诊断研究可做进一步随访，以确定在指标试验完成之时，是否目标疾病即已出现。

和预后模型一样，只要研究者通过逆采样分数法对病例组和对照组的样本做了调整，以此得出原始队列中结局指标流行率的正确估值，采用巢式病例对照设计的诊断模型也可是低偏倚风险的（105-109）。相似地，采用非巢式病例对照设计--这其中，具有晚期症状的病例案例和健康对照研究对象具有过度代表性--会导致疾病流行率的不正确估计，以及被过高估计的模型诊断性能（105-109）。

案例：在 Perel 及同事的研究（88）中，用于建立预后模型的数据是来自于一个随机对照试验（CRASH-2，严重出血的抗纤溶药临床随机研究 2），且这些数据合并了两个治疗组。因为作者将随机分配的治疗措施看作是一个预测因子而纳入到模型中，该信号问题应该回复“是”。

案例：Aslibekyan 及同事（86）采用非巢式病例对照研究设计，但是没有在其数据分析中使用逆抽样分数法对病例和对照组样本的权重做再分配。因此，该信号问题应该被回复为“否”。

1.2 研究对象的纳入和排除标准是否恰当？

那些没有恰当地纳入或排除研究对象的研究，可能会产生出存在一定偏差的模型预测性能评估值，这是由于该模型是基于一个具有选择性的研究对象亚组，它可能无法代表意向中的目标群体。

不恰当的纳入可能是因为纳入了一些在测量预测因子时既已发生了特定临床结局的研究对象。例如，在一项二型糖尿病发生风险预测模型建立的研究中，如果纳入标准是依据患者自报告数据而招募没有糖尿病的研究对象的话，有些研究对象其实可能已经发生了二型糖

尿病。纳入已经发生了特定疾病的研究对象，极可能会产出一个预测性能被过高估计的模型。

相似地，对于一个想要检测带有临床症状的患者中是否患有肺栓塞的诊断模型来说，排除现在即有肺部疾病的患者可能被认为并不太恰当。与那些现在并未患有肺部疾病的人相比，这类患者[译者注：现在患有肺部疾病的患者]会更难给出肺栓塞诊断；因此，如果一个模型在排除这类患者之后，被建立以用于全部疑似肺栓塞的患者，模型的诊断准确性可能会被过高估计。既然那样，作者当时就应该详细说明所建立的模型仅适用于现在没有发生肺部疾病的疑似肺栓塞患者。

注意，该信号问题不是在问原始研究中纳入研究对象之后出现的失访问题（也就是说，这里并不涉及到在研究进行中出现的不恰当排除）；这个问题[译者注：失访]会在领域四处理。该信号问题是关于在招募过程中被不恰当地纳入或排除的研究对象。再者，这里重点是要区分：由纳入排除标准上的约束限制条件所施加到研究对象上的**选择偏倚**，和具有会限制原始研究之于系统评价研究问题适用程度的特征的**研究群体**（见**适用性**）。

总结来看，这里的关键问题是，纳入排除标准或招募策略是不是有可能已经让所纳入的研究对象无法代表意向目标人群。有些偏倚风险评估工具（如 QUADAS-2）包含有这样一个信号问题，问到原始研究是否招募到了一个完全连续的或随机的患者样本。因为对于任何研究而言，这都是很难实现的，我们并没有将其纳入到 PROBAST 中去。

案例：因为他们使用了病例对照研究设计，Aslibekyan 及同事(86)排除了所有罹患致命性心肌梗塞的研究对象。死于心肌梗塞的研究对象，也因为无法从他们中收集到回顾性的自报告数据，而被排除。由此，用于非致命性心肌梗塞的预测模型，是基于所筛选出的较健康的研究对象，这仅包括那些存活的心肌梗塞患者（病例组）或没有发生心肌梗塞的研究对象（对照组）。这很可能会引入偏倚，因为研究对象仅呈现出了具有任何类型心肌梗塞发生风险的原始群体中的一个被筛选出的仅有较低风险的样本。仅声明一下该预测模型仅预测非致命性心肌梗塞并不能解决问题，因为在做出风险预测的那一刻，是不可能找出将来会发生致命性心肌梗塞的研究对象的。该信号问题应该回答“可能否”。

评定领域一的偏倚风险。表格七呈现出，该如何回答信号问题以及如何得出领域一的总体判断。

适用性。该领域的适用性会考虑，原始研究中纳入的研究人群在多大程度上和系统评价研究问题中确定的研究对象匹配（步骤一；表格五）。试想有一个系统评价，想收集以诊断伴有临床症状的儿童中是否发生细菌性结膜炎为目的的所有模型建立和验证研究。该系统评价可以将纳入标准确定成这样：同时有成人和儿童的预测模型研究是符合标准的。仅招募纳入了儿童的模型研究，极可能会得到适用性关切低的判断，而那些招募了成人和儿童的研究可能具有较高的适用性关切。

基于随机对照试验数据的预测模型研究，其外推性和适用性需要许多仔细的考虑。随机试验倾向于采用较严格的纳入排除标准，并且测量较少的预测因子和临床结局，这些因此会降低以这类数据为基础而建立或验证的模型之适用性。相反，来源于常规医疗或健康管理登记的数据，其研究对象特征、预测因子和临床结局有更加广泛的数据分布；因此，使用这类登记数据库去建立或验证模型的预测模型研究倾向于具有更好的外推性。

通常，很难看出来一个原始研究相关的有些问题什么时候会极可能引入偏倚风险，以及这些问题是否会引发适用性的关切。适用性评价完全取决于系统评价研究问题（表格二和五）。试想一下信号问题 1.2 中假设的肺栓塞案例，该案例中系统评价员可以将系统评价的目标人群限定为疑似罹患肺栓塞但现在没有肺病的病人。对于这个目标人群，在原始研究中纳入有肺病的患者，可能会造成一定的适用性关切，但不一定会引发偏倚风险。同样地，试想一个纳入了较广年龄段（18 至 90 岁）患者的诊断模型建立研究。这可能本不会在原始研究中引入任何偏倚，但如果系统评价研究问题关注于年轻（18 至 30 岁）成人的话，这可能会限制模型的适用性。

最后，原始研究有时候使用（对研究者来说）本来就与模型建立研究中的特定人群大相径庭的研究对象数据来验证一个模型。例如，基于一般健康人群而建立的心血管预测模型已经在明确患有二型糖尿病的患者中被验证（110）；一个在二级医疗急诊场所建立的用于诊断深静脉血栓的模型，也在初级医疗场所做了验证（85）。在这两个案例中，建立研究和验证研究间模型性能上的异质性是可想而知的（40）。

领域二：预测因子

该领域涵盖和预测因子定义及测量相关的潜在偏倚来源和适用性关切。预测因子是那些

被评价和临床结局之间是否存在关联的变量；它们最终可合构成预测模型。

在判断依据文本框中，系统评价员可以罗列并描述预测因子是如何被定义，其测量的时间点，以及在测量预测因子时是否已知其他信息。

注意，对关注某个特定预测模型的系统评价而言，仅仅罗列并描述待验证模型中包含的预测因子即可。

偏倚风险。三个信号问题可以帮助判断该领域的偏倚风险（表格八）。

2.1 是否对所有研究对象均通过相似的方法来定义和评测预测因子？

为了降低偏倚风险，应该对所有研究对象采用同样的方法来定义和测量预测因子。如果针对同一个预测因子，研究对象之间使用了不同的定义和测量方法的话，可以想象该预测因子和临床结局之间的相关关系可能会[译者注：在研究对象之间]存在差异。例如，下消化道活动性出血可能会被视为一个预测因子，用于建立直肠癌诊断模型。对于一些研究对象来说，预测因子“便血”可以依据粪便上可见的血迹进行评判，而对于另外一些研究对象，可以使用大便潜血试验进行检查。然而，如果这些检测方法（及不同的最小可检测量水平）在同一个预测因子上互换使用的话，“便血”可能会带来偏倚，特别是当检测方法的选取又是依据先前检查或症状的时候。

对于牵涉到主观判断的预测因子而言，该偏倚出现的可能性会更高，比如影像检查结果，这可能会导致转而去研究观察者（而不是预测因子）的预测能力。当需要特殊技巧或培训[译者注：来判断预测因子时]，说明一下由谁（比如，是经验丰富的主任医师还是缺乏经验的住院医师）评定预测因子也显得很重要。

案例。Perel 和同事（88）测量了以下预测因子：人口学特征（年龄和性别）、损伤特征（类型和损伤已发时间）和生理学变量（Glasgow 昏迷评分、收缩压、心率、呼吸频率和毛细血管再充盈时间）；所有这些都记录在 CRASH-2 随机试验入组登记表上。因为用于建立该预测模型的数据来自于随机试验的一个亚组，并且预测因子都来自于研究入组登记表，极有可能所有预测因子在全部研究对象间经由相同的方法而被定义和测量—尽管文中没有对此做特别说明。该信号问题因此可以被回答成“可能是 PY”。如果数据取自多个数据源（例如，常规医疗数据登记，其中极可能使用不同版本的 Glasgow 昏迷评分量表或者不同的损伤

类型定义），该信号问题可被回复为“可能否 PN”。

2.2 预测因子的评测是否是在不知晓临床结局数据的前提下做出的？

当预测因子的评测时在不知道临床结局事件状态的情况下做出的（常称为盲法），偏倚风险即较低。对于牵涉到需要主观解释或判断的预测因子来说，比如那些基于影像学检查、组织学检测、病史、或体格检查的因子，在测量预测因子时掩蔽临床结局数据是尤其重要的。盲法的缺乏可增加将临床结局信息带入到预测因子测量的可能性，这极可能会拉高它们[译者注：因子与结局之间]的相关关系，并引入带有偏差的、过高的模型预测性能估值（1，111-119）。

在使用前瞻性队列设计的预后研究中，预后因子是早于临床结局发生而被测量的，自然而然即导致预测因子测量员不可能会知道临床结局信息。该偏倚更可能出现在那些回顾性记录预测因子的研究（易受回忆偏倚的影响）或横断面研究（比如诊断模型研究），它们的预测因子和临床结局是在一个相对接近的时间段内而测量的（1，111-120）。

大多数预测模型研究并不报告预测因子测量相关的盲法信息，即测量预测因子时，是否已知晓临床结局数据（121，122）。在预后研究中，该信号问题因此应该被回答为“缺少信息（NI）”（表格八）。然而，该领域总体偏倚风险仍可被定级为低风险，因为预测因子的测量如果早于结局指标发生很长时间的话，可以推断预测因子测量是盲于临床结局信息的。需要注意的是，即使在预后研究中，预测因子有时也可能是在临床结局信息已经被收集之后才被测量，例如，当预测因子采集自对已存影像信息的重新解释，或当使用回顾性随访设计时。一个案例是，再次启用冷冻组织或肿瘤样本来测量新的预测因子（生物标志物）；这种样本可能已经和研究对象随访信息链接，因此新的预测因子测量可能发生在临床结局事件出现之后，且可能没有被盲于临床结局信息。

*案例。*Oudega 及同事（85）说到：“在获取知情同意之后，初级卫生保健医师使用标准的表格系统呈现了患者病史和体格检查信息，该表格中详述有信息条目以及可能的回答。患者病史包括性别、既往深静脉血栓史、深静脉血栓家族史、肿瘤病史（此前 6 个月内出现过活动性肿瘤）、制动超过三天、近期实施过手术（此前四周内）和三种主要症状（腿部疼痛、发红或水肿）的持续时间。体格检查包括沿着深静脉系统出现的压痛、侧肢浅表静脉发生的

非静脉曲张型扩张、凹陷性水肿、患肢水肿和两腿腓部周长差值……在完成病史询问和体格检查之后，所有患者都被送到医院去做 D 二聚体检测和腿部超声检查。”（85）因为该研究报告说，在 D 二聚体检测之前，即已完成了所有患者的病史和临床信息（即预测因子）收集，[译者注：预测因子的]测量因此是[译者注：在不知晓临床结局的情况下]盲测的；该信号问题应该被回答为“是（Y）”。

2.3 在想要使用预测模型的时间节点上，模型中所有预测因子的信息是否都可以获取？

对于一个在真实医疗场所中可用的预测模型来说，所有模型中纳入的预测因子均要在模型即将运用的时刻（即，做预测时）可以测量得到（表格二）。这听上去很理所当然，但不幸的是，有些模型会纳入一些在模型预期运用的那一刻并不可知的预测因子。

例如，一个想要在*术前*使用以预测术后 24 小时内恶心和呕吐发生风险的预后模型，不应该纳入诸如术中用药等预测因子，除非该药的使用是提前确定、且术中不会改变用药方案。纳入一些在模型运用时并不可得的预测因子是不恰当的，这会让一个模型失去临床实用价值。这也会夸大模型的表面预测性能，因为那些预测因子在更接近临床结局评估的时间点上才被测量，并且极可能与临床结局之间有更强的相关性。对于那些无论在什么时候都稳定不变的预测因子而言（如性别和基因相关的因子），并不会存在这些方面的问题。

那些想对某个已知预测模型做外部验证的研究，如果它们在做验证时预测因子数据有缺失，但研究者却仍坚持将这些有缺失数据的预测因子剔除掉，才去验证该模型的话，这些研究即会有较高的偏倚风险。这是在验证研究中较常见的一个缺陷，且实际上会生成另一个模型（而不是原本所建立的、那个意向中有待验证的模型）的验证结果。在这种情况下，该信号问题应该被回答为“否（N）”。

*案例。*Rietveld 和同事们（89）想建立并验证一个预测模型，用于诊断在初级医疗场所就诊的、有急性结膜炎临床症状的儿童急性结膜炎患者其细菌来源，以指导抗生素管理的决策制定。所有预测因子应该在全科医生初步问诊时即能获取知晓。这个研究中的预测因子确实都可以在病史采集和体格检查时获取得到，因此该信号问题应该被回答为“是（Y）”。如果这个研究将实验室检查（如显微镜检查）纳入到预测因子之中的话，该信号问题很可能要被回答为“否（N）”。因为获知显微镜检测结果牵扯到时间上的延迟，全科医生极不太可能

在初步问诊时即能获知相关检测结果。

领域二偏倚风险评级。表格八说明了应该怎样回答信号问题，以及如何对领域二偏倚风险做出一个总体判断。

适用性。在该领域上，导致适用性关切的一个常见原因是预测因子的定义、测量或评测时机与系统评价研究问题不一致。预测因子应该经由适用于系统评价探讨的医疗场所的测量方法来测量（表格二和五）。将特殊测量技术用于测量预测因子的原始研究，可能会得出对系统评价的目标场所来说较偏乐观的预测。例如，如果一个模型应该被用于无法获得高级影像技术的医疗场所，一项纳入了正电子发射计算机断层显像结果的模型建立研究可能显得并不怎么适用，因此可以被认为具有较高关切。

和领域一一样，该领域的偏倚风险评价和适用性评价之间存在一个微小的区别。想一下信号问题 2.1 中提到的那个案例，将下消化道活动性出血考虑为结直肠癌诊断的一个预测因子。这个出血可以根据粪便上肉眼可见的血迹或者大便隐血试验来判断。系统评价员可能将其系统评价关注到仅纳入肉眼评估作为一个预测因子的结直肠癌诊断模型，意即使用大便隐血试验的一个原始研究可能会导致适用性关切。

相似地，和领域一一样，在想要评估某特定模型平均预测性能的系统评价中，模型建立研究和模型验证研究之间，由于[译者注：研究间的]预测因子定义和测量的差异，出现模型性能上的异质性是可想而知的（17，40，44）。当不同的定义或测量方法被应用的时候，一些验证研究可能会得出不同于其他研究的预测性能，因此应该可以被认为存在适用性关切。有时候，[译者注：当用自己的数据验证一个已知模型时，]研究者有意地使用不同[译者注：不同于原建模研究中]的定义或者测量方法 — 比如对于某些血液学指标，使用床旁即时检查而非实验室检查方法。如果系统评价的具体目的是想纳入一个特定模型的所有验证研究，但忽略其预测因子的定义和测量方法的话，这可能不会是一个问题。

领域三：临床结局

这个领域涵盖了和临床结局的定义与判定相关的潜在偏倚来源和适用性关切。一个理想的临床结局判定方法，可以无差错地区分全部研究对象的临床结局。

在*诊断模型研究*中，临床结局是目标症状的有或无。临床结局的判定或者验证需要用参

照标准来测定（文本框二）。在*预后模型研究*中，所预测的临床结局在将来才会发生，即在给出临床预测的时间节点之后。对两类预测模型而言，参照标准或临床结局的判定方法可以是单个的实验检查或操作、多个实验检查的联用（复合型结局），也可以是专家共识（例如临床结局裁决委员会）。

判断依据文本框可供系统评价员用于描述临床结局是怎样以及在何时定义和判读的，以及在判读时，哪些信息是可获取到的。

偏倚风险。六个信号问题可用于该领域的偏倚风险评价（表格 9）。

3.1 临床结局是否被恰当地判定？

该信号问题是试图找出因为不佳或者较差的方法被用于判定临床结局而致的临床结局错误分类所引起的潜在偏倚。在临床结局的区分上出现的差错，可以导致有偏倚的回归系数，有偏倚的截距值（逻辑回归和参数生存模型）或基线风险（Cox 回归模型），因此也会导致有偏倚的预测模型性能评估。

当预测模型研究使用来自于常规医疗登记库的数据，或使用来自那些本来是为了回答另一个研究问题而设计开展的既有研究的数据，评价员需要 — 有时基于该研究较早发表的研究报告中的细节 — 仔细评判临床结局判定方法的适用程度。在常规医疗登记库中，临床结局数据可能根本不会被记录，抑或是，所用的判定方法相对不够好，且可能无法判定（或错误区分）该临床结局。在诊断研究中，由不可靠参照标准所致的目标症状误判而引起的问题和偏倚已经被广泛探讨过（112，116，123-127）。

和预测因子的测量（信号问题 2.1）一样，对于涉及主观判断（比如，影像学检查，手术探查或者甚至病理检查结果）的临床结局而言，偏倚出现的可能性会比较高。在有赖于特殊技巧或者培训的时候，讲明由谁（例如，经验丰富的主任医师，还是缺乏经验的住院医师）判读临床结局也是很重要的。

案例。在 Han 和同事的研究（87）中，“每个模型都有两个已定义的临床结局：一个是 14 天的死亡率，另一个是六个月的不利临床结局” — 由作者基于 Glasgow 结局量表，将其[译者注：不利临床结局]定义为“严重残疾、植物人状态或者死亡。”因为临床结局（死亡率和由 Glasgow 结局量表而定义的三个结局类别）都采用完善且恰当的方法给出临床结局判

读，该信号问题应该被回答为“是（Y）”。如果由不曾接受培训的测量员来测量该 Glasgow 结局量表得分，有些问题可能就会出现。尽管临床结局的分类个数并不多，对于 Glasgow 结局量表来说，其错误分类的情况并不少见（128，129）。由缺乏经验的测量员[译者注：来测量 Glasgow 结局量表得分的话]，将会使该信号问题得到较不利的答案（即“可能否（PN）”或“缺少信息（NI）”）。

3.2 是否使用了预先设定的或标准的临床结局定义？

该信号问题旨在找出这种情况下的潜在偏倚风险：即模型性能因选用了可得出更有利结果的临床结局定义而被夸大，亦即选择性结果报告（130）。

当使用预先确定的或标准的临床结局定义，并有来自临床指南、此前发表的研究或已发表的研究方案的结局定义作为补充时，偏倚风险可以被判定为低。在一个连续性测量尺度上，如果使用不常见的阈值来判定一个临床结局发生与否的话，偏倚风险就会比较高。如果为了找出最有利的临床结局定义以获得模型性能最优评估值，作者检验多个阈值的话，反而可能会产生带有偏倚的模型性能估计。例如，如果作者使用一个取值在 3 至 15 的连续量表（如 Glasgow 结局量表），并通过是否能得到最优的模型预测性能来选择阈值，去定义临床结局的“好”和“坏”的话，可能就会得到一个带有一定偏倚的模型性能。

使用复合型临床结局也可能会引入偏倚风险。例如，作者可能会引入一定的偏倚，通过调整一个复合型临床结局定义 — 经[译者注：从一个复合型临床结局中]排除常见结局组分或添加不常见结局事件 — 以便于得到更优的模型性能。

许多临床结局允许使用基于专家共识的定义，这包括阈值以及首选的复合型临床结局定义。“有效性试验核心结局指标测量”（COMET）工作组（www.comet-initiative.org）即被成立以推进被广泛认可的或标准化的临床结局指标集的建立。判断被使用的定义是标准的还是不标准的，这可能要求具备专业的临床知识。

*案例。*在 Han 和同事的研究（87）中，“每个模型都有两个已定义的临床结局：一个是 14 天的死亡率，另一个是六个月的不利临床结局” — 由作者基于 Glasgow 结局量表，将其[译者注：不利临床结局]定义为“严重残疾、植物人状态或者死亡。”因为临床结局（死亡率和由 Glasgow 结局量表而定义的三个结局类别）都采用完善且恰当的方法给出临床结局判读，该信号问题应该被回答为“是（Y）”。相反，如果作者不是使用标准的定义，而是基于

他们自己的临床经验或院内指南对 Glasgow 结局量表的结局分类做出修改；此时，应该基于临床判断来决定是不是修订后的 Glasgow 结局量表仍能形成一个标准的临床结局判定措施。如果不是的话，该信号问题可被判定为“可能否（PN）”或“否（N）”。

3.3 预测因子是否从临床结局定义中被排除掉？

理想情况下，临床结局应该在不知晓预测因子信息的情况下，而被定义/判读（见信号问题 3.5）；但是在有些情况下，不太可能能够避免预测因子的影响 — 例如，当临床结局要求由专家共识小组使用尽可能多的可获取信息而做出判读的时候。如果模型的一个预测因子是该模型待预测的临床结局定义或评判的一部分时，预测因子和临床结局之间的相关关系极可能即会被过高估计，并且模型性能的估计值也会是偏乐观的；在诊断研究中，该问题通常被称之为合并偏倚（incorporation bias）（104，111，115，117，119，131-134）。

当临床结局很难由单个操作（如单个参照测试）来确定的时候，判断临床结局事件是否出现可以基于多个组分或测试（像世界卫生组织心肌梗塞诊断标准一样），或者甚至基于所有已知信息，包括在研的预测因子。后者还被称为共识或专家小组临床结局测量，它也容易受到合并偏倚的影响（135）。

*案例。*基于预测因子（如饮食构成、身体活动、吸烟状况、酒精摄入、社会经济状态和肥胖超重变量）的预测能力水平，Aslibekyan 和同事（86）想建立一个心血管风险评分用于预测非致命性心肌梗塞。该研究报告称，心肌梗塞是按照世界卫生组织标准而被定义的，包括心脏标志物、心电图、影像检查、或尸检验证。由于 Aslibekyan 和同事用于建模的生活方式和社会经济学预测因子并不是该心肌梗塞定义的组成部分，在该信号问题上，该研究可以被判定为“是（Y）”。如果该研究在待评价的预测因子中纳入了某个心脏标志物（如入院时的高敏肌钙蛋白 T 初始测定），该信号问题很可能要被判定为“否（N）”。这是因为高敏肌钙蛋白 T 初始测定本来就可能已经是判定临床结局（心肌梗塞）所用信息的一部分。

3.4 对所有研究对象来说，临床结局是否均经相似的方法而被定义和判定？

与预测因子（信号问题 2.1）相似，应该对所有研究对象采用相同的方法来定义和判定临床结局。

临床结局的定义和评判，应该在研究对象间用相同的阈值和结局类别去定义结局事件的

发生与否。当用到复合型临床结局的时候，[译者注：该复合结局的]所有构成组分的结果应该[译者注：对每个研究对象]总是经同样的方法进行合并，以判定临床结局的出现与否。当用到专家共识或专家小组临床结局判定委员会时，应该[译者注：对所有研究对象]采用同样的方法（例如，多数票）来判定临床结局（131，135，136）。

当临床结局判定方法在研究对象之间存在着差异的时候 — 例如，因为多中心研究不同研究网点之间存在的差异，偏倚风险可能会出现。在以其他用途为目的而收集的数据为基础 — 例如，常规医疗登记数据，这其中本来就相差各异的临床结局定义和测量方法极可能会被使用 — 而不是以预先设计的研究为基础的预测模型研究中，偏倚风险也会较高。再者，当各测量方法之间判定临床结局发生与否的准确度存在较大差异（差异性临床结局确认），且偏倚的方向不易预测时，偏倚风险亦会较高。例如，一项预后模型研究想要预测健康成年人将来发生糖尿病的风险，个体水平的糖尿病发生与否可以经由多种方法（比如空腹血糖水平，口服糖耐受测试，或自报告）来判定，而这些判定方法在糖尿病发生与否上的判别能力却各不相同。当临床结局要求做出较主观的解释时，偏倚风险也会较高。相似地，需要多次测量（如多次门诊就诊）的临床结局也存在相当程度的偏倚风险，尤其是在研究对象之间测量频率各不相同的情况下；较频繁的测量能提升检测出临床结局的可能性。

在诊断研究中，研究者有时非常清楚并没有或者无法在每个研究对象都使用同一种临床结局测量方法。例如，在癌症筛查研究中，只有在先前检验指标（如影像检查）结果为阳性的研究对象，病理结果才极可能会用作一个参照标准。这时，有两种情况可能会发生：*部分验证*，即在检验指标为阴性且没有参照标准结果的那部分研究对象中，临床结局数据完全丢失；或者*差异化验证*，即未被分诊去做首选的参照标准检查的研究对象，使用一个准确度与之[译者注：参照标准]不同的 — 往往准确度较低 — 替代标准而做评价（106，111，117，119，131-134，137）。这些临床结局判定上的差异，会影响预测因子和临床结果间相关关系评估，因此也会影响诊断模型的预测准确度。用于处理部分和差异化验证问题的方法已有过介绍（138-141）。

案例。Han 和同事（87）验证了一个用于预测严重创伤性脑损伤患者“六个月后不良临床结果”发生风险的模型。对所有被纳入到该单中心研究中的患者，均使用 GOS（5 分制量表中得分 1 至 3 即为不良结局）来判定临床结局。该信号问题应该被判断为“是（Y）”。如果该研究中的一家医院采用了不同的工具（如功能状态检查）来测量其临床结局，因为结局

测量工具[译者注：在参研医院之间]不一致，这会导致潜在的偏倚风险。该信号问题因此可以被判断为“可能否（PN）”或“否（N）”，以反映可能存在的偏倚风险。

3.5 结局指标是否是在不知晓预测因子信息的情况下而被判定的？

理想情况下，临床结局应该在不知晓预测因子信息的情况下，而被判定。这类似于随机对照试验，其中临床结局应该在不知道治疗措施分组的情况下而被判定。知道预测因子的测量结果可能会影响临床结局的判定，并导致带有偏倚的模型预测准确度 — 通常是因为高估预测因子和临床结局之间的相关关系（111，115，117，119，132-134）。对于客观结局指标来说，比如全因死亡，或新生儿是自然分娩还是剖腹产，该偏倚风险相对较低；但对需要主观解释的临床结局（如特定原因所致的死亡），该偏倚风险常常较高。

有些临床结局本来就很难通过一个测量或检查方法来判定。像在信号问题 3.3 中讨论过的，诊断和预后研究常常不可避免地需要采用专家共识小组或终点事件判定委员会[译者注：来决定临床结局]，用这些方法，临床结局判读可能需要知晓预测因子的信息。如果研究的确切目的是评价特定预测因子[译者注：能带给一个模型]的预测增加值，或比较相互竞争的模型的预测性能（如，基于同一个数据集验证多个模型），盲测临床结局会变的更加重要，以防过高估计该特定预测因子所带来的增加效应，或者防止对一个模型带有偏袒的喜爱。

系统评价作者应该仔细评价预测因子信息对判读临床结局的人是否是已知的。如果该信息在结局判读时即可获知，抑或是如果这个问题不够清楚，在判断该领域的总体偏倚风险时，应该考虑这些的潜在后果。该总体偏倚的判断应该考虑到临床结局存在的主观性特征以及系统评价研究问题。

*案例。*在 Rietveld 和同事的诊断预测模型研究（89）中，临床结局是以细菌培养为参照标准而确定的眼部细菌感染。细菌培养结果的判读具有一定的主观性。因此，该文作者明确告知读者其研究的盲测情况：“全科医生没有收到细菌培养结果，且分析细菌培养的微生物学家不会知道检验指标的结果[该研究的候选预测因子]”（89）。该信号问题因此应该被判断为“是（Y）”。

3.6 预测因子评测和临床结局判定之间的时间间隔是否恰当？

该信号问题是为了找出这样的情况，即预测因子评估和结局指标判定之间的时间间隔不够合理（要么过长，要么过短）。这个判断需要一定临床知识去决定恰当的时间间隔，并且也取决于临床背景。

在诊断研究中，其模型是预测临床结局（即基于参照标准确定的目标疾病）是否在做出预测的时候即已出现（文本框 2），理想情况下，预测因子（检验指标）和临床结局的评判应该在同一个时间节点上就发生。在实际中，预测因子和临床结局的测量之间难免会经过一定的时间，在此之间，诊断结果的分类可能会变好或变糟。有时，判定该临床结局的发生与否需要一段时间的临床随访，因此，预测因子和临床结局测量之间的延迟，是该研究设计所固有的一个重要特征以降低偏倚（如 Oudega 和同事[85]一文）。

对于慢性病来说，预测因子测量和结局指标判定之间存在几天的延迟可能不会有问题，然而对于急性感染性疾病，甚至很短的延迟都可能是不合适的。反之，当参照标准涉及到临床随访时，可能需要稍微间隔一点时间，才能观察到临床症状或体征上的改变，由此显示在测量预测因子时疾病即已出现。有时候，用于预测因子测量和临床结局判定的生物样本可在同一时间节点上被采集到，因此即使该样本上的参照标准检查在稍晚的时候才能够做完，这时实际上并不存在足够疾病状态发生变化的时间间隔。

在预后研究中，预测因子测量和临床结局判定之间的时间间隔也可能会是过短或者过长，以至于无法获取到临床相关的结局。

对于诊断和预后模型来说，偏倚可以通过两种方式出现。第一，如果临床结局被过早确定，这个时候，相关临床结局事件无法被检查出来或者结局指标事件数目不具有代表性，偏倚即会产生。例如，如果一个模型想在手术切除结直肠癌肿块时，就可判断出有无发生癌症转移，其参照标准的随访时间点可能会对转移的检测造成一定偏倚。因为现有探查手段的局限性，选取一个过早的时间节点，会在探查到的转移灶数目上造成一定偏倚；在较早的随访时间点上，转移灶可能尚未长到足够大，无法探查出来。第二，[译者注：在各研究之间，]临床结局类型可能会因时间间隔而不同。例如，较早探查到的转移灶可能主要是肝脏转移，然而在随访一年的时候，可能会探查得到更多骨转移。因此，如果预测因子测量和临床结局判定之间的时间间隔，导致出现了没有充分代表性的临床结局类型（即转移部位）或数量的话，偏倚风险即会发生。

显然，一个系统评价可能想专门评价某临床症状或短期、或长期的临床预后；因此的话，预测因子测量和临床结局确定之间的时间间隔，也和纳入研究之于该系统评价研究问题的适用性相关。

案例。Rietveld 和同事（89）建立了一个诊断模型，想用于预测眼部结膜炎感染的微生物起因；由于在患者问卷和体格检查中收集预测因子以及采集结膜样本去判定细菌感染，是在同一次门诊就诊中完成的，其中时间间隔相关的偏倚风险显然已经被最低化。尽管需要 48 小时以上的菌落培养才能读取参照标准结果，然而这无需和偏倚挂钩，这是因为菌落培养结果其实可以反映样本采集时的疾病状态。该信号问题可以被判断为 “是（Y）”，表示较低的偏倚风险。

案例。在 Aslibekyan 和同事的研究中（86），一个模型被建立用于预测心肌梗塞；因为缺少因子测量和结局判定间时间间隔的相关信息，该信号问题应该被回答为“缺少信息 NI”。不同的时间间隔会对应于不同数目的可检测出的心肌梗塞结局事件。

领域三偏倚风险评级。表格九呈现了该如何回应领域三的信号问题以及得出总体评级。

适用性。该领域的适用性问题需要考虑：已建立或验证的模型中所预测的临床结局多大程度上与系统评价研究问题相匹配。如果[译者注：系统评价和预测模型中]用到了不同的临床结局定义、测量时间或判定方法，应该认为这存在一定适用性关切。例如，一个原始研究可能会使用一个复合型临床结局，而该结局指标的组成部分，却与系统评价研究问题中临床结局定义所包含的那些指标存在着差异（142）。

像领域一和二中讨论的那样，在想要评估某个特定模型在所有纳入验证研究中的平均预测性能的系统评价中，由于临床结局定义和测量上的差异，验证研究间出现预测性能上的异质性是可想而知的（17,40,44）。有些时候，研究者有意地使用不同的临床结局定义或者测量方法。此时，如果系统评价很确定是想纳入该模型的所有模型验证研究——不管什么临床结局的定义以及测量方法——这样做或许并不是问题。

领域四：数据和分析

使用不恰当的统计方法或者没能考虑重要统计问题，都会增加模型预测性能评估上的出

现偏倚的可能性。领域四即评价是否正确处理了关键的统计学问题。这些方面中的有些地方要求具备一定专业知识，我们建议该领域应该由至少一名在预测模型研究方面具有统计专长的研究人员去评价。判断依据文本框应该罗列并描述在处理该领域时需要考虑的重要问题。

九个信号问题可以帮助做出该领域的偏倚风险判断（见表格十）。

4.1 有特定临床结局事件的研究对象例数是否足够合理？

和所有医学研究一样，样本量越大就越好，因为这可以得出更加精确的统计结果——即更小的标准误和更窄的置信区间（CIs）。在预测模型研究中，总体样本量确实重要，但发生临床结局的研究对象例数更加重要。对于一个二分类结局指标来说，有效样本量是两类结局事件（即“发生结局事件”和“无结局事件”）中发生频次较小的那个。对于事件发生时间结局指标而言，需要考虑的关键点是截止到模型所预测的主要时间节点为止，发生结局事件的研究对象总例数。更重要的是，在预测模型研究中，发生结局事件的研究对象例数不仅会影响到统计结果的精确度，还会影响预测性能——也就是说，会是一个潜在的偏倚来源。模型建立和模型验证研究，分别所需要的（对应于低偏倚风险的）发生结局事件的研究对象例数并不相同。

模型建立研究。当模型建立和性能评价使用同一个数据集的时候，任何预测模型的预测性能在一定程度上都可能会被高估（49,50,146,147）。当样本量较小，且发生有临床结局的研究对象明显很少时，这种高估会更严重。如果最终模型中纳入的预测因子是从很多——相对的，有少量的研究对象发生临床事件——候选因子中筛选而来，并且如果预测因子的筛选依赖于单因素统计分析的时候（见信号问题 4.5），偏乐观的预测性能问题会更加让人担心。对模型建立研究样本量大小的判断，过去一直是依据每个变量所对应的事件发生数（the number of events per variable, EPV）。更准确地说，对于候选预测因子，需要估计临床事件发生例数和回归系数数量的相对比值。例如，一个有六个类别的候选因子，会要求五个自由度（即需要计算五个回归系数）。“候选”一词也是很重要的：它不是指纳入到最终模型中的预测因子数，而是指在预测模型建立的任意阶段所考虑过的全部预测因子总数。

尽管 EPV 大于等于 10 已经被广泛采用作为一个可以最小化过度拟合的[译者注：确定样本量的]参考标准（148-150），近期研究已经显示，该阈值缺乏科学基础（145），因此许多研究人员已经建议使用更高的 EPV 阈值（即 EPV 至少是 20）（145,151,152）。总体来说，

EPVs 低于 10 的研究极可能会出现过度拟合，而 EPVs 超过 20 的研究很少会发生过度拟合问题。然而，需要用于最小化过度拟合的样本量，具有一定的临床背景特异性，并取决于临床结局流行率、模型总体性能（R^2）、和预测因子的数据分布（143-145）。 因此，判断[译者注：一个预测模型研究]是否使用了一个恰当的样本量可能是很困难的，尤其是当 EPV 处于 10 至 20 之间的时候。想要最小化过度拟合的话，使用机器学习技术所建立的预测模型通常需要相当高的 EPVs（往往大于 200）才能最小化过度拟合（153）。

因此，有效样本量越小，且 EPV 越低时，最终预测模型纳入可疑预测因子的风险就越高（过度拟合的模型），抑或是，最终预测模型没有纳入重要预测因子（拟合乏适的模型）的风险就越高。过度拟合和拟合乏适都很可能会得出具有一定偏倚的模型表面预测性能估计值（49-51,146,147,154）。在 EPV 较小的情况下，作者有必要量化其所建预测模型的拟合不当程度（例如，通过使用内部验证技术）。通过使用内部验证，可以得出调整乐观偏差后的模型预测性能估计值，模型参数也会被调整（即，缩减回归系数）以减少其偏倚（见信号问题 4.8）。

模型验证研究。模型验证研究的目的，是基于不同于模型建立所用的数据集，量化一个已知模型的预测性能（文本框一）（8,49,50,155-157）。模型验证研究中的关注点在于对模型预测性能做准确且精准的评估，由此可以得出有意义的结论。一般建议，验证研究要纳入至少 100 例有结局事件的研究对象；否则，很可能性会得出一个带有偏倚的模型预测性能估计值（77,78,158）。

案例。Aslibekyan 和同事（86）建立了两个预后模型，以预测心肌梗塞的发生风险：一个仅纳入易获取的预测因子，另一个还额外考虑了各种饮食和血液生物标记物。尽管作者采用了病例对照设计，且设定有许多纳入和排除标准，其最终样本有 839 个心肌梗塞病例用于建立预测模型一，以及 696 例用于建立模型二。候选预测因子的确切数量并没有明确提及，但从论文方法学和附件表格一和二部分，我们可以估计作者极可能使用了 20 至 30 个预测因子，或者说自由度，因为作者将多个连续型预测因子分类成为五组。这意味着 EPV（若用最小的事件发生例数）在 696/20（即 35）和 696/30（即 23）之间。因为在任一一种情况下，EPV 都大于 10 甚至超过 20，该信号问题应该被判断为“是（Y）”，即低偏倚风险。

案例。Oudega 和同事（85）验证了一个诊断模型，可用于检查那些因其可疑深静脉血

栓症状而咨询社区医师的患者中发生深静脉血栓的风险。该验证研究的总样本是 1295 例带有临床症状的患者，其中 289 人发生了深静脉血栓（由 D-二聚体检测和腿部 B 超确诊）。因为该研究具有超过推荐用于验证的 100 例结局事件，该信号问题应该被判断为“是（Y）”，即低偏倚风险。如果该数值较低（如，仅有 80 或者 40 个患者发生了深静脉血栓），该案例信号问题应该判断为“可能否（PN）”或“否（N）”。

4.2 连续型预测因子和分类型预测因子是否被恰当地处理？

应该避免对连续型预测因子（如，年龄和血压）进行二分类化（159-161）。二分类化通常需要选取一个较武断的界值，例如，高于某些数值即划分为高（或不正常），低于某些数值即划分为低（或正常）。[译者注：对于此，]一个常见的错谬论据是，这种方法可有助于[译者注：结果的]临床解释，且保持了简约性。然而，这[译者注：其实会]导致信息丢失，并且一个纳入了二分类化连续型预测因子的预测模型，可能会有明显被拉低了的预测能力（159-162）。

例如，以中位数为界值来二分类化地处理一个变量，已经被证实会降低统计效能，等同于丢弃多达三分之一的数据量（163）。此外，可以对应[译者注：连续性]预测因子整个取值范围的、由模型所预测的风险概率，便不复存在：刚好略低于界值的人可能会被认为和恰好略高于界值的那些具有不同的风险，尽管他们的预测因子取值相差无几。相反，预测因子取值相差许多——但无论如何都在界值以上（或以下）——的两个人，会被认为有相同的风险值。预测因子和结局事件发生风险之间的线性（或非线性）关系因此[译者注：二分类化处理连续性数值]消失。当一个预测因子是经被广泛认可的界值而被分类化处理（即，并不是基于已获取的数据），即使信息业已缺失，其偏倚风险也是偏低的，因为界值事先已经被提前确定。

模型建立研究。当所纳入的预测因子都被维持其连续性时，其建立的模型会有较低的偏倚风险。应该仍需通过使用，比如，限制性立方样条或分式多项式技术，来检验预测因子和结局事件风险之间的相关性是线性的还是非线性的（49,50,164）。

如果二分类化的连续型预测因子被纳入进来时，所建模型即有高偏倚风险，尤其是当界值是经挖掘该数据集（例如，试图找出能最大化预测因子效应、或最小化相关 P 值的最优界

值）才被选取确定的时候（159-162），且为了找出[译者注：具有统计学]显著性的阈值而采取某种筛选策略时（49,50）。

当一个模型将连续型预测因子分类成为四组或者更多组别，而不是简单的二分类时，偏倚风险则会被降低，尤其是当该分类是基于被广泛认可的界值时（160,162）。然而，对于一个期待有较低偏倚风险的模型来说，需要清楚的一点是，界值的取值数目和取值点应该在数据分析之前就要被确定，即是预先明确的。出于像信号问题 4.1 中所讨论的同样的原因，内部验证之后通过调整模型性能的乐观偏差和预测模型参数，也可降低偏倚风险（另见信号问题 4.8）。对于在数据分析时经二分类化处理连续型预测因子，但没有经内部验证和缩减技术去调整该分析的模型建立研究，该信号问题应该被判定为“否（N）”。

模型验证研究。在模型验证研究中，该模型应该像起初在建模数据集中所拟合的那样，在验证数据集中，去评价其预测准确性。这意味着，原来所呈现的原始截距（或基线风险值）以及回归系数要被用于完全一致样式的相应预测因子。例如，如果体质指数（BMI）在起初被纳入时，在模型中即是二分类化处理的，验证研究应该照例用在相同界值上二分类化的 BMI 值，而不是用 BMI 的连续型数值、或另取界值做二分类化处理。如果在模型的建立和验证之间，预测因子出现了任何形式上的差异，其验证可能即会有高偏倚风险，因为源自模型建立研究中的 BMI 的预测因子——结局指标之相关性（回归系数），从效应上看，在验证研究中是被用在该预测因子的另一种形式。

案例。Oudega 和同事（85）验证了用于鉴别深静脉血栓患者的 Wells 量表。然而，作者说“该量表的最后一个条目（其他临床诊断的发生）并不曾被明确定义过，且常常在该量表的使用者之间引发争议。在我们的研究中，医师被要求对患者发生深静脉血栓的风险，按以下分值给出他们自己的评价：分值一是指深静脉血栓发生高风险，分值二指中等风险，或可能发生的其他临床诊断，三分即低风险或明确发生的其他临床诊断。为了规范医师在该条目上的评判，研究信息收集表中提供了可用于深静脉血栓疑似患者的七个其他常见临床诊断。如果一个病人被评为低或中等风险，我们即从 Wells 量表评分中减去二分用于数据分析。”因为这并没有真的偏离于原来的定义，该信号问题应该被判断为“是（Y）”。

案例。Perel 和同事（88）建立了一个预测模型（CRASH-2），用于预测创伤性脑损伤患者早期死亡的风险。在模型建立的过程中，他们将一个三分类变量“损伤类型”（插入伤、钝

挫伤或钝挫以及插入伤）视做一个二分类变量（插入伤组，钝挫伤和插入伤联合组）去进行数据分析；却不曾给出这样处理的原由。不管怎样，在建模时，连续型变量是被视作连续数据而进行分析的，因此对于该变量而言，从三分类压缩到二分类，可能是由于钝挫伤组中有较少的研究对象或结局事件。再者，损伤类型后来也没有被纳入到最终模型中去，因此减少预测因子分类组别，极不可能是为了去提高该预测因子的统计学显著性。所以，我们应该将该信号问题判断为“是（Y）”。当对 CRASH-2 模型做外部验证的时候，作者“使用了 CRASH-2 中所建模型的回归系数”，并且也采用了原来的预测因子及量纲（像原本所编制的那样）。因此，在评价模型验证时，判断“是（Y）”也是恰当的。

4.3 是否所有参与研究的研究对象均被纳入到数据分析？

和所有其他医学研究一样，在一个研究中招募到的所有研究对象，都应该被纳入到数据分析中去，否则可能会出现偏倚风险（48,111,165,166）。该信号问题牵涉到的议题是：从起初**符合纳入标准的研究样本**中排除一些研究对象。这并不是在讨论不恰当的纳入排除标准（这在信号问题 1.1 中即被处理），也不是讨论该如何处理预测因子或结局指标中的缺失数据（这包括在信号问题 4.4）。

已入组的研究对象通常会出于以下原因而被排除掉：无法解释（不清楚）的研究结果、离群值、或者预测因子或结局指标上出现了缺失数值。离群值、无法解释的数值或缺失值可以发生在任何类型医学研究中。如果剩下的被分析的研究对象不是完全随机，而相反是一个具有选择性的样本亚组的话，从数据分析中移除一些已入组的研究对象，可以导致带有偏倚的预测因子——结局指标相关关系，并导致所建或所验证的模型，其模型预测性能出现一定偏倚。对比纳入分析的研究对象和被删除的研究对象的话，预测因子和结局指标之间的相关关系，也会出现差异。例如，排除带有不清楚的预测因子数值（比如，影像或试验室检查结果）的研究对象，极可能会生成一个由预测因子数值位于其取值范围两极端的研究对象组成的样本[译者注：该样本中的研究对象预测因子取值要么明显较低，要么明显较高]。这反过来可能导致带有一定偏倚的、被过高估计的模型区分度（166）。当仅有较低比例的研究对象未被纳入统计分析的时候，偏倚风险可能比较低。然而，我们很难定义一个最低的、可接受的百分比，因为这取决于哪些研究对象被排除了，以及它是否是一个具有选择性的样本亚组。

偏倚风险会随着被排除研究对象的比例升高而增加。

基于常规医疗数据库或者登记数据的预测模型建立或验证研究，其研究对象并不会像预先设计的研究那样，而被正式招募入组，甚至数据的收集也是出于不同的原因，这类研究尤其易于发生该条目中的偏倚风险。当这类数据源被用在模型建立或模型验证时，用在数据分析中的研究对象之筛选，应当需要基于清晰的标准。对于使用这种常规医疗数据的预测模型研究，并不清楚其潜在偏倚的严重程度是多少，因为它们不曾充分地报告符合标准的相关信息，且没有阐明排除研究对象的原因是什么。

案例。在 Han 和同事的研究中（87），所有 300 个研究对象，均满足为验证三个不同版本的 IMPACT 创伤性脑损伤模型（即核心模型、扩展模型和实验室模型）而设的符合标准。之后，研究人员因血糖水平的数据缺失，而从验证实验室模型的样本中排除了 36 个（12%）研究对象；但所有研究对象都可被纳入用于核心模型与扩展模型的验证。对于核心模型和扩展模型的评价而言，该信号问题应该被判断为“是（Y）”，因为所有研究对象均纳入了数据分析。对于实验室模型来说，取决于从数据分析中排除的 36 个研究对象（12%）所造成的关切程度，该信号问题应该要么判定为“可能否（PN）”或者判定为“可能是（PY）”。这[译者注：种判断]有赖于一定的临床知识，和对丢失的血糖水平检测数据是否极可能与创伤性脑损伤严重程度有关这一问题所做出的判断。

4.4 出现缺失数据的研究对象是否被恰当地处理？

像此前条目提到的那样，当所分析的研究对象属于一个具有选择性的样本，而不是一个原始全样本的完全随机抽样，简单地从数据分析中排除已入组的但有丢失数据的研究对象，会导致带有一定偏倚的预测因子——结局指标因果关系，以及有偏差的模型性能（167-177）。如果一个研究报告没有提及缺失数据，存在缺失值的研究对象很可能已经被从数据分析中移除出去（“可用资料分析”，或“完整资料分析“），因为如果没有采用其他方式处理的话，统计工具包会自动排除在所分析的任一数据上有缺失值的研究对象。大量的文献综述都显示，可用资料分析或完整资料分析是预测模型研究在处理丢失数值时最常采用的分析方法（68,178-186）。

用于处理缺失值的最恰当方法是多重插补法，因为该方法可以得到偏倚最少的、有正确

标准误和 P 值的统计结果（167-173,175-177）。在预测模型研究中，不管是模型建立研究（173,176,187）还是模型验证研究（176,188-190），多重插补法在偏倚和精度问题上都表现地比其他方法更具优越性。相比于无法解释的或者离群的数值，用另一个变量类别来囊括缺失值，并不是一个恰当的方法；在预测模型研究中，这种设立缺失值指示变量的方法，会造成有偏差的分析结果，因此该信号问题应该判断为“否（N）”（172,177）。由缺失值所导致的偏倚风险，会随着缺失值所占比重的增加而增高，但很难明确规定一个可接受的、可用作低偏倚风险判定界值的最低缺失值比重（173）。如果作者提供了（被排除样本和被分析样本）两组间的预测因子和结局指标数据分布（百分比、平均值或中位数），或者呈现了在纳入和排除丢失值之后所得出的预测因子——结局指标因果关系和模型预测性能之比对，这会有助于判断可能存在的偏倚风险。纳入和排除缺失值之后，如果仍能得出相似的分析结果，这即有力地说明：数据分析结果极可能不存在偏倚。如果该比较未被呈现出来，且研究人员也不曾使用插补技术，我们建议将该信号问题判断为“可能否（PN）”或“否（N）”，尤其是在相当比例的研究对象因为丢失值而被排除的时候。

有时候，当一个模型在用其他数据做验证，但其中却系统性地缺失了模型中的一个预测因子数据（如，不曾测量过该因子的数据）时，研究者可能会在简单地将该预测因子从模型中移除出去之后，再去验证原始模型（即，原始预测因子权重或回归系数）。这会导致高偏倚风险，这样的研究应该在该问题上被判断为 “否（N）”。[译者注：这是因为]如果这个模型起初不曾考虑这个被移除了的预测因子，而被拟合的话，其余所有预测因子的回归系数都会变的不再一样。

案例。基于一个缺失值很少的数据集，Perel 和同事（88）建立了一个预后模型，因此，他们做了完整资料分析。在同一篇文章中，作者还报告了该模型的外部验证，此时他们使用了多重插补法。在模型建立研究中，存在缺失值的研究对象的比例有多低是并不清楚的，数据完全可获取的研究对象和被排除的研究对象之间也不曾做过对比，这让判断该模型建立是否存在一些偏倚风险变得比较困难。在模型验证研究中，作者使用了多重插补法，意味着他们知晓该技术；如果模型建立过程中需要用到多重插补法的话，他们极可能也会在建模中就使用该方法。据此，严格地说，对模型建立研究应该将该信号问题判断为 “缺少信息（NI）”，在模型验证研究应该判断为 “是（Y）”，然而模型建立也可以被判定为“可能是（PY）”。

案例。Aslibekyan 和同事（86）声称，他们在模型建立过程中使用了完整资料分析，并排除了 10%的研究对象。他们没有提供更多信息，可去证实完整资料分析是一个可信的方法——即，纳入和排除的研究对象是否足够相似，以至于被纳入的研究对象可以近似等同于一个原始样本的完全随机抽样。据此，对于模型建立，该信号问题应该判定为“否（N）”。对于模型验证而言，丢失数据和其处理方法均未被提及，因此严格地说该信号问题应该被判定为“缺少信息（NI）”，但考虑到其对模型建立部分的报告，且所有临床研究往往都会有一些丢失数据，该问题甚至也可以给“可能否（PN）”。

4.5 是否避免依据单因素分析来筛选预测因子？（仅适用于模型建立研究）

一个数据集中往往会有许多变量可被用作候选预测因子；在许多研究中，研究人员会希望减少模型建立过程中的预测因子数量，以得到一个较简单的模型。

在一个单因素分析中，每个预测因子都会被检查其和临床结局之间的相关性。科研人员通常选择那些具有统计学显著性的单因素相关性（例如，$P < 0.05$）的预测因子，以纳入到最终预测模型的建立。这个方法可能会导致不恰当的预测因子选留，因为这些预测因子是作为单个独立因子基于其统计学显著性结果，而被筛选的，而没有同时考量到其他预测因子的情况（49，50，191）。当单因素分析促使从模型中移除一些变量的时候，偏倚即可发生，因为一些预测因子只有在同时调整分析其他因子时才会显出其重要性；有些因子在先前的研究中，即已明确其重要性；在一个特定的建模数据集中，有些因子无法达到其统计学显著性（例如，由于该数据样本量较小）。并且，预测因子也可能是由于其在该建模数据中和结局指标之间的虚假（偶然）的相关性，才被筛选出来的。

一个更好的用于判断多因素建模分析中移除、合并或纳入候选预测因子的策略，是用非统计学方法——即，对候选因子和结局指标之间的相关性不做任何统计学单因素预筛选的方法。这类较好的方法，可以基于先前已经确定的预测因子相关的现有知识，结合对适用于目标场合的预测因子测量的可靠性、一致性、适用性、可获取性和测量花费之考量。已被认可的预测因子和那些有临床可信度的因子，无论其统计学显著性如何，均应该被纳入并保留在一个预测模型中（49，50，192）。再者，一些事先不在预测因子和临床结局之间做统计检验的统计方法（例如，主成分分析），其实也可以被用于缩减建模中的预测因子数量。

在建模过程中，可以采用预测因子筛选策略：去移除预测因子（例如，向后选取法）；并借此拟合出一个比较精小简单的最终模型（49，50，192）。然而，采用上述这种多因素因子筛选策略，对预测模型在现有建模数据上的潜在过度拟合的影响，应该用内部验证和乐观偏差调整策略去做检验，这些在信号问题 4.8 中会讨论到。

当模型建立恰当地避免了对候选因子做单因素分析筛选，并且没有在多因素建模之前对预测因子做过单因素筛选的证据，预后研究应该被判断为 “是（Y）”或“可能是（PY）”。如果在多因素建模之前使用了单因素分析对预测因子做过筛选，该信号问题应该被判定为 “否（N）”。

案例。在 Perel 和同事（88）建立其模型之前，他们即咨询了模型的潜在用户，为的是基于[译者注：预测因子]已知的重要性和其在院前、急救中和急诊室这些临床场合使用的便捷性，来找出候选预测因子以及它们的相互联系。之后，研究者在多因素分析中纳入了所有由此找到的候选因子。至于哪些预测因子需要保留在最终模型中，作者基于下面几个考虑做出决定：临床推理、在模型使用的时候测量预测因子能否可知、以及在临床场合使用设备采集预测因子数据的可操作性。尽管还有其他预测因子可能也可以被认为是很重要的，在这里，预测因子的选取并没有采用可能存在偏差的单因素变量筛选策略。在该信号问题上，该研究因此可以被判断为 “是（Y）”。

案例。Rietveld 和同事（89）使用以单因素分析（$P \leq 0.10$）为基础的预测因子筛选方法，为多因素模型筛选预测因子。该信号问题因此可以被判断为 “否（N）”。如果所有预测因子都被纳入到多因素分析中，而事先没有做单因素筛选，可以判断为“是（Y）”。

4.6 是否恰当地考量处理了数据中出现的复杂问题（如，删失数据、竞争风险数据、对照组研究对象的抽样）？

对于预测模型的建立和验证，一定要确保所用的统计方法和相应的统计假设，是适用于研究设计和所分析的结局指标数据类型的。在此，我们想让你们注意一些与数据复杂性相关的重要问题，如果在数据分析中没有被恰当地处理的话，它们可以会对模型预测性能评估值造成一定偏倚风险。

像信号问题 1.1 所讨论过的，如果一个预测模型使用了病例队列或巢式病例对照设计，分析方法必须将（源自原始队列的）抽样比考虑进去，以能够恰当评估结局事件绝对发生风险（97，99，105，109）。例如，在一个使用了巢式病例对照设计的诊断预测模型（建立或验证）研究中，其全部对照组研究对象中的一部分是从原始队列中抽取的，这时所使用的逻辑回归中，需要将其对照组研究对象用抽样比的倒数做加权处理，而不是使用常规的逻辑回归分析；否则，该模型所预测的事件发生风险会存在一定偏倚。当对抽样比做了恰当的调整之后，这会缓解一些信号问题 1.1 中提到的偏倚风险。如果没有做调整的话，系统评价员应该将信号问题 1.1 或该部分的信号问题判断为 “否（N）”，且仅需在一处做此判断。

对于想要预测长期结局事件，且其中出现数据删失的预后模型，需要使用事件发生时间分析（例如 Cox 回归），来纳入有删失数据的研究对象，一直到他们随访的终点。使用逻辑回归建模，其中简单地排除有删失数据、随访不完整的研究对象，是并不恰当的。用这种错误的逻辑回归方法，会产生一个具有选择性的数据集，其中仅纳入了未发生临床事件的研究对象中的少数，这会使预测风险出现偏倚，因为发生临床结局的研究对象会因此占到更大比重。事件发生时间分析却可以恰当地处理这些有数据删失的个体。

如果出现有重要的竞争风险，在一个预后模型建立时，它们也应该被考虑进事件发生时间数据分析。竞争风险的一个例子是，对一个用于预测再次髋关节置换发生风险的模型来说，其首次髋关节置换的老年患者中，在再次髋关节置换发生之前可能就出现的死亡事件。如果没有正确处理竞争风险的话，绝对风险预测会被高估，且是存在一定偏差的，因为发生竞争事件的患者被简单粗略地做删失处理了（193）。

另外，当每个人可以发生超过一次结局事件时，也需要使用正确的建模方法，比如在一个癫痫发作模型中，一些人会有多于两次的癫痫发作。这时需要使用多水平或随机效应（逻辑或从生存）建模法，以避免在预测因子的效应上出现低估和偏倚（194-197）。

我们需要具备一定统计专长，才能在具体研究中，识别出上述这些和其他问题。我们在这里突出讲解的这些问题，通常是预测模型研究中需要注意的最重要的问题。如果系统评价员觉得，一项研究没有注意到重要的统计复杂性，这意味着该信号问题即是高偏倚风险。

案例。Aslibekyan 和同事（86）使用条件逻辑回归模型，来建立一个用于心肌梗赛的预后预测模型。所纳入的研究对象提供有 1994 至 2004 年的数据；然而，不清楚的是：是否所

有研究对象都记录到该时间段起始时（相比于，在 1994 年之后，会因此有更短的随访时间）的预测因子数值。如果所有研究对象在 1994 年入组时就记录有预测因子数值，该模型可以预测 10 年期的（即，截至到 2004）心肌梗塞发生风险，并且是容易解释的。然而，如果一些研究对象在 1994 年之后入组，逻辑回归的可解释性和偏倚即会是一个需要注意的问题，因为模型的预测不是针对于一个特定时间段，且随访时间长度问题正在被忽略。如果研究对象有不同的随访时间，更好的是拟合一个生存分析模型，这样可以做不同时间上的风险预测，并且也可以纳入延迟入组的研究对象。再者，即使研究人群中包括有达 86 岁高龄的患者，由非心肌梗塞状况所致的死亡事件，其竞争风险的流行率却并不可知。这些问题，可能是该研究病例对照设计（而非队列设计）的自然属性所导致的后果。因此，由于这些统计学复杂性，偏倚风险是不可避免的，该信号问题应该被判定为“否（N）”或 “可能否（PN）”。

案例。在 Rietveld 和同事的研究（89）中，用标准逻辑回归建立一个诊断模型，是相对直观的，因为通过采用全队列（没有用到抽样），所建立的模型是为了预测细菌性结膜炎的风险。因此，这里不涉及随访、数据删失或者竞争事件等问题。在此，该信号问题应该被判断为“是（Y）”。

4.7 是否恰当地评估了相应的模型性能测量指标?

文本框四提供有多个适用于多因素预测模型的性能评价指标。PROBAST 是被设计用于评价涉及多因素模型的研究，这些模型可被建立或验证以对个体做出诊断或预后性预测，即个体化预测（文本框一）。据此，想要完整地评判一个模型的预测性能，系统评价员一定要评估模型的校准度和区分度（如 c-指数），探讨模型所预测风险的整体范围（7,8）。如果校准度和区分度没有被评价，该研究即存在偏倚风险，因为该模型能否做出精准个体化风险预测的能力或性能不是完全可知的（文本框四）。

当观察到校准图或表仅有少数几个组的时候（例如，由于一个小样本仅有很少的结局事件），为了恰当地判断该信号问题，需要评估校准图。当不存在对比了预测和观察结局事件风险的校准图或表时，对于该信号问题，那些仅报告校准度统计检验的研究，应该被判定为“否（N）”。

另外，用于评价模型校准度和区分度的方法，要适用于模型预期想预测的临床结局。用

于评价经逻辑回归所建立的预测二分类临床结局的模型的方法，并不适用于评价由 Cox 回归而建立的预测长期结局事件发生风险（例如，五年死亡率或生存率）的模型，因为删失数据要被考虑进[译者注：Cox 回归]。不管是在模型建立研究还是在模型验证研究中，当评价预后模型校准度和区分度的时候，没有考虑删失数据，即意味着该信号问题应该被判断为“否（N）”或“可能否（PN）”。

有些研究还会提供分类指标，包括敏感度、特异度、预测值，或再分类指标（如净再分类指数），来呈现模型预测性能，通常此时不再报告模型校准度和 c 统计量（文本框四）。分类指标最常出现在诊断模型研究。评价分类性能以及再分类性能时，性能参数会要求在模型所预测的风险范围中引入一个（或更多）阈值。使用阈值，可以让研究者能依照具有潜在临床相关性的风险阈值，来报告模型预测性能，而不是模型所预测的完整风险值范围。但是不管怎样，阈值的使用通常会导致信息的丢失，因为模型所预测的完整风险范围并没有被充分地利用；并且阈值的选择也可能是由数据决定的，而非出于临床原由被事先拟定（另见信号问题 4.2）。这种操作可导致在所评价的分类（再分类）指标上出现重要偏倚，尤其是当阈值的选用是为了最大化模型表面性能的时候（83,198）。当阈值的选择不是预先确定的时候，这些方法[译者注：为最大化模型性能而选定阈值的方法]容易导致偏倚风险，该信号问题应该被判断为“否（N）”。该信号问题也应该被判断为“否（N）”，如果报告分类和再分类指标而不是模型校准度的时候。在模型所预测风险值被分组处理之前，需要用校准度来帮助了解所预测的风险值是否正确（文本框四）。

案例。Rietveld 和同事（89）用 Hosmer-Lemeshow 检验评价了校准度，该检验的 *P* 值是 0.117；他们将此解释成模型已经被很好地校准了。如果这是唯一用于评价模型校准的指标的话，该信号问题应该被判断为“否（N）”，因为这样一个 *P* 值既不能说明说明校准错误的出现与否，也不能说明错误校准的程度。然而，在其表格四，作者呈现了各个亚组的平均预测风险和置信区间，以及相应的结局事件发生频率观察值。该校准表呈现了模型校准度，因此，该信号问题可以被判断为“可能是（PY）”。

案例。在验证一个预测创伤性出血患者早期死亡事件的模型时，Perel 和同事（88）通过呈现一个比对了风险观察值与预测值（按预测风险的每十分之一为一个组别分组）的校准图，来评价模型校准度。通过这种形式呈现校准度，可以让读者判断模型在横跨整个风险值范围上的精确度。通过加上一个非参数（LOWESS，局部加权回归散点平滑法）平滑线，该

图可以得到进一步改善。作者还报告了一个 c 统计量，可以让读者判断模型的区分能力，即使没有 95%置信区间去反映该[译者注：指数]估计值的不确定性。这个研究可能是低偏倚风险，该信号问题可以被判断为“是（Y）”。

4.8 模型过度拟合和模型性能上的乐观偏差是否被恰当地考量处理？（仅适用于模型建立研究）

像在信号问题 4.1，4.2 和 4.5 所讨论的，用建模数据来量化一个模型的预测性能（表面性能），往往会得到偏乐观的预测性能估计值，这是由于可能会存在过度拟合问题——即，模型过度适应建模数据集。当以下任何情况出现时，乐观偏差出现的可能性便会更高：过小的结局事件发生总数；相比于候选因子个数，结局事件发生数较小（即 EPV 较小）；连续型预测因子的二分类化；使用以单因素分析为基础的预测因子筛选策略；或在较小数据集（EPV 较小）的多因素分析中，使用传统的逐步因子筛选策略（例如，向前或向后筛选）（49，50）。

所以，模型建立研究应该都要进行某种形式的内部验证，比如自助抽样法和交叉验证法。如果样本量和 EPV 并不是极其得大，通过内部验证来量化所建模型的过度拟合和其在预测性能上所表现出的乐观偏差是非常重要的。内部验证意味着仅使用[译者注：建模所用]原始样本的数据，即验证是基于[译者注：与建模数据]相同的研究对象数据。如果乐观偏差出现的话，很重要的进一步处理是调整或缩减模型预测性能估值（例如，c 统计量）和最终模型中的预测因子效应量。不幸地是，这通常很少会被做到。当未缩减的模型用在其它个体[译者注：建模所用数据之外的人]时，使用不曾被缩减或不曾调整乐观偏差的回归系数，会导致带有一定偏差的（通常偏于极端）预测值。例如，一个均匀（或线性）缩减因子——可以从自助抽样过程中得到，可以用于[译者注：缩减]每个预测因子的估计效应量。惩罚回归法也正变得越来越常用，如岭回归和套索回归，它们可以让每个预测因子的效应量能被不同程度地缩减，甚至可以允许完全删除一些预测因子（199）。有些研究人员说，缩减技术彼此之间差别不大（200，201），但其他人倾向于支持惩罚类策略（49，199）。

在建立一个预测模型的时候，对于那些样本量较小以及 EPV 偏低的研究和那些采用逐步预测因子筛选策略的研究而言，调整其模型的过度拟合和乐观偏差显得更加有必要。当已经用了内部验证和缩减技术的时候，该信号问题应该被判定为 “是（Y）”。恰当调整过度拟

合，可以减少因低 EPV （信号问题 4.1）、连续型因子二分类化（信号问题 4.2）、和预测因子筛选策略（信号问题 4.5）等问题，而引起的偏倚风险。那些建立预测模型却忽视或没有检验模型的错误拟合的研究，在该信号问题上应该被判定为 “否（N）”；尤其是当存在小样本量、低 EPV、连续型预测因子的分类化、或使用预测因子筛选策略的时候。但有一个特例，即 EPV 高且样本量极大的建模研究，其中无需过于担心过度拟合问题。

有些研究可能会使用不恰当的方法来检验或调整乐观偏差。研究人员常常随机将一个数据集在个体水平上拆成两个组（一个用于模型建立，一个用于内部验证），这已经被证实为一个不太恰当的检测乐观偏差的方法（154，202）。研究人员有时也会用自助抽样法和交叉验证技术去检验乐观偏差，但却不会重复整套的模型建立流程（例如，在单因素和多因素分析中，用预测因子筛选策略），因此他们可能会低估其模型的实际乐观效应（203，204）。在该问题上，这些不恰当的方法，都可以导致 “否（N）” 的判定。

案例。Perel 和同事（88）使用自助抽样法，检查了其模型建立中的乐观偏差。作者写到 “我们用置换法从原始数据中抽取了 200 份、和原始建模数据具有相同样本量的样本。在每个自助抽样样本，我们都重复一次完整的建模流程，包括变量筛选。我们然后取自助抽样样本得到的 200 个模型的 c 统计量平均值。此后，200 个模型中的每一个都在原始样本运行一次，我们再估计出一个平均 c 统计量。两个 c 统计量平均值之间的差值，就是我们预后模型 c 统计量的乐观偏差” （88）。然而，尽管 c 统计量的乐观偏差已被检查，却没有考虑到绝对风险预测上的乐观偏差，并因此没有对预测因子回归系数用收缩因数[译者注：去做调整]。虽然如此，该研究所报告的 c 统计量乐观偏差非常小（0.001），因此该信号问题应该被判断为“可能是（PY）”或“是（Y）”。

案例。Rietveld 和同事的研究（89）应该被定级在 “可能否（PN）”或“否（N）”，因为他们没有使用统计学方法来处理过度拟合。作者使用的预测因子筛选策略，首先是以单因素分析的 *P* 值为依据，然后看多因素分析的 *P* 值；他们还考虑了纳入预测因子之间的交互效应；因此，乐观偏差出现的可能性很大。然而，作者没有检查过度拟合，并且也未因乐观偏差而尝试做收缩。作者确实报告说曾经使用自助抽样技术。然而，这似乎是为了检查离群值的影响和计算置信区间，而不是为了检测过度拟合和区分度与校准度的乐观偏差。

4.9 最终模型中的预测因子及其权重分配是否和多因素分析中的结果相符？（仅适用于模型建立研究）

最终模型中的预测因子和回归系数（包括截距或基线成分），应该被完整地报告出来，以使其他人可以在其他个体上正确地使用该模型。所呈现的最终模型和所报告的多因素分析结果（比如，截距和预测因子回归系数）之间出现不匹配问题，也是很常见的。2010 年的一个肿瘤预测模型综述发现，在 38 个最终预测模型方程式中，只有 13 个用到了和最终报告的多因素分析一样的预测因子和回归系数；8 个使用了相同的预测因子但回归系数却不同；11 个既没有使用相同的回归系数，也没有使用相同的预测因子；还有 6 个其使用的从多因素分析结果推导最终预测模型的方法并不十分清楚（121）。

当所呈现的最终模型和多因素分析所报告的分析结果之间不匹配时，偏倚即可出现。出现这个[译者注：匹配错误]问题的一种情况是，当从一个较大[译者注：预测因子个数较多]的模型中删除无统计学显著性的预测因子，近而构建出最终模型时，仍然使用较大模型中的预测因子回归系数去确定最终模型[译者注：而不是重新拟合回归系数]，这[译者注：模型中留下的回归系数]其实已经不再正确。当从较大模型中剔除预测因子时，很重要的一步是，重新估算较小模型中的所有因子回归系数，因为后者这时已经成了最终模型。这些新近估算的预测因子回归系数很可能与此前不同，尽管从较大模型中移除的是无统计学显著性或不相关的预测因子。

当一项研究呈现出这样一个最终模型，其预测因子和回归系数均能与多因素回归分析或建模结果相匹配，该信号问题应该被判定为“是（Y）”。如果最终模型仅是以从多因素回归分析中选取的预测因子作为基础[译者注：来构建较小的模型]，却没有重新拟合较小的模型，该问题应该被判定为“否（N）”或“可能否（PN）”。如果没有报告可推导预测因子和回归系数的多因素模型的相关信息，该问题应该被判定为“缺少信息（NI）”。

该信号问题并不是要检查那些为最终模型去筛选预测因子的不恰当方法；这些方法由信号问题 4.5 处理。

案例。Perel 和同事（88）报告了一个有每个预测因子的 OR 值和交互效应量的最终模型，以及有预测因子回归系数的模型方程式。该全模型可以被判定为“可能是（PY）”或“是（Y）”，因为最终多因素分析中的所有预测因子和由多因素分析得出的回归系数，都被纳入

到了模型中去。Perel 和同事还报告了一个另行建立和验证的简化模型，以及在简化模型中重新拟合的回归系数。假设简化模型未曾被重新拟合，去矫正含有更少预测因子的简化模型中的回归系数，该研究在该问题上理应被判定为“否（N）”。

*案例。*Rietveld 和同事（89）将所有最终模型中的预测因子都纳入到了简化临床量表中，该量表用整数来提升其可用性。这些凑整的数字评分，是基于最终模型预测因子的原始权重而推导出的：每个多因素回归中估算出的回归系数，都除以其中最小的回归系数（即预测因子“瘙痒”相应的回归系数数值 0.61），然后将结果凑整为最临近的整数。然而，对于预测因子“双眼粘合”，回归系数 2.707 却被四舍五入成 5 而不是 4（因为 2.707/0.61＝4.4）。该信号问题应该被判定为“否（N）”，因为分配给预测因子的权重，并不能完全对应到最终多因素分析的结果。

领域四偏倚风险评级

表格 10 呈现了应该如何回答该领域的信号问题，以及如何得出领域四的总体判断。

添加额外信号问题以调整 PROBAST

我们鼓励研究人员也用 PROBAST 去评价那些涉及二分类或事件发生时间结局指标之外的、其他临床结局（如，有序多分类、无序多分类或连续型临床结局）的预测模型研究，以及那些使用除回归技术之外的、其他分析方法（如，基于树的分析技术、机器学习或人工智能技术）的预测模型研究。为了处理其他结局指标类型或建模技术相关的偏倚，系统评价员可能需要通过添加额外信号问题，来微调一下 PROBAST。例如，当处理预测连续型结局的模型时，变量平均事件发生数（领域四）相关的信号问题，可以被修改用于处理变量平均总研究对象数（49）。当研究使用机器学习或人工智能技术时，大部分（尽管不是全部）信号问题依然适用。额外问题可能需要被增加进来，因为这些技术会用到不同的预测因子筛选策略、不同的因子——结果相关性估算方法、和不同的过度拟合调整方法。

当调查分析有关某特定预测因子在一个现存模型上所能增加的预测价值的研究时，PROBAST 使用者可以增加一个针对用于量化增加价值分析方法（如，净再分类指数，或决

策曲线分析）的信号问题（84，205）。相似地，如果调查那些针对另一临床场合而对某现有模型再校准或更新的研究时，PROBAST 使用者可以增加一个讨论再校准或更新方法（例如，再校准基线风险或危险，更新原始回归系数，或整个模型的再拟合）的信号问题。

不管什么时候若系统评价员想修订或增加信号问题，为了和现有信号问题保持一致，它们都要做恰当地遣词造句，使得 “是（Y）”即对应于低偏倚风险，“否（N）”即意味着高偏倚风险。也应该制定出该如何评价每个新增信号问题的具体指导性说明。

我们不建议从该工具中删除信号问题，除非它们很显然和系统评价研究问题无关。如果对于某个特定信号问题，所有研究得到的都是 “是（Y）”或“否（N）”，在该工具中留下此问题仍然是有必要的。这会呈现出，对那个系统评价来说，一个特定的偏倚来源或适用性关切是否会是一个潜在的问题。

第四步：总体评价

表格 11 呈现了一个预测模型评估之偏倚风险和适用性的总体评价。如果一个预测模型评估的所有偏倚风险和适用性相关的领域都被判定为低风险，给出“低偏倚风险”或“适用性关切低”的总体评价是很恰当的。如果一个预测模型评估在至少一个领域被判断为高风险，它就应该被判定为具有“高偏倚风险”或“适用性关切高”。如果该预测模型评估在一个或多个领域上是“不清楚”，而剩余领域被定级为“低”，它总体上应该被判断为“偏倚风险不清楚”或“适用性关切不清楚”。

表格 11. 偏倚风险和适用性的总体评价

得出预测模型评估偏倚风险的总体评价	**标准**
低偏倚风险	如果所有领域均被判定为低风险； 如果一个预测模型被建立起来却没有对其做外部验证，尽管所有领域均判定为低风险，也建议考虑将总体评价结果降级为高偏倚风险。除非，该模型的建立是基于极

	大样本数据集而实现的，且已经进行过某种内部验证，才可认为该模型存在低风险
高偏倚风险	如果至少一个领域被判定为高风险
不清楚	如果至少一个领域被判定为不清楚，且其他领域均为低风险
得出预测模型评估适用性关切的总体评价	
低度关切	如果所有领域均被判定为低度关切，该模型预测评估可被认为有低适用性关切。
高度关切	如果至少一个领域被判定为高度关切，该模型预测评估可被认为是适用性关切高。
不清楚	如果至少一个领域被判定为不清楚，且不存在高度关切

对于一项研究而言，因广为人知的、涉及到评分得分的问题，PROBAST 不应该被用于推导总体“质量评分”（206，207）。PROBAST 使用者应该判断及讨论每个领域中所出现问题的影响，而不是费力去计算这样一个总得分。

PROBAST 评价在系统评价中的呈现和使用

报告偏倚风险和适用性评价，是解释说明系统评价证据强度的一个重要途径。所有的系统评价都应该包括偏倚风险和适用性关切的叙述性总结，并联系到这会怎样影响对研究结果的解释和推论强度。另外，也应该呈现一个表格，展示对偏倚风险和适用性的所有评价结果。表格 12 即是一个范例，可帮助找出全部纳入预测模型和其研究中的关键问题所在。一个总结图（见图）可以呈现出每个领域的偏倚风险和适用性，并可按评价结果划分纳入研究占比，

该图可以有效总结所有研究。这符合 PRISMA 声明（系统综述和 Meta 分析优先报告的条目）条目 22 的要求（34，35）。这些总结其实还是不够的，即：并没有由此去讨论这些可见的评价结果程式，对系统评价研究问题有关的证据基础意味着什么。

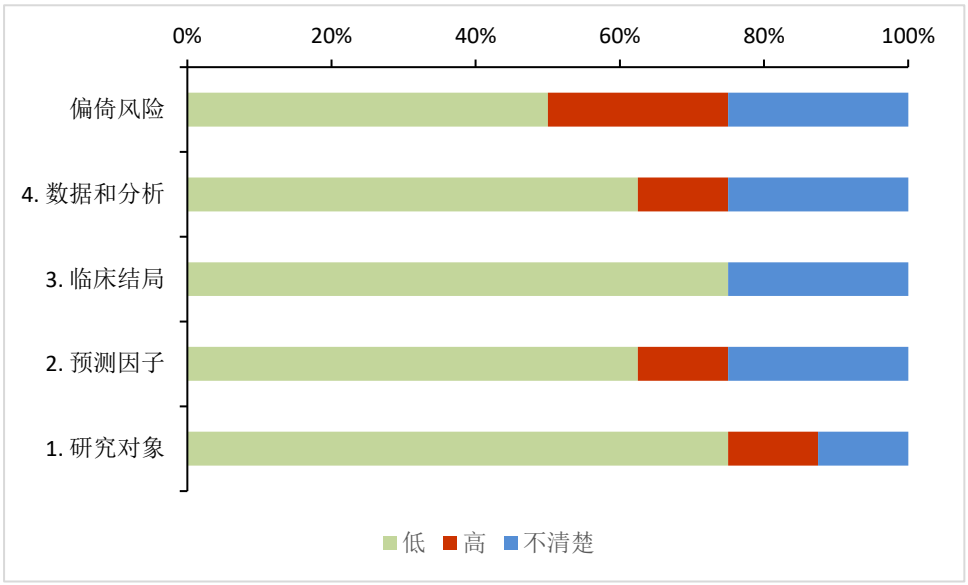
表格 12. PROBAST 评价结果的表格呈现建议*

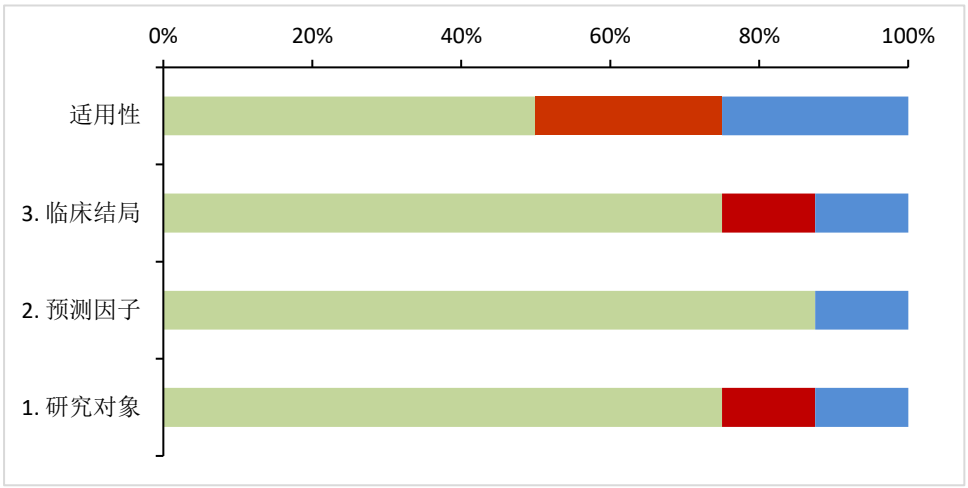
研究	ROB				适用性			总体评价	
	研究对象	预测因子	临床结局	数据分析	研究对象	预测因子	临床结局	ROB	适用性
1	+	-	?	+	+	+	+	-	+
2	+	+	+	+	+	+	+	+	+
3	+	+	+	?	-	+	+	?	-
4	-	?	?	-	+	+	-	-	-
5	+	+	+	+	+	?	+	+	?
6	+	+	+	+	?	+	?	+	?
7	?	?	+	?	+	+	+	?	+
8	+	+	+	+	+	+	+	+	+

PROBAST = 预测模型偏倚风险评价工具；ROB = 偏倚风险。

* +意味着低偏倚风险或适用性关切低；—意味着高偏倚风险或适用性关切高；？意味着偏倚风险或适用性关切不清楚。

图示. PROBAST 评价结果的推荐图示





应该在系统评价规划阶段或系统评价研究方案中就要说明，[译者注：如何]对偏倚风险和适用性关切做更进一步的整合应用。PROBAST 使用者，可以将 PROBAST 评价结果纳入到数据分析中去：通过设计敏感性分析，将分析限制在总体评价或具体领域上得到“低偏倚风险”或“适用性关切低”评价的研究；或者基于 PROBAST 关切评级，使用亚组分析来调查研究间的异质性（17，40，44）。

本文总结

据我们所知，PROBAST 是第一个严格制定出的工具，专门设计用于评价那些建立、验证或更新（包括拓展）以个体化预测为目的的预测模型的原始研究的偏倚风险和适用性关切。PROBAST 涵盖了诊断和预后模型，不管它们的医疗领域、临床结局类型、预测因子或用到

的统计分析方法。

该说明和详述文件提供一个详尽的关于如何使用 PROBAST 的指导性说明（39），其中包括怎样解释每个信号问题，怎样对每个领域以及总体评价定级，怎样在一个系统评价中呈现和整合应用 PROBAST 评价结果；这些还同时带有有关诊断和预后预测模型研究的简要说明。这个详解的 PROBAST 说明和详述文件，使我们有一个针对性强且透明的方法，去评价那些建立、验证或更新以个体化诊断或预后预测为目的的预测模型之研究。五个填好的 PROBAST 评价案例，可以在我们的网站上找到（www.probast.org），它们涵盖了建立研究、验证研究或者两者兼有的情况，并且评价了诊断以及预后模型。我们也鼓励并提供 PROBAST 的所有翻译。

PROBAST 的使用，要求预测模型研究者和临床医师具备一定专长和知识。相比于随机对照干预研究和诊断准确性研究，预测模型研究的方法学指导仍处于相对早期阶段。我们认识到，评价偏倚风险和适用性所需的信息往往未被报告，因此我们希望杂志和作者可以遵循 TRIPOD 报告指南来改善这类问题。

和医学研究的其他偏倚风险和报告指南一样，PROBAST 和其指南也会随着预测模型研究方法学的发展而需要更新。我们建议您要时常到网站（www.probast.org），去下载最新版的 PROBAST 和指南文件。

作者单位

Julius 健康科学和初级卫生保健中心和 Cochrane 荷兰，乌得勒支大学医学中心，乌得勒支大学，乌得勒支，荷兰（K.G.M., J.B.R.）；

Kleijnen 系统评价，约克，英国（R.F.W., M.W.）；

预后研究中心，初级卫生保健和健康科学研究所，吉尔大学，吉尔，英国（R.D.R.）；

布里斯托大学布里斯托医学院和英国国立卫生研究院（NIHR）应用健康研究和护理西部领

袖合作机制（CLAHRC West），英国国家健康服务体系基金信托布里斯托大学医学集团，布里斯托，英国（P.F.W.）；

医学统计学中心，牛津大学，牛津，英国（G.S.C.）；

Kleijnen 系统评价，约克，英国；公共卫生和初级卫生保健学院，马斯特里赫特大学，马斯特里赫特，荷兰（J.K.）；

应用健康研究所，英国国立卫生研究院伯明翰生物医学研究中心，医学和牙医科学学院，伯明翰大学，伯明翰，英国（S.M.）。

通讯作者

Karel G.M. Moons, 博士, Julius 健康科学和初级卫生保健中心，乌得勒支大学医学中心，乌得勒支大学，邮政信箱 85500,3508 GA 乌得勒支，荷兰；

电子邮箱 K.G.M.Moons@umcutrecht.nl

免责声明、致谢、资金支持、利益冲突声明、作者现住址、作者贡献和参考文献见原文

时春虎翻译，马捷、王俊峰审校。

本文是根据发表于 *Ann Intern Med*. 2019;170(1):W1-W33.的全文翻译而成。本文的翻译符合 PROBAST 的规定，感谢 Karel G M Moons 为此提供的帮助。