临床预测模型的内外部验证指标及举例说明

刘岳鹏* 杨艳君[1]

*[1]徐州市医学科学研究所

临床预测模型的构建及其验证已经成为临床研究的一种重要的形式。在临床流行病学研究，也就是临床研究中的观察性研究，为临床提供了某种疾病结局的大量的危险因素之后，临床预测模型研究通过构建数学模型而综合这些危险因素，进而实现对某种疾病结局的评估和预测，更可通过制作列线图或者网页计算器，落实到实际的临床应用，从而搭建起一个临床研究到临床应用的桥梁。

成熟的临床预测模型需要经过构建、内部验证和外部验证三个必要的阶段[1,2]。内部验证是采用与构建预测模型的数据来源相似的数据对模型进行评价，而外部验证是以空间上（多中心）不同于模型构建的数据对模型进行评价。有些研究采用同一个研究中心但是时间上不同的数据进行外部验证，严格说只能算“半个”外部验证。在评价指标上，目前共识是从预测模型的区分度和校准度两方面进行评价，评价指标常用的是 C 统计量（c-statistics）和校准度曲线，显得单薄并且也不够系统。本文将系统介绍适合于内部验证和外部验证的评价指标，并举例对这些指标的含义。

1、综合评价指标

Brier 评分，该评分是对模型区分度和校准度的综合评价，缺点是不能区分其中区分度和校准度各自对评分的贡献[3]。Brier 评分取值范围为 0~1。在预测概率为 50% 时，无论事件发生或者不发生，Brier 评分为 0.25，所以 0.25 成为 Brier 评分的一个分界点，小于 0.25 代表模型可以正确预测事件的发生，大于 0.25 说明模型错误预测了事件的发生。在正确预测的前提下，即模型的评分位于 0~0.25 之间时，评分越接近 0 说明模型效能越好。

2、区分度评价指标

C 统计量和分组 OR/HR 值。区分度多用在分类模型中，可解释为随机抽取一个发生事件的个体和一个未发生事件的个体，前者得分高于后者得分的概率，通俗的理解是在模型所建立的这个体系中是否能找出一个值来区分发生该疾病结局和不发生该疾病结局的人群。C 统计量是最常用的指标，在逻辑回归中等同于 ROC 曲线下面积（AUC）。此外，专家还推荐将概率或其它参数不等分组为 3~4 组并计算组间 OR 值（逻辑回归）或者 HR（Cox 回归）的形式来直观体现模型的区分度[4]。如果模型区分度佳，则组间 OR 或 HR 值就大，反之则小。这个评价指标一方面丰富了模型区分度的评价形式和角度，另一方面用于分组的参数如果是概率或者诺莫图的总得分，那么同时提供参数的界值使分组评价具有一定的临床应用意义。

DOI: 10.14218/MRP.2021.019
通讯作者：刘岳鹏 Email: liuyp2080@163.com

3、校准度评价指标

校准度曲线及其斜率和 Spiegelhalter’s 检验的 P 值。校准度通俗的理解是，模型所构建的体系所预测的概率和事件实际发生概率的符合程度。校准曲线用斜率是校准度曲线的参数表示，可以和校准度曲线图相互参照，共同说明预测模型校准度的优劣。另外，Spiegelhalter’s 检验是基于 Brier 评分的一种检验，单纯反映模型的校准度，其 P 值不显著（P>0.05）则认为校准度较好 [5] 。

4、举例

4.1. 背景

杨瑞等 [6]收集较大规模的国内患者数据，以性别、年龄和肿瘤尺寸构建了甲状腺微小癌中央淋巴结转移的临床预测模型，该模型的内部验证表明该模型具有一定的预测效能（内部验证区分度的 C 统计量为 0.706）。以下以 SEER 数据库 2004~2015 年的甲状腺微小癌的临床流行病学数据作为外部验证的数据集对杨瑞等构建的甲状腺微小癌局部淋巴结转移的预测模型应用以上评分系统进行外部验证。

4.2. Brier评分

Brier 评分为 0.201，小于 0.25，表明模型可以正确地预测微小甲状腺癌患者局部淋巴结转移的发生，如果距离 0 还有较大的距离，表明模型整体表现欠佳，需要在区分度和校准度方面进行进一步提高。Brier 评分的计算过程是首先用构建模型的参数计算出验证数据集中每个个体的事件发生的概率，然后使用 R 软件的 rms 文库中的 val.prob 函数计算。

4.3. 模型区分度评价

根据 C 统计量的判定标准，其值低于 0.5 为模型不佳，等于 0.5 为模型区分能力与随机相似，位于 0.6~0.7 之间为模型有一定的预测价值，大于 0.7 为模型有较好的临床应用价值。在逻辑回归模型中区分度指标为 C 统计量与 ROC 曲线下面积（AUC）一致（图 1），此处为 0.647（95% CI: 0.638~0.655），代表构建的此模型有一定的区分局部淋巴结转移和非转移的人群的能力。C 统计量的计算过程是首先用构建模型的参数计算出验证数据集中每个个体的事件发生的概率，然后使用 R 软件的 pROC 文库绘图和计算。

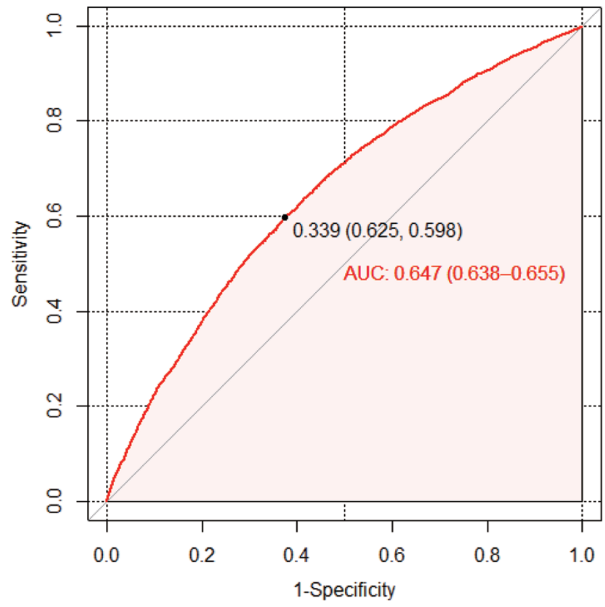

图1 在SEER数据集中依据模型预测概率制作的ROC曲线。最佳阈值0.339，此时特异度0.625，灵敏度0.598，AUC0.647（95%CI：0.638~0.655）。

4.4. 将预测概率分低中高风险三组考察区分度

根据预测概率来划分组别并展示各组中局部淋巴结实际转移率、预测概率取值范围、诺莫图总得分取值范围、各组的 β 和 OR 值（表 1）。高风险组中局部淋巴结转移概率是低风险组的 3.4 倍，中风险组中局部淋巴结转移概率是风险组的 2.01 倍，证明模型有一定的区分度。表中同时给出了预测概率和诺莫图总得分的取值范围，临床上可以据此来估计患者发生局部淋巴结转移的风险。以上统计量使用

表1 SEER数据集各组间参数比较（以预测概率分组）

分组	预测概率取值范围	诺莫图总得分取值范围	β(95%CI)	OR(95%CI)
低风险	0.021~0.266	0~77.0	Reference	1
中风险	0.269~0.509	77.3~108	0.698(0.623~0.773)	2.010(1.865~2.166)
高风险	0.513~0.938	108~186	1.224(1.142~1.307)	3.402(3.133~3.695)

python 软件 scikit-learn 文库完成。

4. 5. 模型校准度评价

结果显示校准度曲线与对角线（Ideal 线）贴合不紧密，代表构建的这个模型系统中所计算的事件发生的概率，与事件发生概率不一致，存在低估或高估的现象（图 2），其斜率为 0.499，与校准度曲线相互对照，共同表明该预测模型校准度不佳。另外，Spiegelhalter's z 检验的评分为 14.486（P=0.0001），P 值显著同样表明预测模型校准度不佳。以上图的制作和参数的计算是首先用构建模型的参数计算出验证数据集中每个个体的事件发生的概率，然后使用 R 软件的 rms 文库中的 val.prob 函数计算。

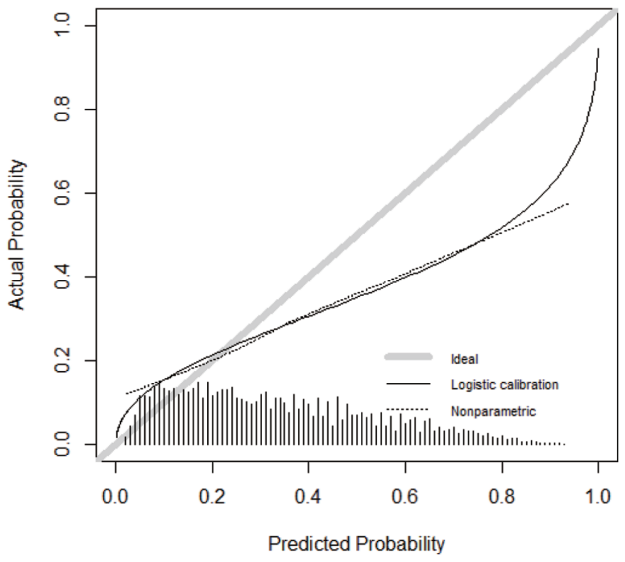

图2 在SEER数据集中依据模型预测概率制作的校准度曲线。校准度曲线贴合理想曲线表明模型预测概率和实际概率一致，可据此判断模型具有良好的校准度。此处，校准度曲线偏离理想曲线，比如，预测概率为0.6时，实际概率为0.35，高估了实际概率。

综上所述，以上评价体系可以从多角度对构建的临床预测模型进行系统的评价，有助于全面评价构建的预测模型。

参考文献

[1] 谷鸿秋，王俊峰，章仲恒，周支瑞. 临床预测模型:模型的建立. 中国循证心血管医学杂志 2019;11(01):14–16+23.
[2] 王俊峰，章仲恒，周支瑞，谷鸿秋. 临床预测模型:模型的验证. 中国循证心血管医学杂志 2019;11(02):141–144.
[3] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128–138.
[4] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 2013;13:33.
[5] Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020;27(4):621–633.
[6] 杨瑞，张守鹏，黄韬，明洁，杨鹏，朱俊玲，瞿芳. cN0期甲状腺微小乳头状癌淋巴结转移模型的构建和验证以及手术方式探讨. 临床耳鼻咽喉头颈外科杂志 2021;35(02)137–140.