研究方案 v1.0 · 专家阅读版

中国临床与住培场景下静脉血栓栓塞症中文测评集的构建与初步验证

方案日期 2026 年 11 月 22 日 | 阅读版 2026 年 11 月 23 日

研究目的

在中国大陆住培、跨科会诊、护理交接和患者教育场景中，构建覆盖常见及高风险决策的中文测评集，逐题定位参考答案的证据，并检验独立评审和评分者能否按一致规则工作。静脉血栓栓塞症（venous thromboembolism，VTE）包括深静脉血栓形成（deep vein thrombosis，DVT）和肺血栓栓塞症（pulmonary thromboembolism，PTE，亦称 pulmonary embolism，PE）。一期研究验证题集和评分规则，系统间问答比较另行开展。

设计与主要终点

环节	方案要点
构建	多来源取题、按蓝图补题、逐题记录证据和评分点。计划准备约 450 道候选母题，审定目标为 360 道母题。
审定与验证	经两位专家独立审题和分歧裁决形成预冻结题库，再进行独立抽查和评分一致性验证。主要终点为预冻结母题独立抽查的首次无需实质性修订通过率。
拟定用途	审定后计划分为 60 道公开开发题和 300 道封存测试题。开发题用于说明接口、校准评分和调试。封存测试题用于后续盲测。

> 本次为首轮内容审阅，共 32 道试写题，八个领域各 4 道。审阅范围包含题面、候选答案、关键证据和评分规则。

测评蓝图与母题计数

题目以中国临床工作任务为单位，每题设一个主要领域、一个认知层级和一个题型。认知层级是设计复杂度，取得真实作答后可另行估计实测难度。

主要领域	L1	L2	L3	总数
院内与围手术期风险评估及预防	15	33	12	60
疑似 DVT 的诊断、排除及鉴别	12	22	10	44
疑似 PE 的诊断、排除及鉴别	12	24	10	46
PE 危险分层、急性处置及升级转诊	10	23	15	48
DVT 急性治疗、血管外科干预及适用条件	10	21	11	42
抗凝选择、疗程、调整及出血处理	14	26	14	54
特殊人群与并存风险影响决策	8	16	12	36
出院交接、随访、远期并发症及患教	9	15	6	30
合计	90	180	90	360

认知层级

L1 提取一条明确推荐。L2 将规则用于给定情景。L3 处理多项条件、冲突证据或信息不足。若肿瘤、妊娠或产褥期、肾功能不全或透析、高龄等因素改变主要决策，归入特殊人群领域，其余情况作为交叉标签。

题型和作答对象

计划设置 120 道单项选择、60 道结构化任务和 180 道开放题，并按领域和层级交叉报告。结构化任务区分因子抽取、计算和最终风险层级。开放题涵盖病例决策、证据解释、追问、跨科交接和患者沟通。逐题注明作答角色和接收者。

计数单位

每题族选一个独立决策任务作为母题。同一患者或基础情景的追问、译文、同义改写或轻微参数变体归入同一题族。独立病例、附加追问和评分点另计。

题目来源、许可与临床证据

来源	取得与使用
公开原题及改编	检索中文医考和病例题（CNMLEQA、CMExam、MedQA、CMB）、英文医考及摘要问答，以及 PE 指南问答、ChatVTE、LLMEval-Med、DOACT.Vasc 等专科资源。按题保存原题定位、原答案、证据版本及许可。获准发布并审定合格者可进入开发集，其余保留为外部对照或仅记出处。
指南原创	从中国指南的推荐、适用条件和决策点命题，覆盖选择、执行、复评和补充信息等任务。计划约 180 道或更多，可纳入开发或封存题。
合成教学病例	由临床教师提出典型问题，或依据证据独立构造。体现中国常用检查、转诊层级、药物通用名和计量单位。封存题中计划约 120 道来自该路径。

筛选与许可

先校验来源、原件版本与许可，再记录题干、选项和解析中的命中位置。以答对核心任务是否需要 VTE 知识判断相关性，按原题、近义和共享病例归并，留存原答案冲突与排除理由。内部审阅、改编和再发布权限分别登记。外文改编须核对否定、程度、单位、数值及其条件。

MedQA-US 和 PubMedQA 用于补充召回。德语 cpgQA-DE 提供字段设计参考，孕产期 VTE 与 CHIPS-VTE 研究用于完善任务蓝图。未取得原题或不能确认使用权限的材料登记访问状态。

证据截止日为 2026 年 11 月 22 日

优先采用中国 DVT 第四版（2026）和中国 PTE 指南（2025），按主题补充急性 PE、检验诊断指南及预防质量目标。药物用法如决定答案，核对相应适应证与国内说明书。逐项登记发布机构、版本、推荐号或页码及表图定位、适用与排除条件。合理替代路径可并列纳入评分。

DVT 指南第四版 中国 PTE 指南 急性 PE 指南 VTE 实验室诊断指南

候选题记录

题面、来源与题族、参考答案、逐条评分点和证据互相连接。原题答案与中国版候选答案分别保存。证据暂无法定位时保留待核状态，补证、修订和淘汰均记录理由。

专家审定与评分规则

独立初审及分歧处理

每道拟入选题由两名具备相关专长的临床评审独立核对，其中至少一位熟悉住培或对应岗位。评审回避本人参与命制的题。记录任务相关性、题意、标准答案及替代答案、证据适用性、临床安全、认知层级和可重复判分性。

项目组保留两份首次独立意见，由第三位专家裁决重要分歧。影响正确答案、适用条件、关键安全判据或得分的修订保留前后版本，并复核修订后的内容。待审或未返回的判断保持空值。

临床内容评分

题型或维度	判分要求
单项选择	按最终选项记 0 或 1 分，核查是否有唯一可判定选项。
结构化任务	分开核对输入因子、计算和最终分类，保留分项得分和整题全对标记。涉及数值时注明单位与容差。
开放题	每题设 2 至 5 个可观察评分点。逐点设零分、部分分和满分锚点及预设权重，按适用项归一至 0 至 100 分。患教题另核对受众能否理解。
信息不足	在题面只支持追问或依据不足时，为指出缺失信息及给出合适下一步评分。
安全	触发题内预先列明的重大伤害性建议时，保留临床原分，另记危险错误，并判定该题安全合格失败。

引用质量独立计分

检查所引文献能否定位、能否支持相应判断，及需要证据的关键主张被覆盖的比例。没有可追溯引文时登记缺引用。时效和适用性相当、能支持同一判断的证据可被接受。

独立验证、报告与发布管理

预冻结题库的独立抽查

全量双审和裁决完成后锁定预冻结版本。按固定种子 20261122 简单随机无放回抽取 72 道母题，不足 72 道则全查。两名未参与相应题目命题与初审的临床专家独立复核，分歧另行裁决。

首次通过率的分子为无需实质性修订的母题数，分母为实际抽中数。仅修改措辞、拼写或排版可判通过。核心答案、关键适用条件、主要证据、安全判据或可判分性须改时，计入首次需改。报告通过、需改、结论缺失的原数与完成率，并给出 Wilson 95% 置信区间。返修通过另列。按预期通过率约 90% 估算，抽查 72 题的普通二项近似区间半宽约 7 个百分点。

开放题评分一致性

开发集预留至少 24 道开放题。每题准备完整、部分完成和含预设重要错误的匿名作答各 1 份，24 道题对应 72 份作答。两名未参与编写这些作答的评分者独立评分。报告双向随机、绝对一致、单次测量组内相关系数（intraclass correlation coefficient，ICC(A,1)）、平均绝对分差及安全标志的一致率和 Cohen κ，按母题重采样计算区间。校准评价预设作答的判分可重复性，系统实际作答中的危险错误发生率留待后续测评。

报告内容及伦理安排

报告题源纳排、领域与层级覆盖、证据支持、原答案保留或修订、首次缺陷和两位评审的仲裁前判断。分类判断报告原始一致率和 Cohen κ，有序层级报告加权 κ。住培医师真实作答属于可选验证，须先落实相应伦理和参与者授权。项目伦理安排待按院内要求确认。

按题族分集与保护材料

公开原题和已展示的题族进入开发候选或外部对照。其他合格题族按领域、层级和题型分层分配，同一题族只进入一个集合。封存前检查题干、答案及病例与开发材料和已知索引的重合，逐一复核可疑匹配。公开开发题、受控测试题面与封存答案分别管理。测试输入仅含题号、题面、选项和作答要求。保留题族快照、分集记录、证据与评分版本，后续变更另起版本。