**北京大学人民医院临床研究方案撰写要素**

**（预测模型研究）**

**撰写说明：**

**1.提交完整版研究方案删除此页，研究方案需双面打印，封面课题负责人签字；**

**2.撰写中请注意删除提示语及斜体标注的案例；**

**3.请研究者根据自身研究特点删除不适用项并重新排序，即标有（如果适用）的项目，如果不适用，均可删除，其余不建议删除；**

**4.书写格式：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **标题** | **字体** | **字号** |  | **行间距** |
| **一级标题** | **宋体** | **四号** | **加粗** | **2.0** |
| **二级标题** | **宋体** | **小四** | **加粗** | **1.5** |
| **三级标题等** | **宋体** | **小四** | **不加粗** | **1.5** |
| **正文（中文）** | **宋体** | **小四** | **不加粗** | **1.5** |

**一般格式要求：**

页眉、页脚完全，并包括研究题目、方案版本号（如1.0//2.0等）、版本日期（如20200101）和页码等信息。

**研究方案封面页及研究人员列表要求：**

研究方案封面课题负责人签字，研究人员列表包括课题负责人，研究人员列表中如有课题负责人科室以外的人员，课题审批表中需要相应合作科室主任签字。

研究题目

|  |
| --- |
| 本单位课题负责人： |
| 科室： |
| 版本号：1.0 |
| 版本日期：YYYYMMDD（如20200101） |
| 联系人： |
| 联系电话： |
| 本单位课题负责人签字： |

研究人员列表

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 单位/科室 | 职务/职称 | 分工 |
| 例：XXX | 北京大学人民医院/XX科 | 科主任/主任医师 | 方案设计 |
|  |  |  |  |
|  |  |  |  |

**正文要求：**

**一、研究摘要**（不超过500字）

**原则**：概述研究目标、研究设计、研究设定、研究对象、样本量、预测因子、结局指标，统计分析方法、结果和结论等核心信息。

**二、研究背景以及立项依据**（1000字及以上）

**原则：**充分说明研究的医学背景（包括是诊断还是预后）以及建立或验证多因素预测模型的理由，包括对现有模型的引用与参考。

**三、研究目的**

**原则：**描述研究预测的目的，预测的结局及模型适用的人群，还应说明本研究是模型开发还是模型验证，或两者都有。

***书写举例：****“开发并验证新版预测模型QRISK3，……，以改进某些患者组的心血管疾病风险预测。”*

**四、研究方法和设计**（1000字及以上）

**1、研究设计类型**

**原则：**研究设计类型选择依据：预测模型包括诊断模型和预后模型，对研究设计要求不同。诊断模型的预测因子与结局变量是在同一时点或很短时间窗内测量，可利用横断面研究数据；预后模型的预测因子与结局存在因果关系，可利用队列研究、随机试验、巢式病例对照等。

**2、研究对象及选择标准**

**原则：**分别描述建立或验证模型（如适用）的数据集中研究对象来源和选择时间范围。针对研究场所说明研究人群来源于初级医疗机构、二级医疗机构或普通人群），包括研究中心的数量和位置；时间范围应涵盖数据收集的开始时间、结束时间、随访结束时间等关键时间点。诊断模型不涉及随访时间，预后模型需给出随访时长的计算方法和预测的时间范围。

1. **研究对象的选择标准**

 **原则：**包括研究对象的纳入标准、排除标准等。其中纳入标准是否合理会影响研究对象的纳入过程及模型的预测准确度，对模型的验证及应用也会造成困难。用来制定纳入标准和剔除标准的因素包括：①疾病严重程度；②有无并发症和伴发症；③病人的年龄、性别及居住区域等；④病程长短和既往治疗史等。

***书写举例：***

***入选标准：***

*a. 年龄在XX之间，性别不限；*

*b.*

*c.*

*d.*

*e. 签署知情同意书。*

***排除标准****：*

*a.；*

*b.；*

*c.；*

**4、结局指标**

**原则：**明确定义每个结局指标，并从2个维度上进行说明：一是结局指标的说明，如定义，测量工具，测量方法；二是测量时间点或时间段、频率等。

**5、预测因素**

**原则：**定义建立或验证多因素预测模型所使用的所有预测因素，包括如何以及何时测量。其中，对于模型开发研究，所有预测因素是指所有可能进入最终预测模型中的变量，即所有备选变量，而不仅仅是最终模型中包含的变量；而对于模型验证研究，所有预测因子则是指被评估模型中所包含的变量。

预测指标一般选择基线指标，例如将治疗作为起点，那么收集治疗起点及治疗开始前，可能与待研究疾病相关的因素，这些因素除了罗列，对于除社会人口学信息外的信息，建议明确列出该指标，指标定义、测量方法、测量时间点/时间段等信息。

**6、预测模型的开发**（含统计分析方法）

**原则：**明确报告预测模型开发时所用的所有统计方法，应至少包括：①统计模型类型，如logistic模型、Cox模型等；②备选预测因素的选定：指在建模前所有潜在预测变量是通过什么方式选定的；③进入模型预测因素的筛选方法：指在建模中进入最终回归模型中的预测因素的筛选过程，如单因素分析、逐步回归等。

***书写举例（不要全部照搬，根据自身研究修改）：***

*分别采用二分类logistic回归和机器学习方式构建**XXX预测模型。*

*（1）基于二分类logistic回归构建预测模型*

*采用均数±标准差或中位数、四分位数间距对计量资料进行描述，采用率、构成比对计数资料进行描述；采用两独立样本t检验（或符号秩检验）对两组间计量资料进行比较；采卡方检验对两组间计数资料进行比较；采用LASSO回归对预测指标进行筛选，采用多因素二分类logistic回归模型构建XXX预测模型，并对建立预测模型的拟合优度进行报告。*

*（2）基于机器学习方法构建预测模型*

*将回顾性队列研究数据以8:2的比例随机拆分为模型训练集和测试集，针对预测指标，基于预测模型变量的高预测能力和广泛可获取性原则，结合受试者特征曲线（ROC）和数据缺损率筛选出用于构建XXX预测模型的候选变量，候选预测变量需满足以下条件：①AUC>0.6，②数据缺损率<50%。将初步筛选后的预测变量作为输入项，带入XGBoost中进行训练，将患者治疗转归作为输出。将原始训练集数据平均分成5组，将每个子集数据分别当作一次验证集，其余的4组子集数据作为训练集，会由此得到5个模型，用这5个模型最终的验证集的分类准确率的平均数作为五折交叉验证模型下的性能指标。*

**7、预测模型的验证及效果评价**（含统计分析方法）

**原则：**描述内部验证方法，如重抽样、交叉验证等；外部验证可采用空间验证或时间外部验证法；模型预测的性能评价主要包括区分度和校准度，评估区分度通常采用ROC曲线下面积（AUC），也叫C统计量（C-statistics），评估校准度通常Hosmer-Lemeshow good of fit test（拟合优度检验）。所有预测模型均应报告这两者，也可增加、灵敏度、特异度、决策曲线等多维度反映模型性能，若要评价引入新预测因子后的预测效能增加值，也可使用综合判别改进指数（IDI）及净重分类改进指数（NRI）等指标。

***书写举例(不要全部照搬，根据自身研究修改）：***

*本研究构建的XXX预测模型模型分别进行内部验证和外部验证，验证方法说明如下：*

1. *内部验证*

*采用bootstrap法对构建预测模型的开发队列数据进行有放回抽样，构造一个相同样本量大小的bootstrap重抽样样本，作为训练集，整个模型开发队列数据作为验证集，评价模型性能，重复此过程200次。*

1. *外部验证*

*外部验证采用空间验证法，来评价预测模型的可移植性和可泛化性。以xx、xx医院的XX患者数据为外部验证数据，对构建的预测模型进行外部验证。*

1. *预测效果评价*

*将基于二分类logistic回归构建的预测模型和基于XGBoost构建的预测模型分别应用于验证队列数据，以C-index指标比较两者的预测效果。*

**8、样本量的计算依据**（建议咨询科研处统计方法学老师）

**原则：**务必明确样本量计算方法和参考依据，不能仅说明本研究拟纳入的研究对象例数。针对现有数据库样本量较大的情况下，可直接基于现有数据建立；针对样本量较小的回顾性研究或者前瞻性数据的研究，需要提供规范的样本量计算方法。

1. 传统方法：10倍EPV这一经验法则（rules-of-thumb）[即10 EPV原则]，即以发生率为10%的二分类结局事件为例，如（备选）预测因子15个，则至少需要150例阳性事件，总样本量至少应达到1500例。模型验证时，如果单中心验证，一般要求至少100例阳性事件；如果多中心验证，保证每个中心至少50例阳性事件。
2. 新观点：基于具体的场景和研究问题予以确定，样本量计算应综合考虑预测的结局类型（二分类指标、生存结局或连续性变量）、涉及参数包括总体结局风险/平均结局值、备选预测参数个数及预期模型表现等，目前R语言已有pmsampsize包用于计算连续、二元或生存（时间到事件）结果模型开发所需的最小样本量。

**10、研究进度**

**原则：**请考虑研究方案获得伦理审查批准后方可开展的因素，分别说明不同时间段的项目进度安排，可用甘特图等展示。

**五、前期研究基础**（如果适用）

**原则：**主要研究者在本研究内容方面有何前期研究基础和研究结果。

**六、研究场地及数据管理**

**原则：**（1）说明标本收集、贮存场所，实验室研究场地；（2）数据管理：说明数据管理的工作流程应包含数据采集/管理系统建立（如 病例报告表及数据库的设计）、数据接收与录入、数据核查、数据审核、数据库锁定、 数据导出及传输、数据及数据管理文件的归档等流程。

**七、受试者保护**

**原则：**说明知情同意、受试者利益与风险、隐私保护措施、保密、利益冲突等伦理考虑；特殊人群保护（妇女、儿童、老人等，其他特殊人群如在押人员等的风险评价及保护措施）

**八、研究管理**

**原则：**围绕整个研究流程说明质量控制措施，包括研究准备阶段、研究实施中、数据收集和管理等阶段；研究提前中止：定义提前终止研究的原则及处理办法。

**九、多中心研究参加单位**（如果适用）

组长单位及各分中心基本情况介绍，课题任务分工说明。

**十、参考文献**

格式规范，统一。

**十一、附件**