从文献中学习：整合大语言模型与贝叶斯分层建模开展肿瘤临床试验设计

Guannan Gong[1], 博士；Satrajit Roychoudhury[2], 博士；Allison Meisner[3], 博士，Lajos Pusztai[1], 医学博士、哲学博士；Sarah B Goldberg[1], 医学博士，Wei Wei[1], 博士*

[1]耶鲁癌症中心，耶鲁医学院，美国康涅狄格州纽黑文[2]辉瑞公司，美国纽约州纽约市[3]弗雷德·哈钦森癌症中心公共卫生科学部

*通讯作者：Wei Wei，博士

耶鲁癌症中心，耶鲁医学院

美国康涅狄格州纽黑文，邮编06520

电子邮箱：wei.wei@yale.edu

关键词：肿瘤临床试验设计、贝叶斯分层建模、证据综合、大语言模型

摘要

目的

设计现代肿瘤学试验需要综合既往研究的证据，以进行假设生成和样本量计算。基于不完整信息或不精确摘要构建的试验设计可能导致假设设定错误和研究效能不足，从而错失机会，并产生假阳性或假阴性的试验结果。为应对这一挑战，我们开发了LEAD-ONC（肿瘤学分析与设计的文献证据），这是一个人工智能辅助平台，可纳入已发表试验的结果，将非结构化数据转换为定量信息，并开展贝叶斯证据综合，以支持未来试验做出充分知情的设计决策。由于LEAD-ONC仍在积极开发中，本文所呈现的分析旨在初步展示该框架在试验设计方面的潜力，而非确定性的临床结论。LEAD-ONC的当前版本可在开发者网站获取（http://34.55.33.152/step1#）。

方法

针对符合预先设定入选标准、经专家筛选的文献，本框架使用大语言模型（LLM）从“表1”中提取基线特征摘要，并根据 Kaplan–Meier（KM）曲线重建个体患者数据（IPD）。随后，根据基线特征的相似性对试验进行聚类，以界定更适合开展证据综合的亚群。在与目标试验人群最为相似的聚类中，我们对重建的个体患者数据拟合贝叶斯层级模型，并生成目标试验生存曲线的预测分布。

结果

我们利用五项近期完成的非小细胞肺癌（NSCLC）III期试验结果，展示 LEAD-ONC 的实用性。这些试验评估了 PD-L1/PD-1抑制剂单用或联合 CTLA-4抑制剂（即双联或单药免疫检查点抑制剂（ICIs）对比单独化疗）。这些试验因关键入选标准存在重叠而被选中。根据从所选文献中提取的基线特征，LEAD-ONC 确定了三种不同的患者人群，主要由组织学类型界定，分别为全非鳞状（KEYNOTE-189）、全鳞状（KEYNOTE-407）和混合组织学类型（POSEIDON、CheckMate-227、CheckMate-9LA）。对于一项在混合组织学类型人群中比较单药与双药免疫检查点抑制剂的前瞻性随机试验，LEAD-ONC 预测双联治疗带来具有临床意义获益的可能性较低：总生存期（OS）中位数差异为2.8个月（95%置信区间，−2.0至7.6），双药免疫检查点抑制剂使中位总生存期至少延长3个月的概率约为0.45。

结论

将大语言模型与贝叶斯分层建模相结合，能够有效地从文献中学习，并为未来试验的规划提供循证方法。

引言

设计临床试验需要系统地综合大量且迅速增长的既往证据。在试验设计阶段，研究人员通常会审阅大量文献，并借鉴历史试验的结果，为关键决策提供依据，包括预期治疗效应量、为积累足够事件所需的随访时间，以及界定目标人群的入选标准。在实践中，受限于个体患者数据 (IPD) 的获取渠道有限，[1] 从已发表文献中提取定量信息既耗费人力又容易出错，以及跨研究选择和整合证据时存在主观性，从既往试验中有效学习往往受到阻碍。这些局限共同导致对关键设计参数的假设不够精确或过于乐观；当预期结局目标未能实现时，这会增加研究效能不足以及肿瘤药物研发遭遇高昂失败的风险。

为应对这些挑战，我们开发了 LEAD-ONC，这是一种人工智能辅助方法框架，可将研究者筛选整理的肿瘤试验文献转化为定量证据，为未来一项试验（下文称为目标试验）的设计提供参考。首先，研究者根据预先设定的入选标准（人群、治疗策略、试验实施等）确定相关研究，并将选定的文献上传至 LEAD-ONC。基于这些经过筛选整理的输入，LEAD-ONC（图 1）利用大语言模型和光学字符识别，从非结构化来源中提取定量信息，包括 Kaplan–Meier（KM）曲线和基线协变量汇总（“表1”），并将其转化为可分析的数据集，其中包含重建的个体水平生存数据和协调后的基线协变量。

即使试验符合相似的入选标准，异质性仍可能存在。因此，LEAD-ONC 聚焦于所选历史试验中普遍报告的基线特征，以量化试验间相似性，并将试验聚类为更具可比性的亚群。在与目标试验人群最为相似的聚类中，利用 KM 曲线重建的个体患者数据（IPD）通过贝叶斯层级模型（BHM）进行拟合，以生成目标试验生存曲线的预测分布。该预测分布综合了经过筛选整理的历史试验中的生存结局和基线特征，代表了与目标试验生存终点相关的全部证据。

与依赖手动裁剪图形并反复通过点击操作对 KM 曲线进行数字化的现有工具相比，[2, 3] LEAD-ONC 支持批量导入、多图解析，并能自动将曲线与相应的风险表关联起来。与证据综合领域中大多忽略协变量信息的现有方法相比，[4–7] LEAD-ONC 基于已发表的基线协变量汇总数据进行聚类，从而降低试验间的异质性，并使贝叶斯分层模型的估计结果能够结合特定目标人群的背景进行解读。相比之下，基于单一贝叶斯分层模型、并将其应用于

所有历史试验的证据综合方法，可能会产生不针对所关注临床试验的先验知识。LEAD-ONC通过以可扩展的方式自动从先前发表的文献中提取证据，改进了对先验证据的综合；这对于成功的试验设计至关重要。

本文介绍了 LEAD-ONC 的架构，包括借助人工智能从试验报告中提取定量信息，以及采用有针对性的贝叶斯层级模型框架，从选定的历史试验中学习生存经验，为未来试验的设计提供依据。我们展示了这一方法如何为非小细胞肺癌前瞻性随机试验的关键设计决策提供指导。

案例研究的动机

针对 CTLA-4 和程序性死亡配体1或 PD-1 的双重免疫检查点阻断，作为转移性非小细胞肺癌患者的一线治疗，已显示出优于铂类化疗的疗效。然而，除程序性死亡配体1或 PD-1 单药治疗外，双重免疫检查点抑制能否进一步改善总生存期，仍是一个尚无定论且具有重要临床意义的问题。[8] 为回答这一问题，Federico等人开展了一项系统综述，筛选标准聚焦于纳入未经治疗的晚期非小细胞肺癌患者的3期随机对照试验，并比较 PD-1或程序性死亡配体1抑制剂（单独使用或联合铂类化疗或 CTLA-4抑制剂）与铂类化疗的疗效。[8]他们确定了几项入选标准相互重叠的随机对照试验 ，包括 POSEIDON、CheckMate-227、CheckMate-9LA、KEYNOTE-189 和 KEYNOTE-407。[9–15] 我们在**表1**中总结了这些试验的试验层面患者特征。个体患者层面的基线特征数据无法获取。

假设我们希望在一项随机对照试验（RCT）中，评估双药与单药免疫检查点抑制（ICI）对总生存期（OS）的影响，该试验招募的是转移性非小细胞肺癌（NSCLC）初治患者。在设计阶段，我们希望利用相关历史试验中的所有可用信息（“表1”中的患者基线特征、患者总生存期的KM曲线），回答以下关键设计问题：（1）这些入选历史试验的患者特征是否存在有意义的差异，这是否意味着其中一些试验比其他试验更适用于试验设计？（2）对于未来试验中随机分配至单药与双联抑制剂组的患者，总生存期有哪些合理的假设？（3）预计双药与单药免疫检查点抑制剂治疗之间的生存获益有多大？（4）基于这些假设，计划开展的试验成功概率是多少？本文将展示 LEAD-ONC 如何回答这些问题。

方法

借助人工智能从已发表文献中提取定量试验结果

IPDfromKM [2] 和 WebPlotDigitizer [3] 等工具使得以合理精度重建个体患者数据（IPD）日益可行。在实践中，由于工作流程和可扩展性方面的障碍，这些工具在试验设计中的应用仍未得到充分探索。首先，大多数流程仍要求在开始数字化之前，先从源 PDF 中手动裁剪图表。这一步容易出错（坐标轴或图例可能被裁切），耗时且难以在不同用户或研究之间标准化；当需要处理数十张图表时，处理能力也因此受到限制。其次，许多工具依赖逐条曲线的点击式数字化，以及风险 表的手动转录；它们通常一次只能处理一张图中的一条曲线，几乎不支持批量导入、多图解析，或将曲线与相应风险表自动关联。因此，要扩展到大量试验，需要投入大量协调人员的时间，还会增加转录错误的可能性，并产生难以复现的审计追踪。

相比之下，LEAD-ONC 提供了一套自动化程度更高的人工智能辅助框架，用于从已发表的肿瘤学试验报告中提取定量信息。该平台旨在通过统一的工作流程处理多个治疗组和多项试验的 KM 曲线，从而减少人工干预并提升可扩展性。除 KM 曲线外，LEAD-ONC 还会从已发表的“表1”数据中提取基线特征摘要，以便后续评估试验可比性，并开展更有针对性的证据综合。图 1 总结了 LEAD-ONC 的端到端工作流程，涵盖从研究人员筛选的文献输入到为后续贝叶斯分析生成结构化数据的全过程。

面向用户的工作流程。研究人员上传试验 PDF，选择 Kaplan–Meier（KM）图和相关表格（风险人数；基线特征/表1），核验自动识别的数值，并完成简短的三点 KM 校准和追踪。每次上传都会获得一个持久的研究标识符，并保留子图限定符（例如“图 1A”“CheckMate-227-B”），以防不同子图面板的数据混淆。

人工智能辅助风险表识别（主要采用大语言模型路径，并提供光学字符识别回退方案）。使用多模态大语言模型（Google Gemini 2.0 Flash），通过 Base64 图像提交和结构化 JSON 模式（治疗组名称须完全一致；数组须对齐），提取治疗组名称、x 轴时间点和风险人数。采用指数退避重试（≤5 次；3–48 秒），以应对暂时性错误。返回的 JSON 会经过模式验证（类型、数组长度相同、风险人数单调递减或持平），并提交给人工快速确认或更正。如果未通过验证或未达到置信度阈值，则启动光学字符识别回退方案：(i) 检测 x 轴下方的风险表区域（OpenCV），并在检测失败时使用关键词引导的回退方案；(ii) 对感兴趣区域进行预处理（灰度化、自适应阈值、倾斜校正、伪影清理）；(iii) 使用 PaddleOCR（针对数字优化；英语）进行识别，并以 Tesseract 作为次选；(iv) 通过将列与刻度中心或投影峰值对齐，并按行高对行分组，重建表格；以及 (v) 应用领域约束（风险人数单调递减或持平；漂移较小时将时间吸附到网格；数组长度相同）。当两条路径均成功时，融合与判定步骤

会逐单元格计算差异；对于轻微且孤立的差异（≤±1），优先采用大语言模型的输出；对于较大冲突或违反规则的情况，则优先采用光学字符识别的结果。并排差异视图可辅助用户裁定。

三点校准和人工监督数字化。在进行数字化之前，需要先由人工完成校准，因为自动坐标轴检测（OCR/LLM + CV）在各种异构KM图布局中仍不够稳定——坐标轴可能因扫描而发生倾斜、没有标签或仅部分带有标签、使用非正交网格、以周而非月绘图，或显示重叠曲线和删失刻度；此外，可见的x轴最大值并不总是等于风险表所暗示的末端月份。三点校准使用户能够快速确定缩放比例和方向，无需完整的坐标轴解析，并且适用于这些不同情况。在描绘曲线之前，用户点击三个带标签的刻度，以定义像素→数据变换：原点 (0, 0)；最右侧带标签的x轴刻度 (max, 0)，其中max是根据风险表核实的末端月份；以及顶部y轴刻度 (0, 100)。这些不共线的对应点可确定一个二维仿射映射

$$t = a\,u + b\,v + c,\;\; s = d\,u + e\,v + f$$

该映射涵盖平移、旋转、缩放和轻微剪切（能够适应倾斜）。随后，用户在HTML5画布上描绘每条KM曲线；像素轨迹会立即转换为校准后的 (t,s) 对，并以JSON格式存储。支持多臂面板。

标准化和KM约束。 所有描绘出的曲线均通过分段线性插值重新采样为500个均匀间隔的时间点，并强制满足KM典型性质：(1) 在t=0 时的生存率为= 100%（若缺失则插入）；(2) 时间严格递增；(3) 生存率单调不增。若发现违规，应用内会发出警告，以便在最终确定前进行针对性修正。

治疗臂映射和去重。 使用分层匹配器将曲线与风险表治疗臂关联：(1) 用户确认的映射；(2) 精确的规范化字符串匹配；(3) 带长度感知约束的模糊部分匹配（±5 个字符，以避免假阳性，例如将“化疗”匹配到“纳武利尤单抗 + 化疗”）；(4) 回退至曲线颜色标签。治疗臂名称按输入图中的原文保留。

质量控制和验证。 自动检查会验证：(i) 坐标文件（xy.csv，每条曲线500行）与风险表（risk_table.csv）是否按研究–治疗臂键一一对应；(ii) 时间点数组与风险人数数组长度是否相同；(iii) 是否满足KM约束（初始化和单调性）；以及(iv) 研究标识符（包括子图后缀）是否唯一。任何检查失败都会阻止导出，并提示进行针对性修正。

个体患者数据（IPD）重建。以提取的XY坐标（xy.csv）和风险表（risk_table.csv）为输入，可使用modified-iKM方法[2]生成与所选历史试验的历史KM曲线相关联的个体患者数据（IPD）。

评估所选历史试验的同质性

大多数临床试验的发表文献都会包含表1，用于概述关键基线特征的分布：连续尺度变量的均值和标准差、分类变量的计数和百分比，以及每个治疗组的样本量。假设在经过严格的筛选过程后，我们选出了*H*项与计划开展的临床试验相关的历史试验，并希望通过评估这些*H*项历史试验在共同的*p*个基线协变量方面的相似性，来判断它们的同质性并识别亚群。

令 x_{ij} 和 x_{ik} 分别表示两个不同试验中第 i 个基线变量（例如性别、ECOG体能状态）的样本均值，两个试验分别以 j 和 k标记。我们使用标准化差异来衡量协变量差异度[16]。 对于连续变量，试验j 和 k之间的标准化差异定义为

$$d_{ijk} = \frac{|x_{ij} - x_{ik}|}{\sqrt{\frac{S_{ij}^2 + S_{ik}^2}{2}}}$$

相应的样本方差通常会在已发表的S_{ij}^2 S_{ik}^2临床试验结果“表1”中报告。对于仅报告连续变量中位数和范围的试验，我们使用报告的中位数近似样本均值，并将最大值与最小值之差除以四，以近似样本方差。

对于二元变量，我们考虑

$$d_{ijk} = \frac{|x_{ij} - x_{ik}|}{\sqrt{\frac{x_{ij}(1-x_{ij}) + x_{ik}(1-x_{ik})}{2}}}$$

标准化差异常用于匹配和加权相关文献中衡量协变量平衡，也可作为不同单位协变量的平衡度量[16]。我们通过j 和 k 两个试验间不同协变量的标准化差异的平均值，来定义患者基线特征的不相似性

$$D_{jk} = \frac{\sum_{1}^{p} d_{ijk}}{p}.$$

或者，我们可以取标准化差异的最大值，来定义两项试验之间患者基线特征的不相似性。

$$D_{jk} = max(d_{1jk}, \ ..., d_{pjk}).$$

我们可以使用 K-中心 medoid 算法对试验进行聚类[17]，也可以根据 D_{jk} 历史试验两两之间的不相似性，采用其他选定的方法。通常使用 *0.2*、*0.5*和 *0.8* 分别表示d_{ijk}小、中和大的效应量[18]，，因此我们将（通过计算标准化差异的平均值或最大值获得的）小于 0.2 视为试验间具有良好可比性的指标；较大的值则表明D_{jk}试验人群之间的差异可能更大。

使用贝叶斯分层建模进行证据综合

贝叶斯分层模型（BHMs），例如荟萃分析预测先验（MAP）方法，常用于试验规划中的证据综合。[6] 以往关于贝叶斯分层模型的研究主要侧重于对不同试验结局建模，而所得证据往往难以解读，因为：(1) 不同试验的结局可能碰巧相似或不同；(2) 尚不清楚综合证据来自哪类患者人群，以及这些综合证据与目标人群的相关性如何。

我们选择以特定簇方式构建贝叶斯分层模型，以确保能够结合特定目标人群来解读模型结果。对于所在簇与目标试验人群最为相似的试验，我们构建贝叶斯分层模型，并假设这些试验的生存函数是来自同一贝塔–斯泰西过程（BSP）的随机样本。BSP模型[19]由精度函数和参数生存函数构成；后者是$c(t)$　　$G(t)$，即同一簇中不同试验所对应生存函数的均值。Deborah等人利用约500个III期肿瘤试验治疗组的重建个体患者数据表明，双参数威布尔模型足以描述这些治疗组的生存曲线[20]。因此，我们假设 $G(t)$服从威布尔模型。我们假设等于一个常数，使得一个$c(t)$　　c贝叶斯分层模型所代表的信息量相当于c名患者。我们对c及的参数假设独立非信息先验。贝叶斯分层模型的详细信息见附录。$G(t)$

根据既往临床试验所得的全部证据，BHM 的后验预测分布代表了对目标临床试验潜在生存曲线的先验知识。**图3**展示了在一项纳入鳞状和非鳞状组织学类型患者的目标临床试验中，与单药和双药免疫检查点抑制剂相关的生存曲线预测分布。这种有针对性的证据综合方法使用户能够选取最相关的聚类，并利用该聚类的 BHM 结果为目标临床试验的规划提供依据。

结果

基线特征与试验相似性

表1总结了入组五项III期试验、针对未经治疗的晚期非小细胞肺癌患者的基线特征。在CheckMate-227中，患者根据

PD-L1表达≥1%被前瞻性地分为两个试验部分。PD-L1表达 <1%的患者入组第1a部分，并随机接受纳武利尤单抗联合伊匹木单抗（双联治疗）、纳武利尤单抗单药治疗或化疗（CT）；而PD-L1表达 <1%的患者入组第1b部分，并随机接受双联治疗、纳武+化疗或单独化疗。由于第1a部分纳武利尤单抗组与第1b部分纳武+化疗组的中位总生存期相近（15.7个月 vs. 15.2个月），因此将这两个治疗组汇 pooled for 后续分析。然而，CheckMate-227未按组报告第1b部分（PD-L1 <1%）入组患者的基线特征。因此，我们采用双联治疗组的基线特征进行聚类分析。

我们在**图2**中展示了五项试验两两之间的患者特征差异度。差异度量是通过计算这些试验共同报告的协变量所对应的标准化差异的平均值（**图2a：** 左面板）或最大值（**图2b：** 右面板）获得的。K中心点聚类得到三个组别，主要与组织学类型相一致：非鳞状组织学类型（KEYNOTE-189）、鳞状组织学类型（KEYNOTE-407）以及由CM227、CM9LA和POSEIDON组成的混合组织学类型。无论采用何种方式计算差异度量（平均值或最大值），这些试验的聚类归属均相同。图2a显示，混合组织学聚类中两项试验的协变量之间的标准化差异平均小于0.20；而图2b显示，混合组织学聚类中两项试验的协变量之间的标准化差异最大为0.22，表明基线协变量之间仅存在较小差异。KEYNOTE-189完全由非鳞状组织学类型患者组成，而KEYNOTE-407的患者均具有鳞状组织学类型。此外，KEYNOTE-189纳入的ECOG=0 患者多于其他试验。

新型混合组织学类型临床试验的生存期预测

图3中的每条灰色曲线都代表未来一项比较单药与双联免疫检查点抑制剂的随机试验中可能观察到的一条生存曲线，假设患者基线特征与混合组织学聚类中各试验的患者相似。这些曲线共同刻画了在设计阶段，基于对既往单药与双联抑制剂试验重建的个体患者数据进行贝叶斯分层建模后，单药与双联抑制剂组生存函数的不确定性。通过对这些合理的生存曲线取平均，我们可以得到各治疗组的预期总生存率，以及预设时间点的预期总生存差异，并计算相应的95%贝叶斯可信区间（置信区间）。**表2**汇总了多个时间点预期总生存差异（单药与双联免疫检查点抑制剂）的95%置信区间。由于所有评估时间点的总生存差异贝叶斯可信区间均包含零，因此，在假设的目标人群中，未来试验显示双重免疫检查点抑制具有临床意义的优效性的概率有限。

基于对重建的个体患者数据拟合的贝叶斯层级模型，模型预测单药和双药免疫检查点抑制剂的中位总生存期分别为14.0个月（95% 置信区间：11.2-17.1）和16.8个月（95% 置信区间：13.3-20.8）。预测的中位总生存期差异为2.8个月（95%置信区间：-2.0，7.6）。与单药治疗相比，预测双联治疗可使总生存期至少延长3个月的概率约为0.45。与这一发现一致，**图4**显示两组预测中位数的95%置信区间存在较大重叠。

对前瞻性随机对照试验的启示

综合来看，这些结果表明，基于从单药和双重免疫检查点抑制相关历史试验中综合得到的证据，对于在与混合组织学聚类相似的人群中开展比较双重与单药免疫检查点抑制的随机试验，其预期成功概率有限。

讨论:

通过整合大语言模型与贝叶斯分层建模，我们开发了 LEAD-ONC，这是一个端到端框架，可将已发表的肿瘤学试验转化为定量且与设计相关的证据。具体而言，LEAD-ONC：(i) 使用基于常见基线特征构建的不相似度指标，量化历史试验之间的可比性；(ii) 通过大语言模型辅助提取和个体患者数据重建，将已发表的 Kaplan–Meier 生存曲线转化为可分析的数据；以及 (iii) 使用基于聚类的贝叶斯层级模型，为预先指定的目标人群生成生存函数的后验预测分布。在应用于一线非小细胞肺癌时，基于成对不相似度的聚类揭示了具有临床可解释性的结构，按组织学构成区分试验，并明确哪些研究最能为未来试验设定假设提供信息。

在混合组织学聚类（POSEIDON、CheckMate-227 和 CheckMate-9LA）中，LEAD-ONC 预测单药免疫检查点抑制与双重免疫检查点抑制的生存曲线差异有限。估计的总生存期中位数差异不大，且后验可信区间高度重叠，这表明若在未来一项纳入类似人群的随机试验中评估双重阻断，其获得具有临床意义获益的可能性有限。这些发现说明，将从已发表文献中提取的证据与贝叶斯建模相结合，可用于将设计阶段的假设形式化，并通过明确量化不确定性和试验成功的概率，为关键决策提供支持。

肿瘤学试验日益面临成本不断攀升、周期延长和患者入组困难等挑战，在患者人群因生物标志物定义的亚组而分散的情况下尤为如此。下一步，我们将评估如何扩展所提出的证据综合框架，以构建信息性先验，用作早期阶段

单臂试验的合成对照，或用于增强随机研究中的对照组，旨在提高效率、减少所需样本量，同时保持推断严谨性。尽管本研究聚焦于生存结局，LEAD-ONC 框架具有通用性，可轻松扩展至采用二元或连续终点的试验。

目前，**LEAD-ONC** 的元数据提取和证据综合模块分别实现，证据综合所需的贝叶斯分层建模目前在本地计算环境中进行。目前正在持续开发并拓展 LEAD-ONC 的功能，包括集成先进的文献检索能力、增强文本与图表处理，以及加强提取模块与贝叶斯综合模块之间的衔接。这些改进旨在打造一个更流畅、可扩展且面向用户的平台，以支持循证试验设计。

1. Danchev V, Min Y, Borghi J 等：实施国际医学期刊编辑委员会数据共享声明要求后数据共享情况的评估。《JAMA Netw Open》4:e2033972，2021年

2. Liu N、Zhou Y、Lee JJ：IPDfromKM：根据已发表的 Kaplan-Meier 生存曲线重建个体患者数据。《BMC医学研究方法学》21:111，2021年

3. Rohatgi A：automeris.io：使用 WebPlotDigitizer 从图表中提取数据的计算机视觉辅助工具 [互联网][引用于2025年11月 2] 网址： https://automeris.io

4. Roychoudhury S、Neuenschwander B：针对生存时间终点临床试验的历史对照数据贝叶斯借力。《医学统计学》39:984–995，2020年

5. Fu C、Zhou S、Huang X 等：基于重建的生存数据进行证据综合 [互联网],2022[引用于2025年11月 2] 网址： http://arxiv.org/abs/2201.00281

6. Neuenschwander B、Capkun-Niggli G、Branson M 等：总结临床试验中的历史对照信息。《临床试验》7:5–18，2010年

7. Lang L、Zhao Y、Gao Q 等：RESOLVE-IPD：高保真个体患者数据重建与不确定性感知亚组荟萃分析 [互联网]. arXiv [statME] ,2026年可从以下网址获取：http://dx.doi.org/10.48550/arXiv.2511.01785

8. Di Federico A、Stumpo S、Mantuano F 等：晚期非小细胞肺癌患者接受 CTLA-4 与程序性死亡配体1或 PD-1 双重阻断治疗后的长期总生存期及基于生物标志物的亚组分析：系统综述与重建的个体患者数据

荟萃分析。《柳叶刀·肿瘤学》26：1443–1453，2025年

9. Novello S、Kowalski DM、Luft A 等：帕博利珠单抗联合化疗治疗鳞状非小细胞肺癌：III期KEYNOTE-407研究的5年更新 [网络]。《临床肿瘤学杂志》，2023[引自2025年11月 2] 可访问：https://ascopubs.org/doi/10.1200/JCO.22.01990

---

10. Brahmer JR、Lee J-S、Ciuleanu T-E 等：在CheckMate 227中，纳武利尤单抗联合伊匹木单抗与化疗作为转移性非小细胞肺癌一线治疗的5年生存结局。《临床肿瘤学杂志》41:1200–1212，2023年

11. Reck M、Ciuleanu T-E、Schenker M 等：随机CheckMate 9LA试验中，转移性非小细胞肺癌患者接受一线纳武利尤单抗联合伊匹木单抗及2个周期化疗，与单独接受4个周期化疗的5年结局。《欧洲癌症杂志》211:114296，2024年

12. Peters S, Cho BC, Luft AV 等：度伐利尤单抗联合化疗，无论是否加用曲美木单抗，用于一线转移性非小细胞肺癌：3期POSEIDON试验的五年总生存结局。《胸部肿瘤学杂志》20:76–93，2025年

13. Hellmann MD, Paz-Ares L, Bernabe Caro R 等：纳武利尤单抗联合伊匹木单抗治疗晚期非小细胞肺癌。《新英格兰医学杂志》381:2020–2031，2019

14. Paz-Ares L, Ciuleanu T-E, Cobo M 等：一线纳武利尤单抗联合伊匹木单抗，并联用两个周期化疗治疗非小细胞肺癌患者（CheckMate 9LA）：一项国际性、随机、开放标签3期试验。《柳叶刀·肿瘤学》22：198–211，2021年

15. Gandhi L, Rodríguez-Abreu D, Gadgeel S 等：帕博利珠单抗联合化疗治疗转移性非小细胞肺癌。《新英格兰医学杂志》378：2078–2092，2018

16. Austin PC：倾向评分匹配样本中治疗组基线协变量分布比较的平衡诊断。医学统计学 28:3083–3107，2009

17. Schubert E、Rousseeuw PJ：更快的 K 中心点聚类：改进 PAM、CLARA 和 CLARANS 算法，载于《相似性搜索与应用》。沙姆：施普林格国际出版社，2019，第 171–187 页

18. Cohen J：《行为科学统计功效分析》。劳特利奇。2013

19. Walker S、Muliere P：Beta-Stacy 过程与 Pólya urn 方案的推广。统计学年刊 25:1762–1780，1997

20. Plana D、Fell G、Alexander BM 等：癌症患者生存率可通过参数化来提高临床试验精度并揭示时间依赖性治疗效果。自然通讯 13:873，2022

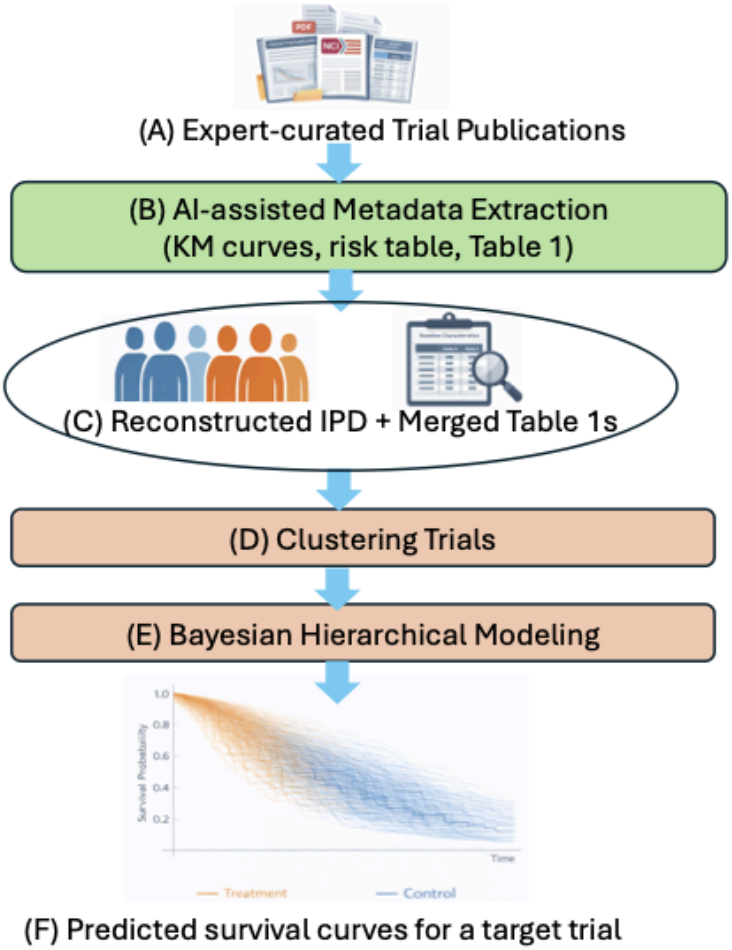
表1：五项针对未经治疗的晚期非小细胞肺癌患者的临床试验中入组患者的基线特征。从这些试验中提取基线特征汇总数据，并根据共同可用的协变量进行合并。

	第1簇（混合组织学类型）				第2簇（非鳞状）	第3簇（鳞状）
临床试验	CheckMate-9LA	CheckMate-227	POSEIDON		KEYNOTE-189	KEYNOTE-407
Arm	纳武利尤单抗+伊匹木单抗（N=361）	纳武利尤单抗+伊匹木单抗（N=583）	曲美木单抗+度伐利尤单抗+CT (N=338)	度伐利尤单抗+CT (N=338)	帕博利珠单抗+化疗 (n=410)	帕博利珠单抗+化疗 (N=278)
Age (中位数，范围)	65(59-70)	64(26-87)	63(27-87)	64.5(32-87)	65(34-84)	65(29-87)
女性(%)	30	32.6	20.4	25.1	38	20.9
ECOG体能状态（%）						
0	31	35	32.5	32.2	45.4	26.3
1	68	64.7	67.5	67.8	53.9	73.7
组织学类型（%）						
鳞状	31	28	36.7	37.9	0	97.5
非鳞状细胞	69	71.9	63.3	61.8	100	0
吸烟状态 (%)						
从不吸烟者	13	13.6	17.5	24.9	11.7	7.9
既往/目前吸烟者	87	85.2	82.6	75.1	88.3	92.1
PD-L1 >=1% (%)	60	67.9	63	66.3	63.4	63.3

表2： 新试验中单药与双药免疫检查点抑制剂的预计总生存期（OS）及其在关注时间点的预计差异。这些预测基于以下假设：新试验的基线患者特征与混合组织学簇中的试验（POSEIDON、CM227 和 CM9LA）相似。

时间/月 onth	总生存期（单药免疫检查点抑制剂）			总生存期（双药免疫检查点抑制剂）			总生存差异		
	估计	2.5% 置信区间	97.5% 置信区间	估计	2.5% 置信区间	97.5% 置信区间	估计	2.5% 置信区间	97.5% 置信区间
12	0.54	0.48	0.59	0.58	0.52	0.63	0.04	-0.04	0.12
24	0.35	0.30	0.41	0.41	0.36	0.46	0.06	-0.02	0.13
36	0.24	0.20	0.29	0.30	0.25	0.36	0.06	-0.01	0.13
48	0.17	0.13	0.22	0.23	0.19	0.28	0.06	-0.01	0.12
60	0.13	0.09	0.17	0.18	0.14	0.23	0.05	-0.01	0.12
72	0.09	0.06	0.13	0.14	0.10	0.19	0.05	-0.01	0.10

图 1：LEAD-ONC 工作流程。LEAD-ONC 将研究者筛选的试验发表文献转化为定量且考虑不确定性的证据，以支持预先指定的目标试验设计。(A) 研究者选择相关的试验发表文献。(B) 使用大语言模型和光学字符识别，从非结构化来源中提取定量信息，包括 KM曲线和基线特征表。(C) 将提取的信息转化为重建的个体患者数据和协调后的基线协变量。(D) 根据基线协变量的相似性对试验进行聚类。(E) 在与目标试验人群最相似的聚类中，使用贝叶斯分层模型综合多个试验的证据。(F) 得出的目标试验生存曲线预测分布可为治疗效应量等设计参数的选择提供依据，并估计试验成功的概率。



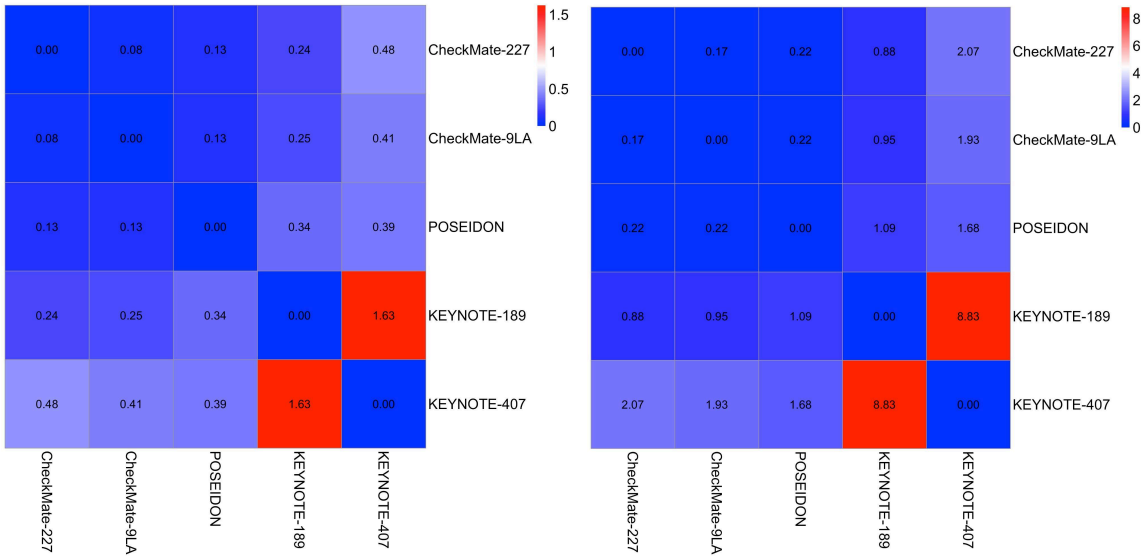


图2：所选非小细胞肺癌试验之间基线患者特征的不相似性，其中0.2、0.5 和 0.8 分别代表小、中和大差异。不相似性数值通过计算所有可用协变量的标准化差异的平均值（图2A：左面板）或最大值（图2B：右面板）得出。

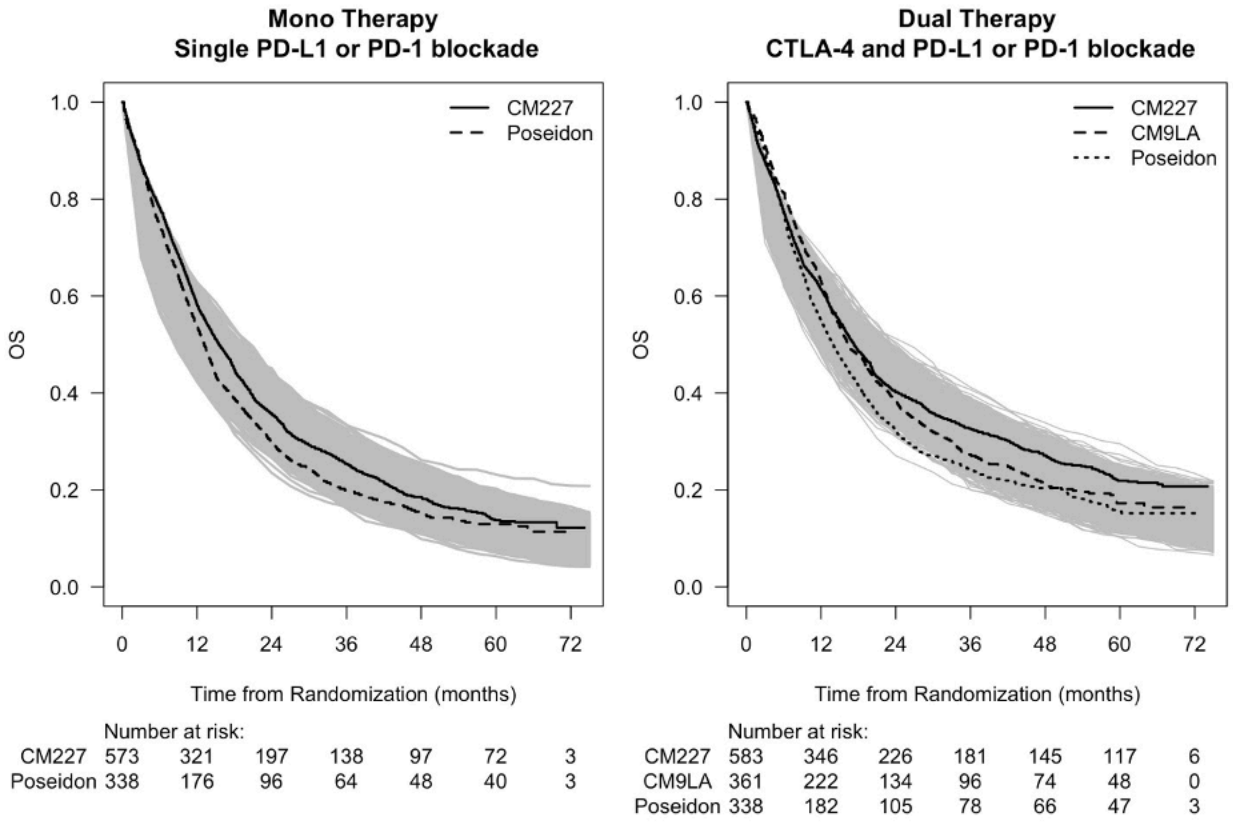


图3：假设新试验的基线协变量特征与混合组织学簇中的试验（POSEIDON、CM227 和 CM9LA）相似，则该新试验中单药和双药抑制剂组对应的生存曲线预测结果。每条灰色曲线代表未来试验中可能观察到的一条生存曲线；这些预测基于混合组织学簇中试验所发表的单药治疗（左面板）和双药治疗（右面板）的 Kaplan-Meier 曲线（黑色）。

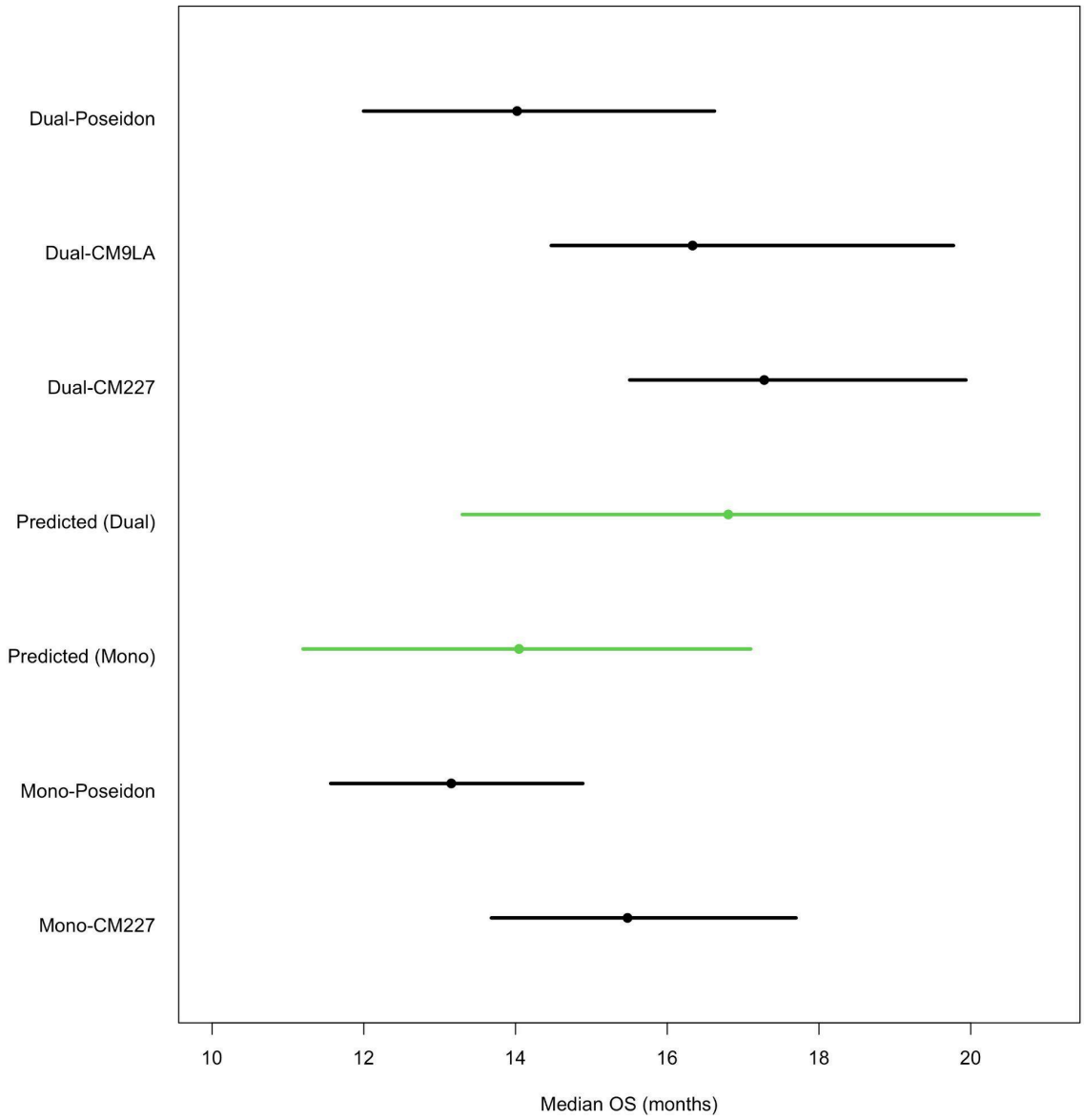


图4：模型预测的新试验中单药和双药免疫检查点抑制剂的中位总生存期和95%置信区间；该试验的患者特征与 POSEIDON、CM227 和 CM9LA 类似。各临床试验中位数的 Kaplan-Meier估计量以黑色显示。

补充附录：技术细节

2026年2月9日

目录

1 IPD重建 **2**

2 生存证据综合 **2**2.1 beta-Stacy过程综述 .22.2 层次贝叶斯生存模型 .3

3 新试验生存曲线的预测分布 **5**

1 个体患者数据重建

我们希望综合关于接受相同治疗或同一治疗类别C的患者生存结局的证据。假设经过全面的文献综述和严格的筛选过程，我们从具有相同关键入选标准的历史试验中，确定了 J 个不同的治疗组，这些治疗组中的患者均接受C治疗。我们将这些历史治疗组记为 $H_1, \ldots, H_J$，对应的样本量记为 $n_1, \ldots, n_J$。

既往研究的 Kaplan-Meier（KM）曲线通常会与风险表一同发表，风险表显示在等间隔时间网格上处于风险集中的患者人数。modified-iKM方法以风险表和提取的KM曲线XY坐标作为输入[1]，为每位 H_j生成个体患者数据（IPD），其中 $j = 1, \ldots\ldots, J$。令 i 为患者索引。按照标准生存分析记号，我们将 Y_i 记为该患者的随访时间，并令$\delta_i = 1$ 表示观察到感兴趣事件，$\delta_i = 0$ 表示该患者被删失。我们将 H_j 的重建个体患者数据记为 $D_j = \{(Y_i, \delta_i); i = 1, \ldots\ldots, n_j\}$。

2 生存率证据综合

我们使用贝塔–斯塔西过程（BSP）对重建的个体患者数据进行建模，该过程与右删失观察数据共轭 [2]。我们先简要回顾BSP；有关详细介绍，请参见 Walker 和 Muliere [2]。

2.1 beta-Stacy过程综述

令 $F(t)$ 表示定义在预先指定的时间网格 $\{t_1, \ldots, t_K\}$ 上的离散化事件时间的未知累积分布函数（CDF）。定义 CDF 在 t_k 处的增量为

$$\theta_k = F(t_k) - F(t_{k-1}),\ t_0 = 0,\ F(0) = 0.$$

在给定 $\theta_1, \ldots, \theta_{k-1}$ 的条件下，随机增量 θ_k 服从 beta-Stacy分布 [3]，其概率密度函数为

$$f(\theta_k|\alpha_k, \beta_k, \theta_1, \ldots, \theta_{k-1}) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k - 1} \frac{(1 - \theta_1 - \cdots - \theta_k)^{\beta_k - 1}}{(1 - \theta_1 - \cdots - \theta_{k-1})^{\alpha_k + \beta_k - 1}},$$

其中 $\alpha_k > 0$ 且 $\beta_k > 0$。

作为一个特例，θ_1 服从参数为 α_1 和 β_1 的 beta分布。

我们可以将 F 定义为随机增量之和，从而构造其先验分布

$$F(t_k) = \sum_{j=1}^{k} \theta_j, \tag{2.1}$$

假设每个 θ_j 在给定累积分布函数此前增量的条件下，服从参数为α_j 和 β_j 的beta-Stacy分布

的分布。

按照 Walker 和 Muliere[2], ，得到的关于 $F(t)$ 的先验分布是在 $\{t_1, \ldots, t_K\}$上定义的离散时间 Beta–Stacy 过程（BSP），记作

$$F(t) \sim BSP(\{\alpha_k, \beta_k\}).$$

BSP先验对于右删失生存数据具有共轭性。设 d_k 和 r_k 分别表示在 t_k时的事件数和风险集患者数。记作 $m_k = r_k - d_k$。 $F(t)$的后验分布仍是一个BSP，其参数更新为$\{\alpha_k^*, \beta_k^*\}$，并在$\{t_k\}$处发生跳跃。我们将此后验分布表示为

$$F(t)|\boldsymbol{m}, \boldsymbol{d} \sim BSP(\{\alpha_k^*, \beta_k^*\}),\ \alpha_k^* = \alpha_k + d_k,\ \beta_k^* = \beta_k + m_k, \tag{2.2}$$

其中 $\boldsymbol{m} = (m_1, \ldots, m_K)$和 $\boldsymbol{d} = (d_1, \ldots, d_K)$。

设 $G(t)$表示预先指定的生存函数， $c(t)$表示一个已知的正函数，用于表示时间 t先验知识的强度。按照 [2], ，我们将BSP先验参数化为

$$\alpha_k = c(t_k)\{G(t_{k-1}) - G(t_k)\}, \tag{2.3}$$

$$\beta_k = c(t_k)G(t_k). \tag{2.4}$$

经过重新参数化后， $S(t) = 1 - F(t)$的先验分布以 $G(t)$为中心。也就是说，

$$E\{S(t)|G(t), c(t)\} = G(t).$$

2.2 层次贝叶斯生存模型

对于每个历史治疗组 H_j，令 F_j 表示定义在预先指定的时间网格 $\boldsymbol{\tau}^* = (t_1, \ldots\ldots, t_K)$ 上的离散化生存时间的累积分布函数。也就是说，我们假设离散化个体患者数据只能取自 $\boldsymbol{\tau}^*$。由于重建数据的准确性低于原始数据，因此有必要对重建的生存时间 $Y_1, \ldots\ldots, Y_{n_j}$进行离散化。不同试验的风险表可能具有时间分辨率各异的时间网格。为尽量减少离散化造成的信息损失，我们建议在 J 个历史治疗组中，选择时间分辨率最高的网格作为时间网格 $\boldsymbol{\tau}^*$ 。虽然从技术上讲， t_K 可以趋于无穷大，但我们将 t_K 视为不同试验的最大随访时间。例如，如果一个历史试验的风险表采用时间网格（6，12，……，60），而另一个历史试验的风险表采用时间网格（3，6，……，72），我们将选择后者作为 $\boldsymbol{\tau}^*$。

令 F^* 表示新试验中接受 $\mathcal{C}$ 治疗的患者的累积分布函数。我们构建一个贝叶斯层级模型（BHM），假设不同试验的累积分布函数服从离散时间的 BSP先验：

$$F^*, F_1, \ldots, F_J \sim BSP(\{\alpha_k, \beta_k\}), \tag{2.5}$$

定义在网格 τ^* 上。

记 η_{jk} 为 H_j 在 t_k 时的风险率。我们可以通过乘积极限估计量将风险率与 F_j 联系起来

$$F_j(t) = 1 - S_j(t) = 1 - \prod_{t_k \le t} (1 - \eta_{jk}). \tag{2.6}$$

根据 Walker 和 Muliere[2], 公式 2.5 可知，不同临床试验的风险率服从相互独立的分布，其先验为 贝塔分布 先验

$$\eta_k^*, \eta_{1k}, \ldots, \eta_{Jk} \sim \text{Beta}(\alpha_k, \beta_k). \tag{2.7}$$

令 d_{jk} 和 r_{jk} 分别表示基于 H_j 的离散化个体患者数据，在 t_k 时 k =1、.....、K 的事件数和风险集患者数。对于每个 t_k，我们考虑一个二项模型

$$d_{jk} \sim \text{Bin}(r_{jk}, \eta_{jk}). \tag{2.8}$$

按照 Walker 和 Muliere[2], ，不同时间点的参数 $\{\alpha_k, \beta_k\}$ 可以通过生存模型 $G(t)$ 和精度函数 $c(t)$ 联系起来。我们定义

$$\begin{aligned} \alpha_k &= c(t_k)\{G(t_{k-1}) - G(t_k)\}, \\ \beta_k &= c(t_k)G(t_k). \end{aligned}$$

近期一项针对 500 个试验组（约 220,000 个事件）的实证研究表明，肿瘤学试验中的生存率可由双参数威布尔模型[4]很好地描述。因此，我们考虑

$$G(t) = exp(-\lambda t^{\kappa}).$$

不同于先前的研究将 $G(t)$ 和 $c(t)$ 视为固定值[2, 5, 6], ，我们认为二者都是随机变量，并为 λ 和 κ 指定相互独立的弥散先验，其中

$$\lambda \sim Gamma(0.01, 0.01), \tag{2.9}$$

and

$$\kappa \sim Gamma(0.01, 0.01). \tag{2.10}$$

令 $c(t)$ 等于正数常量 c，可得 $\beta_k = \beta_{k-1} - \alpha_k$，且 $t = 0$ 时的风险集患者数为 $\alpha_1 + \beta_1 = cG(0) = c$。因此，我们可以将 c 解释为 BSP先验的先验有效样本量（ESS）。

历史患者的总样本量为 $N = \sum_{j=1}^{J} n_j$。由于 c 是 BSP先验的有效样本量，因此 c 不应超过 N，而对于 c，一个合理的先验为

$$\frac{c}{N} \sim \text{Beta}(1, 1). \tag{2.11}$$

3 新试验生存曲线的预测分布

利用第 2.2节所述的层次化 Beta–Stacy 过程模型，可以基于以下内容推导出新试验中生存函数的先验预测分布

$$S^*(t) = \prod_{t_k \le t} (1 - \eta_k^*), \tag{3.1}$$

其中考虑了估计 $G(t)$ 和 c 时的不确定性。

我们使用另一个贝塔–斯塔西过程近似 $S^*(t)$ 的预测分布，并假设

$$S^*(t) \sim BSP(\{\alpha_k^*, \beta_k^*\}),$$

其参数与（2.3）中的通常贝叶斯斯塔西过程参数化相匹配:

$$\alpha_k^* = c^* \left\{ G^*(t_{k-1}) - G^*(t_k) \right\},$$

and

$$\beta_k^* = c^* G^*(t_k),$$

其中

$$G^*(t) = \exp\{-\lambda^* t^{\kappa^*}\}, \tag{3.2}$$

并且 λ^* 和 κ^* 分别是 λ 和 κ 的后验均值。

Walker 和 Muliere [2] 证明，遵循 beta-Stacy过程的生存函数所对应的风险率是相互独立的贝塔随机变量。因此，我们可以假设新试验中的风险率 $\eta_1^*, \ldots, \eta_K^*$相互独立，且分别服从 Beta($\alpha_1^*, \beta_1^*$) …,Beta($\alpha_K^*, \beta_K^*$)，那么 $S^*(t)$ 的先验方差可以用闭式计算:

$$V^*(t) = \prod_{\tau_k \le t} u_k^* - \prod_{\tau_k \le t} v_k^*, \tag{3.3}$$

其中

$$v_k^* = \left(\frac{\alpha_k^*}{\alpha_k^* + \beta_k^*} \right)^2,$$

and

$$u_k^* = v_k^* + \frac{\alpha_k^* \beta_k^*}{(\alpha_k^* + \beta_k^*)^2 (\alpha_k^* + \beta_k^* + 1)},$$

令 $\hat{}V(t)$ 表示基于 $S^*(t)$ 的预测分布的大样本得到的估计方差，该预测分布为 $S^*(t)$ ，由式（3.1）给出。给定 λ^* 和 κ^*，我们通过找到使 c^* 最小的 c^* 来求解

$$\sum_{k=1}^{K} (V^*(t_k) - \hat{V}(t_k))^2. \tag{3.4}$$

我们可以将先验预测分布的有效样本量（ESS）和有效事件数（ENE）定义为

$$ENE = \sum_{k=1}^{K} \alpha_k^*,\ ESS = c^*. \tag{3.5}$$

参考文献

[1] Liu N、Zhou Y、Lee JJ。IPDfromKM：从已发表的 Kaplan-Meier 生存曲线重建个体患者数据。《BMC 医学研究方法学》。2021年；21(1):111。

[2] Walker S、Muliere P。Beta-Stacy 过程与 Pólya urn 方案的推广。《统计年刊》。1997：1762-80。[3] Mihram GA、Hultquist RA。双变量预警时间/失效时间分布。《美国统计学会会刊》。1967；62(318)：589-99。[4] Plana D、Fell G、Alexander BM、Palmer AC、Sorger PK。癌症患者生存可通过参数化提高试验精度，并揭示时间依赖性治疗效果。《自然·通讯》。2022；13(1)：873。[5] Rigat F、Muliere P。使用 beta-Stacy过程进行非参数生存回归。《统计规划与推断杂志》。2012；142(9)：2688-700。

[6]Ventz S，Trippa L. 贝叶斯多臂减量设计。《贝叶斯分析》。2024年;1(1):1-23。