面向层级医疗数据的贝叶斯不确定性加权自适应保形预测

Marzieh Amiri Shahbazi[1],**Ali Baheri**[2],**Nasibeh Azadeh-Fard**[1]

[1]罗切斯特理工学院工业与系统工程系，纽约州罗切斯特 14623，美国 [2]罗切斯特理工学院机械工程系，纽约州罗切斯特 14623，美国 ma7684@g.rit.edu, akbeme@rit.edu, nafeie@rit.edu

arXiv:2601.01223v1 [cs.LG] 3 Jan 2026

摘要

临床决策需要量化不确定性，既提供无分布假设的覆盖保证，又具备风险自适应精度，而现有方法无法同时满足这两项要求。我们提出一种混合贝叶斯—保形框架，以解决医疗预测中的这一根本局限。该方法将贝叶斯层级随机森林与考虑群组的保形校准相结合，利用后验不确定性对符合度分数加权，同时保持严格的覆盖有效性。在涵盖美国 3,793 家医院、4 个地区的 61,538 例住院病例上进行评估后，我们的方法实现了目标覆盖率（94.3%，目标值为 95%），并具备自适应精度：对于低不确定性案例，区间窄 21%；对于高风险预测，则适当扩大区间。至关重要的是，我们证明，仅依靠经过良好校准的贝叶斯不确定性会导致严重的覆盖不足（14.1%），凸显了混合方法的必要性。该框架可支持风险分层临床方案、针对高置信度预测的高效资源规划，以及对不确定案例进行保守分配并加强监督，从而在多样化医疗场景中提供具备不确定性感知能力的决策支持。

引言

住院时长（LOS）预测对于资源管理、出院规划和护理协调至关重要，但当前机器学习方法缺乏临床应用所需的不确定性估计（Rajkomar 等，2018；Stone 等，2022；Beam 和 Kohane，2018）。可靠的不确定性界限有助于在资源压力日益增大的情况下更好地分配床位、配置人员并管理患者流转（McCarthy 等，2021；Harris 等，2018）。然而，医疗保健数据具有层级结构，患者隶属于医院和医疗系统，由此产生的统计依赖违反了标准方法中的独立性假设（Goldstein，2011；Raudenbush 和 Bryk，2002；Cressie 等，2009；Gelman 等，2013；Werner 等，2023；Bollmann、Groll 和 Havranek，2023；Austin，2024；Berta、Levantesi 和 Marta，2024）。这些由机构实践和病例组合所驱动的依赖关系会削弱现有方法，并形成当前方法无法捕捉的系统性模式（Snijders 和 Bosker，2011；Iezzoni，2012；Dimick 等，2013；Freeland 等，2013；Kaboli 等，2012）。

版权所有 © 2026，人工智能促进协会（www.aaai.org）。保留所有权利。

应对这些挑战，需要构建兼顾统计可靠性与对机构差异响应能力的模型。分层随机森林（HRFs）能够高效捕捉多层级结构（Shahbazi 和 Azadeh-Fard，2025；Hajjem，2010；Wager 和 Athey，2018），但其不确定性估计依赖渐近假设和启发式方法，缺乏形式化的覆盖保证（Ovadia 等人，2019；Mentch 和 Hooker，2016）。这可能削弱临床信任并危及患者安全（Guo 等人，2017；Sendak、Balu 和 Schulman，2020；Liu 等人，2020）。

保形预测无需分布假设即可提供严格的覆盖率保证（Vovk、Nouretdinov 和 Gammerman，2009；Lei 等人，2018；Angelopoulos、Bates 等人，2023），但其生成的区间一律相同，忽略了预测难度，并且依赖可交换性假设，而层级结构会违反这一假设（Barber 等人，2023；Tibshirani 等人，2019）。贝叶斯模型能够捕捉层级依赖关系并提供自适应不确定性（Gelman，2006；Carlin 和 Xia，1999），但只能提供渐近覆盖率，在实践中可能无法实现（Gelman 等，2013；Betancourt，2017）。这一差距是根本性的：保形方法缺乏适应性，而贝叶斯方法缺乏覆盖有效性；然而，临床部署需要两者兼备（Amodei 等人，2016；Varshney，2016）。

医院住院时长预测必须兼顾患者安全所需的可靠性与资源规划所需的精确性（Mee-hinkong 和 Ponnoprat 2025；Arthur 等人 2022）。管理者需要能够支持风险分层决策的不确定性估计，并且这些估计应能推广至各类医院，无需针对具体院区重新校准（Helm 等人 2011年；Proudlove 等人 2003；Sendak、Balu 和 Schulman 2020 年；Liu 等人 2020年）。具体而言，我们采用分割式层级保形预测，这是一种不加权的绝对残差方法，可生成宽度一致的区间。我们不采用自适应变体（方差缩放、加权、分位数保形），因为这些方法需要可靠的逐点代理指标，而我们的 HRF 实现无法提供此类指标；此外，它们的保证通常假设数据可交换，这与聚类数据的特性不相容。

我们提出将保形预测与贝叶斯方法相结合，以增强分层随机森林中的不确定性量化，并设计了一个框架，该框架（i）通过考虑群组的保形校准实现无分布边际覆盖率；（ii）恢复实例级

自适应性，并通过贝叶斯后验预测标准差进行评估，应用于医院住院时长预测。

贡献。 本文作出以下贡献：

1. 我们提出一种混合框架，将分组感知的保序校准与贝叶斯后验不确定性相结合，从而生成自适应预测区间，在医院聚类情况下保持覆盖率，并根据实例级不确定性进行调整。
2. 我们提出一种标准化校准策略，无需针对特定站点再训练，即可推广至具有不同特征的医院；同时系统评估覆盖率稳定性，并推荐用于监测和重新校准的框架。
3. 临床评估阐释了如何在医院环境中实际应用不确定性引导的预测区间，分析预测置信度如何映射至资源分配决策，并概述将统计性能转化为临床工作流程时需要考虑的实施因素。

相关工作

研究者已广泛使用临床评分系统、回归模型和机器学习（ML）方法（包括集成方法、梯度提升和深度学习）开展住院时长预测研究（Stone 等，2022；Chrusciel 等，2021；Ciaroni 等，2023）。尽管相关研究已有数十年，大多数研究仍侧重点预测准确性，而非临床决策所必需的不确定性量化。人们早已认识到，住院时长预测中的不确定性是有效开展医院排程和容量规划的一大障碍（Young，1973；McCarthy 等，2021；Harris 等，2018），但严谨的方法仍然有限。

医疗保健领域的标准不确定性量化方法（自助法区间、集成方差和贝叶斯后验）能够提供有用的摘要，但缺乏有限样本下、无分布假设的覆盖保证（Guo 等人 2017；Sendak、Balu 和 Schulman2020）。临床部署要求不确定性估计同时具备有效性（在不依赖分布假设的情况下维持名义覆盖率）和自适应性（对难以预测的对象提供更宽的区间），而传统方法鲜少能兼顾这两种特性（Liu 等人 2020；Rajkomar 等，2018）。在全局可交换性假设下应用时，现有方法未能考虑医疗保健数据固有的层级依赖关系，因而会产生校准失真的区间和系统性的医院层面残差模式（Snijders 和 Bosker，2011；Iezzoni，2012；Dimick 等，2013；Freeland 等，2013；Kaboli 等，2012）。

层级随机森林在保留基于树的集成方法计算效率的同时，显式地对多层级数据结构进行建模（Breiman 2001）。Hajjem 等人率先提出了适用于聚类数据的混合效应随机森林（Hajjem 2010），后续研究提升了其可扩展性（Sela 和 Simonoff 2012；Shahbazi 和 Azadeh-Fard 2025），并将其扩展到多重嵌套层级（Pellagatti 等人，2021）。然而，现有的所有分层随机森林实现都缺乏严谨的不确定性量化；文献已明确指出这一不足（Mentch 和 Hooker 2016；Tyralis 和 Papacharalampous 2024），这也正是我们提出方法贡献的动因。

尽管保形预测能够提供有力的覆盖保证（Vovk、Nouretdinov 和 Gammerman 2009；Shafer 和 Vovk 2008；Angelopoulos、Bates 等人 2023；Lei 等人 2018；Baheri 和 Shahbazi 2025），但其假设与医疗保健数据的嵌套结构相冲突，限制了其在临床环境中的有效性（Barber 等人 2023；Tibshirani 等人 2019）。近期的方法学进展通过组感知校准方案（Principato 等人 2024）、针对双层层级结构的理论框架（Dunn、Wasserman 和 Ramdas 2023），以及对可交换性假设的扩展（Barber 等人 2023）来应对这一局限。尽管取得了这些进展，尚无先前研究将层级保形方法与随机森林结合，用于连续型临床结局，而这类场景中机构聚类和自适应不确定性都至关重要。

贝叶斯层级模型通过在嵌套结构间进行部分汇聚，自然地刻画多层次不确定性（Gelman 等人，2013；Gelman，2006；Murphy，2012），并因其可解释性而广泛应用于医疗保健领域（Carlin 和 Xia，1999；Raudenbush 和 Bryk，2002）。在树模型场景中，贝叶斯加性回归树（BART）及其扩展通过模型平均来量化不确定性（Chipman、George 和 McCulloch，2010；Tan 和 Roy，2019；Linero，2018；Pratola 等人，2020）。然而，贝叶斯方法只能提供渐近覆盖保证（Gelman 等，2013；Vehtari、Gelman 和 Gabry，2017），因此在安全关键型医疗保健应用中可靠性有限，而这类应用尤为重视有限样本有效性（Angelopoulos、Bates 等人，2023）。

混合贝叶斯-保形方法是一种新兴方法，将保形覆盖保证与贝叶斯自适应性相结合（Fong 和 Holmes，2021；Stanton、Maddox 和 Wilson，2023；Zhang、Park 和 Simeone，2024），并已应用于优化（Stanton、Maddox 和 Wilson，2023）、医学成像（Ekmekci 和 Cetin，2025）和物理建模（Podina、Rad 和 Kohandel，2024）。然而，这些方法尚未适用于层级医疗保健数据；在此类数据中，机构聚类要求采用专门的校准方案，并确保满足临床部署所需的有限样本保证。

除了住院时长预测之外，层级（多层）模型也广泛用于在相关决策问题之间共享统计强度（Baheri 2025年）。在控制和强化学习中，层级结构是表示多层决策和时间抽象的标准方式，也见于优先级／层级控制设计中（Bafandeh 等，2018年；Wong 和 Wonham，1996）。优化领域也有类似思想：多任务和多保真度优化采用层级／多输出高斯过程代理模型，在不同任务和保真度之间传递信息（Anandalingam 和 Friesz，1992年；Baheri 2023）。

论文结构第3节介绍方法论、算法和实验设计。第4节报告实证结果及性能比较。

第5节讨论研究意义、局限性和未来方向。第6节总结主要贡献。

方法论

我们提出一种混合不确定性量化框架，将保形预测的覆盖有效性与贝叶斯建模的自适应精度相结合，用于层级医疗数据。核心思想是，这两种方法各有互补优势：保形方法能够在不依赖分布假设的情况下提供覆盖有效性，而贝叶斯方法则支持针对具体实例进行自适应。我们的框架分为三个阶段。首先，通过依次分解残差，利用层级随机森林捕捉多层级依赖关系。其次，我们通过组感知保形校准处理层级结构，并利用反映实例特定不确定性的贝叶斯后验分布，获得不确定性估计。第三，我们整合这两种方法，在保持保形覆盖率的同时，实现自适应区间宽度。

问题表述

我们针对层级数据开展医院住院时长预测，其中患者（$n = 61,538$）嵌套于医院（$H = 3,793$）和地区（$R = 4$）之中。每条患者记录包括人口统计学和临床特征、医院及地区标识符，以及以天为单位的观测住院时长。这种结构会产生院内依赖关系，违反标准的独立性假设，因此需要专门的不确定性量化方法。我们的设计旨在构建满足以下三个目标的预测区间：(i) 尽管存在层级依赖关系，仍保持名义覆盖概率（95%）；(ii) 尽可能缩小区间宽度，以提高运营效率；(iii) 根据预测难度调整区间大小。

层级随机森林基础

我们采用 HRF 作为基础预测模型（Shahbazi 和 Azadeh-Fard，2025），将预测依次分解为三个层级：患者层级效应、医院特定调整和地区模式。该模型依次训练三个随机森林，首先仅使用患者特征（f_0），然后使用患者-医院组合来捕捉医院特定残差（$\hat{}\eta$），最后使用患者-医院-地区组合来捕捉地区残差（ξ）。由此得到最终预测：

$$\hat{y}_i = \hat{f}_0(x_i) + \hat{\eta}_{h_i}(x_i, h_i) + \hat{\xi}_{r_i}(x_i, h_i, r_i) \tag{1}$$

这种分解确保每个层级都能捕捉不同的变异来源，同时对患者特征与机构因素之间的复杂交互进行建模。

层级保形预测框架

我们采用拆分保形预测来提供分布无关的覆盖有效性。该方法将数据划分为训练集、校准集和测试集，并在校准数据上计算非符合度分数 $C_i = |y_i - f(x_i)|$ $\hat{}$ 数据，并构造预测区间 $[f(x) \pm q]$，其中 q 是校准分数的（$1 - \alpha$）分位数。我们选择标准拆分保形方法，而非自适应变体，原因有三：（i）自适应方法会带来高昂的$\mathcal{O}(n)$ 单次预测计算成本，不适用于大规模临床部署；（ii）这类方法需要 HRF 原生不提供的条件不确定性估计；（iii）我们通过贝叶斯后验不确定性实现针对具体实例的自适应，而非依赖复杂的保形机制，从而构建模块化框架，由保形方法确保覆盖有效性。

层级医疗数据违反了保形预测的可交换性假设，因为同一医院内的患者会呈现相关的结果。我们通过 CDF 池化来解决这一问题，即利用所有校准数据对全局分位数进行校准。尽管这种方法在聚类情况下无法保证有限样本覆盖率，但在我们的超大数据集中（3,793家医院），它具有更优的经验覆盖率、更小的变异性和确定性的可复现性。我们将 CDF 池化与两种能够保留保证的替代方法进行比较：单医院子抽样和重复子抽样。

贝叶斯分层随机森林

HRF 通过其三层分解（公式 1）提供点预测，但无法估计不确定性。我们采用贝叶斯分层模型来量化残差变异和参数不确定性：

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2) \tag{2}$$

$$\mu_i = \beta_0 + \beta_1 \hat{f}_{\text{HRF}}(x_i, h_i, r_i) + \alpha_{h_i} + \gamma_{r_i} \tag{3}$$

其中 β_0，β_1 用于校准 HRF 预测，$\alpha_h \sim \mathcal{N}(0, \sigma_h^2)$ 和 $\gamma_r \sim \mathcal{N}(0, \sigma_r^2)$ 则捕捉 HRF 未能充分建模的医院和区域残差效应，而 σ^2 表示患者层级方差。我们使用马尔可夫链蒙特卡洛拟合该模型，以获得后验样本。对于每位新患者，后验预测标准差 $\sigma_{\text{pred}}(x, h, r)$ 作为不确定性估计，涵盖参数不确定性、残差机构效应和患者层级变异。

混合 HRF（贝叶斯保形）框架

我们通过依据后验不确定性对符合度分数加权，将保形覆盖保证与贝叶斯自适应精度相结合：

$$S_i^{(w)} = \frac{|y_i - \hat{f}(x_i, h_i, r_i)|}{\max(\sigma_{\text{pred}}(x_i, h_i, r_i)^\gamma, \epsilon)} \tag{4}$$

其中 $\sigma_{\text{pred}}(\cdot)$ 为后验预测标准差，$\gamma \in [0, 2]$ 控制自适应性（默认值为 $\gamma = 1$），而 $\epsilon = 10^{-6}$用于防止除法不稳定。我们根据校准集上的这些类 z 分数残差校准单个分位数 $\hat{}q$，然后依据每个测试点的不确定性$\sigma_{\text{pred}}(x)^\gamma$对 $\hat{}q$ 进行缩放，从而得到自适应区间：对于不确定性较高的病例，区间会变宽，同时在各医院之间保持覆盖率稳定。

这种模块化设计将覆盖有效性（保形校准）与不确定性估计（贝叶斯

算法1：混合 HRF（贝叶斯保形）预测

输入： 训练集 $\mathcal{D}_{\text{train}}$、校准集 $\mathcal{D}_{\text{calib}}$、测试点 $(x_{\text{new}},\ h_{\text{new}},\ r_{\text{new}})$、失覆盖 α、自适应 γ、方法 M**输出：** 预测区间 $[L, U]$，覆盖率为 $\geq 1-\alpha$

阶段1：训练

训练 HRF：$\hat{f}_{\text{HRF}} \leftarrow \text{HierarchicalRF}(\mathcal{D}_{\text{train}})$

训练贝叶斯：$\theta \leftarrow \text{BayesianHRF}(\mathcal{D}_{\text{train}},\ \hat{f}_{\text{HRF}})$

第 2 阶段：加权校准分数

对每个 $(x_j,\ y_j,\ h_j,\ r_j) \in \mathcal{D}_{\text{calib}}$ **执行**

$\hat{y}_j = \hat{f}_{\text{HRF}}(x_j,\ h_j,\ r_j)$; 从后验分布中获取 σ_j

$S_j^{(w)} = |y_j - \hat{y}_j| / \max(\sigma_j^{\gamma}, \epsilon)$

结束循环

阶段 3：分层保形校准

如果 $M = \text{CDF 池化}$ **则**

$\mathcal{S} = \{S_j^{(w)}\}_{j=1}^{n_{\text{calib}}}$

否则如果 $M =$ 单次子抽样**则**

$\mathcal{S} = \{S_{j_h}^{(w)} : h \in \text{医院}\}$

否则如果 $M =$ 重复子抽样**则**

$\hat{q} =\ B$ 子抽样分位数的中位数

结束条件判断

如果 $M \neq$ 重复子抽样**则**

$\hat{q} =$ 分位数$(\mathcal{S}, (1-\alpha)(|\mathcal{S}|+1)/|\mathcal{S}|)$

结束条件判断

阶段 4：测试预测

$\hat{y}_{\text{new}} = \hat{f}_{\text{HRF}}(x_{\text{new}}, h_{\text{new}}, r_{\text{new}})$

从贝叶斯后验分布获取 σ_{new}

$w = \hat{q} \cdot \max(\sigma_{\text{new}}^{\gamma}, \epsilon)$

返回 $[L, U] = [\hat{y}_{\text{new}} - w,\ \hat{y}_{\text{new}} + w]$

建模）：经验覆盖率仅取决于校准方案，而非贝叶斯模型的正确性。我们将贝叶斯分层残差与分层保形校准及分层随机森林结合起来；这种组合通过算法 1 有别于此前的加权保形方法。

验证指标与实验框架

我们针对住院时长五分位数采用五折分层交叉验证，训练/校准/测试集划分比例为 64%/16%/20%。缺失值（2%）采用均值/中位数填补，连续变量进行了 z 标准化。超参数通过网格搜索选取。HRF 包含三个层级：患者（100 棵树，深度 15）、医院（75 棵树，深度 12）和地区（50 棵树，深度 10）。贝叶斯模型采用马尔可夫链蒙特卡洛（2 条链，预热 500 次，预热后 250 次，目标接受率 = 0.99），并通过 $R < 1.01$ 和有效样本量 > 100 检查收敛性。

经验覆盖率（目标为 ≥95%（针对 $\alpha = 0.05$））和不确定性五分位数内的条件覆盖率作为主要指标。次要指标包括区间宽度、不确定性-误差相关性（皮尔逊 r）以及自适应指数（高不确定性与低不确定性下的宽度比）。稳健性分析考察了交叉验证折、置信水平（80%、90%、95%、99%）、层级变化以及医院亚组中的性能。

实验

我们使用2019年医疗费用与利用项目（HCUP）全国住院患者样本（医疗保健研究与质量局（AHRQ）2025年）的数据，分析了来自美国四个普查地区3,793家医院的61,538条住院记录。该数据库是美国规模最大的公开可用全支付方住院数据库。所有分析均使用符合HIPAA法规的去标识化数据（医疗保健研究与质量局（AHRQ）2024年）。

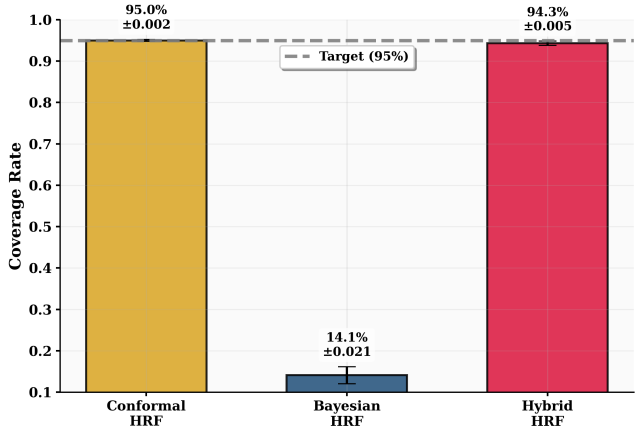
18 个预测变量涵盖三个层级：患者层级（人口统计学特征、临床严重程度评分、诊断、医疗操作）、医院层级（床位规模、教学状态、所有权）和区域层级（美国人口普查地区）。这一三级层级结构呈现出显著的聚类现象。组内相关系数显示，住院时长方差的 12.5% 可归因于医院层级因素，40.8% 可归因于区域因素，46.7% 可归因于患者层级因素（总方差：50.6 天²）。这种显著的区域聚类（ICC= 0.408）表明存在机构差异，违反了标准的独立性假设。住院时长呈现典型的医疗保健数据特征：中位数为 3 天（四分位距：2–6 天），均值为 4.9 天（标准差：7.1 天），且呈右偏分布。

覆盖有效性。在五折交叉验证中，各方法的边际覆盖率存在显著差异（图1A）。保形 HRF 达到 95.0%（±0.2 个百分点），符合目标（95%）。混合 HRF 达到 94.3%（±0.5 个百分点），仅降低 0.7 个百分点。贝叶斯 HRF 仅为 14.1%（±2.1 个百分点），表明存在严重的覆盖不足。保形 HRF 和混合 HRF 的覆盖率稳定性较高（变异系数（CV）分别为 0.21% 和 0.53%），各折的曲线在 ∼0.95 附近近似水平（图1B）。贝叶斯 HRF 的折间变异显著更大（CV：14.9%）。

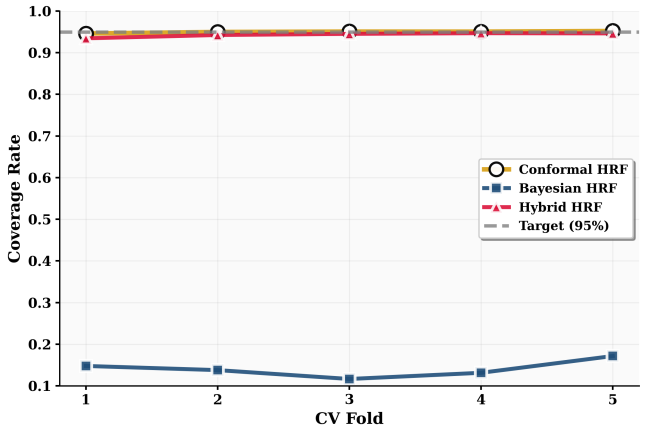
条件覆盖率分析按不确定性五分位数对预测结果进行分层（Q1 =不确定性最低/最容易的案例，Q5 = 不确定性最高/最困难的案例），揭示出不同的模式（图2）。贝叶斯 HRF 的覆盖率始终较差（12.7%–15.5%）。保形 HRF 在 Q1–Q4 中实现了近乎完美的覆盖率（96.2%–99.7%），但在 Q5 中骤降至 81.7%（较目标低 −13.3 个百分点）。混合 HRF 在 Q1–Q4 中保持较高的覆盖率（93.9%–97.1%），且 Q5 的覆盖不足较小（90.9%，较目标低 −4.1 个百分点）；与保形 HRF 相比，最困难案例的覆盖缺口减少了 69%。

区间效率与自适应。五折交叉验证揭示了区间宽度特征存在显著差异（表 1）。贝叶斯 HRF 生成了极窄的区间（0.75 ± 0.02 天），但覆盖率可靠性不足。保形 HRF 生成了均匀区间（16.32 ± 0.36 天，按设计适应比为 = 1.00）。混合方法在保持覆盖有效性的同时，实现了略窄的区间（15.99± 0.59 天，减少 2.0%），并具有显著的自适应区间调整能力（适应比 =1.29，$P < 0.001$）。

关键的是，混合方法通过真正的优化而非牺牲覆盖率来提升效率。总体缩减 2.0% 看似幅度不大，但效率提升集中在最关键之处：低不确定性案例（Q1）的区间缩窄了 21%（13.21 天，相比 16.71 天），从而支持精准的资源规划；而



（a）五折的总体覆盖率（均值 ±标准差）。虚线：95%目标值。



（b）各折的覆盖率稳定性。保形方法和混合方法稳定在 ≈0.95；贝叶斯方法则在 0.11–0.17 之间波动。

图1：各方法的覆盖率性能

表1：交叉验证区间宽度分析

宽度指标	保形	贝叶斯	混合	差值†
平均宽度（天）	16.32 ± 0.36	0.75 ± 0.02	15.99 ± 0.59	-0.33
相对于保形方法的宽度比	1.00 ± 0.00	0.045 ± 0.001	0.98 ± 0.04	-0.02
适应比	1.00 ± 0.00	1.01 ± 0.01	1.29 ± 0.01	+0.29
交叉验证稳定性（%）	2.2	2.8	3.7	+1.5

注： ± = 5 折的标准差。 †差值 = 混合方法与保形方法之间的差异。适应比 = Q5/Q1 宽度比。

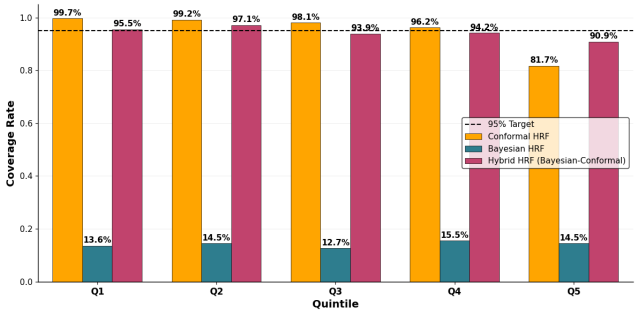


图2：不同不确定性五分位数的条件覆盖率。(A) 贝叶斯 HRF：整体覆盖率极低（12.7%–15.5%）。(B) 保形 HRF：Q1–Q4 接近完美（96.2%–99.7%），Q5 覆盖不足（81.7%）。(C) 混合 HRF：Q1–Q4 表现良好（93.9%–97.1%），Q5 有所改善（90.9%）。

高不确定性案例（Q5）则得到适当的保守性扩宽（16.98 天，相比 15.99 天， +6%），以支持安全优先的规程（表2）。这种风险响应式区间调整的适应比为 1.29，能够实现按不确定性分层的临床工作流程，而统一的保形区间无法做到这一点。

交叉验证稳定性证实了稳健的性能（图3）。保形和混合方法在各折中保持了一致的宽度（交叉验证 $<$ 2%），混合方法在每一折中均表现出正向效率提升（范围：0.31–0.37 天，未出现效率损失）。自适应指数保持稳定（1.29 ± 0.01），95%置信区间为 [1.28, 1.30]，排除了区间均匀的原假设（1.00, $P < 0.001$），证实自适应区间调整是真实的方法学特征，而非数据伪象。

这些结果表明，混合方法实现了真正的优化：在维持覆盖率有效性的同时，通过自适应区间宽度实现风险分层资源分配。这一能力对于不确定性引导的临床方案至关重要，而统一的保形区间无法提供这种能力。

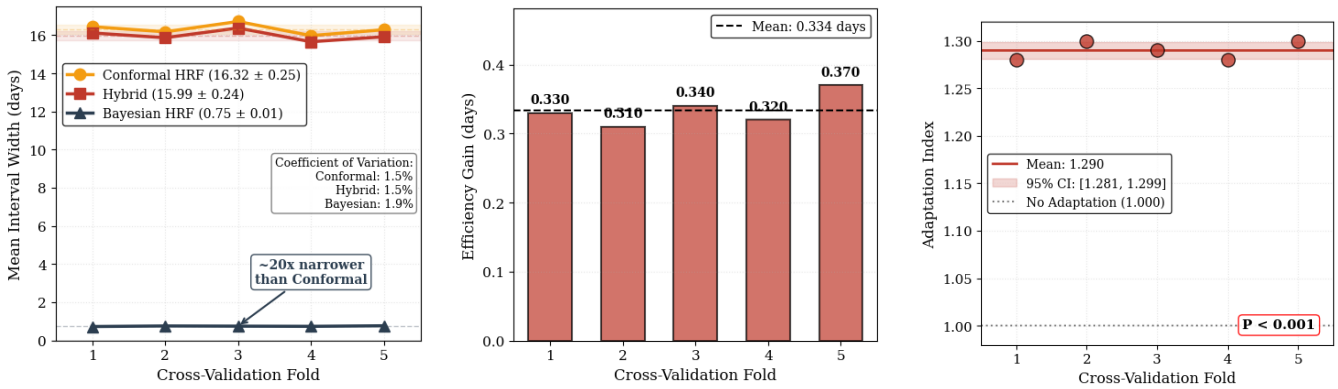
不确定性校准。 原始贝叶斯后验分布的不确定性与误差相关性较弱（皮尔逊 $r = 0.203$、斯皮尔曼 $\rho = 0.043$），且存在严重的校准偏差。等渗回归校准大幅提升了不确定性质量（表3）：皮尔逊相关系数提高至 0.476（+134%），斯皮尔曼相关系数提高至 0.243（+465%）；校准斜率从 91.4 提高至 0.926；期望校准误差从 2.52 天降至 0.037 天（降低 98.5%）；均值不确定性从 0.19 天增至 2.71 天（+1326%）。QQ图（图4）证实分布得到改善，并趋近正态分布。

尽管不确定性校准良好，覆盖率表现却大相径庭（表4）。贝叶斯 HRF 的覆盖率仅为 14.1%，区间很窄（0.75 天）；而混合 HRF 的覆盖率达到 94.3%，区间宽度适当（15.99 天）。

表2：按不确定性五分位划分的区间宽度

五分位	均值 σ	保形	贝叶斯	混合	混合/保形
Q1（低）	0.168	16.71 ± 0.06	0.69 ± 0.01	13.21 ± 0.10	0.79
Q2	0.176	16.24 ± 0.10	0.74 ± 0.01	14.86 ± 0.12	0.92
Q3	0.179	16.06 ± 0.17	0.73 ± 0.02	15.14 ± 0.15	0.94
Q4	0.186	16.62 ± 0.15	0.78 ± 0.01	16.48 ± 0.31	0.99
Q5（高）	0.195	15.99 ± 0.10	0.81 ± 0.04	16.98 ± 1.38	1.06

注：均值 σ = 后验预测不确定性（五分位数划分）。相对于保形基线的宽度比。



(a) 宽度稳定性：各方法与各折。

(b) 混合效率提升（所有折均为正值）。

(c) 自适应指数一致性（P ¡ 0.001）。

图3：5 折交叉验证稳定性

表3：保序校准对不确定性质量的影响

状态	皮尔逊 r	斯皮尔曼 ρ	斜率 β	ECE	Mean σ
Raw	0.203	0.043	91.4	2.52	0.19
校准后	0.476	0.243	0.926	0.037	2.71
改进	+134%	+465%	−98.9%	−98.5%	+1326%

表4：具有校准后不确定性的覆盖率性能

方法	覆盖率 (%)	宽度 (天)	CRPS	Winkler 评分
贝叶斯HRF	14.1	0.75	2.04	95.02
混合HRF	94.3	15.99	2.04	35.52

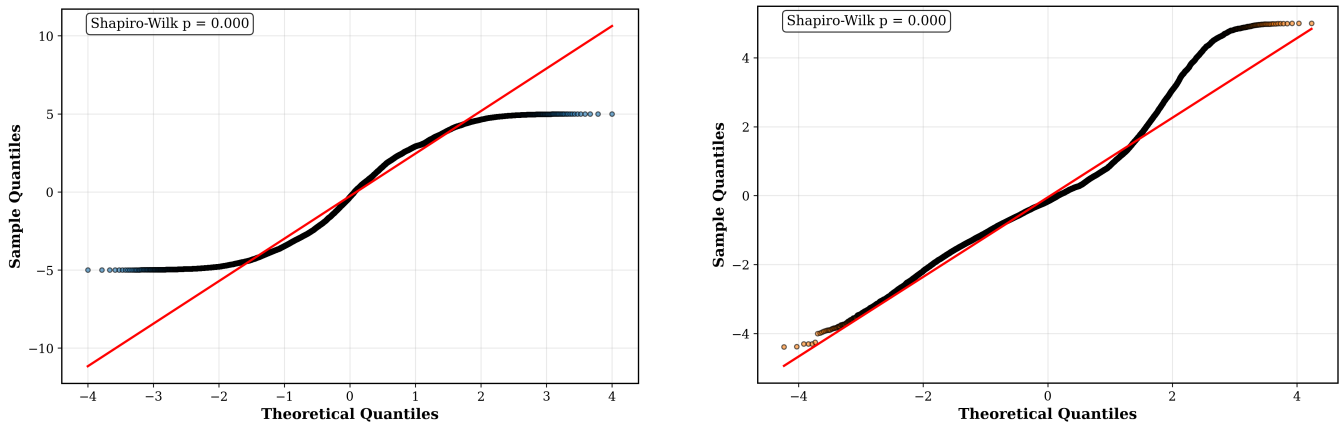
相同的连续排序概率评分（2.04）证实了两者的区分能力相当。Winkler 评分更优（35.5 对 95.0），表明可靠的覆盖率胜过狭窄的区间，这也验证了该框架的设计。

临床解读。 94.3% 的覆盖率意味着，医院规划人员可以预期，每 1,000 例住院病例中约有 943 例的实际住院时长落在预测范围内，从而为容量管理提供可量化的可靠性。知道每 1,000 例中约有 57 例会超出区间，便可提前规划升级容量——例如分配备用病床、安排出院规划资源，或与急性后续护理网络协调——将不可预测的需求转化为可管理、已纳入预算的应急容量。与保形方法相比，高置信度预测（Q1–Q2）对应的区间更窄（Q1：13.21 天，对比 16.71 天），在常规监测下可将缓冲时间缩短 1.4–3.5 天。高不确定性案例（Q5）则对应适当保守的区间，从而触发强化监督。

稳健性与泛化能力。全面的稳健性分析证实，该方法在多个维度上均表现稳定（表5）。各方法的预测准确度相当（均方根误差：6.0–6.4天，图A）。保形方法和混合方法能够在不同置信水平（80–99%）下恰当地跟踪名义目标，而贝叶斯方法的覆盖率不足（图B）。在所有医院特征（床位规模、所有权、地理位置和地理区域）下，覆盖率均保持稳定；保形方法和混合方法达到93.5–95.6%，而贝叶斯方法仅为11.0–40.6%（图C）。其他层级结构设定对混合方法性能的影响很小（覆盖率：94.2–94.8%；宽度：15.80–16.30天，图D）。

讨论

本研究表明，混合层级（贝叶斯保形HRF）不确定性量化能够实现可靠的覆盖率，同时提供自适应精度，适用于医疗保健——



(a) 原始不确定性　　(b) 校准后不确定性

图4：展示分布改进的QQ图：(a) 存在系统性偏差的原始不确定性；(b) 趋近于正态性的校准后不确定性。

表5：跨多个维度的综合稳健性分析

维度	条件	保形	贝叶斯	混合
	A. 预测准确度（均方根误差，天）			
	交叉验证（5 折，均值 ±标准差）	6.05 ± 0.59	6.39 ± 0.60	6.01 ± 0.65
	B. 置信水平敏感性（覆盖率 %，均值 ± SD）			
	$\alpha = 0.01$ （目标 99%）	99.1 ± 0.1	22.1 ± 2.1	99.0 ± 0.2
	$\alpha = 0.05$ (95%目标值)	95.0 ± 0.4	14.1 ± 2.1	94.4 ± 0.3
	$\alpha = 0.10$ (90%目标值)	90.0 ± 0.6	12.0 ± 1.0	88.9 ± 0.7
	$\alpha = 0.20$ （目标值为 80%）	80.0 ± 0.6	8.0 ± 0.6	79.0 ± 1.6
	C. 医院特征（覆盖率 %，均值 ± 标准差）			
床位规模				
	小型（n=13,534）	95.1 ± 1.0	11.0 ± 1.9	95.3 ± 1.0
	中型 (n=17,735)	95.3 ± 0.2	13.9 ± 2.4	95.2 ± 0.1
	大型 (n=30,269)	94.8 ± 0.6	15.4 ± 1.7	94.3 ± 1.1
所有权				
	公立（n=7,023）	93.5 ± 2.2	12.7 ± 1.4	93.5 ± 3.0
	私立非营利（n=45,896）	95.0 ± 0.3	15.8 ± 2.2	94.7 ± 0.6
	私营营利性（n=8,619）	94.9 ± 1.7	13.6 ± 1.8	94.0 ± 0.8
地理位置				
	农村（n=5,259）	94.8 ± 2.1	14.2 ± 0.8	94.7 ± 1.3
	城市非教学型（n=10,868）	95.1 ± 0.3	15.6 ± 4.1	94.8 ± 1.3
	城市教学 (n=45,411)	94.7 ± 0.7	13.7 ± 1.4	94.8 ± 1.0
地区				
	美国东北部（n=11,223）	95.2 ± 1.1	37.9 ± 4.4	95.4 ± 0.8
	美国中西部（n=13,654）	95.6 ± 0.1	38.7 ± 1.3	94.6 ± 0.6
	南部（n=24,643）	94.9 ± 0.2	38.7 ± 0.5	94.1 ± 0.5
	西部（n=12,018）	95.0 ± 0.4	40.6 ± 3.3	94.7 ± 1.1
	D. 层级规范敏感性（混合 HRF）			
	三级（患者→医院→地区）	94.3 ± 0.1% / 15.99 ± 0.59 天		
	2层（患者→医院）	94.8 ± 0.3% / 16.30 ± 0.34 天		
	2层（患者→地区）	94.2 ± 0.4% / 15.80 ± 0.42 天		

覆盖率和宽度值以均值 ± 标准差表示。图A：不同方法之间的点预测等效性。图B：不同置信水平下的目标跟踪。图C：所有医院亚组的性能。图D：不同层级结构下的覆盖率/宽度。

护理预测。我们使用来自美国四个地区3,793家医院的61,538名患者的数据，该框架在留出测试集上实现了94.3%的覆盖率，使低不确定性案例的区间缩短了21%，并对高不确定性案例进行保守扩展（适应比：1.29，95%置信区间[1.28，1.30]）。在医院特征、验证折和置信水平方面，性能均保持稳定。

主要贡献在于证明，无论底层不确定性模型的质量如何，无分布校准对于安全关键应用都至关重要。尽管贝叶斯区间与混合方法具有同样经过良好校准的不确定性判别能力（连续排序概率评分：两者均为2.04），但仅凭贝叶斯区间只能达到14.1%的覆盖率，这表明不确定性校准无法替代正式的覆盖有效性。这验证了我们的模块化设计：使用贝叶斯分层模型进行自适应加权，使用保序校准确保覆盖有效性。我们的方法不同于现有的自适应保形方法。局部加权保序预测和分位数回归森林同样能够提供自适应区间，但我们的框架通过离线贝叶斯训练而非逐点重新加权来提高计算效率，将覆盖率保证与不确定性估计模块分离，并利用天然适用于聚类医疗数据的层级结构。然而，与这些方法进行系统比较仍是未来一项重要工作。

局限性与缓解措施。主要局限在于最高不确定性五分位组（Q5）的覆盖不足（90.9%，低于95%目标值，差4.1个百分点），这影响了最需要可靠界限的患者。出现这一问题的原因是，尽管不确定性判别能力有所改善（校准后皮尔逊相关系数：0.203 →0.476），单一全局保序乘数仍无法完全适应异方差误差模式。虽然保序理论保证了可交换性条件下的边际覆盖率，但无法确保不确定性分组子集内的条件覆盖率。

缓解策略包括按不确定性分层进行校准，即为每个五分位组分别使用保序分位数；对高不确定性案例保守地扩大区间（例如，超过第75百分位数时采用1.2×倍乘数）；以及加强临床监督，要求对标记案例进行高级人员审查。部署决策必须权衡覆盖率缺口与临床风险承受度。优先考虑绝对安全的场景，在Q5覆盖率问题解决之前，可能需要采用标准保序方法的保守做法；而普通内外科人群则可通过加强监督方案来缓解轻度覆盖不足。

结论

本研究表明，将贝叶斯层级不确定性量化与无分布假设的保形校准相结合，可为聚类医疗数据实现可靠且自适应的预测区间。在来自3,793家医院的61,538名患者数据上进行评估后，该框架在留出测试数据上实现了接近目标覆盖率的表现（94.3%，95%目标值），并具有有意义的自适应精度；同时也表明，在安全关键应用中，经过良好校准的不确定性模型无法取代正式的覆盖有效性保障。在多样化临床环境中开展前瞻性验证，是迈向实际部署必不可少的下一步。

致谢

本材料基于美国国家科学基金会通过 DGE-2125362 号奖项资助的研究工作。材料中表达的任何意见、研究结果、结论或建议均属于作者观点，并不一定反映美国国家科学基金会的观点。

参考文献

医疗保健研究与质量局（AHRQ）。2024年。《全国数据使用协议——医疗费用与利用项目》。https://hcup-us.ahrq.gov/team/NationwideDUA.jsp。访问日期：2025年7月22日。

医疗保健研究与质量局（AHRQ）。2025。《医疗费用与利用项目（HCUP-US）全国住院患者样本概览》。https://hcup-us.ahrq.gov/nisoverview.jsp。访问日期：2025年7月22日。

Amodei, D.；Olah, C.；Steinhardt, J.；Christiano, P.；Schulman, J.；以及 Man´e, D. 2016年。人工智能安全中的具体问题。*arXiv*预印本 *arXiv:1606.06565*。

Anandalingam, G.；以及 Friesz, T. L. 1992年。层次优化：导论。《运筹学年鉴》，34(1)：1–11。

Angelopoulos, A. N.；Bates, S.；等。2023年。保形预测：简明介绍。《机器学习基础与趋势®》，16(4)：494–591。Arthur, R.；等。2022年。规划医院容量，以挽救新冠疫情及其后续时期的生命。《欧洲运筹学杂志》，295(3)：1202–1214。Austin, P. C. 2024年。在《神经病学》临床研究中使用多层模型和广义估计方程模型处理聚类。《神经病学》，103(8)：e209947。Bafandeh, A.；Bin-Karim, S.；Baheri, A.；以及 Vermillion, C. 2018年。时空变化系统分层控制结构的比较评估及其在空中风能中的应用。《控制工程实践》，74：71–83。Baheri, A. 2023年。基于学习的自主系统的安全验证：一种多保真方法。载于《*AAAI*人工智能会议论文集》，第37卷，15432–15432。Baheri, A. 2025年。多层约束多臂老虎机：一种具有安全保证的分层上置信界方法。《数学》，13(1)：149。Baheri, A.；以及 Shahbazi, M. A. 2025年。跨尺度保形预测：具有分层效率的有限样本覆盖率。《应用数学成果》，26：100589。Barber, R. F.；Candes, E. J.；Ramdas, A.；以及 Tibshirani, R. J. 2023年。超越可交换性的保形预测。《统计年鉴》，51(2)：816–845。Beam, A. L.；以及 Kohane, I. S. 2018年。医疗保健中的大数据与机器学习。《美国医学会杂志》，319(13)：1317–1318。

Berta, P.；Levantesi, S.；以及 Marta, G. 2024年。揭示COVID-19心力衰竭患者的死亡率模式和医院效应：一种新颖的多层逻辑聚类加权建模方法。*arXiv*预印本 *arXiv:2405.11239*。Betancourt, M. 2017年。哈密顿蒙特卡罗的概念性介绍。*arXiv*预印本 *arXiv:1701.02434*。Bollmann, S.；Groll, A.；以及 Havranek, M. M. 2023年。使用医院数据进行自动变量选择时处理聚类问题：不同LASSO方法的比较。*BMC*医学研究方法学，23(1)：280。Breiman, L. 2001年。随机森林。机器学习，45：5–32。Carlin, B. P.；以及 Xia, H. 1999年。使用贝叶斯层级模型评估环境正义：两个案例研究。《暴露分析与环境流行病学杂志》，9(1)。Chipman, H. A.；George, E. I.；以及 McCulloch, R. E. 2010年。BART：贝叶斯加性回归树。《应用统计年鉴》，4(1)：266–298。Chrusciel, J.；Girardon, F.；Roquette, L.；Laplanche, D.；Duclos, A.；以及 Sanchez, S. 2021年。使用非结构化数据预测医院住院时长。*BMC*医学信息学与决策制定，21(1)：351。Ciaroni, S.；Maggiotto, G.；Moro, J.；Calza, S.；Martino, G. P.；Benvenuto, M.；De Santis, M.；以及 Scaglione, F. 2023年。基于机器学习的急诊科延长住院时长入院预测：梯度提升算法分析。《人工智能前沿》，6：1179226。Cressie, N.；Calder, C. A.；Clark, J. S.；Hoef, J. M. V.；以及 Wikle, C. K. 2009年。生态分析中的不确定性处理：分层统计建模的优势与局限。《生态学应用》，19(3)：553–570。Dimick, J. B.；等。2013年。医院业务量与手术结局：越多是否总是越好？《新英格兰医学杂志》，369：1073–1075。Dunn, R.；Wasserman, L.；以及 Ramdas, A. 2023年。双层层级模型的无分布假设预测集。《美国统计学会杂志》，118(544)：2491–2502。Ekmekci, C.；以及 Cetin, M. 2025年。保形生成式贝叶斯成像：一种面向计算成像的不确定性量化框架。*arXiv*预印本 *arXiv:2504.07696*。Fong, E.；以及 Holmes, C. C. 2021年。保形贝叶斯计算。《神经信息处理系统进展》，34：18268–18279。Freeland, K. N.；等。2013年。医院医学中的住院时长决定因素。《美国医学质量杂志》，28(5)：357–365。Gelman, A. 2006年。层级模型中的方差参数先验分布（对 Browne 和 Draper 文章的评论）。Gelman, A.；Carlin, J. B.；Stern, H. S.；Dunson, D. B.；Vehtari, A.；以及 Rubin, D. B. 2013年。贝叶斯数据分析。*CRC*出版社。

Goldstein, H. 2011. 多层统计模型。John Wiley & Sons。Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. 关于现代神经网络的校准。收录于国际机器学习大会，1321–1330。PMLR。Hajjem, A. 2010. 用于聚类数据的混合效应树与森林。*ProQuest LLC*：蒙特利尔大学。Harris, S. L.; et al. 2018. 床位管理与住院时长。*International JournalofHealthCareQuality Assur-ance*，31(8)：1014–1027。Helm, J. E.; et al. 2011. 提升医院效率与患者流转。《生产与运营管理》，20(3)：385–397。Iezzoni, L. I. 2012. 用于衡量医疗保健结果的风险调整。*Health Administration Press*。Kaboli, P. J.; et al. 2012. 住院时长缩短与30天再入院率及死亡率之间的关联。《内科学年鉴》，157(12)：837–845。Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. 用于回归的无分布假设预测推断。《美国统计学会杂志》，113(523)：1094–1111。Linero, A. R. 2018. 用于高维预测和变量选择的贝叶斯回归树。《美国统计学会杂志》，113(522)：626–636。Liu, X.; et al. 2020. 涉及人工智能干预的临床试验报告指南。《柳叶刀·数字健康》，2(10)：e537–e548。McCarthy, K.; et al. 2021. 通过机器学习提升住院时长预测。《医疗保健管理杂志》，66(4)：272–285。Meehinkong, P.; and Ponnoprat, D. 2025. coverforest：使用Python和随机森林进行保形预测。*arXiv*预印本 *arXiv:2501.14570*。Mentch, L.; and Hooker, G. 2016. 通过置信区间和假设检验量化随机森林中的不确定性。《机器学习研究杂志》，17(26)：1–41。Murphy, K. P. 2012. 机器学习：概率视角。麻省理工学院出版社。Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. 你能信任模型的不确定性吗？在数据集偏移下评估预测不确定性。《神经信息处理系统进展》，32。Pellagatti, M.; Masci, C.; Ieva, F.; and Paganoni, A. M. 2021. 广义混合效应随机森林：一种预测大学生辍学的灵活方法。《统计分析与数据挖掘：美国统计学会数据科学杂志》，14(3)：241–257。Podina, L.; Rad, M. T.; and Kohandel, M. 2024. 保形化的物理信息神经网络。*arXiv*预印本 *arXiv:2405.08111*。Pratola, M. T.; Chipman, H. A.; George, E. I.; and McCul- loch, R. E. 2020. 通过乘法回归树实现异方差贝叶斯加性回归树。《计算与图形统计学杂志》，29(2)：405–417。

Principato, G.; Stoltz, G.; Amara-Ouali, Y.; Goude, Y.; Ham- rouche, B.; and Poggi, J.-M. 2024. 层级数据的保形预测。*arXiv*预印本 *arXiv:2411.13479*。Proudlove, N. C.; et al. 2003. 良好的床位管理能否解决急诊科的拥挤问题？《急诊医学杂志》，20(2)：149–155。Rajkomar, A.; Oren, E.; Chen, K.; Dai, A. M.; Hajaj, N.; Hardt, M.; Liu, P. J.; Liu, X.; Marcus, J.; Sun, M.; et al. 2018. 利用电子健康记录实现可扩展且准确的深度学习。*NPJ*数字医学，1(1)：18。Raudenbush, S. W.; and Bryk, A. S. 2002. 层级线性模型：应用与数据分析方法，第1卷。Sage。Sela, R. J.; and Simonoff, J. S. 2012. RE-EM树：一种针对纵向和聚类数据的数据挖掘方法。机器学习，86：169–207。Sendak, M. P.; Balu, S.; and Schulman, K. A. 2020. 医学中的机器学习：实用方法。*Gastroen-terology*，158(2)：2070–2084。Shafer, G.; and Vovk, V. 2008. 保形预测教程。《机器学习研究杂志》，9(3)。Shahbazi, M. A.; and Azadeh-Fard, N. 2025. 层级数据建模：对统计、基于树和神经网络方法的系统比较。《机器学习及其应用》，100688。Snijders, T. A.; and Bosker, R. 2011. 多层次分析：基础与高级多层次建模导论。Stanton, S.; Maddox, W.; and Wilson, A. G. 2023. 基于保形预测集的贝叶斯优化。收录于人工智能与统计国际会议，959–986。PMLR。Stone, K.; Zwiggelaar, R.; Jones, P.; and Mac Parthal´ain, N. 2022. 住院时长预测的系统综述：迈向统一框架。《*PLOS*数字健康》，1(4)：e0000017。Tan, Y. V.; and Roy, J. 2019. 贝叶斯加性回归树与通用BART模型。《医学统计》，38(25)：5048–5069。Tibshirani, R. J.; Foygel Barber, R.; Candes, E.; and Ram- das, A. 2019. 协变量偏移下的保形预测。《神经信息处理系统进展》，32。Tyralis, H.; and Papacharalampous, G. 2024. 机器学习预测不确定性估计综述。人工智能评论，57(4)：94。Varshney, K. R. 2016. 构建机器学习安全性。收录于*2016*年信息论与应用研讨会（*ITA*），1–5。IEEE。Vehtari, A.; Gelman, A.; and Gabry, J. 2017. 使用留一法交叉验证和WAIC进行实用的贝叶斯模型评估。《统计与计算》，27：1413–1432。Vovk, V.; Nouretdinov, I.; and Gammerman, A. 2009. 用于保形预测的线性时间高效非符合性度量。《数学与人工智能年鉴》，56(1)：83–99。Wager, S.; and Athey, S. 2018年. 使用随机森林估计和推断异质性处理效应. 美国统计学会杂志, 113(523): 1228–1242. Werner, E.; Clark, J. N.; Hepburn, A.; Bhamber, R. S.; Ambler, M.; Bourdeaux, C. P.; McWilliams, C. J.; and Santos-Rodriguez, R. 2023. 用于患者分型和风险预测的可解释层次聚类. 实验生物学与医学, 248(24): 2547–2559. Wong, K. C.; and Wonham, W. M. 1996. 离散事件系统的层次控制. 离散事件动态系统, 6(3): 241–273. Young, W. J. 1973. 住院时长预测.卫生服务研究, 8(4): 287–295. Zhang, Y.; Park, S.; and Simeone, O. 2024年. 通过在线保形预测实现具有形式化安全保证的贝叶斯优化. *IEEE*信号处理专题期刊。