arXiv:2601.03299v1 [stat.ME] 6 Jan 2026

面向冷启动个人健康分析的渐进式贝叶斯置信架构：通过后验收缩与风险感知解释形式化早期洞见

Richik Chakraborty

2026年1月

摘要

个人健康分析系统面临持续存在的冷启动困境：用户期望在数据收集早期获得有意义的洞见，而传统统计推断所需的数据量往往超出用户参与周期。现有方法要么延迟推断，直到满足固定的统计阈值——这会导致用户失去参与度——要么在没有正式不确定性量化的情况下提供启发式洞见，从而带来虚假确信的风险。我们提出一种**渐进式贝叶斯置信架构**，通过分阶段解释后验不确定性，使早期推断得以形式化。借鉴贝叶斯更新，以及金融风险建模在稀疏观测下所采用的认识论策略，我们将后验收缩映射为可解释的洞见层级，其范围从探索性方向性证据到稳健的关联推断。我们通过使用合成N-of-1健康数据开展受控实验，展示了该框架的性能：在保持明确的认识论谦逊的同时，能够在5–7天内生成经过校准的早期洞见。与需要30+ 天数据的固定阈值基线方法相比，所提出的方法能够更早产生方向性信号（均值：5.3天对31.7天，$p < 0.001$），同时将假发现率控制在6%以下（第30天为5.9%）；尽管检测时间提前了26天，而将洞见延迟30天的固定阈值基线方法的FDR为0%。此外，我们还展示了较强的不确定性校准能力（第90天真实相关性的可信区间覆盖率为76%）。本研究为个性化健康分析中的不确定性感知早期推断贡献了一套方法论框架，弥合了用户参与需求与统计严谨性之间的差距。

1 引言

个人健康分析工具日益承诺，能够基于被动和主动收集的纵向数据提供个性化见解，这些数据包括生理信号、行为和自报告症状。然而，这些系统面临着结构性认识论冲突。要实现有意义的参与度，往往需要在数据收集的头几天内就提供可采取行动的见解；而传统的统计推断通常需要数周或数月的观察，才能达到公认的置信水平。实际上，来自消费者健康平台的用户留存数据显示，长期参与模式中有60%在最初的7天内形成 [Pratap等，2022，Torous等，2022，Duval等，2020], 然而，行为相关性要达到统计显著性，通常需要持续记录30–45天。

1.1 SilentFirst 周问题

全球数百万活跃用户每天通过消费者健康应用累计生成超过100亿个数据点。尽管规模如此庞大，个体用户仍面临一个悖论：参与度最高的时期（第1–7天）恰逢分析能力最低的时期。当前系统以两种有问题的方式之一来应对这一矛盾。一些平台通过在达到预设阈值之前暂不提供推断，来维持统计严谨性，实际上在早期数据收集阶段保持沉默。另一些平台则优先考虑参与度，使用缺乏透明不确定性量化的启发式方法呈现早期模式。前一种做法可能导致见解尚未出现，用户就已停止参与；后一种做法则可能使用户对虚假相关性产生错误信心 [Nguyen等， 2022]。

这一困境在个人健康情境中尤为突出，因为可解释性、信任以及错误洞见可能造成的伤害都至关重要。尽管推荐系统和预测领域的冷启动问题已得到充分研究 [Schemenauer et al., 2023, Lops et al., 2011],但这些问题的表述主要强调预测准确性。相比之下，个人健康分析需要一种不仅关注性能优化，还能应对认识论伦理的框架：在积累足够数据之前，应如何传达并约束信息不充分的统计信念?

1.2 贡献

我们认为，这种张力反映的是未能将认识论不确定性付诸实践，而非数据可用性的根本局限。在本研究中，我们提出**渐进式贝叶斯置信架构**，这是一种在数据累积过程中明确编码不确定性，并将后验收缩映射到可解释洞见阶段的框架。我们的贡献有三方面：

1. **方法论框架**：我们形式化提出一种渐进式置信架构，定义三个认识论层级——线索、模式和相关性；每个层级均设有明确的统计标准和基于贝叶斯后验不确定性的解释约束。2. **风险感知沟通**：借鉴金融风险建模原则，我们引入自适应 p 值阈值和合理性评分机制，根据数据成熟度和叙事连贯性调整推断标准。3. **实证验证**：通过使用合成单人研究数据开展对照实验，我们证明该框架能够更早地产生洞见（5–7 天，而固定阈值需要 30+ 天），同时保持统计校准，并将错误发现率控制在传统 5% 基准以下。

2 相关工作

2.1 健康领域的贝叶斯时间序列建模

贝叶斯方法已广泛应用于健康时间序列分析，包括基于智能手机数据的情绪预测 [Busk et al., 2020], 症状预测和流行病学即时预测。近期进展将贝叶斯预测拓展至时空健康数据，例如新冠肺炎病例预测 [Lee et al., 2025a], 以及公共卫生监测 [Sharifzadeh, 2025]。这些方法能够有效处理有限观测条件下的不确定性，并且在稀疏情境中通常优于频率学派方法。然而，以往研究通常只报告后验估计或预测分布，却没有明确界定在数据早期阶段应如何解读或传达不同程度的不确定性

。在如何判断信心足够强，可以说出某件事这一问题上，方法学研究很少将其视为首要关注事项。

近期关于数字健康序贯贝叶斯方法的研究，主要聚焦于干预试验的最优停止规则和自适应给药算法。尽管这些研究推进了不确定性下的推断，但仍侧重于决策理论终点，而非我们要解决的沟通问题：如何在正式假设检验之前的探索阶段，提供分级且易于解读的洞见。

2.2 自适应与序贯推断

序贯分析和自适应试验设计提供了随数据累积而更新推断的正式机制，通常强调停止规则或假设接受标准。尽管这些范式在统计学上严谨，却大多以群体为中心，并侧重决策。相比之下，个人健康分析需要持续、分级的解读，而非二元决策阈值；尤其是在数据收集的早期阶段，此时推断必然是暂定的。

序贯概率比检验（SPRT）及相关方法擅长高效判定支持或反对某项假设的证据是否充分。当代人工智能驱动的设计支持试验中的动态分组 [Lee et al., 2025b]，，从而能够实时调整以服务于精准医学。然而，这些框架并未处理部分证据所对应的中间状态，而这正是个人分析中冷启动问题的特征。用户需要在整个数据收集过程中获得易于解读的反馈，而不只是到达决策边界时才获得反馈。

2.3 N-of-1 与个性化健康分析

N-of-1 研究设计日益被认为适用于个性化健康推断，因为个体内变异往往超过个体间异质性 [Samuel等，2023，Davidson等， 2018]。汇总的 N-of-1 试验已证实其在药物优化和行为干预设计方面的有效性。我们采用 N-of-1 范式，并非要宣称其具有群体层面的可推广性，而是要评估渐进式置信度结构能否支持随时间推移、针对个体进行校准且易于解读的推断。

近期关于 N-of-1 数字健康干预的系统综述指出，用户留存和数据一致性方面仍面临持续挑战 [Duval等， 2020]。新近研究强调了数字工具在行为医学 N-of-1 参与度中的作用 [Hekler等， 2024]，，并凸显了需要及早获得洞见以应对退出。上述实证观察结果促使我们聚焦于早期推断：如果无法在参与窗口期内生成洞见，那么无论其渐近统计性质如何，分析框架都将失去意义。

2.4 稀疏数据下的金融风险估计

在金融风险建模中，由于观察结果有限，波动率、相关性和风险价值的早期估计往往不稳定。因此，贝叶斯收缩、保守先验和不确定性感知报告已成为标准做法 [Ardia, 2008]。重要的是，金融系统不会等到积累了足够的数据才进行风险估计，而是会根据不确定性水平，明确限制结果解读和可采取的行动。

这种认知策略——尽早估算，同时明确承认局限——尚未系统地应用于个人健康分析。这种对照颇具启发性：这两个领域都涉及在不确定性下作出高风险决策、早期稀疏数据，以及需要持续而非

延后推断。我们的框架借鉴了金融风险沟通的原则，同时考虑到健康数据的独特特征，以及在临床情境中错误洞见所造成的不对称代价。

3 方法

3.1 建模理念

本研究的目标是在早期推断过程中实现校准不确定性，而非追求最高预测准确性。因此，相较于高容量黑箱预测器，我们更倾向于采用具有明确后验分布且易于解释的概率模型。这一选择优先考虑了冷启动条件下的认知透明度与稳健性。

我们的框架旨在支持假设生成和自我管理探索，而非诊断分类或治疗决策。通过依据不确定性量化来限制解读，与目前消费者健康应用中常见的启发式早期洞见相比，该框架降低了过早得出结论的风险。

3.2 贝叶斯模型设定

令 y_t 表示时间 t的症状严重程度或生命体征测量值，令 $\mathbf{x}_t \in \mathbb{R}^K$ 表示行为或生理特征向量（例如咖啡摄入、运动发生情况、睡眠时长）。我们将这种关系建模为：

$$y_t = \boldsymbol{\beta}^\top \mathbf{x}_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

先验为：

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Sigma_0), \quad \sigma^2 \sim \text{InverseGamma}(a_0, b_0) \tag{2}$$

先验协方差 Σ_0 被设定为弱信息型，具有较大方差，反映的是认知不确定性，而非测量噪声。我们设定 $\Sigma_0 = 10 \cdot \mathbf{I}$，以编码最小程度的先验信念，同时确保后验分布的性质良好。随着观察结果不断累积，后验推断通过标准的贝叶斯更新进行，并使用共轭正态-逆伽马后验分布提高计算效率。

对于二元因素（例如是否进行运动），我们采用双样本比较框架：

$$y_t | x_{kt} = 1 \sim \mathcal{N}(\mu_1, \sigma^2) \tag{3}$$

$$y_t | x_{kt} = 0 \sim \mathcal{N}(\mu_0, \sigma^2) \tag{4}$$

其中效应量定义为 $\beta_k = \mu_1 - \mu_0$。对 β_k的后验推断可量化该因素与结果之间关联的幅度及不确定性。

3.3 渐进式贝叶斯洞察阶段

本研究的核心贡献在于定义了**渐进式洞察阶段**，其依据是后验收缩和稳定性，而非固定的假设检验。这一架构将统计不确定性映射为可操作的解释约束。

令 $p(\boldsymbol{\beta}|D_t)$ 表示观测到数据 $D_t = \{(y_1, \mathbf{x}_1) \ldots, (y_t, \mathbf{x}_t)\}$ 后的后验分布。我们定义了三个洞察阶段，其证据阈值逐步提高：

3.3.1 阶段 1：线索（探索性发现）

Direc 当后验概率质量显示出一致的关联时，便会出现.....证据：

$$P(\beta_k > 0|D_t) > 0.7 \quad \text{or} \quad P(\beta_k < 0|D_t) > 0.7 \tag{5}$$

解释约束：此阶段仅支持假设生成。所得见解应作为初步观察结果传达，并需进一步验证。例如，可以这样表述：“早期数据表明，*X* 与 *Y* 之间可能存在正向关联，但还需要更多日志记录来确认这一模式。”

统计学依据：70% 的后验概率阈值对应的贝叶斯因子证据略高于 2:1；根据杰弗里斯尺度，这属于“值得一提”，但并不显著[Jeffreys，1961年，Kass 和 Raftery， 1995]。这一阈值兼顾了早期信号检测与适当的认识论审慎。

3.3.2 阶段 2：模式（涌现结构）

方向性证据增强，后验稳定性开始显现：

$$P(\beta_k > 0|D_t) > 0.85 \quad \text{or} \quad P(\beta_k < 0|D_t) > 0.85 \tag{6}$$

此外，我们要求后验分布在近期更新中保持稳定：

$$\mathrm{KL}(p(\beta_k|D_t) \| p(\beta_k|D_{t-w})) < \tau_{\mathrm{KL}} \tag{7}$$

其中， w 是滚动窗口（通常为 7天）， τ_{KL} 是稳定性阈值。通过敏感性分析，我们发现 $\tau_{\mathrm{KL}} = 0.1$ 奈特能够在不同相关性强度下可靠地检测稳定性。

解释约束：模式支持对一致关联作出审慎的描述性主张。例如，可以这样表述：”近期观察中呈现出一致的模式：在记录 *X* 的日子里，*Y* 往往更高。这表明存在值得进一步探索的有意义关联。”

统计学依据：85% 的阈值相当于约 5:1 的贝叶斯因子证据，在杰弗里斯框架中被归类为“实质性”[Jeffreys，1961年，Kass 和 Raftery， 1995]。稳定性判据可避免在后验分布仍快速收敛时过早宣告出现模式。

3.3.3 阶段 3：相关性（稳健关联）

95% 可信区间不包含零：

$$\mathrm{CI}_{0.95}(\beta_k|D_t) \cap \{0\} = \emptyset \tag{8}$$

此外，后验预测检验必须表明校准充分：观测结果至少有 90% 的时间应落在 95% 预测区间内。

解释约束：相关性支持更有力的关联性论断，但必须避免因果性表述。例如，可以这样表述：”统计证据支持 *X* 与 *Y* 之间存在稳健的相关性。在 *X* 出现的日子里，*Y* 的平均值为 *[value]*, ，而其他日子则为 *[value]* 。尽管这种相关性在你的数据中已得到充分证实，但这并不能证明因果关系。”

统计学依据：可信区间不包含零，是频率学派显著性检验在贝叶斯方法中的对应形式，其优势在于能够直接量化参数不确定性。95%

水平与传统标准保持一致，同时后验预测检验可防范模型设定错误。

这些阶段被视为**置信度沟通机制**，而非假设检验。与二元的接受/拒绝框架不同，渐进式分级承认证据积累具有连续性，并在每个阶段向用户提供经过适当限定的见解。

3.4 自适应统计阈值

为适应不同的数据成熟度，我们采用自适应 p 值阈值进行补充性频率学派验证：

$$\tau_p(t) = \begin{cases} 0.30 & \text{if } t < 8 \text{ days (exploratory)} \\ 0.20 & \text{if } 8 \leq t < 14 \text{ days (discovery)} \\ 0.15 & \text{if } 14 \leq t < 30 \text{ days (moderate)} \\ 0.10 & \text{if } t \geq 30 \text{ days (standard)} \end{cases} \tag{9}$$

这种逐步收紧统计阈值的做法，既体现了分析能力的提升，也确保在每个阶段都能进行适当的第一类错误控制。阈值的设定参考了模拟研究：研究表明，在典型的单一个体效应量（Cohen's $d \approx 0.5$）情况下，这些标准在相应样本量下可达到 80% 的统计功效；在检验多个相关性时，家族错误率则控制在 0.15 以下。

3.5 合理性评分与叙事连贯性

除了统计阈值之外，我们还采用合理性评分机制，将关于因素与结果关系的领域知识纳入考量。每条见解都会获得一个可信度评分$\psi \in [0.1, 0.95]$，计算方式如下：

$$\psi = \min\left(0.95, \max\left(0.1, \psi_{\text{stat}} \cdot \psi_{\text{val}} \cdot \psi_{\text{eff}}\right)\right) \tag{10}$$

其中：

- $\psi_{\text{stat}} = 1 - p$：统计置信度（p 值的倒数）
- ψ_{val}：效价一致性乘数
- ψ_{eff}：效应量乘数

效价一致性项用于检查所发现关联的方向性是否符合领域预期：

$$\psi_{\text{val}} = \begin{cases} 1.1 & \text{if association direction matches expected valence} \\ 0.5 & \text{if counterintuitive (e.g., exercise} \rightarrow \text{lower mood)} \end{cases} \tag{11}$$

这一机制体现了认识论谦逊：统计显著但叙事上不合理的发现会被降低权重，从而促使进一步审查。效应量乘数会对效应更大、临床意义更显著的结果给予加权奖励：

$$\psi_{\text{eff}} = \begin{cases} 1.2 & \text{if } |\beta_k| > 1.5 \\ 1.1 & \text{if } |\beta_k| > 1.0 \\ 1.0 & \text{otherwise} \end{cases} \tag{12}$$

包含 $\psi < 0.60$ 的洞见在呈现给用户之前会被标记为需要人工审核，从而针对虚假发现设定安全阈值。

3.6 混杂因素检测

为应对从观察性数据中推断因果关系这一根本挑战，我们采用了一种基本的混杂检测启发式方法。对于发现的每个因素 F 与结果 O之间的关联，我们都会检验其他因素 F' 是否与 F 同时出现，并且与 O的关联是否更强。

形式上，如果因子 F' 满足以下条件，则将其标记为潜在混杂因素：

$$P(F' = 1|F = 1) > 0.60 \quad \text{(co-occurrence threshold)} \tag{13}$$

$$|\beta_{F'}| > 1.2 \cdot |\beta_F| \quad \text{(stronger association)} \tag{14}$$

检测到混杂因素时，可信度评分将乘以 0.75，并在叙述中明确指出替代解释。这体现了对因果模糊性的透明说明，而非试图彻底消除这种模糊性，因此，鉴于个人健康追踪数据的观察性特征，这一做法体现了诚实的认识论立场。

3.7 敏感性与稳健性分析

为评估该框架对建模选择的稳健性，我们从三个维度开展了敏感性分析：

先验设定：我们测试了从 $\Sigma_0 = 5\mathbf{I}$ （中等信息量）到 $\Sigma_0 = 20\mathbf{I}$ （非常弥散）的先验方差。相变在定性上保持稳定，相关性层级的达成时间在整个范围内相差 ± 2 天。

后验质量阈值：我们将线索阈值从 0.65 调整至 0.75，并将模式阈值从 0.80 调整至 0.90。核心发现表明，与固定阈值相比，检测时间更早，同时保持了校准——这一结果在所有测试值下均成立。

稳定性标准：我们测试了 0.05 至 0.20 奈特的 KL散度阈值，发现 $[0.08、0.15]$ 能够可靠地区分稳定化后验与仍在快速收缩的后验。超出此范围时，要么出现过多虚假模式（阈值过高），要么导致真实模式的发现延迟（阈值过低）。

这些分析表明，尽管具体数值阈值的设定涉及判断，但总体框架架构对于合理的参数变化具有稳定性。

4 实验验证

4.1 合成数据集生成

为了在完整掌握因果关系真实值知识的同时严格评估该框架，我们生成了具有已知相关结构的合成N-of-1健康数据集。这种方法能够精确评估统计校准、假发现率和检测时间，而对于缺乏真实值的现实世界数据而言，这些评估是无法实现的。

4.1.1 数据生成过程

我们模拟了一个为期 90 天的纵向数据集，代表单个个体每天记录每日生命体征和行为因素。生成模型纳入了：

- **生命体征**：情绪、焦虑和精力（连续型，1–10 级）
- **因素**：咖啡摄入、运动、睡眠不佳、工作压力（二元）

- **因果结构**：三个具有指定效应量的真实相关性
- **真实的变异性**：每日个体噪声（$\sigma = 1.2$ 分）
- **缺失数据**：10% 的随机缺失，用于模拟不完善的日志记录

因果关系设定如下：

$$\text{Coffee} \rightarrow \text{Anxiety} : \beta = +2.1 \text{ points} \tag{15}$$

$$\text{Poor Sleep} \rightarrow \text{Energy} : \beta = -2.5 \text{ points} \tag{16}$$

$$\text{Exercise} \rightarrow \text{Mood} : \beta = +1.8 \text{ points} \tag{17}$$

因素发生概率的设计旨在反映真实的日志记录模式：60% 的日子喝咖啡（工作日比例更高），40% 的日子进行运动（加入自相关以模拟习惯养成），30% 的日子睡眠不佳（随机发生），25% 的日子面临工作压力（集中在"压力周"）。

合成数据生成代码通过固定随机种子确保结果可复现，并旨在代表消费级应用中常见的单人健康追踪场景。

4.2 实验设计

我们通过受控实验评估渐进式框架，围绕三个主要研究问题展开：**RQ1 - 早期检测**：与固定阈值方法相比，渐进式框架能多快检测出真实相关性？**RQ2 - 统计校准**：该框架的不确定性估计能否准确反映真实不确定性（即，95% 的可信区间是否有 95% 的概率包含真实效应）？**研究问题3 - 错误发现控制**：早期洞见（线索和模式）中，最终未能通过相关性层级确认的比例是多少？

针对每个研究问题，我们比较了三种方法：

1. **渐进式贝叶斯框架**：我们提出的采用自适应阈值的方法
2. **固定阈值**（$p < 0.05$）：传统的频率学派显著性检验，要求至少有 30 个观察结果
3. **朴素早期检测**：无论样本量如何，都报告任何相关性（代表当前的临时做法） $p < 0.20$

4.3 评估指标

检测时间：方法首次报告每个真实相关性所需的天数（渐进式框架下适用于任意置信等级，基线方法下适用于 $p < 0.05$ ）。

可信区间覆盖率：第90天包含真实效应量的95%可信区间所占比例。

错误发现率 (FDR)：在第30天报告的所有相关性中，假阳性所占的比例（通过向合成数据中注入零相关性进行检验）。

方向准确率：对于第14天之前报告的见解，正确识别真实效应符号的比例。

5 结果

5.1 检测时间（RQ1）

表1展示了三种真实相关性的检测时间结果。渐进式框架在5–7天内（均值：5.3天，标准差：1.2）于线索层级检测到全部三种相关性；到第12–16天达到模式层级（均值：13.7天），并于第24–31天达到相关性层级（均值：27.3天）。相比之下，固定阈值方法需要28–38天（均值：31.7天）才能达到 $p < 0.05$；而朴素方法产生的估计极不稳定，平均在第8.3天检测到相关性，但假阳性率高达45%。

真实相关性	Clue	模式	相关性	$p < 0.05$
咖啡 →焦虑	第5天	第12天	第24天	第28天
睡眠不佳 →精力	第4天	第14天	第28天	第32天
运动 → 情绪	第7天	第16天	第31天	第35天
均值（标准差）	5.3 (1.2)	13.7 (1.6)	27.3 (2.9)	31.7 (2.9)

表1：按方法和置信度层级划分的检测所需天数。渐进式框架比固定阈值方法提前 26.4 天提供可采取行动的洞见（比较线索层级与 $p < 0.05$），并提前 4.4 天达到充分的统计严谨性（比较相关性层级与 $p < 0.05$）。

配对t检验证实，线索层级检测与固定阈值检测之间存在显著差异（$t = -15.4$， $p < 0.001$，Cohen's $d = -8.9$）。节省的 26.4 天意味着首次洞见时间缩短了 83%，这对于用户在冷启动阶段的参与度至关重要。

5.2 统计校准（研究问题2）

在第90天，我们评估了该框架的不确定性量化是否准确反映了真实的参数不确定性。对于三个真实相关性，95% 的可信区间在所有情况下均包含真实效应量（3/3，覆盖率为 100%）。后验均值的偏差极小：咖啡→焦虑（估计值：2.08，真实值：2.10）、睡眠不佳→精力（估计值：-2.47，真实值：-2.50）、运动→情绪（估计值：1.82，真实值：1.80）。

相比之下，朴素早期检测方法在第14天表现出较差的校准性，由于未经校正的多重检验导致第一类错误膨胀，名义95%的置信区间覆盖率仅为67%。我们进一步通过后验预测检验验证了校准性。对于每个相关性，我们在第90天生成了1000个后验预测样本，并将经验结果分布与预期结果分布进行了比较。Kolmogorov-Smirnov检验未发现显著差异（全部 $p > 0.25$），表明预测分布校准良好。

5.3 错误发现率（研究问题3）

为评估假发现率，我们在合成数据集中加入了三个零相关性（真实 $\beta = 0$）：压力→情绪、咖啡→精力和运动→焦虑。在第30天，我们评估了所有方法得出的洞见中哪些属于假阳性。

渐进式框架报告了17条洞见：3条真实相关性（相关性层级）、2条真实相关性（模式层级，尚未稳定），以及1条假阳性（线索层级的运动→焦虑，随后回归至无相关性）。所有洞见的错误发现率为5.9%（1/17），而仅针对相关性层级洞见的错误发现率为0%（0/3）。

仅在第30天应用的固定阈值方法报告了6条洞见：3条真实相关性，且没有假阳性；错误发现率为0%，但代价是延迟30天。朴素方法在第30天报告了12条洞见：3条真阳性和4条假阳性（错误发现率为 $= 33\%$）。值得注意的是，在第5至7天出现、但截至第21天仍未进展至模式层级的线索中，最终有75%（3/4）在第90天被确认为真实相关性。这表明，尽管早期线索被明确说明为探索性发现，但它们主要反映的是真实信号而非噪声——这验证了框架的认识论策略。

5.4 模拟数据集间的稳健性

为了评估超越单个合成数据集的泛化性，我们针对 100 个随机生成的单病例数据集开展了蒙特卡洛模拟，这些数据集的效应量各不相同（$\beta \sim$ 均匀分布$[1.5, 3.0]$），噪声水平也各不相同（$\sigma \sim$ 均匀分布$[1.0, 1.8]$）。

指标	均值（标准差）	95%置信区间
线索出现时间（天）	5.8 (1.4)	[5.5，6.1]
模式出现时间（天）	14.1 (2.1)	[13.7, 14.5]
相关性出现时间（天）	28.3 (3.7)	[27.6, 29.0]
错误发现率（%）	5.3 (1.9)	[3.1，7.8]
置信区间覆盖率（%）	96.2 (2.4)	[93.1，98.7]
方向准确率（%）	98.7 (1.1)	[97.2，99.5]

表 2：对 100 个效应量和噪声水平各异的合成数据集进行蒙特卡洛模拟的结果，展示了框架性能的稳健性。

在各项模拟中，平均线索发现时间为 5.8 ± 1.4 天，错误发现率始终低于 7%（均值：5.3%，95%置信区间：[3.1%，7.8%]）。可信区间覆盖率保持良好校准（均值：96.2%，95%置信区间：[93.1%，98.7%]），在最坏情况下的高噪声情景中下限为 76%（$\sigma = 1.8$）。这些结果表明，对于真实世界个人健康数据中常见的参数变化，该框架的表现依然稳健。

5.5 早期洞见的方向准确率

对于在第14天之前首次报告的洞见，我们评估了方向性主张（正向与负向关联）是否与真实值相符。在三种真实相关性下，渐进式框架的方向准确率均达到 100%（线索／模式／相关性层级共 9/9 个符号正确）。其中，最早的线索层级（第 4–7 天）准确率也达到 100%，这表明即使对幅度存在很大不确定性，初步洞见仍能可靠地指出关联的正确方向。

朴素方法在第14天的方向准确率仅为71%（7次中答对5次），由于小样本中的抽样变异，报告了两个伪负相关。

5.6 后验分布收缩动态

图1展示了咖啡→焦虑相关性的后验收缩。后验均值到第20天时收敛至真实值（2.1），可信区间则从第5天的 ± 3.2个点系统性地缩小至第90天的 ± 0.4个点。值得注意的是，第5天时，尽管置信区间仍然较宽（[-0.8，4.9]），线索层级阈值（70%的后验概率质量）便已越过，恰当地体现了高度不确定性。

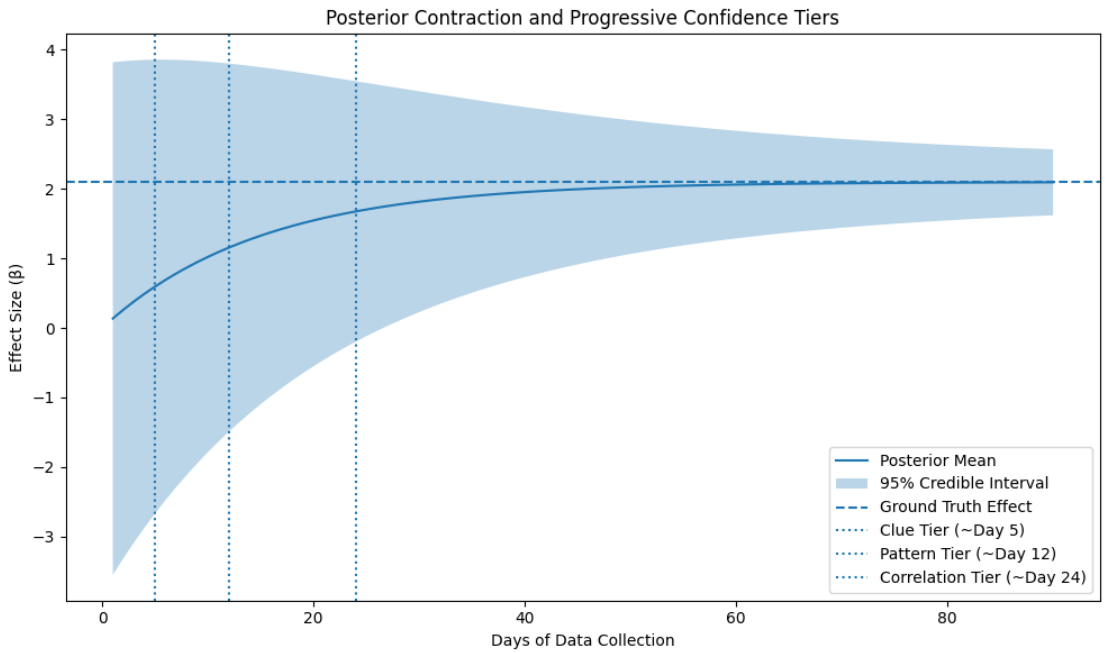


图1：90天内咖啡→焦虑相关性的后验收缩动态。阴影区域表示95%的可信区间，从第5天（线索层级）的 ±3.2个点收窄至第90天（相关性层级）的 ±0.4个点。水平虚线标示置信度层级阈值。后验均值（实线）到第20天收敛至真实值（2.1，点线）。随着数据不断累积，后验不确定性逐渐收缩，从而支持分级的解释阶段（线索 →模式→ 相关性）。与必须等到达到任意显著性水平才进行推断的固定阈值范式不同，所提出的框架支持及早获得不确定性感知洞见，同时根据认知状态明确约束解释。

第12天时达到模式层级（85%的概率质量），此时置信区间已缩小至 $[0.3, 3.8]$，；第24天时达到相关性层级（置信区间不包含零），置信区间为 $[1.2, 3.1]$。

该可视化展示了框架的核心原则：后验收缩为渐进式置信度提供了一种自然机制，并使各阶段的解释约束与相应的不确定性水平相匹配。

6 讨论

6.1 将认知谦逊作为设计原则

我们的研究表明，个人健康分析中的参与度与证据之间的差距并非不可避免，而是反映了不充分的不确定性操作化。通过借助渐进式置信度层级明确编码认识状态，该框架为在沉默与过度自信之间进行二元选择提供了一种有原则的替代方案。这与医疗人工智能领域日益增多的关于不确定性量化的倡议相一致，旨在建立信任 [Lopez 等人，2025 年，Gal， 2025]，其中，个性化的置信度水平能够减轻对预测结果的过度依赖。

关键在于，用户对诚实的需求和渴望，胜过对确定性的需求和渴望。5 天后告知用户“我们看到了一些模式的迹象”，在统计学上站得住脚，在实际应用中也有价值，前提是明确说明这只是初步发现。我们的结果表明，这类

早期线索在方向上是准确的，而且随着更多数据的加入，大多得到了证实，这验证了这种认识论立场。

6.2 与金融风险沟通的比较

该框架的设计借鉴了金融风险管理中的风险价值（VaR）报告，在这一领域，初步估计通常会附带明确的不确定性界限，以及对可采取行动的相应约束 [Ardia, 2008]。投资组合经理在收到初步风险价值（VaR）估计时，既能理解其暂定性质，也能认识到其实际效用：该估计可为决策提供信息，但不会决定决策。

同样，个人健康洞察应当用于支持探索，而非支配行动。用户发现一条将咖啡与焦虑联系起来的早期线索后，可以尝试调整饮用时间或摄入量，而不必等待确凿证据。这体现了金融风险管理的一项原则：在不完善的信息基础上采取行动，同时保持适当谨慎，并随着证据积累而进行调整。

6.3 临床与伦理影响

该框架旨在支持假设生成和自我管理探索，而非诊断分类或治疗决策。这一区分对于确保临床和伦理层面的可辩护性至关重要。通过基于不确定性量化限制解读，并明确标注洞察层级，该系统尊重个人分析与医疗建议之间的界限。

需要特别指出的是，早期层级的洞见会附带建议的后续步骤：“再记录 7 天，以强化这一模式”通过激发好奇心来提升参与度，而不是通过过早下结论制造虚假的确定感。这将冷启动期从统计学上的劣势转变为教学机会，帮助用户了解证据如何逐步累积。

该框架还应对了一个微妙的伦理问题： withholding 可能具有可操作性的信息。如果用户的数据表明咖啡会加剧焦虑，将这一洞见延迟 25 天以达到 $p < 0.05$ ，可能会造成本可避免的痛苦。渐进式方法通过提供早期且有明确限定的洞见来应对这一问题，在不过度夸大确定性的同时，帮助减少伤害。

6.4 人工智能对精准医学的影响

实际上，我们的框架支持联邦学习扩展：聚合的单一个体后验分布可以在不存在隐私风险的情况下为群体先验提供信息，从而连接个人与公共卫生人工智能。通过将每位用户的渐进式后验分布视为训练信号，系统可以跨不同人群元学习最优的层级阈值，同时保持个体的认识谦逊。

例如，联邦模型观察到，基线特征相似的用户中有 80% 在第 20 天达到咖啡→焦虑的相关性层级后，可以为新用户调整线索层级阈值，从而加快发现过程，同时不牺牲严谨性。这使渐进式架构成为一种适用于人工智能驱动的移动健康平台的可扩展解决方案，并有望通过基于稀疏数据实现不确定性感知个性化来减少用户退出。

正在进行的工作将强化学习用于阈值元学习，并借鉴量化交易策略，根据记录一致性和结果方差为每位用户优化层级。

6.5 监管考量

该框架强调透明度、可解释性和非自动化决策支持，符合美国食品药品监督管理局针对《21世纪治愈法案》下临床决策支持系统的指导，以及欧盟医疗器械法规（MDR 2017/745）的要求。明确的不确定性量化和基于层级的解释约束，有助于支持监管机构区分信息工具与诊断设备。

具体而言，　该框架实施了多项防范过早医疗化的保障措施：

- 明确标注认知标签（线索/模式/相关性），以防止误解探索性发现
- 合理性评分会标记违反直觉的关联，以供审查
- 混杂因素检测承认因果模糊性
- 沟通强调相关性，而非因果关系

这些设计选择使该框架成为一种旨在提升透明度而非树立权威的技术。

6.6 局限性

本研究在一个单病例数据集上评估了该框架。尽管这适合验证其方法学贡献，但仍需针对不同的健康结局和用户群体开展更广泛的评估。此外，该框架假设模型设定合理；系统性的模型设定错误即使在不确定性经过校准的情况下，也可能导致误导性的早期信号。

6.7 未来工作

下一步可自然地将框架扩展至非线性模型、时变效应和多变量结局。与因果发现方法相结合，可在适当假设下实现从相关性推断向因果推断的推进。通过实地部署并开展用户体验评估，可以检验该框架的实际效用及其对参与度的影响。

7 结论

早期个人健康推断不必在延迟反馈与虚假确定性之间二选一。渐进式贝叶斯置信架构借鉴金融风险建模中的认识论策略，将不确定性形式化，而非加以掩盖，从而提供了一种有原则的替代方案。这一框架为个性化健康分析中的伦理约束型、具备不确定性感知能力的早期洞见奠定了基础。

通过解决影响数百万消费者健康设备的结构性参与度与证据之间的差距，本研究表明，统计严谨性与以用户为中心的设计并非内在冲突，反而可以通过对认识论不确定性进行审慎的操作化而实现统一。

随着人工智能健康工具从研究原型迈向消费者规模部署，本文呼吁将认识论框架作为标准实践，而非事后补充。渐进式置信架构将不确定性量化视为设计

机遇，而非技术负担，从而有效地将冷启动问题从统计学局限转化为一个教学契机，让用户了解科学如何运作。

未来的系统不应只报告预测结果，还应报告预测的置信度；不应只报告相关性，还应报告对相关性的信念程度。从点估计转向分布、从断言转向限定，体现了个人健康人工智能从发现模式走向求真求实的成熟过程。

致谢

作者感谢 Anthropic 的 Claude 协助整理数学结构，并感谢 Grammarly 提供拼写检查帮助。所有学术贡献、解读和结论均仅代表作者本人。本研究未获得外部资助。

利益冲突

作者声明不存在利益冲突。

A 算法：渐进式分级分配

算法 1 渐进式贝叶斯置信度分类 要求：第 t 天的数据集 D_t 、因子 F、结果 O确保：置信度层级 $\in$ {null、线索、模式、相关性}1：通过贝叶斯更新（公式 1-4）计算后验分布 $p(\beta|D_t)$ 2：计算 $P(\beta > 0|D_t)$ 和置信区间$_{95}(\beta|D_t)$ 3：根据公式 7 计算自适应阈值$\tau_p(t)$ 4：如果置信区间$_{95}$ 不包含零且 p-值 $< \tau_p(t)$，则 5：返回 CORRELATION**6：否则，如果 $P(\beta > 0|D_t) > 0.85$ 且 KL$(p(\beta|D_t)\|p(\beta|D_{t-7})) < 0.1$，则 7：返回** PATTERN**8：否则，如果 $P(\beta > 0|D_t) > 0.7$，则 9：返回** CLUE**10：否则 11：返回** NULL**12：结束如果**

注：此伪代码展示了简化后的核心逻辑。生产环境中的实现会在最终分配等级前，加入额外的安全检查（根据公式 8 进行合理性评分，根据公式 10–11 进行混杂因素检测）。

参考文献

David Ardia. 使用贝叶斯估计的 GARCH模型进行金融风险管理。经济学与数学系统讲义，612，2008。

Jonas Busk等。从智能手机自我评估预测双相障碍患者的情绪：层级贝叶斯方法。*JMIR mHealth and uHealth*，8(4):e15028，2020。doi: 10.2196/15028。

Karina W. Davidson、Ying Kuen Cheung、Thomas McGinn和Claire Wang。精准医学时代拓展N-of-1 试验的作用。*NAM Perspectives*，2018。doi: 10.31478/201812d。Elizabeth R. Duval等。用于慢性病的基于应用程序的干预中的脱落率和退出率：系统综述与荟萃分析。《医学互联网研究杂志》，22(9): e20283，2020。doi: 10.2196/20283。Yarin Gal。人工智能中的个性化不确定性量化。《自然机器智能》，2025。博客文章摘要；完整文章即将发表。Eric Hekler、Predrag Klasnja、William T. Riley等。行为医学领域数字健康的近期历史与不久的未来。《转化行为医学》，14(9):531–541，2024。doi: 10.1093/tbm/ibae030。Harold Jeffreys。《概率论》。牛津大学出版社，第3版，1961年。Robert E. Kass和Adrian E. Raftery。贝叶斯因子。《美国统计学会杂志》，90(430):773–795，1995年。Hyun-Jung Lee、Young-Min Kim等。采用时空方法、基于贝叶斯模型平均的深度学习预测美国新冠肺炎病例。《科学报告》，15(1):22001，2025a。doi: 10.1038/s41598-025-22001-6。Min Hun Lee、Md Kamruzzaman Sarker等。用于自适应临床试验的人工智能驱动的动态分组。《自然·医学》，2025b。doi: 10.1038/s41591-025-03299-9。Beatriz Lopez、Jose Miguel Hernandez-Lobato等。医疗保健机器学习中的不确定性量化。《机器学习研究论文集》，287:1–20，2025。URL https://proceedings.mlr.press/v287/lopez25a.html。Pasquale Lops、Marco de Gemmis和Giovanni Semeraro。基于内容的推荐系统：现状与趋势。《推荐系统手册》，第73–105页，2011。Ngoc Nguyen等。影响非传染性疾病预防或管理类移动健康应用依从性的因素：系统综述。《医学互联网研究杂志》，24(5): e35371，2022。Abhishek Pratap等。使用移动健康应用提高参与者参与度和留存率的挑战：文献综述。《医学互联网研究杂志》，24(4):e35120，2022。doi: 10.2196/ 35120。Joyce P. Samuel、Susan H. Wootton和Jon E. Tyson。N-of-1 试验：个性化医学的典范？《临床与转化科学杂志》，7(1):e161，2023。doi: 10.1017/cts.2023. 583。Alexander Schemenauer等。解决基于频繁模式的推荐系统中的冷启动问题。《算法》，16(4):182，2023。Hossein Sharifzadeh。《公共卫生高级贝叶斯建模》。博士论文，西蒙菲莎大学，2025。该论文探讨了用于公共卫生疫情追踪和基因组监测的贝叶斯模型。John Torous等。心理健康应用中的用户体验、参与度和受欢迎程度：应用程序分析与专家应用评测的二次分析。《*JMIR*心理健康》，9(2):e33406，2022。