超越传统监测：充分利用专家知识进行公共卫生预测

Garrik Hoyt[1], Eleanor Bergren[2], Gabrielle String[3,4], and Thomas McAndrew[5]

***Abstract*— 缩小美国公共卫生 workforce 在 2025 年的规模会放大在公共卫生危机期间的潜在风险。来自公共卫生官员的专业判断代表了一个重要的信息来源，这与传统的 surveillance infrastructure 相区别，应该被重视——而不是抛弃。了解专家知识如何在约束条件下发挥作用对于理解减少 capacity 的潜在影响是至关重要的。**

为了探索专家预测能力，2024 年 CSTE 研讨会上的 114 名公共卫生官员生成了 103 个关于住院高峰的预测及 102 条理由说明，以及 114 个关于 2024/25 季节宾夕法尼亚州 H3 与 H1 优势的预测。我们将专家预测与计算模型进行比较，并利用理由说明分析推理模式，采用潜在狄利克雷分配方法。专家更好地预测了 H3 的优势，并将不合理情景的概率赋值低于模型。专家的理由基于历史模式、病原体相互作用、疫苗数据和累积经验。

专家公共卫生知识构成了一种关键的数据来源，应该与传统数据集同等重视。我们建议开发一个国家工具包，系统地收集和分析专家预测和理由，将人类判断视为可量化的数据，并结合监视系统以增强危机应对能力。

I. 介绍

有效的公共卫生决策需要严格的群体水平流行病学和生物统计学培训，遵循基于证据的决策而非轶事推理，并在卫生机构和官员网络中进行协作领导 [1] 。自 2025 年 2 月 14 日起，美国政府通过取消 COVID 时期的拨款、政策调整和重新预算，减少了与公共卫生服务相关的职位 [2], [3] 。美国卫生部长终止了卫生与公众服务部内的 10,000 个职位。免疫实践咨询委员会的所有十七位专家均被撤职 [4] 。国家过敏和传染病研究所的主任被停职 [5] 。持续减少的从业者、科学家和流行病学家与公共卫生相关，被认为可能对美国公共卫生产生巨大的影响 [6], [7] 。这些削减显著减少了公共卫生从业人员、科学家和流行病学家的人力资源，并将资源从长期优先事项（如数据收集）转向即时需求 [2], [8] 。

有效的决策结合了数据、模型，以及可能最重要的是流行病学家、传染病建模专家和公共卫生官员（我们在本研究中称其为专家）的经验 [9]。这种合作专长的重要性在考察成功的国际卫生紧急事件应对措施时变得清晰。在 2010 年海地地震之后，疾病控制与预防中心（CDC）支持了海地卫生部（MSPP），加强了疾病监测系统和由美国总统艾滋病救济紧急计划已经支持的医疗机构中的实验室检测能力 [10]。当九个月后发现霍乱时，美国外国救灾办公室、MSPP、CDC 以及地方公共卫生官员迅速协调了向医院分发霍乱治疗物资的工作，并在流离失所者营地和社区推广使用点用水处理和卫生设施，并建立了一个仍在使用的国家霍乱监测系统 [11], [12]。

国际上的响应如上所述，说明了专家判断如何在不确定性下实现快速决策——官员迅速评估风险模式、预测资源需求并协调干预措施。然而，鉴于减少的人力容量带来的限制，要保持过去危机应对中展示的合作专业知识，将需要创新的方法来优化现有资源，并更深入地理解专家判断在压力下的运作方式。为了最大限度地提高较小团队的有效性，必须考察专家判断和计算建模的互补优势，确定如何系统地结合这两种方法以增强预测准确性和决策质量。

以往的研究调查了专家对未来做出良好预测的能力的优势和劣势 [13] 。过去的研究表明，专家在评估与人口健康状况恶化的因素和机制相关的方面表现出色 [14], [15] 。除了评估因素与健康状况之间的联系外，在面对挑战（例如爆发）时，专家通常比新手做出更好的决策。这种决策能力被称为识别启动的决策。给定一个挑战，专家可以迅速识别过去的类似经验，并从一组表现良好的决策中选择 [16], [17] 。尽管如此，

[1]Department of Computer Science and Engineering, PC Rossin College of Engineering and Applied Sciences, Lehigh University, Bethlehem, PA, USA

[2]Council of State and Territorial Epidemiologists, Atlanta, Georgia, USA

[3]Department of Population Health, College of Health, Lehigh University, Bethlehem, PA, USA

[4]Department of Civil and Environmental Engineering P.C. Rossin College of Engineering and Applied Sciences, Lehigh University, Bethlehem, PA, USA

[5]Department of Biostatistics and Health Data Science, College of Health, Lehigh University, Bethlehem, PA, USA

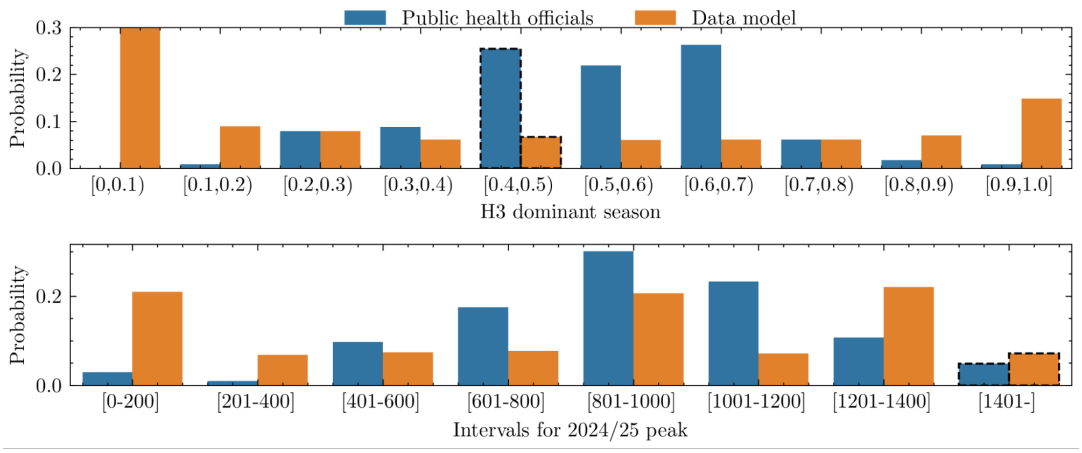

Fig. 1. 由模型生成的预测（橙色）和公共卫生官员群体生成的预测（蓝色），针对 2024/25 流感季节在 PA 相关的两个问题：（上图）确诊的大多数流感病例会被归类为 H3（而非 H1），以及（下图）住院人数的峰值会是多少。两幅图中的柱状高度表示分配给每个潜在未来观察的概率。虚线包围的柱子标识了赛季结束时计算出的真实情况。公共卫生官员对这两个季节性目标进行了预测，这些预测与传统的计算模型相当竞争。

专家并不需要成为主要的决策者，也可以补充统计模型。专家被要求设计数据稀疏模型的概率密度，对未来传染病爆发的具体方面（例如峰值强度、每周病例等）进行预测，并对未来疫苗功效做出长期预测 [18], [19], [20] 。然而，专家判断像所有人一样容易受到群体思维和锚定等偏见的影响 [21] 。专家的预测并不总是如我们所期望的那么准确，可能过于自信 [22] 。重要的是，专家判断在向公众传达信息方面至关重要。

以下，我们提供了一个案例研究，比较专家预测与两个计算模型的能力，以及专家生成合理机制以解释其预测的能力，并建议一个可用于收集专家预测和理由的工具包所需具备的属性，类似于如何收集监控数据。

我们的案例研究探讨了专家在时间限制下如何对季节性流感进行预测。2024 年 11 月 20 日，在州和地方流行病学家关于传染病预测的工作坊上，我们向参加会议的公共卫生官员、流行病学家和传染病建模者提出了两个问题：在宾夕法尼亚州（PA），在 2024/25 季节中，(1) 大多数实验室确认的流感病例会被归类为 H1 还是 H3（H3 通常会导致更严重的症状）？(2) 由于流感而确认住院的峰值人数是多少，原因是什么（例如他们预测的理由）？我们收到了 114 位专家的 217 份答复（原始数据见补充材料）。会议上的专家们被提供了简短的背景信息以帮助他们形成预测，并且只有五分钟时间回答每个问题（问题格式可以在补充材料中找到）。快速的回答时间强调了公共卫生经验的重要性，而不仅仅是对背景数据进行更繁琐的研究。除了定量预测外，我们还使用潜在狄利克雷分配分析专家的理由，以识别他们在推理中的主题。

II. 结果

2025 年 5 月 31 日，2024/25 流感季节结束时，宾夕法尼亚州报告的 H3 型流感病例占总病例的比例为 40.0%，而全美这一比例为 47.3%（参见图 1A）。过去三个季节中，宾夕法尼亚州 H3 型流感病例所占平均百分比为 49%（2021/22 赛季为 84%，2022/23 赛季为 53%，2023/24 赛季为 20%）。专家们（与模型相比）对本季节真实 H3 型流感病例比例的估计概率为 0.25（与模型的 0.06 相比）。此外，专家们（与模型相比）对一个不太可能的情况——低于 20%的 H3 型流感百分比的概率估算为 0.01（与模型的 0.38 相比）。然而，相比于专家们的预测，该模型报告了较大的方差（模型方差为 11%，而专家们为 2%）。

在宾夕法尼亚州，2024/25 季节的入院人数峰值为 4,318 人（见图 1B）。这一入院人数是与过去三个季节相比最高的数值（2021/22 赛季为 200 人；2022/23 赛季为 1,299 人；2023/24 赛季为 933 人）。模型对观察到的医院住院峰值数分配的概率高于专家综合预测的概率（模型概率 = 0.07 对比 专家概率 = 0.05）。不过，该模型也将较高概率分配给了极不可能出现的小于 200 人的住院峰值（模型概率 = 0.21 对比 专家 = 0.03）。

当被要求为其预测的住院高峰提供理由时，专家的理由集中在：历史流感模式；其他病原体如 COVID-19 可能如何调节流感强度；情景假设，比如季节是否

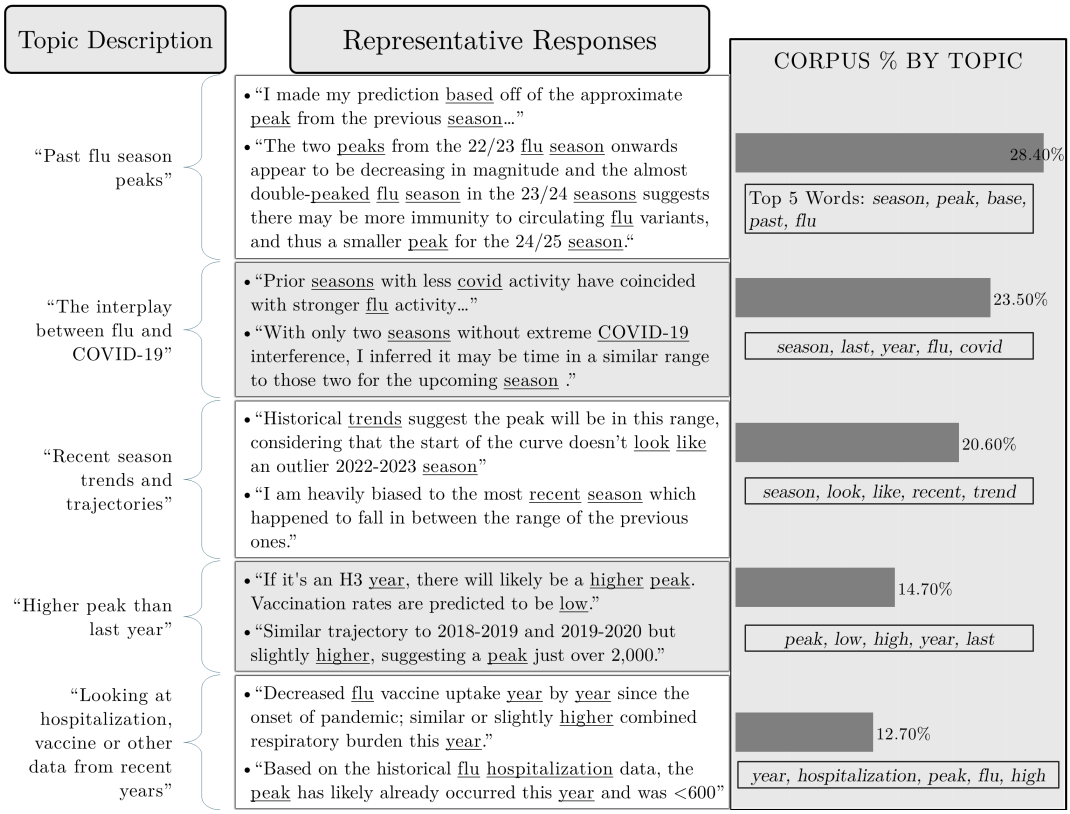

Fig. 2. 潜在狄利克雷分配从 102 位专家的回应中识别出五个主题，其出现频率以语料库的百分比表示。这些主题展示了专家们对流感历史模式、COVID-19 相互作用、近期趋势、比较评估和疫苗接种数据的考量——这说明了在传染病预测中，多样化的推理方法如何补充计算模型。

（或不是）以 H3 为主导的季节；以及疫苗的有效性和接种数据。这些主题是通过潜在狄利克雷分配分析专家提供的理由识别出来的（见图 2）。

几位专家指出了 COVID-19 与流感之间的相互作用，增加了诸如仅在两个没有极端 *COVID-19* 干扰的季节中，我推断对于即将到来的赛季可能需要类似的时间范围。和此前季节中新冠活动较少的时期与流感活动较强的时间相吻合。*[...]* 我猜我们不会遇到一种能够取代流感病毒的高负担新冠变异株。等理由。类似于模型，专家们还考虑了流感的先前峰值数据：查看最近的流感季节，高峰值从大约 *1200* 降至约 *1000*。我假设存在一种趋势，即在报告变化后回归到较低的峰值发病率。和此前的流感季节在 *1* 月达到约 *1K* 的峰值。然而，与模型不同的是，专家能够利用大量关于影响流感因果机制的知识：疫苗株匹配不佳将导致比去年观察到的更严重的流感季节。、自大流行开始以来，流感疫苗接种率逐年下降；今年的综合呼吸道负担相似或略高。和去年平均值 + 鸟流感。一些专家甚至实时敲击了一个数学模型，写下了数学模型现在实时运行中。值得注意的是，一些专家在做预测时能够识别出自己的潜在偏见。一位专家评论了当我看到链接中的历史数据时，我本能地倾向于这些数据，因为我的数据科学家眼光喜欢模式，但：如果我错了怎么办？另一位则写道我严重倾向于最近的那个赛季，它恰好落在前几个赛季的范围内。。

III. 讨论

我们的案例研究表明，在传染病预测方面，专家判断与计算模型之间存在互补优势。公共卫生专家对 H3 型流感主导地位的概率估计与实际观察结果更为一致。尽管模型和专家生成的峰值住院人数预测表现相似，但模型将大量概率分配给了一些专家正确识别为不太可能的情景，例如低于 200 人的峰值住院人数。从专家那里收集预测的一个主要优势在于他们能够描述其推理过程，而计算模型只能提供数值输出。这些发现表明，最优的预测方法应该同时收集计算和专家预测（包括理由），并可能利用自然语言处理工具如小型语言模型系统地将专家的理由融入计算框架中。仅给定五分钟来生成一个预测和理由，这项工作突显了公共卫生实践中经验的价值。

本研究有几个重要的局限性。我们缺乏与非专家预测的直接比较，限制了我们量化公共卫生专长相对于一般预测能力的具体价值的能力。尽管大多数参会者参与了，但并非所有会议参加者都参与其中，这可能会使预测偏向于那些在预测任务中更加自信或熟练的个人。我们的分析仅集中在某一地理区域内一个流感季节的两个预测目标上，限制了对专家在各种场景或病原体中的表现进行评估的范围。此外，我们简单的计算模型并不能代表所有可用的复杂建模方法，如集成方法或机器学习技术，这些可能与专家判断的表现不同。

人类判断在预测中的优势和脆弱性与关于不确定性下专家决策的现有文献相一致 ~ [13]。正如我们的结果所示，并得到先前研究的支持，专家们擅长基于认知的决策，迅速利用积累的经验来评估可能的情景和因果机制，而纯粹的数据驱动模型可能会忽略这些 ~ [16], [17]。然而，专家预测仍然容易受到认知偏见、过度自信和锚定效应的影响，这可能会损害准确性。这一点在我们的研究中得到了体现，与计算模型相比，专家的预测方差较小。当被要求在新环境或快速变化的情况下做出决策时，这些偏见可能尤为显著，因为过去的经历在这种情况下可能不太适用 ~ [21], [22]。

过去的预测系统整合了人类的预测和理由，在公共卫生实践中取得了成功，值得进一步探索。现代预测越来越依赖于利用集体智慧来提升决策过程中的各种领域的预测平台和专业工具。Metaculus 是一个在线预测平台，直接向其全球社区征求概率预测，并使用复杂的机器学习加权预测聚合这些个人概率，其表现优于传统的预测市场。像 IDEAcology 这样的专业工具通过设计用于生态学和生物安全领域等定量和概率估计的协议来简化严格的专家征询过程 [23]。现有预测平台的成功突显了建立正式系统以收集专家知识的价值。具体而言，对于公共卫生而言，开发一个专门的网络工具包来协调国家层面的专家理由的收集和分析，将把人的判断视为与传统监控系统同等重要的数据来源。我们主张建立一个包含专家预测、理由及它们所处背景的专用数据库可以改善危机应对。

为了支持公共卫生反应，特别是在资源紧张时期，我们建议未来的工作应致力于构建放大专家判断、专业知识、所做决策及其结果的工具。专家推理应被视为与通过监测系统收集的传统数据源同等重要的可量化数据来源。

IV. 建议

我们建议开发一个工具包，以应对在公共卫生预测中系统地捕捉和利用专家判断的挑战。尽管计算模型提供了定量预测，但它们往往缺乏人类预报员在复杂流行病学场景中所具备的情境推理和领域专业知识。提议的方法应将专家预测与其背后的推理相结合，从而创建改进的预测模型并洞察决策过程。

这样一个工具包需要三个组成部分。首先，该系统应系统地收集公共卫生专家的预测，并对他们的推理、假设和事实真相进行结构化的记录。其次，该系统应该使用语言模型根据第一部分收集的数据来识别有效与无效推理的特点。第三，一个仪表板将展示综合预测、历史数据以及专家之间常见的推理模式的可视化。

这样一个我们提出的工具包应该通过迭代进行修订，并收集公共卫生官员的反馈。这种以用户为中心的方法将确保数据收集程序、调查工具和仪表板界面与公共卫生实践中的现有工作流程和决策需求保持一致。

致谢

作者感谢汤玛斯 · 马丁 · 莱昂博士（加州公共卫生部建模部门负责人）和贾斯汀 · 克劳，MPA（弗吉尼亚州卫生部预见与分析协调员）提供的宝贵反馈。

References

[1] Hank Aaron, Robert F. Kennedy Jr, and the Public's Health. AJPH, Vol. 115 Issue 2. Accessed July 23, 2025. https://ajph.aphapublications.org/doi/10.2105/AJPH.2024.307945

[2] S. H. Woolf, S. Galea, and D. R. Williams, "The potential impact of the Trump administration policies on health research in the USA," *The Lancet*, vol. 405, no. 10495, pp. 2114-2116, 2025.

[3] Y. Takakazu, "The Trump Administration's Domestic Health Policy and Global Health," *Asia-Pacific Review*, vol. 32, no. 1, pp. 35-53, 2025.

[4] C. J. R. Daval and A. S. Kesselheim, "The Advisory Committee on Immunization Practices—Legal Roles, Challenges, and Guardrails," *JAMA*, June 26, 2025.

[5] "Trump administration purges U.S. health agency leaders," Accessed July 8, 2025. https://www.science.org/content/article/trump-administration-purges-u-s-health-agency-leaders

[6] J. Liu and K. Eggleston, "The Association between Health Workforce and Health Outcomes: A Cross-Country Econometric Study," *Soc Indic Res*, vol. 163, no. 2, pp. 609-632, 2022.

[7] T. McAndrew, A. A. Lover, G. Hoyt, and M. S. Majumder, "When data disappear: public health pays as US policy strays," *The Lancet Digital Health*, vol. 0, no. 0, 2025.

[8] C. P. Duggan and Z. A. Bhutta, "'Putting America First' — Undermining Health for Populations at Home and Abroad," *New England Journal of Medicine*, vol. 392, no. 18, pp. 1769-1771, 2025.

[9] R. C. Brownson, J. G. Gurney, and G. H. Land, "Evidence-based decision making in public health," *J Public Health Manag Pract*, vol. 5, no. 5, pp. 86-97, 1999.

[10] S. Juin, N. Schaad, D. Lafontant, et al., "Strengthening National Disease Surveillance and Response—Haiti, 2010–2015," *Am J Trop Med Hyg*, vol. 97, no. 4 Suppl, pp. 12-20, 2017.

[11] E. J. Barzilay, N. Schaad, R. Magloire, et al., "Cholera Surveillance during the Haiti Epidemic — The First 2 Years," *New England Journal of Medicine*, vol. 368, no. 7, pp. 599-609, 2013.

[12] J. W. Tappero and R. V. Tauxe, "Lessons Learned during Public Health Response to Cholera Epidemic in Haiti and the Dominican Republic," *Emerg Infect Dis*, vol. 17, no. 11, pp. 2087-2093, 2011.

[13] M. Zellner, A. E. Abbas, D. V. Budescu, and A. Galstyan, "A survey of human judgement and quantitative forecasting methods," *Royal Society Open Science*, vol. 8, no. 2, 201187, 2021.

[14] A. Verwiel and W. Rish, "Multidisciplinary perspectives on cumulative impact assessment for vulnerable communities: expert elicitation using a Delphi method," *Integrated Environmental Assessment and Management*, vol. 21, no. 2, pp. 301-313, 2025.

[15] C. C. Hammer, J. Brainard, and P. R. Hunter, "Risk factors for communicable diseases in humanitarian emergencies and disasters: Results from a three-stage expert elicitation," *Global Biosecurity*, vol. 1, 2019.

[16] P. R. Falzer, "Naturalistic Decision Making and the Practice of Health Care," *Journal of Cognitive Engineering and Decision Making*, vol. 12, no. 3, pp. 178-193, 2018.

[17] "Collaborative Activities During an Outbreak Early Warning Assisted by a Decision-Supported System (ASTER)," *International Journal of Human–Computer Interaction*, vol. 26, no. 2-3. Accessed July 9, 2025. https://www.tandfonline.com/doi/abs/10.1080/10447310903499062

[18] C. J. Cadham, M. Knoll, L. M. Sánchez-Romero, et al., "The Use of Expert Elicitation among Computational Modeling Studies in Health Research: A Systematic Review," *Med Decis Making*, vol. 42, no. 5, pp. 684-703, 2022.

[19] T. McAndrew, J. Cambeiro, and T. Besiroglu, "Aggregating human judgment probabilistic predictions of the safety, efficacy, and timing of a COVID-19 vaccine," *Vaccine*, vol. 40, no. 15, pp. 2331-2341, 2022.

[20] T. McAndrew, A. Codi, J. Cambeiro, et al., "Chimeric forecasting: combining probabilistic predictions from computational models and human judgment," *BMC Infect Dis*, vol. 22, no. 1, 833, 2022.

[21] A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science*, vol. 211, no. 4481, pp. 453-458, 1981.

[22] G. Recchia, A. L. J. Freeman, and D. Spiegelhalter, "How well did experts and laypeople forecast the size of the COVID-19 pandemic?" *PLOS ONE*, vol. 16, no. 5, e0250935, 2021.

[23] S. K. Courtney Jones, S. R. Geange, A. Hanea, et al., "IDEAcology: An interface to streamline and facilitate efficient, rigorous expert elicitation in ecology," *Methods in Ecology and Evolution*, vol. 14, no. 8, pp. 2019-2028, 2023.

[24] "Respiratory Virus Dashboard," Accessed July 25, 2025. https://www.pa.gov/agencies/health/diseases-conditions/infectious-disease/respiratory-viruses/respiratory-virus-dashboard.html

[25] "Weekly Hospital Respiratory Data (HRD) Metrics by Jurisdiction, National Healthcare Safety Network (NHSN)," Data, Centers for Disease Control and Prevention. Accessed July 25, 2025. https://data.cdc.gov/Public-Health-Surveillance/Weekly-Hospital-Respiratory-Data-HRD-Metrics-by-Ju/ua7e-t2fy/about_data

V. 方法

A. 数据收集和确定真实值

参与者是州和地方流行病学家委员会（CSTE）和疾病控制与预防中心（CDC）举办的传染病预测研讨会的参会者。会议时间为 2024 年 11 月 19 日至 21 日。参会者是公共卫生、流行病学和传染病建模领域的专家。几乎所有的州都有活跃的公共卫生官员代表参加此次会议。我们将这群人称为专家。

2024 年 11 月 20 日，我们在一次受邀演讲中向专家提出了两个问题。第一个问题是："请为即将到来的季节在宾夕法尼亚州被描述为 H3 季节的概率进行赋值。"专家们能够给出以下答案：季节成为以 H3 为主导的概率分别为 10%，20%..... 到 100%。第二个问题是："2024/25 季节中，宾夕法尼亚州的流感住院人数峰值将会是多少?"专家可以从一组范围中选择答案：[0-200]，[201-400]，[401-600]，[601-800]，[801-1000]，[1001-1200]，[1201-1400]，[1401-]。给专家提供了一些简要的背景信息以帮助他们形成预测（见补充资料）。

两个问题的真实情况是在 2025 年 5 月 31 日确定的，当时北半球的典型流感季节已经结束（在 MMWR 周 2025W22 之后）。H1 与 H3 优势季节的真实情况是从宾夕法尼亚州卫生部的呼吸系统仪表板 [24] 收集的，而由于流感导致的高峰住院人数的真实情况则是从由 CDC 主持的国家医疗安全网络中的每周医院呼吸系统数据集 [25] 中收集的。

B. 评估

专家的综合预测与基于历史观测数据训练的相应计算模型进行了比较。这种比较的前提是，如果专家的综合预测优于合理的计算模型，则这些模型及其所

训练的数据没有捕捉到专家使用的相关信息。这种“未知” 的信息可能包括在该领域多年积累的专业知识。

C. H1 对比 H3 优势流感季节

由于问题要求为 H3（而非 H1）季节分配一个概率，我们可以构建一个聚合密度函数，该函数将概率值分配给未来被归类为 H3 的流感病例的比例。

给定 N 个预测，专家聚合预报为 $x\%$ 个被归类为 H3 的病例分配的概率等于回答 $x\%$ 的专家人数除以 N，我们将其表示为 p_x。

我们可以通过将 p_x 分配到区间 $[x, x + 10\%)$ 来构建一个概率密度函数。

我们提出的计算模型是一个基于宾夕法尼亚州卫生部门呼吸系统监测仪表板上 2021/22 至 2022/23 季节被归类为 H3 的病例比例进行训练的核密度估计（参见补充资料中的数据集）。

D. 确认的流感住院病例的峰值强度

专家的综合预测是对上述标题为数据收集和确定真实值部分概述的八个区间的概率分配。分配给区间 I 的概率是通过选择该区间的专家人数除以所有参与的专家人数来计算的。我们提出的计算模型是一个基于内核密度估计的模型，该模型是在宾夕法尼亚州 2021/22 年至 2022/23 年季节的峰值住院人数数据集上进行训练的（参见补充资料中的数据集）。

用于预测的推理（N=102）通过潜在狄利克雷分配分析来识别常见主题。文本预处理包括去除停用词、词形还原和过滤出现在少于 3%或多于 80%响应中的单词。预处理后，在任何 102 个推理中都有 61 个独特的单词。模型选择涉及训练 LDA 模型，使用 2 到 8 个主题，并选择导致最高连贯性分数的主题数量。最高的连贯性分数（0.38）是在选择了 5 个主题时获得的。最终模型使用 α=0.083,β=0.01，通过 Gensim 4.3.3 训练了 100 次迭代。一位作者（GH）独立分析了这些主题并起草了解释，基于包含的单词和示例响应。这些解释随后由另一位合著者（TM）审查并确认。

VI. 伦理审批声明

在与 Lehigh 大学内部审查委员会（IRB）协商后，这项工作被认为不涉及人类受试者，无需正式评估。

VII. 数据可用性和分析可重复性

用于进行上述分析的所有数据和代码均可在 https://github.com/computationalUncertainty-Lab/cste_predictions 获取。特别提供了一个 Makefile，该文件对数据进行格式化，然后运行本研究的分析。

VIII. 补充材料

预测和专家理由的列表

2. 用于数据收集的表格（即专家征询）