发表于2022年11月在华盛顿特区举行的医疗保健可视分析研讨会（VAHC），该研讨会与美国医学信息学会（AMIA）年度研讨会联合举办

社会脆弱性模式 – 探索美国县级公共卫生风险的交互式仪表板

Darius Coelho*
Akai Kaeru LLC

Nikita Gupta†
石溪大学

Eric Papenhausen‡
Akai Kaeru LLC

Klaus Mueller§
Akai Kaeru LLC 石溪大学

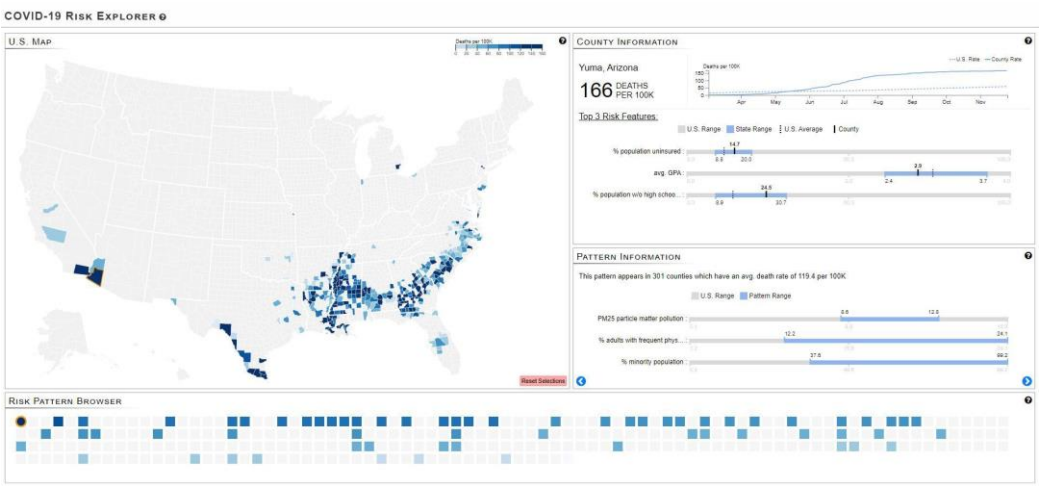


图1：用于探索驱动美国各县新冠肺炎死亡率的风险模式的COVID-19风险探索器仪表板。左上方是美国各县的交互式地理地图，各县根据其新冠肺炎死亡率以不同深浅显示。点击地图上的县所在位置即可选中该县。此处，用户选中了亚利桑那州尤马县，其位置以橙色轮廓标记。底部面板是风险模式浏览器，以平铺块的形式呈现所有风险模式，并根据其新冠肺炎死亡率以不同深浅显示。只有与选定县（亚利桑那州尤马县）相关的模式会以阴影显示。用户可以点击模式平铺块来选择风险模式。此处，用户选中了第一个模式，该模式以橙色轮廓标记。随后，地图上会以阴影显示所有具有该模式的县。右上方面板显示所选县的信息，其下方的面板则提供所选模式的信息。

Abstract

社会脆弱性是指社区受到自然灾害和公共卫生紧急事件（如干旱、地震、洪水、病毒暴发等）不利影响的易感程度。气候变化是近年来许多自然灾害的根源，而新冠肺炎大流行仍然构成持续威胁。社会脆弱性还涉及韧性，即从此类不利事件中恢复的能力。为衡量社会脆弱性的多个方面，美国疾病控制中心（CDC）将社会脆弱性划分为不同主题，例如社会经济地位、家庭构成等。了解社区的'社会脆弱性，有助于政策制定者和应急人员识别社区健康风险，为可能发生的灾害做好准备，或从灾害中恢复。本文研究美国县级的社会脆弱性，并介绍一项研究结果：美国各县可以根据社会脆弱性指标的某些组合或模式进行分组。随后，我们展示一个交互式仪表板，供分析人员以多种方式探索这些模式。我们以COVID-19死亡率作为灾害，展示了所识别模式对大流行局部影响具有较强的预测能力。

* 电子邮箱：dcoelho@akaikaeru.com
† 电子邮箱：nikigupta@cs.stonybrook.edu
‡ 电子邮箱：epapenha@akaikaeru.com §
电子邮箱：mueller@cs.stonybrook.edu

1 引言NTRODUCTION

社会脆弱性衡量影响社区应对危害人类健康的外部灾难性事件（称为压力事件）的韧性的社会经济和人口因素。这些压力事件可能是自然灾害、人为灾害，也可能是疾病暴发。社会脆弱性较高的社区更难以恢复，也更可能因这些压力事件而遭受损失；降低社会脆弱性可以减少人类苦难和经济损失。了解特定社区面临的社会脆弱性，有助于应急响应规划人员和公共卫生官员在特定灾害发生时迅速作出响应，并制定长期居住规划，以减弱灾害的潜在影响。

社会脆弱性的概念已在全球范围内得到研究，并形成了多种衡量方法。在美国，有毒物质与疾病登记署（ATSDR）和疾病控制中心（CDC）利用美国人口普查数据，为每个美国人口普查区确定了社会脆弱性指数（SVI）。人口普查区是美国各县的下级行政区划，美国共有 3,006 个县。CDC/ATSDR 的 SVI 根据 15 项社会因素对各人口普查区进行排名，这些因素可归为四个相关主题：

- 社会经济地位：低于贫困线、失业、收入、无高中毕业文凭
- 家庭构成与残疾：年龄为65岁或以上、年龄为17岁或以下、年龄超过5岁且有残疾、单亲家庭
- 少数族裔与语言：少数族裔，英语"不太流利"
- 住房类型与交通：多单元建筑、活动房屋、拥挤、无车辆、集体居住设施

我们利用 Kaggle 的数据集 [11]，将这一列表扩展为经过精心筛选的 241 项因素。这些因素将上述指标细化为种族和性别等人口特征，并加入吸烟和吸毒习惯、青少年怀孕、睡眠不足、住房债务、疫苗接种率等更多健康与社会风险因素。该 Kaggle 数据集由从 200 个公开的新冠肺炎相关数据集中收集的数据组成，数据来源包括约翰斯·霍普金斯大学、世界卫生组织、世界银行、《纽约时报》等。我们发现该数据集中存在若干重复项，并仔细筛选出 241 项因素；本文所报告的方法和研究均以此为基础。

将社会脆弱性以分级统计地图的形式呈现，有助于规划人员、应急人员和政策制定者比较不同地点的社会脆弱性指数，从而更好地了解哪些社区最容易受到自然灾害和疾病暴发的影响。其中最知名的是 CDC的自有 SVI 交互式地图 [10]，该地图根据各县的 SVI 总分着色；分数范围为 0（脆弱性最低）至 1（脆弱性最高）。点击特定县后，会弹出一张评分卡，以条形图列出各主题的得分。美国联邦紧急事务管理局（FEMA）[6] 提供了一系列分级统计地图，除社会脆弱性和社区韧性地图外，还允许用户生成多种自然灾害风险及预期年均损失地图，涵盖洪水、闪电、龙卷风等。名为县级健康排名的服务 [3] 允许用户生成基于各项因素的详细比较报告，并以可视化形式呈现美国各州县级粒度的数据。此外，还有其他一些机构（如美国国家航空航天局）也会制作基于各项因素的 SVI 分级统计地图。

据我们所知，迄今还没有研究能让用户更直接地探索社会脆弱性的多变量特性，即多个相互作用因素构成的模式。现有地图都要求用户在不同地图之间切换，或并排比较，以评估这些相关性。我们的模式分析根据脆弱性因素对县进行分组；这些因素相互作用，使这些县更容易（或不容易）受到某种特定目标灾害的影响，而这些县在地理上不一定彼此相邻。通过在一张分级统计地图上呈现这些模式以及受影响县的位置，让分析人员能够更深入、更高效地开展分析。本文介绍了我们为绘制此类地图而开展的研究成果。

本文结构如下：第 2 节介绍相关研究，第 3 节阐述我们的模式分析，第 4 节介绍我们的仪表板，用户可借此探索美国各县的这些模式。第 5 节总结全文。附录中介绍了一项案例研究。

2 Related Work

不同于以往大规模的全球卫生危机，新冠肺炎大流行发生在一个机器学习工具无处不在、信息技术基础设施高度发达的时代。新冠肺炎带来的紧迫性不仅推动了生物技术领域以更快速度开发疫苗和医疗治疗，还促进了病毒传播预测建模方面的研究，以及开发可视化信息门户，以便向政策制定者和公众通报死亡人数、感染病例等相关信息。因此，新冠肺炎卫生紧急事件催生了大量近期关于用于公共卫生监测和风险评估的可视化工具的相关研究。

2.1 建模和预测

许多预测建模方法（例如 [13], [24]）都基于机理性的易感者–暴露者–感染者–康复者（易感者-暴露者-感染者-康复者模型）舱室模拟模型，该模型从过程层面模拟新冠肺炎的传播方式。机理模型颇具吸引力，因为它们能让人模拟不同缓解措施（如隔离、社交距离、学校停课等）的影响。然而，该模型的许多参数都需要准确估算各舱室中的人口数量及其转移率，而其中涉及的不确定性使这项工作颇具挑战。

统计预测模型也很受欢迎，例如华盛顿大学'健康指标与评估研究所（IHME）开发的模型 [1]。该模型采用混合效应非线性回归模型，根据全球各地理位置的数据拟合曲线，从而在地方层面生成感染预测、死亡率和卫生资源需求的预测结果。其他方法则采用更为传统的统计模型，例如相关性和线性回归，以了解某些社会经济因素的影响，如县级健康变量、城市密度、贫困、通勤等，同时控制种族等其他因素的影响 [18]。通常，这些结果通过标准的逐步建模方法获得；由于这类方法在因素数量和地方区域数量方面的可扩展性并不强，因此发现具有显著性的统计关系往往相当繁琐。

鉴于模型往往会给出范围广泛的预测，集成方法近来逐渐流行起来（例如，参见 [17]）。集成方法将不同的单个模型结合起来，并对它们的结果进行加权，形成综合预测。这能让预测更加稳健，并使整个过程更加稳定。

2.2 可视化

约翰斯·霍普金斯大学的冠状病毒资源中心一直是一个主要的信息来源 [4]。该中心为美国和全球创建了仪表板，分别展示相应的地理地图，并叠加检测呈阳性者人数的可视化表示，同时提供各类检测和死亡统计数据、排行榜，以及时间增长曲线集成，用于从不同尺度比较各地区检测病例数和死亡人数的增长情况。Tableau [7], 、TIBCO [2], 、开源项目 Nextstrain [9], 、《纽约时报》等媒体也提供了其他仪表板和基于浏览器的交互式新冠肺炎相关数据可视化。这些仪表板和可视化呈现了与疫情暴发相关的特定方面，例如种族、医院过度拥挤、各州检测统计数据、各县口罩佩戴合规情况、失业率和失业申请、经济不平等、病原体演化等。

IHME 的新冠肺炎预测模型也因其简单而有效的交互式可视化仪表板工具而广受欢迎 [5]。不过，IHME 也因其交互式"美国健康地图—Viz中心"而闻名 [8]。该可视化界面提供了一个菜单，供用户从四种结果或风险变量中进行选择，即预期寿命、死亡率、死亡风险等。随后，用户可以选择多种疾病和健康决定因素之一，并在州或县的分级统计地图上显示相应的结果或风险。然而，与其他地图类似，该地图一次也只能显示一个数值，因此只能按单个因素进行比较。将多幅地图并置 [14] 有助于比较，但仍难以识别变量之间的总体相关性或分组。我们的地理可视化显示通过将模式分析与一组信息显示相结合，使这类分析成为可能。

近期的一个可视化界面是 EnsembleVis [21]，它是一个基于 Web 的地理地图界面，用于查看和比较美国县级的模型预测与真实值。用户可以浏览集成模型，从而更好地了解预测范围及其中的不确定性。我们也按县级展示数据，但重点主要在于解释某些健康风险为何会出现，也就是哪些社会脆弱性使某些社区面临更高风险。我们的方法从已经暴露于这些风险的社区中学习这些风险模式，但也有其他社区符合这些模式，只是尚未遭遇同样的命运。因此，我们识别出的模式也可以用于预测这些社区可能面临的风险，或至少向这些社区及相关应急人员发出警示。

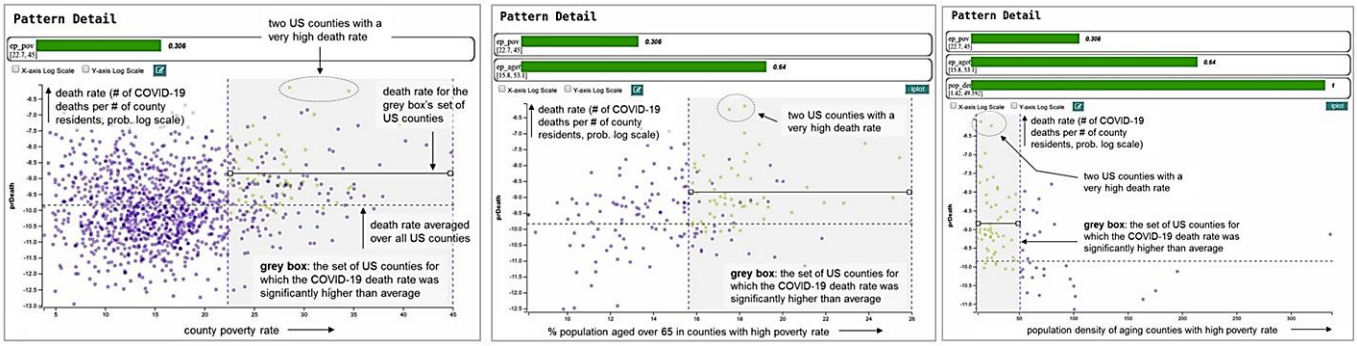


图2：2020年5月发现的一个三维模式的散点图。y轴（目标属性）表示观测到的COVID-19死亡率，每个点代表一个美国县。每幅图展示了模式属性与目标属性在二维投影图中的分布；绘图时，先根据其左侧紧邻图中模式属性所定义的子空间筛选数据点。黄色点位于图中的模式子空间内，紫色点位于其外。可以看出，加入第三个属性后，所有紫色点都被排除。数据点经过筛选，因此从左到右各图中的点数逐渐减少。顶部的绿色条形图显示了各子空间属性（按从上到下、从左到右的顺序分别为：县贫困率、65岁以上人口比例、人口密度）对模式定义的贡献。

3 我们的方法：模式挖掘与仪表板

下面我们将概述模式挖掘方法，然后重点介绍我们设计的、用于探索这些模式的仪表板。

3.1 我们的模式挖掘方法

我们的方法植根于模式分析，这是数据科学和人工智能领域中一个经过充分研究的方向 [19]。模式是指一组具有相似特征或属性的数据点，即一个子群体 [12]。例如，数据可能是一组社会经济构成相似的县。在我们的示例中，每个县都有 241 个特征，例如过度饮酒习惯的成年人比例、频繁遭受精神困扰的成年人比例、失业率等。这 241 个特征构成了一个 241 维特征空间，而该空间通常相当稀疏。

我们设计了一种模式挖掘引擎，可自动在这个稀疏的特征空间中搜索由相似县组成的区域，这些县对特定目标变量都呈现相似的响应。我们在 [20] 中报告了模式挖掘方法的一些结果。在那项研究中，我们特别关注新冠肺炎，所用可视化仅为简单的分级统计地图，无法按特征探索这些模式。如今，我们报告一个专用仪表板，它将人类纳入分析流程，并支持进行详细探查。

在我们此前的工作中，我们试图找出导致COVID-19死亡率高于平均水平的社会经济状况，因此，我们的目标变量是各县的COVID-19死亡率。我们注意到，被视为相似的县不必在地理上相连；它们只需在特征值方面具有相似的特征。模式的一个独特属性是，它位于一个超立方体内，其中描述该模式的特征值范围明确。这一属性及其固有的低维特性[22]，（即使整体特征空间并非如此），使其易于理解和解释。尽管深度神经网络、随机森林等也能学习低维表示，但这些表示不易用原生特征来描述。因此，这类架构通常被称为黑箱人工智能，而模式分析则是一种可解释人工智能方法。

具体而言，给定一个包含属性 $\{A_1 A_2 .... A_m P\}$ 的数据集，其中 P 是一个感兴趣的属性，例如 COVID-19死亡率，模式挖掘的目标是找出一个由约束构成的超立方体（或模式），这些约束的形式为 $A_i \in [v_l, v_r]$ 对于 $i \in [1..m]$ （例如，年龄 > 45，种族 = 亚洲），其中模式中的点是"有趣的。"就我们的用途而言，如果某个县群模式关联的 COVID-19死亡率平均高于美国县平均值，我们就认为它是有趣的。对于何为持续有趣的模式，其定义主要基于

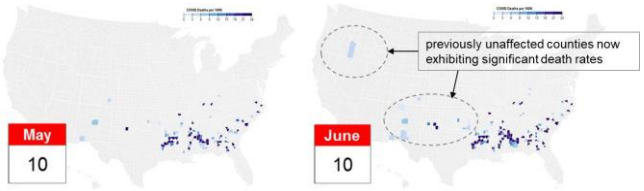


图3：图2中模式所涵盖县的位置，按COVID-19死亡率着色。2020年6月新近受到疾病侵袭的县（位于虚线椭圆内，也包括其他地区的县），要么位于图2中2020年5月散点图的底部，要么当时尚未显示在图中，但都被正确预测为即将受到冲击。

统计假设检验。对于数值属性，我们使用曼-惠特尼检验 [23]，以应对数据通常具有的非参数性质；对于二元目标属性，我们使用 χ^2 独立性检验。提取模式需要进行大量搜索；我们使用 FP-growth算法 [16]，该算法效率较高，因为在挖掘过程中只需扫描整个数据集两次。

3.2 我们的模式挖掘：深入观察

在之前的 COVID-19 研究中，我们使用了引言中提到的 241维数据集，并将美国各县的 COVID-19死亡率作为目标变量。我们在2020年5月发现了297个二维和三维模式。图2展示了其中一个模式，这是一个由高贫困率、老年人占比较高以及人口密度低所定义的三维模式。这三个变量足以确认死亡率均值升高具有统计显著性。图注对该模式的形成进行了更详细的说明。

让我们仔细看看图2最右侧的图，它显示了美国各县在该模式第'三个属性上的投影（黄色点）。我们注意到，该模式中的大多数美国县死亡率都高于美国平均值（而且该模式本身的平均值也高于美国平均值，这使得该模式'颇具"趣味性"）。但我们也注意到，有几个美国县位于美国平均值线下方；它们的死亡率低于美国平均值。这可能意味着还有其他潜在（未测量）的因素保护这些县免于感染病毒。但也可能意味着这些县尚未受到 COVID-19疫情浪潮的冲击–请记住，2020年5月时大流行才刚刚开始。

图3对此做了更深入的说明。图中展示了两幅地图，死亡率显著的县以蓝色标示–蓝色越深，表示死亡率越高。请注意，我们只为符合图2所示模式的县着色（它们也可能符合其他模式，但本图未考虑这些模式）。左侧是

2020年5月的地图，这是我们用于学习模式的月份。右侧是下一个月，即2020年6月对应的地图。可以看到，现在有不少县也被着色了，而这些县在5月时尚未受到影响；请看椭圆圈出的区域。此外，已经受到影响的地区也出现了新的县。

这些新近受到影响的县都位于图2中平均值标线以下，因此该模式成功预测了它们的命运。事实上，我们观察到，在所有模式中，98%的死亡率增幅是美国平均值的2至3倍；其余2%则以平均速度增长，没有任何模式的增长速度低于美国平均水平。这些趋势在7月仍在持续。这表明，我们的模式具有很强的预测能力，同时也能以通俗易懂的方式，用特征所使用的语言解释导致COVID-19死亡率高于平均水平的社会经济状况。

3.3 交互式可视化仪表板

我们的仪表板面向可视化素养水平各异的人群，旨在帮助他们浏览和查看通过我们的方法挖掘出的模式。仪表板由四个相互关联的主要面板组成：地理地图、模式浏览器、模式信息和县信息。下文各节以及图1的说明中均对这些面板进行了介绍。输入数据为包含数据矩阵的标准CSV文件。

风险模式浏览器：如第3节所述模式集合仍可能跨越大量维度。这使得设计一种易于理解的可视化表示来完整解释一个模式变得困难，甚至不可能。因此，我们选择将模式集合表示为平铺块列表，其中每个平铺块代表一个模式，如图1底部面板所示。模式按照死亡率降序排列，从左到右、从上到下依次显示。我们进一步根据新冠肺炎死亡率，将平铺块从深蓝色到浅蓝色进行着色，以强化这一点。只有涉及地理地图中所选县的平铺块会被着色（下文将对此作进一步说明）。

点击这些平铺块中的任意一个后，地理地图视图会更新，以标示该模式所属的县；模式信息面板也会更新，以展示模式的详细信息（详见下文）。此外，我们会将选中平铺块的形状改为圆形，并为其添加黄色边框，以便用户在仪表板的不同元素之间切换焦点时，能够轻松找到所选平铺块。

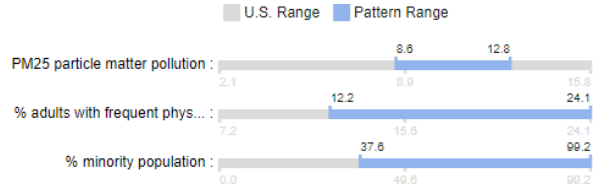


图4：模式由特征范围定义。灰色条表示该特征在美国各地的取值范围，蓝色条表示该模式对应的取值范围。例如，第三个特征“少数族裔人口占比”的取值范围在美国各地为0至99.2，而37.6至99.2是该模式的特征之一，会推动死亡率高于平均水平。

模式信息：此面板向用户传达模式信息。模式本质上是一组具有特定范围的属性。因此，模式信息面板会向用户报告这些范围，并将其置于该维度在所有数据点（本例中为所有县）中的整体范围这一背景下。为了直观呈现这些范围，我们采用了子弹图样式的可视化方式；研究表明，广大受众都能轻松理解这种方式 [15]。图 4 展示了一个示例。此处，每个维度的范围都用一条水平条表示。条形的灰色部分表示美国所有县中该维度的范围，而条形的蓝色部分则表示定义此模式的维度范围。用户可以快速浏览这些条形图，了解定义各个模式的不同范围。

地理地图视图：此面板包含一幅美国县级交互式地图（如图1所示）。地图中的每个县都根据其新冠肺炎死亡率着色。我们使用连续色标，死亡率较高时显示深蓝色，死亡率为零时显示白色。用户可以点击某个县，了解导致该县新冠肺炎死亡率的因素。点击某个县后，风险模式浏览器会更新，突出显示该县所属的模式，并将其余模式灰显。此外，县信息面板也会更新，显示该县的主要风险特征。用户还可以缩放和平移地图，以便选择面积较小的县。

县信息：此面板通过地理地图传达用户选定县的信息。如图1右上角所示，该面板报告当前新冠肺炎县级死亡率以及随时间变化的死亡率。更重要的是，该面板传达该县的前三大风险因素。在此，特征排名是根据这些特征在所有包含所选县的模式中出现的频率计算得出的。我们使用与模式信息相似的子弹图可视化来直观呈现这些特征。这里，灰色条表示美国所有县中该维度的范围，而蓝色条表示所选县’所在州所有县中这些特征的范围。除了图表中显示的范围外，我们还为该县的因素值以及美国平均值添加了标记。图5展示了一个示例。

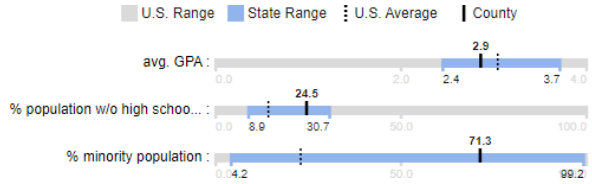


图5：在州和美国范围背景下展示某县前三大特征值的可视化表示。灰色条表示该特征在美国范围内的取值范围，蓝色条表示该特征在该县’所在州所有县中的取值范围。例如，此处所示县的‘平均GPA’为2.9（黑色实心标记），略低于美国平均值（黑色虚线标记）。此外，平均GPA‘的美国范围为0至4，而该特征在该州所有县中的范围为2.4至3.7。

4 结论ONCLUSIONS

我们提出了一种方法，可以将公共卫生的社会经济指标归纳为从观测数据中学习得到的 1—3 个因子模式。这些模式可供政策制定者和卫生官员用于解释和预测某个社区针对特定自然灾害或突发公共卫生事件所面临的潜在风险。为了便于访问这些模式，我们设计了一个交互式可视化仪表板，以便结合各社区的 地理位置探索这些模式。虽然我们使用 COVID-19大流行早期阶段展示了该方法的一种应用，但我们相信，其应用范围要广泛得多；目前正在开展的工作正对此进行探索。

虽然我们为数据提供了时间背景–即 COVID-19死亡率随时间的变化”回 溯”时间，以便用户查看所选时间范围内的模式，但目前用户还无法这样做。这项功能很容易实现，我们计划在未来添加。此外，虽然我们在系统开发过程中通过一小部分用户收集了反馈，但我们计划近期开展规模更大的基于任务的研究，以进一步了解实用性和可用性，

Acknowledgments

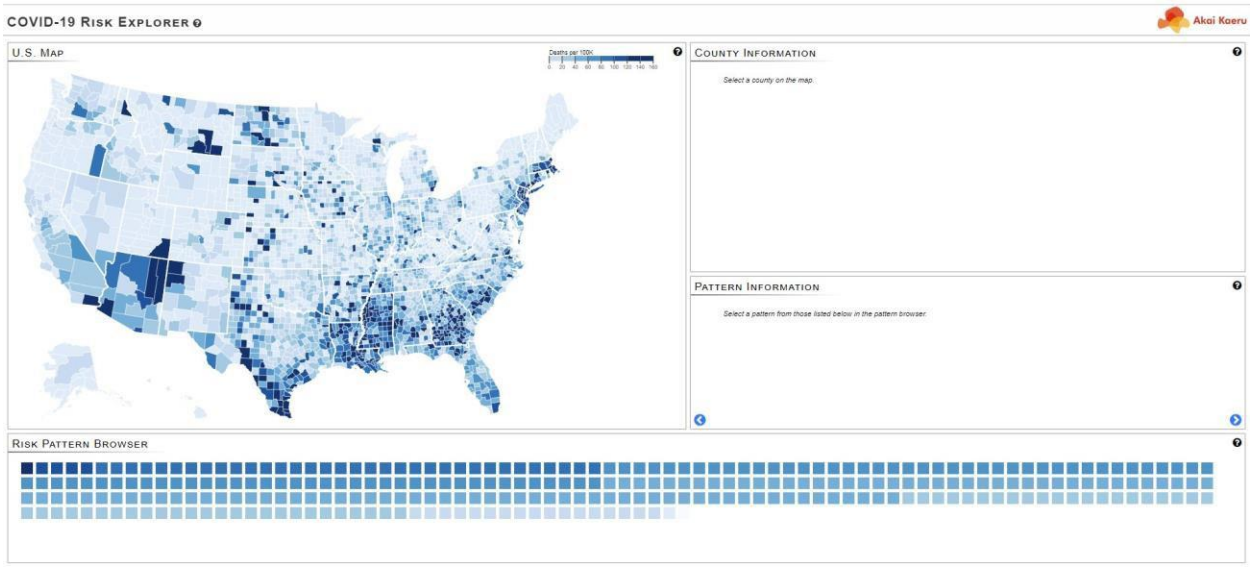
本研究由美国国家科学基金会小企业创新研究项目合同 1926949 资助。

References

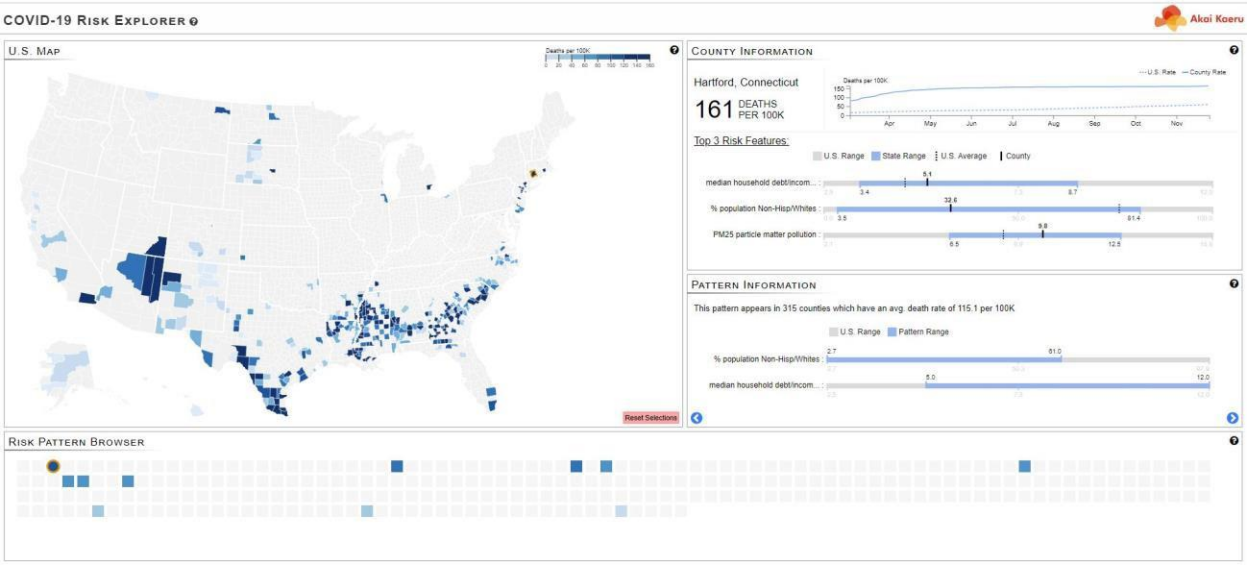
[1] 2020. 华盛顿大学健康指标与评估研究所（IHME）新冠肺炎资源。检索自：http://www.healthdata. org/covid. [2] 冠状病毒可视化分析中心。（访问日期：2022年8月14日）。https://www. tibco.com/covid19.[3] 县级健康排名（访问日期：2022年8月14日）。https://www. countyhealthrankings.org. [4] 约翰斯·霍普金斯大学系统科学与工程中心的新冠肺炎仪表板（访问日期：2022年8月14日）。https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6.[5] 新冠肺炎预测仪表板（访问日期：2022年8月14日）。https: // covid19.healthdata.org/united-states-of-america. [6] FEMA地图（访问日期：2022年8月14日）。https://hazards.fema.gov/nri/social-vulnerability. [7] 全球新冠肺炎追踪器（访问日期：2022年8月14日）。https://www. tableau.com/covid-19-coronavirus-data-resources.[8] 健康指标与评估研究所（IHME）美国健康地图—VizHub。（访问日期：2022年8月14日）。https://vizhub.healthdata.org/ subnational/usa.[9] Nextstrain病原体演化实时追踪。（访问日期：2022年8月14日）。https://nextstrain.org. [10] SVI 交互式地图（访问日期：2022年8月14日）。https://svi.cdc.gov/ map.html. [11] Uncover新冠肺炎挑战赛数据集（访问日期：2022年8月14日）。https://www.kaggle.com/datasets/roche-data-science-coalition/uncover.[12] M. Atzmueller。子群发现。《Wiley跨学科评论：数据挖掘与知识发现》，5(1):35–49，2015年。[13] A. L. Bertozzi、E. Franco、G. Mohler、M. B. Short和D. Sledge。新冠肺炎传播建模与预测面临的挑战。《美国国家科学院院刊》，117(29):16732– 16738，2020年。[14] L. Besanc,on、M. Cooper、A. Ynnerman和F. Vernier。基于分级统计地图的人口统计数据可视化方法评估。arXiv预印本 arXiv:2005.00324，2020年。[15] D. Coelho、H. He、M. Baduk和K. Mueller。问心无愧地饮食：迈向可视化且具情境信息的营养成分标签。2020年。[16] J. Han、J. Pei和Y. Yin。无需候选生成的频繁模式挖掘。《ACM SIGMOD Record》，29(2):1–12，2000年。[17] J.-S. Kim、H. Kavak、A. Z ü fle和T. Anderson。使用代表性聚类的新冠肺炎集成模型。《SIGSPATIAL Special》，12(2):33– 41，2020年。[18] C. R. Knittel和B. Ozaltun。哪些因素与新冠肺炎死亡率相关，哪些无关。技术报告，美国国家经济研究局，2020年。[19] H.-P. Kriegel、P. Krö ger和A. Zimek。高维数据聚类：子空间聚类、基于模式的聚类和相关聚类综述。《ACM知识发现与数据挖掘汇刊》（TKKD），3(1):1–58，2009年。[20] K. Mueller和E. Papenhausen。利用人口统计模式分析预测美国县级的新冠肺炎死亡人数。《数字政府：研究与实践》，2(1):1–11，2020年。[21] S. Srabanti、G. E. Marai和F. Miranda。新冠肺炎EnsembleVis：县级集成预测模型的可视化分析。载于2021年IEEE医疗保健可视分析研讨会（VAHC），第1-5页，2021年–.[22] B. Wang和K. Mueller。3D真的适用于可视化聚类分析吗？是的！载于2014年IEEE VIS国际3DVis研讨会（3DVis），第37–44页。IEEE，2014年。[23] I. B. Weiner和W. E. Craighead。《Corsini心理学百科全书》，第4卷。John Wiley & Sons，2010年。[24] Z. Yang、Z. Zeng、K. Wang、S.-S. Wong、W. Liang、M. Zanin、P. Liu、X. Cao、Z. Gao、Z. Mai等人。改良的易感者-暴露者-感染者-康复者模型与人工智能预测公共卫生干预下中国新冠肺炎疫情趋势。胸部疾病杂志，12(3):165，2020。

附录

在这个案例研究中，我们跟随公共卫生分析师Bob，了解他如何使用COVID-19风险探索器，进一步探究当地社区对COVID-19病毒传播的易感性。现在是2020年12月，期间发生了很多事情。他启动程序，看到下方的屏幕。



Bob 观察到，总体死亡率最高的地区位于得克萨斯州、亚利桑那州、蒙大拿州、北达科他州、爱达荷州、南部地区、佛罗里达州和东北海岸。现在 Bob 想了解时间线。他选择了康涅狄格州的一个深色县（死亡率较高的县）–哈特福德县。

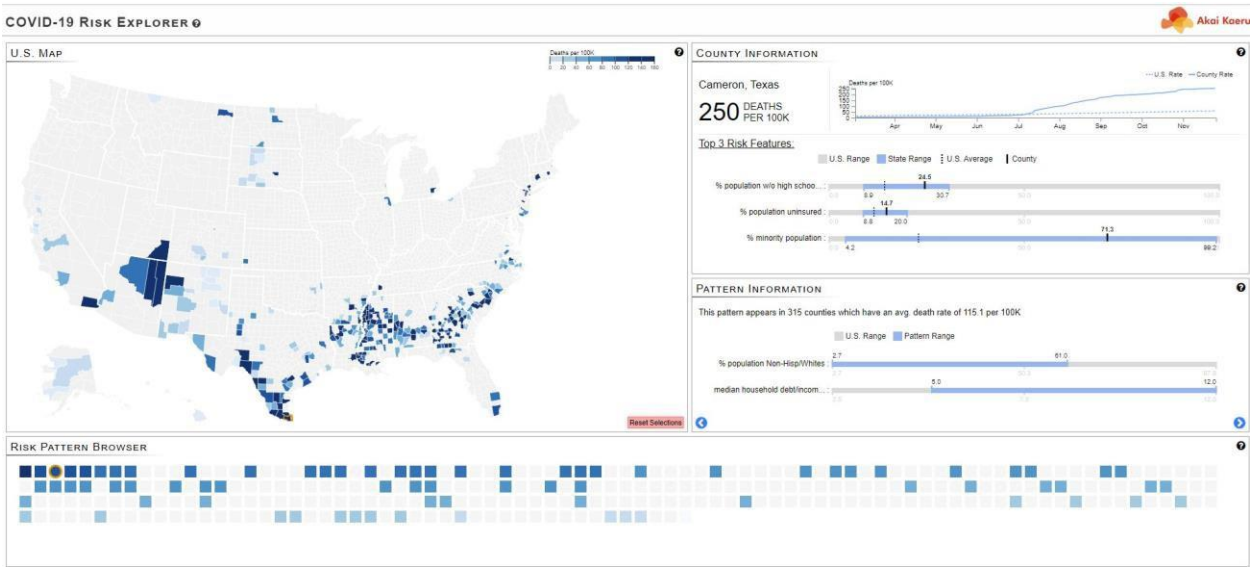


屏幕切换后，Bob 从“县信息”面板中显示死亡率随时间变化的折线图上观察到，该县在 2020 年很早的时候就超过了美国平均死亡率，而且上升得相当迅速，但从 4 月开始便基本保持平稳。显然，该县应对得当，并采取了良好的预防措施来遏制传播。

Bob 查看“模式信息”面板，研究康涅狄格州哈特福德参与其中的最主要模式所涉及的风险因素。“风险模式浏览器”显示，这是数据库中3rd 最主要的模式。这是一个双特征模式，表明在康涅狄格州哈特福德，非西班牙裔/白人的比例处于美国总体范围的低端，而家庭债务与收入中位数之比处于美国总体范围的高端。查看“县信息面板”中哈特福德的“前三大风险因素”后，Bob 发现，除了 PM2.5颗粒物污染外，这两个风险因素实际上也名列哈特福德的前三大风险因素

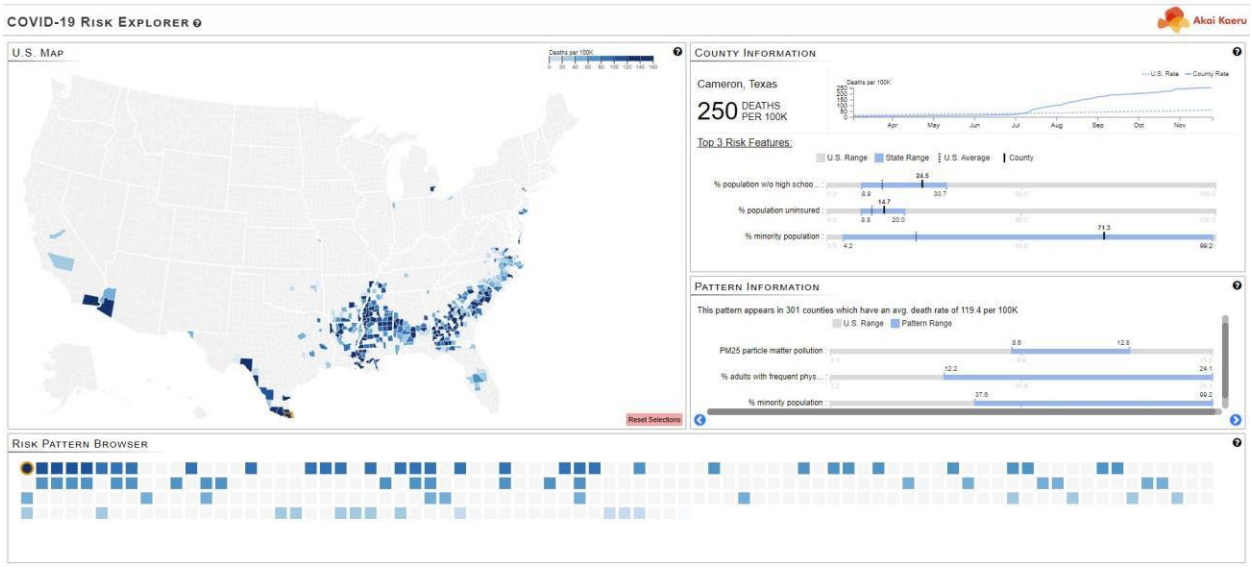
，而且所有这些指标与美国平均值相比都处于不利水平。虽然它们在康涅狄格州并非处于极端值，但往往落在不利的数值范围内。

接下来，Bob 点击得克萨斯州南部一个颜色同样较深的县——卡梅伦县，得克萨斯州，屏幕随即切换为下方所示画面。

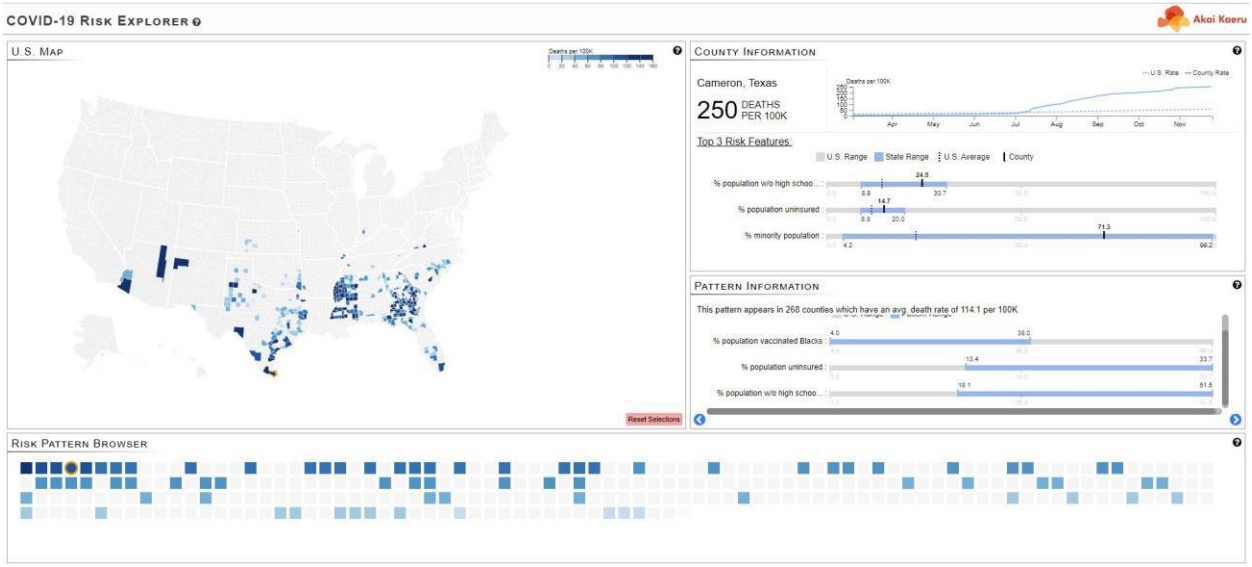


Bob 观察到，该县的死亡率开始上升的时间晚得多，是在2020年7月，而且到12月仍在上升，只是斜率较缓。虽然卡梅伦县，得克萨斯州与康涅狄格州哈特福德共享3rd 风险模式，但它的风险模式比康涅狄格州哈特福德多得多，这从“风险模式浏览器”中许多已填充的方格便可看出。这意味着，与康涅狄格州哈特福德相比，该县具备导致高死亡率的条件更加紧迫。事实上，它的前三大风险因素并非3rd 风险模式中的那些因素。在所有这些方面，卡梅伦县，得克萨斯州的情况无论是在得克萨斯州的取值范围内，还是相对于美国平均值，都更为不利。

接下来，Bob 点击“风险模式浏览器”中的第一种风险模式，并看到下方画面。

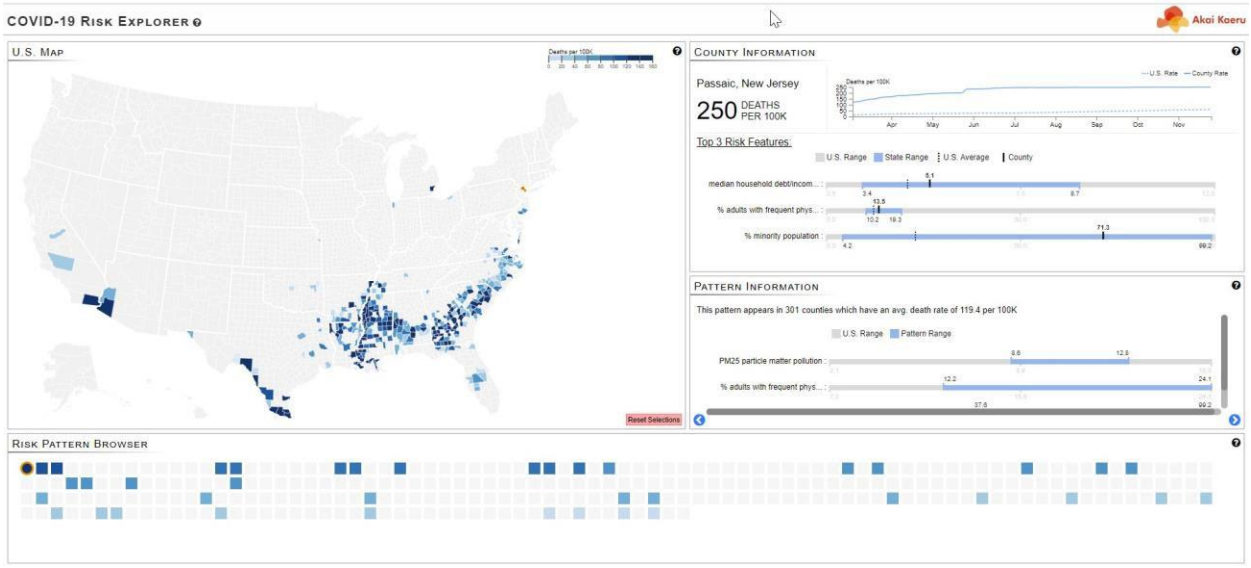


这种模式是一个三特征模式，其三项特征的数值范围都较高且不利。只有“少数族裔人口比例”出现在卡梅伦县，得克萨斯州的前三名列表中。经过进一步调查，Bob 发现另外两项风险特征“未获得高中学历的人口比例”和“未参保人口比例”属于第4种风险模式（见下一页图片）。

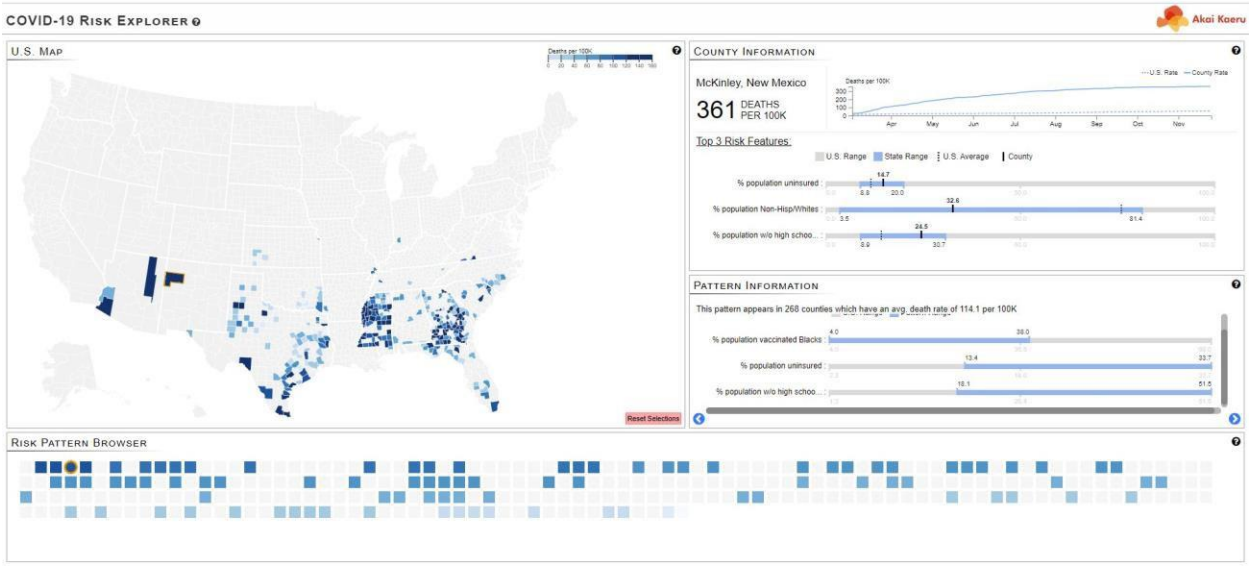


Bob 现在想调查，卡梅伦县，得克萨斯州是否能从其他风险因素相似、但此前 COVID-19死亡率较高的县吸取教训。他寻找具有部分或全部前三名风险特征的县，因此查看了属于第1种风险模式和第4种风险模式的县。

他从风险模式 #1 开始，在地图上点击了几个县，最终了解到帕塞伊克县，新泽西州。



从时间线上，他看到帕塞伊克县，新泽西州最早开始出现死亡率上升，其“少数族裔人口比例”也相近。但这只是卡梅伦县，得克萨斯州前三大风险因素中的一个，因此 Bob 需要继续搜索，才能完成调查。他转而查看第4种风险模式，经过一番搜索后找到了麦金利县，新墨西哥州（见下一页）。



麦金利县，新墨西哥州的死亡率开始上升的时间晚于帕塞伊克县，新泽西州，但仍比卡梅伦县，得克萨斯州早4个月。它的前三大风险因素中包含卡梅伦县，得克萨斯州的另外两个风险因素。因此，如果卡梅伦县，得克萨斯州曾关注麦金利县，新墨西哥州和帕塞伊克县，新泽西州的情况，就能从中吸取教训，并采取它们实施的预防措施。

还有许多类似的案例研究，后发者本可以从先发者的遭遇中吸取教训，做好更充分的准备。从先发者的时间线可以看出，他们都成功控制了传播。后发者本可以采取类似措施，从而有望挽救生命。

总之，我们的研究结果表明，模式分析是公共卫生风险管理的有力工具，而像我们的风险探索器这样的仪表板能让人们轻松识别风险，并了解某些社区发生的疫情暴发和灾害如何迅速为具有相似脆弱性模式的其他社区提供参考，以防止进一步的损失。