arXiv:2303.00617v1 [cs.HC] 1 Mar 2023

Causalvis：因果推断可视化

Grace Guo 佐治亚理工学院 美国佐治亚州亚特兰大
gguo31@gatech.edu

Ehud Karavani I BM Research 以色列特拉维夫吉瓦塔伊姆
ehudk@ibm.com

Alex Endert 佐治亚理工学院 美国佐治亚州亚特兰大
endert@gatech.edu

Bum Chul Kwon IBM Research 剑桥，马萨诸塞州，美国
bumchul.kwon@us.ibm.com

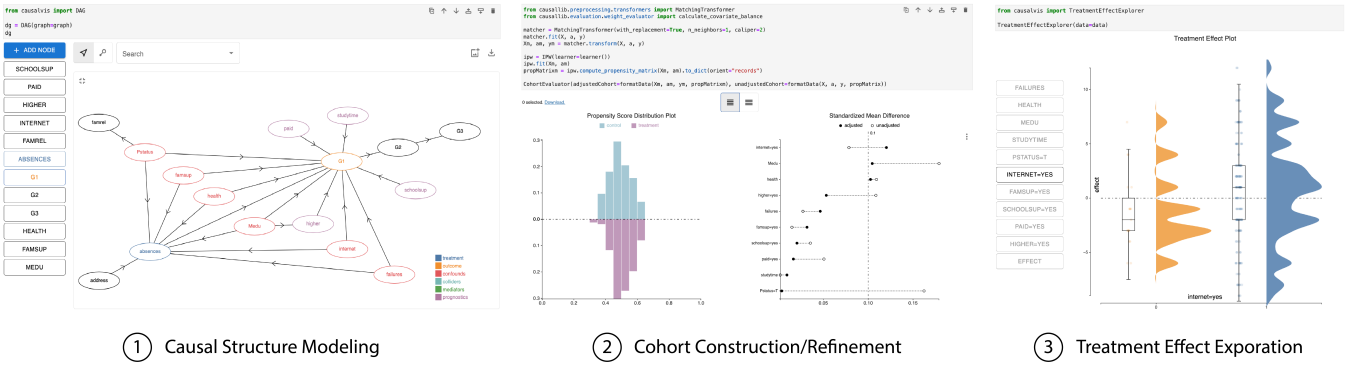

图1：Causalvis 的可视化模块分别对应因果推断工作流程的三个步骤：1）因果结构建模，2）队列构建/细化，以及3）处理效应探索。

摘要

因果推断是一种利用观察性数据量化因果效应的统计范式。这是一个复杂的过程，需要多个步骤、反复迭代，并与领域专家协作。分析人员通常依靠可视化来评估每个步骤的准确性。然而，现有的可视化工具包并非旨在支持分析人员熟悉的计算环境中的整个因果推断过程。本文通过 Causalvis 弥补这一空白：它是一款用于因果推断的 Python 可视化软件包。我们与因果推断专家密切合作，采用迭代式设计过程，开发了四个交互式可视化模块，以支持因果推断分析任务。随后，我们将这些模块展示给专家，征求反馈并进行评估。我们发现，Causalvis 有效支持了迭代式因果推断过程。最后，我们讨论了研究发现对于因果推断可视化设计的启示，尤其是对沟通与协作相关任务的启示。

允许为个人或课堂使用而免费制作本作品的部分或全部内容的数字副本或纸质副本，前提是不得为牟利或商业利益而制作或分发副本，且副本须包含本声明及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*CHI '23，2023年4月23日至28日，德国汉堡*© 2023年 版权所有者/作者所有。ACM ISBN 978-1-4503-9421-5/23/04。
https://doi.org/10.1145/3544548.3581236

CCS 概念

• **以人为中心的计算** → 可视化设计与评估方法；**可视化工具包**；**可视分析**。

关键词

因果推断，因果关系，设计研究

ACM 参考格式：
Grace Guo、Ehud Karavani、Alex Endert 和 Bum Chul Kwon。2023年。Causalvis：因果推断的可视化。载于《*2023*年*CHI*计算系统中的人为因素会议论文集》（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，纽约，NY，美国，共20页。https://doi.org/10.1145/3544548.3581236

1 引言

因果推断是一种利用观察性数据量化因果效应的统计范式。尽管随机对照试验（RCTs）被视为确立因果关系的金标准，但它们并非总是可行，尤其是在操纵处理变量代价高昂、不切实际或不合伦理的情况下。例如，在医学领域，随机选择一组参与者并要求他们吸烟一段时间，以测量吸烟对肺癌的影响，这种做法极不合伦理。因果推断只需使用观察性数据，因此是一种强有力的方法，已广泛应用于医疗保健 [18, 24],经济学 [3],和社会科学[16]。有意研究特定处理暴露影响的研究人员，可以从一个观察性

数据集中选出两组——处理组和对照组——并比较两组的结果。估算处理暴露对结果的影响大小，往往是因果推断分析的目标。

通过用观察性数据替代随机对照试验，必须采取额外步骤，以确保估计的处理效应可靠且无偏。这是一个复杂的过程，需要多个步骤、反复迭代以及与领域专家的协作。研究人员需要准确理解数据集中的因果关系，识别并控制所有混杂变量，进行必要的统计调整，并确保所选的处理组和对照组满足所需的假设。再次考虑这样一种情境：研究人员希望估计香烟对肺癌的影响。研究人员可以先从大型医疗记录数据集中选取两组人群——吸烟者和非吸烟者，然后比较吸烟者（处理组）与不吸烟者（对照组）之间的肺癌发生率（结果）。然而，要使这种比较有效，必须满足若干假设。其中一个假设是，两组之间必须具有一定程度的相似性——例如，如果吸烟组中没有女性，那么对照组中也不应纳入女性。此外，不应存在未测量或未控制的混杂变量——如果吸烟者往往比非吸烟者年龄更大，而老年人本身患癌风险更高，那么年龄就是一个混杂因素；如果不对其进行调整，就会导致对处理效应的不准确估计。

上述场景展示了因果推断的复杂性。分析人员通常依靠可视化来检查并解决流程各个步骤中的错误。然而，尽管已有因果推断专家使用的可视化库，但它们通常并非专为因果推断而设计 [29, 66], ，或仅支持有限的分析任务 [5, 23, 53]。不同的因果推断方法采用彼此不兼容的假设和流程 [33, 61, 62], ，这使情况更加复杂，导致可视化工具只能为分析工作流提供零散的支持，且彼此之间无法协同工作。此外，因果推断通常需要多次迭代来完善分析，但许多可视化库并非为快速迭代而设计，也无法与分析人员熟悉的计算环境和统计软件包集成。综上所述，据我们所知，目前尚无可视化工具包能够协同使用，在整个因果推断工作流中支持因果推断从业者的各种分析任务。

在本文中，我们通过与因果推断专家开展一项设计研究来应对这些挑战。在为期三个月的合作过程中，我们首先通过两轮形成性访谈，了解因果推断工作流和分析任务。随后，我们采用迭代式设计流程，开发支持这些任务的可视化，并最终将可视化呈现给专家，征求反馈并进行评估。这项设计研究的成果是*Causalvis*（图 1），这是一个面向因果推断的 Python 可视化软件包。Causalvis 由四个可视化模块组成，可支持数据分析师完成多项任务，例如理解和传达因果结构、识别混杂变量并通过统计方法加以控制、优化队列、探索异质性处理效应，以及追踪分析溯源。

本文的贡献包括：（i）阐述因果推断过程中用户任务的特点及设计要求；（ii）Causalvis[1]，一个由四个可视化模块组成、用于支持专家开展因果推断的 Python 库；（iii）专家对各模块的设计、功能和改进方向提供的反馈；以及（v）在成熟的计算环境中开发和评估包含多个子模块的 Python 软件包所获得的方法论经验。

2 相关工作

2.1 因果推断

因果关系是以因果关系科学为核心的总体范式。从数据中揭示因果结构的方法可进一步分为因果推断和因果发现。因果发现（也称为反向因果关系 [15]）通常旨在从观察性数据中推断变量之间的因果关系，而不预先指定处理或结果；因果推断（也称为正向因果关系）则旨在量化预先指定的处理对预先指定结果的因果关系强度。例如，研究人员会使用因果发现来理解观察性数据集中各变量之间的因果联系，例如吸烟行为、年龄、生物学性别和肺癌；并使用因果推断，在考虑协变量影响的情况下，估计吸烟（处理）对肺癌风险（结果）的影响大小。

因果推断是我们工作的重点，它依赖非实验性观察数据来估计因果效应。与参与者被随机分配到特定处理的随机对照试验不同，观察性研究不会操控个体的环境或经历。相反，研究人员会记录该个体的相关数据，例如人口统计信息、是否接受了某种处理以及其他相关属性。例如，在医疗保健领域，研究人员可以通过医学数据库获取此类观察性数据，其中包含患者的健康状况和人口统计数据。研究人员可以从这些数据集中定义感兴趣的处理和结果，以及相关协变量。

潜在结果（PO）框架 [2], 也称为 Neyman-Rubin 框架 [26] 或 Rubin 因果模型 [30], ，描述了一个用于因果推断的理论框架。再来看一个情境：我们试图找出吸烟对癌症的因果效应。为了衡量吸烟在总体中的真实因果效应，我们需要为每个人生成两个潜在结果——吸烟时的结果和不吸烟时的结果。然后，我们可以量化每个人两种结果之间的差异，并对总体中的这些差异取平均，从而得到**平均处理效应**（ATE）。然而，现实中我们无法获得每个人的两种潜在结果，因为每个人要么吸过烟，要么没有吸过烟。只有一种潜在结果已经被观察到，而另一种结果永远无法被观察到（即反事实）。因此，因果推断为我们提供了理论和工具，用于确定何时以及如何估计因果效应。通俗地说，为了估计平均处理效应，我们选择一组基线生物学/人口统计特征相同的个体作为队列，根据他们是否吸烟将其分为处理组和对照组，并

[1] https://github.com/causalvis/causalvis

进行统计调整，以消除虚假相关，使两组更具可比性，然后计算两组平均结果（即癌症诊断率）之间的差异。PO框架提供了必要的理论假设，用于判断根据现有数据估计出的这一差异是否确实是因果处理效应。

除 PO框架外，还有一种与之互补的因果推断方法，称为结构因果模型（SCM）。SCM 依赖 Judea Pearl 开创的*do*演算系统 [45, 46], ，该系统倡导使用有向无环图（DAG）表示因果关系。有向无环图由节点和顶点组成，其中节点对应数据中的变量，而连接两个节点的有向顶点则表示这些变量之间可能存在因果关系 [21]。通常，有向无环图可用于完成两项任务：1）确定能否根据数据回答因果问题；2）将其用作计算图来估计因果效应。在结构因果模型框架下，从业者会通过估计有向无环图中所有因果关系（顶点）的效应量，利用有向无环图完成这两项任务。相比之下，PO框架的从业者专注于量化连接处理与结果的那条顶点所代表的效应，并且只会将有向无环图用于识别（第一项任务）。这种方法上的差异使两个框架采用了不同的假设，并使用不同的估计工具。尽管两个框架各有优缺点，但如今许多流行的因果效应估计方法——例如倾向得分方法和匹配——都是在 PO框架下发展起来的 [49]。因此，从经济学、政策到医疗保健和流行病学等领域，因果推断从业者对 PO框架的采用更为广泛。

我们的工具旨在支持PO框架，使用户能够使用常见的统计工具，估计预先指定的处理对预先指定结果的因果效应。在后续各节中，除非另有说明，术语因果推断均指PO方法。

2.2 因果可视化

人们开发了许多可视化技术来传达因果结构和因果关系，形式从静态有向无环图 [47, 48] 到动态可视化 [11, 12, 34]不一。近年来，研究开始探讨将文本叙事与因果图相结合，以帮助用户理解时间事件 [7],，以及如何利用可视化来支持因果推断 [36], 、解读反事实 [38], 、识别中介变量 [69], ，并呈现生物通路中的因果关系 [9]。还有研究探讨了可视化如何可能错误地传达数据中的因果幻觉 [68]。

因果推断的一个重要前提是使用有向无环图（DAG）描绘数据中不同变量之间可能存在的因果关系。传统上，这些因果关系通常由领域专家（例如医生或医师）手动指定，并借助 Dagitty 等工具 [58] 实现 DAG 的交互式指定。近年来，人们开发了自动化的因果发现方法，用于从观察性数据 [19]中恢复因果结构，并学习变量之间的因果关系。这些工具可以是完全自动化的，例如 Causalnex [5],；也可以是交互式的，例如 SeqCausal [33] 和 DOMINO [63], ，让人类能够参与从序列数据中发现潜在的因果结构。然而，在这两种情况下，因果发现工具都不会量化处理对结果变量的影响，因此还需要后续使用因果推断方法。

目前已有各种各样为因果推断开发的可视化工具，例如用于可视化和完善因果结构 [61], 、开展分析 [62], 、解释所用人工智能模型 [27], 以及消除人工智能算法偏差 [17]的工具。尽管这些可视化工具支持多种用户任务，但它们大多基于结构因果模型框架，无法兼容更为流行的PO框架。例如，尽管两个框架都可能使用有向无环图来可视化因果关系，但可视化的分析目标有所不同。此前的方法（如Visual Causality Analyst [61] 和Causal Structure Investigator[62], ）使用多个回归模型来估计有向无环图中所有因果关系的效应量（见2.1）。因此，这些工具使用节点之间的顶点来编码效应量是合理的。相比之下，PO框架侧重于估计单一预定义处理对预定义结果的效应量（即单个顶点）。这一较小的范围使分析人员能够采用统计工具中常见的更多样化估计方法，但也要求他们对可能使估计产生偏差的变量进行调整。因此，为PO框架构建的有向无环图需要突出显示图中变量与处理—结果关系之间的联系（例如，通过引入混杂因素）；此前基于SCM的可视化工具并不支持这一任务。

这个例子仅说明了两种方法之间诸多差异中的一种。这些差异意味着，为 PO框架设计的可视化必须支持截然不同的用户任务和分析过程。为此，我们仅发现了三个为 PO框架开发的可视化软件包——VAINE [23], 、causallib [53], 和 Cobalt [22]。VAINE 和 causallib 均为 Jupyter 笔记本环境而设计。VAINE 是一种交互式可视分析工具，可帮助用户在数据集中识别簇，并估算各簇的平均处理效应；而 causallib 则使用静态可视化来帮助用户评估其因果推断模型。相比之下，Cobalt 是一款为 R 设计的可视化软件包，可帮助分析人员验证所选样本是否适用于因果推断。在这些软件包中，只有 causallib 和 Cobalt 提供静态可视化，因此快速迭代和交互式分析会耗费较多时间。此外，每个软件包仅涉及因果推断过程中的有限任务，而且有些任务（例如识别变量类型）没有任何工具支持。为弥补这一不足，我们与专家开展了一项设计研究，以了解用户在因果推断中的工作流程，并构建和评估一个支持其分析的交互式可视化系统。

2.3 计算环境中的可视化

计算笔记本（例如 JupyterLab、Google Colab 和 Kaggle Notebooks）是一种编程环境，用户可以在同一界面中交错编写代码片段并查看输出。

这些笔记本因能够支持探索性数据分析 [39, 57], 协作 [70], 快速迭代和工作流文档编写 [60]而得到广泛采用。近期研究倡导在此类计算环境中开发交互式可视化，以支持可复现性、简化分析流程并提高可视化系统的采用率 [43, 65]。为实现这些目标，研究人员开发了相关工具，帮助用户在 JupyterLab 中嵌入交互式可视化 [65], 创建仪表板 [64, 67], 并复用工作流程 [14] 。另有研究开发了工具，通过交互式数据漫画 [37] 和演示文稿 [71]来精简笔记本，以促进协作[51] 与沟通。除了这些通用工具之外，在笔记本环境中嵌入流畅、交互式小部件也已成为一种趋势 [40]。在数据可视化领域，已有各种交互式笔记本小部件，适用于生物学 [13], 机器学习 [4, 44], 数据比较 [59], 和编程教育 [28]等领域。针对因果推断，VAINE [23] 和 causallib [53] 均为 Jupyter 笔记本设计。受这些工具启发，我们将 Causalvis 实现为适用于计算笔记本的 Python 软件包，使用户能够轻松地与专家协作，并快速迭代其因果推断工作流程。

3 设计研究

为了了解因果推断的过程、用户的任务及其分析目标，我们与八位因果推断专家开展了一项设计研究 [52]。通过两轮访谈，我们首先进行了形成性用户任务分析，以归纳典型的因果推断工作流及相关分析任务。随后，在第二轮访谈中，我们对研究发现进行了验证；同一批专家帮助我们完善并细化工作流，并通过低保真线框图构思能够支持其工作的可视化设计。在接下来的章节中，我们将简要介绍因果推断框架的背景，并描述参与者招募和设计研究的过程。随后，我们将介绍通过访谈归纳得出的因果推断工作流和用户分析目标。

3.1 参与者

我们系统的目标用户是因果推断专家，他们希望在特定应用领域中使用因果推断来估计某种处理对结果的影响。他们熟悉因果推断过程，并且在以往或当前的项目中有使用因果推断的经验。

我们通过滚雪球抽样法招募了专家。我们首先在 Slack 频道发布消息，并通过我们的企业网络联系项目经理。借助这些联系，我们进一步拓展招募渠道，邀请了其他具备因果推断相关专业知识的参与者。最终，我们从不同领域招募了八位专家（E1-8）。参与者需要自我报告其在以下方面的熟练程度，评分范围为 1（没有经验）至 5（我认为自己是专家）：Python（μ=3.86， σ=0.690）、JupyterLab（μ=3.29， σ=1.38）、因果推断（μ=4.14， σ=1.07）、创建/使用有向无环图（μ=3.71， σ=0.76）以及数据可视化（μ=3.86， σ=0.690）。其中一位参与者拒绝提供上述信息。参与者中有三位是顾问，负责为对因果推断感兴趣的客户提供联系支持这些客户拥有领域专业知识和相关数据集，而顾问的职责是进行因果分析，并将结果反馈给客户。参与者中有两位是数据科学家和研究人员，参与因果推断库的开发。他们同时也是领域专家，曾在医疗健康研究中应用因果推断，并且具备开展和报告研究结果的经验。其余三位参与者是从事因果推断研究的研究生，研究内容包括理论性的（非特定领域）模拟，以及与医疗保健（特定领域）相关的应用。

3.2 研究流程

为了了解因果推断的过程以及专家使用的工具，我们首先对因果推断专家进行了形成性访谈。访谈采用半结构化形式，并在适用的情况下，根据用户的回答及其专业领域提出具体的后续问题。每次访谈不超过一小时，每次通话有 1 至 3 位因果推断专家参加。通过这些访谈，我们初步了解了因果推断过程中通常执行的任务顺序以及面临的挑战。我们还概述了专家所涉足的数据领域，以及支持其工作的现有工具和库生态系统。

接下来，我们创建了一个初步的三步工作流程，用以概述因果推断过程。我们还绘制了低保真线框图，展示可能用于支持工作流程各步骤的可视化工具。这些线框图包括简单的原型、草图，或因果推断研究论文中已发表的可视化截图。对于每种可视化设计，我们都提供了多个备选方案，并在图像上添加注释以说明交互方式。我们还将这些设计与 JupyterLab 中编写的伪代码截图结合起来，以便更好地体现这些设计将如何作为可视化模块集成到计算环境中。

最后，在第二轮访谈中，我们将工作流程和线框图展示给专家，以进行验证并收集反馈。访谈过程中，参与者提出了具体意见，帮助我们加深对工作流程的理解。参与者还指出了可在线框图中实现的特定功能和改动。根据这些反馈，我们进一步完善了三步工作流程，使其更好地体现因果推断过程（见第 3.3 节）。我们还为 Causalvis 制定了一组设计目标，以体现用户的需求和要求（见第 3.4 节）。以下各节将介绍工作流程和设计目标。

3.3 因果推断工作流程

此处介绍的因果推断工作流程概括了我们访谈的所有因果推断专家共同经历的过程（图3）。因果推断工作流程从一个观察性数据集开始，可概括为三个主要步骤：1）因果结构建模；2）队列构建/细化；3）处理效应探索。下文将详细介绍这三个步骤。在适当之处，我们会以引号和斜体呈现专家的评论。

3.3.1 因果结构建模。 因果推断的第一步通常是因果结构建模。这一步的目标是让

因果推断分析人员准确地建模数据集中变量之间的因果关系，并识别在分析过程中必须进行调整的变量。

在这一步中，因果推断分析人员通常“从因果图入手”(E3)。因果图是**有向无环图**（有向无环图），其中节点代表变量，从节点 M 指向节点 N 的有向边表示 M 很可能是 N 的原因。根据该图，分析人员会找出需要调整的变量，以满足不存在未测量混杂因素这一假设。混杂因素是同时影响处理和结果的变量；如果不对其进行调整，就可能导致处理效应估计产生偏差。只有识别并调整所有混杂因素，估计出的处理效应才有效。除了混杂因素外，还必须识别其他几类变量：**中介变量**、**碰撞变量**和**预后因素**，这些是最常见的几类 [25]（这些变量类型的图示见图 2）。并非所有变量类型都应进行调整，而且因果推断分析人员“非常关注所使用的变量”(E2)，因为这些变量会用于后续分析。例如，调整碰撞变量和中介变量（又称处理后变量）反而可能使估计产生偏差。

因此，准确建模因果结构并识别不同的变量类型，是这一步骤的关键结果。在形成性访谈中，因果推断专家提到，他们通常会与领域专家合作；这些领域专家“对什么是相关的有一些直觉”（E3），也了解变量之间的因果关系。然而，领域专家并不总是具备从零开始指定此类图所需的因果推断专业知识。他们可能不了解“什么是混杂因素或中介变量”（E1），而且让他们将自己的直觉表达为有向无环图可能既困难又耗时。因此，因果推断分析人员可能依赖自身知识，或使用 Causalnex [5]，等因果发现软件包，先创建有向无环图的“草稿”，再将其呈现给合作者，以便反复验证连线并完善图。不过，我们访谈的专家希望有一个系统能够更好地支持这一过程，促进“以交互方式构建因果图”（E2），并作为“与领域专家合作者展开对话的基础”（E5）。只有在确认因果结构得到了准确表示之后，分析人员才会使用有向无环图来识别因果推断后续步骤中需要进行调整的变量。



图 2：常见变量类型及其与处理变量和结果变量的关系。

3.3.2 队列构建/细化。 因果推断的下一步是从观察性数据集中选择由处理样本和对照样本组成的队列，以估计处理对结果变量的影响。为了使队列能够产生有效的因果效应估计，它应满足**正值性** 假设。通俗地说，该假设认为，队列中的处理组和对照组不应彼此过于不同。更正式地说，它指出每个样本处于任一处理条件下都必须具有正概率，这意味着处理组和对照组的协变量分布应存在重叠。根据分析人员用于估计处理效应的模型，这一点既可以强制实现，也可以进行检验。

这一步主要有两种建模方法：匹配和逆倾向加权。**匹配** [50] 是一种直观的方法，可用于创建具有可比性的处理组：将处理组中的每个个体与对照组中协变量值相近的个体进行匹配。这种方法通常会在构建过程中使各组具有可比性，但也可能因此舍弃许多样本。另一种常见方法是利用整个数据集估计平均处理效应，即**逆倾向加权** （IPW） [49]。**倾向得分** 是某个单位被分配到处理组的概率。随后可利用倾向得分的倒数构造一个加权伪总体，使各组的特征分布相似。由于我们使用整个队列，且处理分配并非随机，因此无法保证正值性假设成立。

在形成性访谈中，专家提到，无论采用何种方法或研究何种数据领域，他们通常首先会检查是否存在正值性违背，因为“如果存在正值性问题，就无法对这个队列开展分析”（E5）。专家通常通过比较各组的倾向得分分布来进行检查[53]，因为从理论上讲，倾向得分是协变量空间的充分概括 [49]。如果处理组和对照组的倾向得分分布没有重叠，分析人员就会识别并排除队列中的不平衡样本，反复调整队列，直至各组达到良好平衡。

3.3.3 处理效应探索。 在因果推断的最后一步中，分析人员希望探索并识别队列中的异质性处理效应。一些因果效应估计方法，如匹配，能够计算数据集中每个样本的个体处理效应，从而使分析人员能够“针对你所关心的处理组进行汇总，例如男性和女性、老年人和年轻人”（E2）。随后，分析人员可以探索各亚组的平均处理效应与其他亚组有何不同，并“找到这些高度差异化的亚组”。当条件处理效应在数据的不同层级之间存在差异时，这表明存在**异质性处理效应**。识别异质性处理效应是我们访谈过的许多因果推断专家的重要任务，因为估计整个队列的平均处理效应并不总能提供对因果效应的充分洞见。相反，比较不同的亚组可以帮助他们更好地理解数据，例如识别处理对哪些总体最有效，或考察“具有不同特征的总体”（E5）的结果。

3.4 用户分析任务

根据对因果推断专家进行的两轮访谈以及前文所述的用户工作流程，我们发现，参与者经常使用 JupyterLab，并希望留在计算环境中“以方便操作”(E1)。

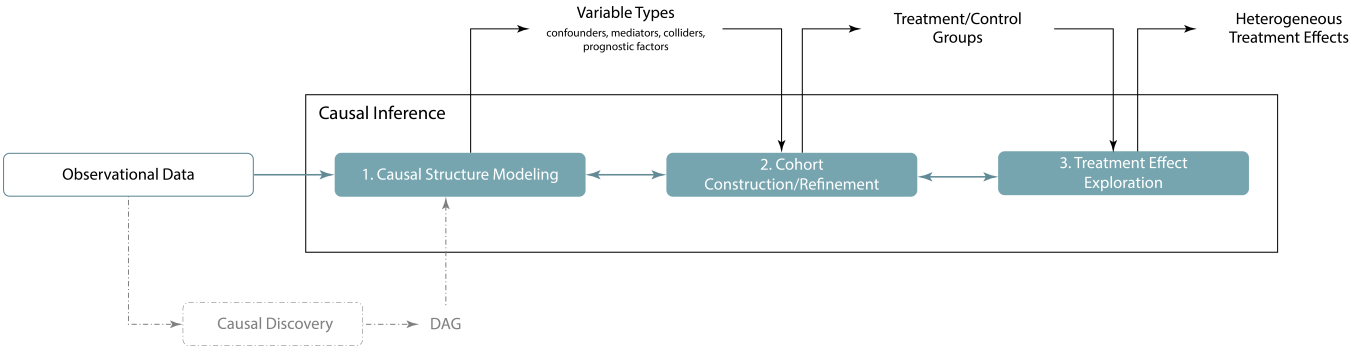

图 3：因果推断工作流程从一些观察性数据开始，包含 3 个主要步骤：1）因果结构建模，2）队列构建/细化，以及 3）处理效应探索。双向箭头表示迭代步骤，分析人员会返回较早的流程以完善分析。各步骤的输入和输出标示在对应步骤上方。对于研究人员而言，每个步骤的目标都是获得相应输出，随后这些输出可作为后续分析的输入。我们还纳入了一个可选的因果发现步骤，有时可用于从数据集中自动生成初始有向无环图，之后领域专家会在因果结构建模过程中对其进行细化。

此外，我们还归纳出八项分析任务，这些任务应得到我们的可视化软件包的支持。对于每项任务，我们都会指出它是否仅与上述工作流程中的某个步骤相关。随后，这些任务被用于指导 Causalvis 可视化模块的设计。

T1 协作创建并交流因果结构。 [因果结构建模]我们采访的许多专家都提到，因果推断通常是一项协作工作，需要熟悉数据的领域专家参与，但他们可能缺乏因果关系方面的专业知识。因此，Causalvis 应便于交流数据集的因果结构。此外，即使没有技术专长或因果关系方面的深入知识，也应能轻松创建和修改因果结构。

T2 保持因果结构独立于特定数据集。 [因果结构建模] 数据科学家与领域专家关于因果结构的讨论有时会在数据集尚未提供之前就开始，因为这些讨论可能有助于确定究竟需要哪些变量。他们也可能希望纳入已知相关的其他变量。因此，Causalvis 应"不受数据集本身的限制"(E1)，并支持独立于任何数据集对因果结构进行建模。

T3 识别不同类型的变量。 [因果结构建模] 鉴于在因果推断过程中选择恰当控制变量的重要性，应着重关注混杂因素、中介变量、碰撞点和预后因素等变量类型，将它们视为"需要留意的事项"（E3）。Causalvis 应帮助用户快速识别不同类型的变量，并对其进行跟踪。

T4] 检查协变量平衡和正值性违背。[队列构建/细化]在队列细化过程中确保处理组和对照组具有可比性，是获得处理效应无偏估计值的一项关键任务。Causalvis 应帮助用户识别何时协变量平衡和正值性假设遭到违背。此外，Causalvis 应允许用户选择需要排除的样本，以满足正值性假设。

T5 估计以某个变量为条件的处理效应。 [处理效应探索] 仅估计总体的处理效应往往是不够的。数据分析师希望以某些变量为条件进行分析，从而深入了解处理效应在不同亚组中的差异。Causalvis 应支持在队列中探索和识别异质性处理效应。

T6 支持灵活且迭代的工作流程。所有受访专家的因果推断过程都具有高度灵活且迭代的特点。专家经常将分析步骤描述为"迭代式"。来自不同领域的专家也可能采用不同的因果推断方法、跳过工作流程中的某些步骤，或优先处理某个阶段，同时在其他阶段运用启发式方法。因此，3.3 节中确定的三个步骤既不是规定性的，也不是单向的。Causalvis 应允许用户反复执行每个步骤，并在需要时返回先前的步骤，以完善其流程。

T7 跟踪分析溯源。 由于使用观察性数据进行因果推断时，通常没有既定的真实基准，分析人员往往会在多个假设、队列以及处理效应值估计之间反复迭代。因此，他们"支持版本控制"（E5）。Causalvis 应提供一种方法，用于跟踪和比较不同分析决策下的结果。

T8 与现有因果分析软件包集成。通过前期用户访谈，我们发现因果推断没有统一的方法，而专家希望 Causalvis 能与他们现有的数据格式（例如 networkx[2] 图）和库无缝集成

[2]https://networkx.org/

（例如他们已经用于因果推断的 CausalNex [5] 和 causallib [53]）。

4 CAUSALVIS

基于设计研究中确定的用户工作流程（见3.3节）和分析任务（见3.4节），我们开发了 Causalvis 可视化软件包，以支持分析人员完成因果推断过程中的三个步骤。该软件包面向 JupyterLab 计算环境开发，因为我们形成性研究中的所有专家都提到，他们通常使用 Python 软件包和 JupyterLab 笔记本。我们采用迭代式设计方法开展开发工作。除所述的两轮形成性访谈外，在实现过程中，我们还向专家展示了各模块的视频，并非正式地收集阶段性反馈，以改进我们的设计。

Causalvis 软件包由四个可视化模块组成，其中前三个模块（DAG、CohortEvaluator 和 TreatmentEffectExplorer）分别对应因果推断工作流程中确定的三个步骤，第四个模块（VersionHistory）则用于追踪分析溯源。各模块均设计为可彼此独立运行（**T6**）。在设计过程中，我们做出的一个关键选择是专注于每个步骤所需的可视化（见图3），而将数据处理和统计调整技术留给用户自行决定。因此，我们转而强调每个模块与现有库和数据结构无缝集成的能力。

Causalvis 前端使用 React 框架以 JavaScript 实现[3]。所有可视化均使用 D3.js 实现[4]。每个模块都集成到 JupyterLab[5] 计算环境中，并使用 Python 和 IPywidget 框架[6]。在接下来的章节中，我们将详细介绍这些模块。在适用之处，我们会说明每个模块提供的函数、接受的参数以及可从模块类访问的变量。对于每个模块，我们还会讨论目前可供分析人员使用的可视化软件包。我们将介绍 Causalvis 如何创新并扩展这些现有解决方案，重点说明我们的模块支持的额外任务；这些任务的设计参考了与因果推断专家进行的形成性访谈。

4.1 使用场景

在介绍以下各节中的可视化模块时，我们将同时演示一个使用场景 [32]，以展示并说明 Causalvis 如何应用于真实的因果推断任务中。我们在该场景中使用 UCI 学生表现数据集 [8] [7]。该数据集记录了两所葡萄牙学校学生全年的数学成绩及相关数据。其中有 30 个属性用于记录学生的人口统计、社会和学业信息（例如年龄、住址和缺勤），另有 3 个属性用于记录学生全年的成绩（*G1*、*G2*、*G3*）。我们从数据集中删除了 7 个敏感或开放式属性（学校、性别、年龄、母亲职业、父亲职业、原因、监护人），最终保留 26 个属性和 395 个样本（行）。

该数据集此前曾被用于构建预测学生学业表现的机器学习模型 [8]，并展示因果发现软件包（如 Causalnex）的使用 [5]。我们选择它作为 Causalvis 的使用场景，是因为其数据领域直观易懂，且已有的使用案例可作为因果推断分析的基线。我们关注的处理属性是缺勤，研究发现它是学生表现机器学习模型中的一个预测因素 [8]。然而，先前的研究尚未估计其因果效应的大小。在该使用场景中，我们根据数据集的中位数将缺勤 转换为二元变量：具有 $\geq median$ 缺勤的学生被编码为频繁缺勤（缺勤 = 1），而具有 $< median$ 缺勤的学生则被编码为很少或没有缺勤（缺勤 = 0）。我们关注的结果属性是*G1*，即学生在当年第一次考试中取得的成绩。

该场景将遵循第 3.3 节介绍的三步工作流程。在整个示例中，我们可能会使用其他软件包（例如 causallib）来执行因果推断所需的统计计算。这些外部软件包不是 Causalvis 的必需项或依赖项。分析人员可以使用自己熟悉的其他统计软件包，只要将数据以预期的数据格式传递给 Causalvis 即可。

4.2 有向无环图

DAG 模块（图 4）旨在帮助用户利用有向无环图（DAG）可视化，快速有效地建模不同的因果结构。在因果结构建模步骤（见第 3.3.1 节）中，用户希望理解因果关系，并识别需要进行调整的变量。他们通常会与领域专家合作，反复完善有向无环图并构建不同的假设。尽管一些现有软件包（例如 Causalnex，见 A.2）允许用户通过指定节点及节点之间的连线来建模因果结构，但这一过程只能通过编写代码手动完成，生成的有向无环图可视化也无法通过直接操作进行交互式编辑。这使得创建和完善有向无环图的过程过于耗时，尤其是在与领域专家进行不频繁的会议协作时更是如此，因为这些专家可能没有足够的时间或技术专长来以编程方式完善有向无环图。DAGitty [58]（见 A.1）是一款支持交互式编辑有向无环图的工具。然而，该应用要求用户了解因果结构术语（例如混杂因素、条件独立性），而这对于通常并非因果推断专家的领域专家来说可能并不熟悉。我们的 Causalvis 有向无环图模块通过支持在 JupyterLab 计算环境中直接进行交互式有向无环图建模，扩展了这些现有工具的功能。通过允许用户直接在可视化界面上创建和完善有向无环图，我们促进了因果推断分析人员所需的交互式协作因果结构建模过程（见 3.3.1 节，**T1**）。此外，我们还实现了自动变量类型识别，以支持后续分析，并提供图像下载功能，以便与主题领域专家轻松共享和沟通 **(T1)**。

正在初始化有向无环图模块。 DAG 模块可以通过多种方式进行初始化，其中最简单的方法是使用DAG() 创建一个空白画布，而无需先获取现有数据

[3]https://reactjs.org/ [4]https://d3js.org/ [5] https://jupyter.org/ [6] https://ipywidgets.readthedocs.io/en/stable/ [7] 我们使用了数学数据集。

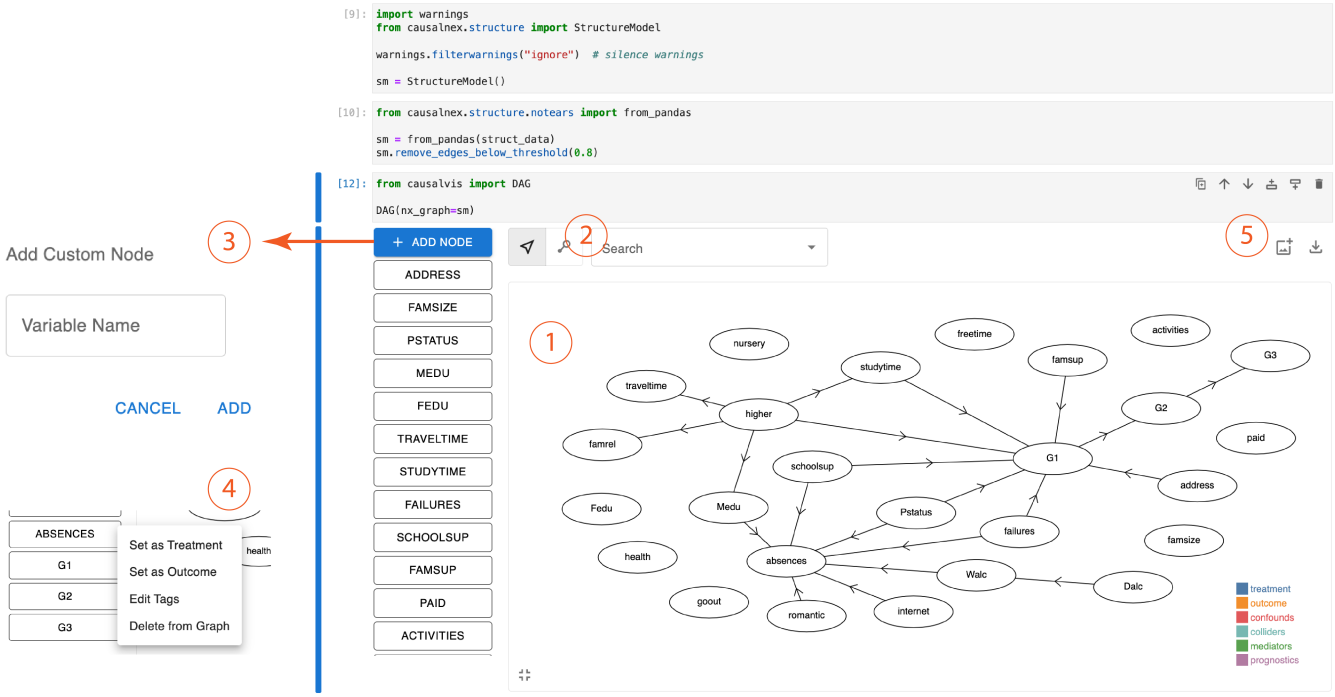

图 4：使用 networkx 图初始化的 DAG 模块。1）加载时会显示该图。2）可以使用切换按钮在布局编辑和连边编辑之间切换。3）可以使用添加节点 按钮向有向无环图中添加自定义节点。4）可以通过每个变量的上下文菜单设置处理变量和结果、编辑标签，以及从可视化中删除变量。5）可以将有向无环图下载并保存为图像或*.json* 文件。

集 **(T2)**。此外，该模块还接受以下参数： 属性、图、数据和 nx_图，让用户能够根据现有数据输入灵活地创建和编辑有向无环图。这些不同的输入格式旨在支持用户协作，使分析人员能够快速加载预先创建的因果图 **(T1)**。某些输入格式还有助于将 DAG 模块与现有 Python 软件包集成 **(T8)**。例如，Causalnex 包会以 networkx 数据格式输出因果结构，可直接通过 nx_图 参数传入该模块。

创建和编辑有向无环图。 创建或编辑有向无环图时，用户可以点击左侧列表中的变量名称来添加节点。如果模块初始化时没有变量，或者用户希望记录其他因素之间的关系，可以使用添加节点 按钮将新节点添加到画布上（图 4 ③）。这使用户能够快速、灵活地记录有关因果关系的领域知识，而不受任何现有数据集的限制**(T2)**。用户还可以在移动 和编辑连线按钮之间切换，从而重新排列节点布局或编辑有向无环图中的连线（图 4 ②）。通过支持对有向无环图的直接交互，用户可以快速迭代不同的假设**(T6)**，而主题领域专家即使没有技术编程知识，也可以协作构建因果结构模型**(T1)**。

识别变量类型。列表中的每个变量都有一个上下文菜单，其中包含以下选项：设为处理变量、设为结果变量、编辑标签和从图中删除。设为处理变量和设为结果变量选项可将特定变量指定为关注的处理变量或结果变量，并更改 DAG 中相应节点的颜色（图 4 ④）。每个 DAG 中只能有一个处理变量和一个结果变量，且同一个变量不能同时充当两者。此外，选定处理变量和结果变量后，其他所有节点都会自动着色，以突出显示不同的变量类型：混杂因素、碰撞变量、中介变量和预后变量。系统会通过递归遍历节点-链接结构中的目标—来源关系（图 2）来识别变量类型；每当用户编辑 DAG 或处理变量、结果变量时，系统都会动态更新这些类型。此类高亮有助于用户识别后续分析**(T3)**中必须进行调整的变量。

保存和分享有向无环图。 创建有向无环图后，用户可以点击下载图像 按钮**(T1)**（图 4 ⑤），将有向无环图下载为*.png* 图像，以分享因果模型。或者，也可以点击下载 *json* 按钮，将节点-链接结构保存为*.json* 文件进行分享。此文件可以自定义，以包含有关已识别的不同变量类型的信息。

在 Python 中访问有向无环图和变量类型。希望在后续分析中使用 DAG 模块输出结果的数据分析人员，也可以在 Jupyter 笔记本中快速访问相关数据变量，而无需下载*.json* 文件**(T8)**。可以使用.DAG获取构建的因果结构，也可以使用以下属性访问不同的变量类型：.confounds、.colliders、.mediators和.prognostics。这些变量类型可以

随后用于分析，以确定需要进行哪些统计调整**(T3)**。

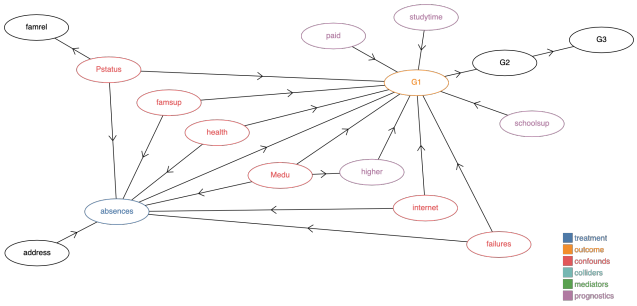

图 5：学生学习数据集中因果关系的有向无环图，由图 4 修订而来。此图像直接通过 Causalvis 有向无环图模块的下载图像 **功能下载，未经进一步处理。**

4.2.1 使用场景。 我们首先构建学生表现数据集的有向无环图模型，以可视化不同属性之间预期的因果关系。为此，我们可以使用 Python pandas 包将数据集加载到 DataFra me 中，并将其传入 DAG 模块，从头开始手动创建有向无环图。不过，更有效的方法是利用 Causalnex 包自动初步探索可能的因果结构。随后，我们将 Causalnex 图加载到 DAG 模块（图 4）中，删除无关节点和连线，根据领域知识添加连接，并识别变量关系。经过反复编辑图后，我们得到修订版有向无环图，它呈现了我们关于影响学生缺勤和考试成绩的因果结构的假设。接着，我们将缺勤 设为处理变量，将*G1* 设为结果变量。有向无环图会自动更新，以突出显示其他相关变量类型（图 5）。由此，我们可以确定有六个混杂变量*(Pstatus*、*famsup*、健康状况、母亲受教育程度、*internet* 和 *failures)* 以及四个预后变量*(paid*、*studytime*、*schoolsup*、*higher)* ，这些变量应在后续因果推断分析步骤中进行调整。

4.3 队列评估器

队列评估器模块（图 6）旨在帮助用户验证所选队列是否满足正值性假设，并在必要时优化队列（见 3.3.1）。进行因果推断分析时，治疗组和对照组的协变量分布应尽可能相似，以减少偏倚协变量的影响。该模块包含的倾向评分图和绝对标准化均差图（aSMD 图，也称 Love 图）已广泛应用于因果推断 [20], , 也是 Python 中的 causallib [53] 和 R 中的 Cobalt [22] 等因果推断工具包的一部分。在 causallib（见 A.3）中，只有使用逆概率加权方法时才支持这些可视化，这可能会排除其他方法的使用，例如匹配（见 3.3.2）。Cobalt（见 A.4）支持这两种方法，但没有对应的 Python 或 Jupyter 笔记本版本。此外，尽管 Cobalt 支持自定义倾向评分和协变量分布可视化，但仅提供静态可视化，且每个图都必须通过单独的函数调用生成。因此，这一过程可能非常耗时，尤其是在需要多次迭代优化队列时。Causalvis 队列评估器模块通过交互式倾向评分图和 aSMD 图弥补了这些不足，用户可以选择不平衡样本，并在后续的队列优化迭代中将其排除。模块还会自动可视化不平衡协变量的详细分布。最后，整个模块均针对 Jupyter 笔记本环境实现。

评估倾向性平衡。 倾向评分图展示了处理组和对照组的倾向评分分布（图 6 ①）。在满足正值性假设的队列中，这两种分布应当有所重叠。缺乏重叠表明，总体中的某些个体无法进行比较，因为他们在另一个暴露组中没有有效的对应个体**(T4)**。

识别违反正值性的样本。 如果倾向评分图中处理组和对照组的分布缺乏重叠，通常需要识别这些个体并将其从队列中排除。可以通过在倾向评分图上拖选来选中违反正值性假设的亚群体。随后可使用 下载 按钮下载所选样本，并将其从队列中排除，从而确保后续所有分析均满足正值性假设 **(T4)**。除了下载所选样本外，还可以在 Python 笔记本中使用 .selection访问这些样本。相反的选择也可以通过 .iselection访问。

评估协变量平衡。 aSMD 图是因果推断专家经常使用的另一种可视化方式，用于验证处理组和对照组是否保持良好平衡（图 6 ②）。在平衡良好的队列中，调整后，每个协变量在处理组和对照组之间的标准化均差应接近于零（即每个协变量的分布相似）。通常，aSMD 图中标示的 0.1 阈值用于识别平衡不佳的协变量 **(T4)**。用户可以使用“排序”按钮对协变量进行排序，以识别需要进一步分析的协变量。aSMD 图上方的切换按钮可用于切换到详情视图，在该视图中，每个协变量的分布会分别显示在单独的图中（图 6 ③ & ④）。为了突出显示平衡不佳的协变量，详情视图默认仅包含调整后的标准化均差大于 0.1 的协变量。随后，用户可以使用“显示/隐藏协变量”对话框手动自定义要显示的协变量（图 6 ⑤）。对于每个协变量，我们在上方显示对照组分布，在下方显示处理组分布。为与 aSMD 图保持一致，未调整分布以黑色轮廓表示，而调整后分布则以填充区域表示。我们还添加了一条虚线来表示每个分布的均值，并在坐标轴上用额外的黑点标示调整后的均值。这有助于快速比较调整后的均值，因为需要确保调整后的处理组分布和对照组分布尽可能相似。处理组和对照组均值之间的较大距离



图6：同时使用调整后队列和未调整队列参数初始化的队列评估器模块。1）处理组和对照组的倾向评分分布。2）绝对标准化均值差（aSMD）图，展示各协变量的 aSMD。3）可使用切换按钮查看各协变量的详细分布。4）详细分布视图会自动显示调整后 aSMD 高于 0.1 的协变量分布。5 ）显示/隐藏协变量按钮可用于自定义要显示的协变量分布，包括已实现良好平衡的协变量。

会向用户表明，对于该特定协变量，这两个组并不平衡**(T4)**。

支持不同的因果推断方法。 通过采访因果推断专家，我们发现数据分析师会采用不同方法进行队列构建，例如匹配或逆概率加权（IPW）（参见 3.3.2）。采用匹配等方法时，数据分析师通常会从观察性数据（未调整队列）中选择样本，以形成平衡的处理组和对照组（调整后队列）。然而，采用 IPW 时，只使用未调整队列，并根据每个样本倾向得分的倒数对其加权，从而创建处理组和对照组达到平衡的伪总体。为兼顾不同方法，队列评估器模块可使用 unadjustedCohort 参数初始化，并可选用 adjustedCohort 参数 **(T6)**。如果只提供 unad-justedCohort ，aSMD 图中的调整后 aSMD 值将通过使用倾向得分的倒数对未调整的 aSMD 值加权计算得出。

4.3.1 使用场景。在前一步中，我们使用 DAG 模块对数据的因果结构进行建模，并识别相关协变量（混杂因素和预后变量）。现在，我们通过匹配对这些协变量进行调整，以获得用于处理效应估计的队列。在此示例中，协变量匹配后得到一个包含 328 名学生的初始队列——处理组（频繁缺勤）有 192 名学生，对照组（很少或没有缺勤）有 189 名学生。随后，我们将选定的队列传递给 CohortEvaluator 模块，以确保处理组和对照组在已识别的协变量方面具有可比性。由于我们使用的是匹配方法，因此我们使用unadjustedCohort 参数将原始数据集传递给 CohortEvaluator 模块，而匹配队列则作为adjustedCohort 参数传入。

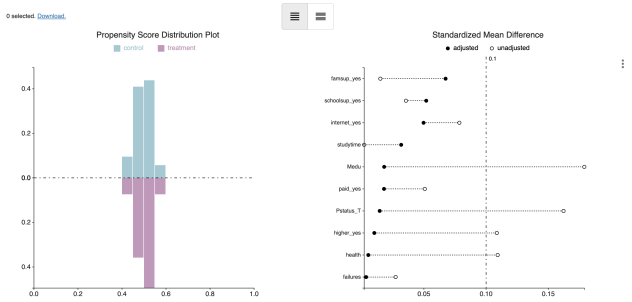

图7：经过优化后选出的学生队列。与图6相比，从 aSMD 图中可以看出，各协变量的平衡性都显著改善，所有协变量的调整后 aSMD 现均低于 0.1（黑点）。

从 CohortEvaluator 模块中，我们可以看到，这些队列并不完全适用于因果分析（图 6）。处理组和对照组的倾向得分在左尾部分具有不同的分布，这表明存在正值性违背。标准化均值差异图还表明，*internet=yes*、母亲受教育程度 和健康状况 变量的调整后 aSMD 大于 0.18。事实上，*internet=yes* 的调整后 aSMD 大于其未调整的 aSMD，这

[8]互联网=是 指学生家中是否有互联网，母亲受教育程度 指母亲的受教育程度，而健康状况 指学生的健康状况。

表明匹配队列扩大了处理组和对照组在*internet=yes*变量上的差异。总的来说，CohortEvaluator 模块中的可视化结果表明，该队列的处理组和对照组并不够相似（即不重叠），这可能会导致处理效应估计出现偏差。

我们回到匹配过程，采用更严格的匹配参数。经过这一优化后，我们获得了一个较小的队列，其中包含200个样本——处理组（频繁缺勤）95个，对照组（很少或没有缺勤）105个。该队列被传递给队列评估器模块（图7）。我们可以看到，处理组和对照组的倾向得分分布现在更加相似。标准化均差图进一步支持了这一点，该图表明所有变量的调整后aSMD均低于0.1。这是一个平衡良好的队列，不存在正性违背，现在可将其用于后续的处理效应估计。

4.4 Treatment Effect Explorer

因果推断的目标通常是估计某一特定处理对关注结果的平均处理效应（ATE）。虽然平均处理效应是针对整个队列计算的，但比较不同亚组之间的效应也可能很有用。如果平均效应在亚组之间存在差异（例如男性和女性），我们称之为存在异质性治疗效应。

识别异质性有助于得出更精确的结论**(T5)**。然而，据我们所知，目前尚无专门为处理效应探索开发的可视化软件包。通过形成性访谈，我们发现，因果推断分析人员目前会使用matplotlib[9]或seaborn[10]等通用可视化创作工具。这些可视化通常是静态的，并针对每项研究临时制作，因此，深入探索亚组的过程会耗费过多时间。因此，TreatmentEffectExplorer 模块（图 8）旨在可视化受不同变量影响的个体处理效应，帮助用户识别处理效应的趋势在不同亚组之间何时发生变化（见 3.3.3）。请注意，该模块只能与某些因果推断方法搭配使用，例如匹配，因为这些方法能够计算个体处理效应**(T6)**。该模块不适用于 IPW 等方法，因为这类方法只能得出整个队列的平均处理效应。

创建和探索亚组。加载时，TreatmentEffectExplorer 模块默认显示一个可视化图表，展示整个队列中个体处理效应的分布（图 8 ①）。该图表使用雨云图 [1]，，这种图表已被证实能够以最小的失真或误导，传达有关数据分布的统计摘要。图表左侧列出了数据集中可用的变量（图 8 ②）。用户可以点击某个变量名称，按该变量对图表进行分面。最多可通过这种方式选择三个变量。第一个选定变量将沿着 x 轴进行可视化，后续变量则用于以矩阵布局创建小多图。如果第一个变量是二元变量，系统会根据该变量自动将队列划分为两组，并使用雨云图分别展示各组的分布（图 8 ③）。如果第一个变量是连续变量，则会使用散点图展示所有个体处理效应值。对于用于创建图表小多图的后续变量，如果选定变量是二元变量，系统会直接根据该变量划分队列。然而，如果分面变量是连续变量，TreatmentEffectExplorer 模块会先计算变量均值，再据此将队列划分为两个子群体。请注意，变量均值仅是一种启发式方法，用于默认划分队列；用户可以通过相应的滑块自定义此阈值（图 8 ④）。此外，我们还会在滑块旁边直接使用蜂群图展示该连续变量的分布，以帮助用户识别数据中自然形成的子群体，从而确定更合适的划分方式。

识别异质性处理效应。TreatmentEffectExplorer 模块中的所有分面图都使用相同的 x和 y 轴范围，以便轻松比较不同的可视化结果。这也有助于用户识别某个子群体的处理效应何时与其他子群体显著不同（**T5**）。当 y 轴的取值范围包含 0 时，我们还会在可视化中添加一条虚线，以突出显示处理效应相反的子群体；这可能是辛普森悖论的征兆。

4.4.1 使用场景。在上一步中，我们从数据集中选取了一个由处理组和对照组构成的队列。由于我们使用了匹配方法，因此可以计算队列中每个处理组-对照组配对的个体处理效应：比较学生某一年频繁缺勤时的*G1*成绩（缺勤 = 1），与很少或没有缺勤时的成绩（缺勤 = 0）之间的差异。这样，我们便可以探索所选队列中个体处理效应的趋势。我们将个体处理效应值传递给TreatmentEffectExplorer 模块，以探索亚组并识别异质性处理效应。

使用 TreatmentEffectExplorer 模块，我们选择根据互联网接入情况（*internet=yes*）和学生健康状况（*health*）对队列进行分组（见图 8， ③）。这些可视化现按学生健康状况分面展示——健康状况较差的学生（*health* < 3.5）显示在左侧分面中，而健康状况较好的学生（*health* > 3.5）显示在右侧分面中。在四个亚组中，我们可以立即看出，健康状况较差且家中没有互联网接入的学生，其成绩因频繁缺勤而出现了明显下降。相比之下，缺勤对其他亚组的影响不那么显著，这些亚组的分布都集中在 0 附近。这一发现可能会引起家长和教育工作者的兴趣，并促使开展后续研究，探讨频繁缺勤如何影响有健康状况问题的学生的表现。

4.5 版本历史

因果推断是一个高度迭代的过程，数据分析师通常需要测试不同的因果结构，并构建可能导致平均处理效应（ATE）估计值不同的队列。因此，跟踪有向无环图和队列是帮助用户回顾其分析溯源的一项关键任务（**T7**）。然而，据我们所知，目前还没有为因果推断开发的数据溯源跟踪工具

[9]https://matplotlib.org/stable/index.html [10]
https://seaborn.pydata.org/

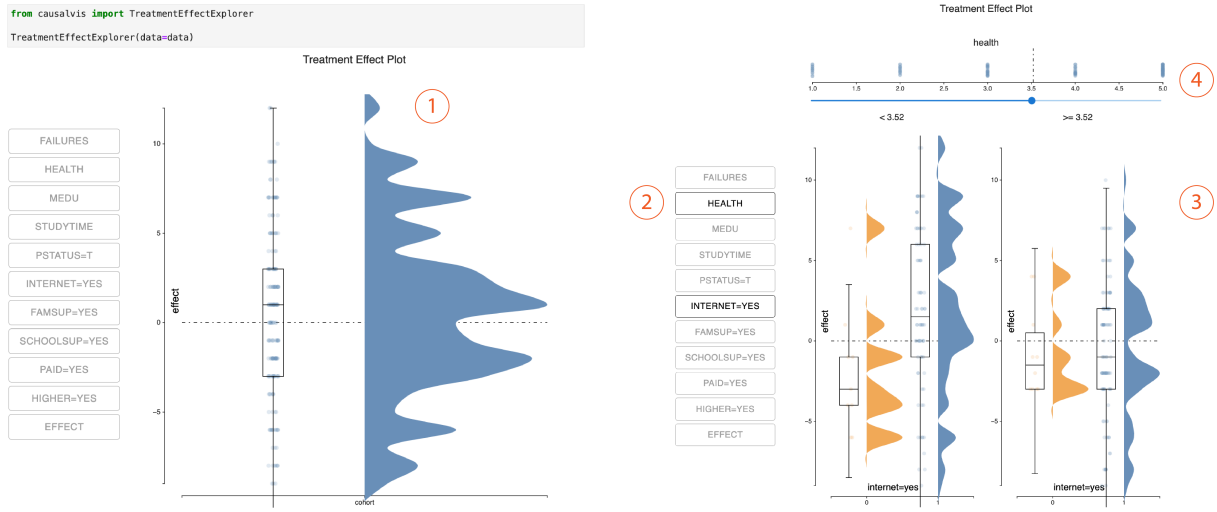

图 8：TreatmentEffectExplorer 模块。1）加载时，模块会显示整个队列中个体处理效应的分布。2）用户最多可以根据三个变量对队列进行分组。3）如果选择了两个或更多变量，可视化将按面进行划分。4）如果选择了非二元变量，可以通过滑块调整分组阈值。

。为此，版本历史模块旨在存储并可视化不同的有向无环图和队列，使用户能够查看其因果推断分析历史，并在必要时恢复以前的版本（图 9）。

使用 VersionHistory 模块跟踪数据溯源。只需调用 VersionHis- tory()，不带任何参数，即可初始化 VersionHistory 模块。这会创建一个空的冰柱图，其中没有有向无环图或队列。随着数据分析师推进因果推断流程，可以调用 .addVersion((DAG, cohort, ATE)) 函数 **(T7)**来跟踪分析溯源。该函数接收一个元组，其中包含有关有向无环图、所选队列以及此轮因果推断工作流程所估计的平均处理效应的信息 **(T6)**。VersionHistory 模块使用冰柱图可视化有向无环图和队列的版本（图 9 ①）。冰柱图下方是一个点图，用于显示每个队列对应的平均处理效应。将鼠标悬停在某个点上，会显示包含有向无环图和队列版本的工具提示（图 8 ②）。每次添加新版本时，VersionHistory 模块中的可视化都会自动更新，用户无需再次运行该组件。

保存和共享版本。 分析结束时，可以使用 . saveVersions() 函数下载所有版本，该函数会将所有有向无环图和队列保存为 *.json* 文件。这样，用户便能轻松恢复分析的早期版本，也能在需要时与合作者共享分析溯源信息 **(T1)**。

5 专家评估

我们通过一项包含 11 位专家的定性研究评估 Causalvis 软件包，以获取关于可视化模块的反馈，并验证这些可视化是否支持已确定的因果推断分析任务（第 3.4 节）。

在未来的工作中，我们还计划通过更深入的部署研究来评估 Causalvis。我们将与因果推断专家合作，考察 Causalvis 在特定应用领域的现实世界分析中会如何使用。这项长期部署将帮助我们更好地了解 Causalvis 如何融入当前的用户实践。就本文的研究范围而言，我们依靠参与者的专业知识来评估 Causalvis模块的可用性，并在实际部署之前发现潜在问题。接下来的章节将介绍研究设计和结果。

5.1 参与者

我们招募了十一位因果推断专家（P1-11，7 名男性/4 名女性）参与评估研究，其中五人也参加了形成性访谈（见第 3 节）。每次研究环节开始前，我们请参与者在 1（毫无经验）到 5（我认为自己是专家）的量表上自我评估其在 Python（μ=4, σ=0.775）、JupyterLab（μ=3.45, σ=1.21）、因果推断（μ=3.64, σ=1.12）、创建/使用有向无环图（μ=3.09, σ=1.04）以及数据可视化（μ=3.36, σ=0.674）方面的熟练程度。所有参与者都是数据科学家，或在工作中承担数据分析任务。他们参与因果推断项目的经验从 3 个月到 4 年不等；其中一位参与者是例外，他没有正式参与过因果推断项目，但曾利用业余时间学习相关内容，并计划在后续项目中应用。十一位参与者中，七位是研究人员，他们开发过支持因果推断分析的工具，或参与过医疗保健领域的因果推断研究项目。两位是顾问，为政府客户提供服务，将因果推断应用于政策相关决策。其余两位是博士生，在博士阶段学习中运用因果推断。

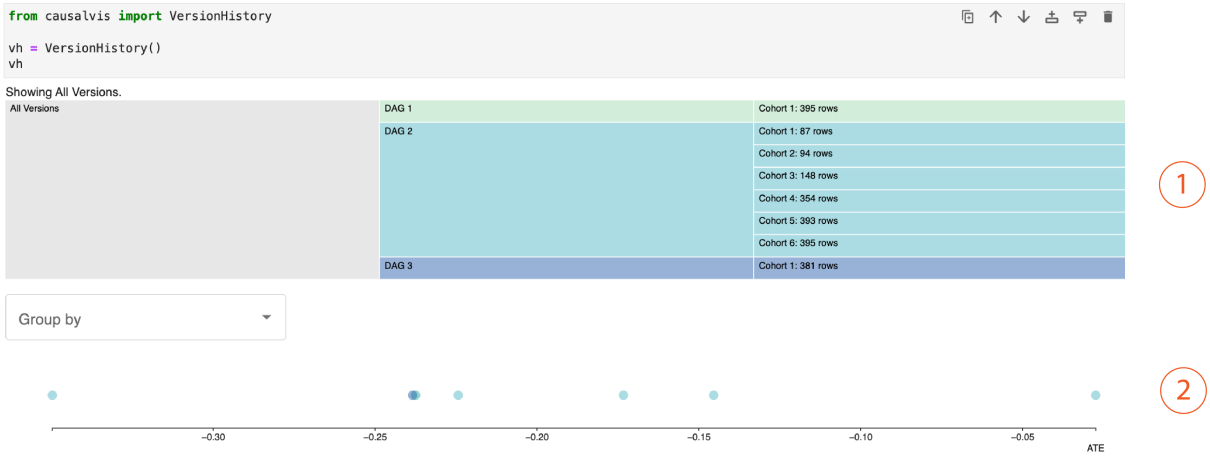

图9：版本历史模块。1）使用冰柱图可视化有向无环图的三个版本，以及为测试每个有向无环图版本中的因果模型而创建的不同队列。2）每个队列的估计平均处理效应。我们看到，估计值大多集中在 −0.25 至 −0.10的范围内。不过，有两个异常队列的平均处理效应非常大或非常小。

5.2 任务

我们在 JupyterLab 中准备了 4 个笔记本，分别演示 Causalvis 软件包的 4 个模块。笔记本中，我们使用同一 UCI 学生表现数据集 [8]，展示了本文所述使用场景的精简版本（见 4.1）。每个笔记本都简要说明了如何在分析中使用相应模块，并提供了示例。在必要时，我们还展示了如何将模块与现有的 Python库 **(T8)** 配合使用。研究期间，我们请参与者依次完成笔记本中的内容，并完成一些引导式任务，例如使用自定义变量列表初始化有向无环图，或在 CohortEvaluator 模块中通过框选倾向得分图来选择样本。这些任务旨在帮助参与者熟悉模块的功能。参与者完成一个笔记本后，我们会征求他们对该模块的反馈以及他们认为可以进行的改进。各场次通过视频会议服务 Webex 远程进行，每场持续一到两个小时。

5.3 结果

总体而言，参与者对 Causalvis 软件包给予了积极反馈。在以下各节中，我们将重点介绍参与者认为 Causalvis 能够有效完成其分析任务（T1-8）的情况，并讨论他们针对每个模块提出的其他反馈和建议。

5.3.1 有向无环图。**DAG 模块促进了沟通与协作（T1）**。参与者认为，将有向无环图保存并分享为图像，有助于传达数据中的因果关系。正如 P4 所说，在为她的因果推断研究撰写论文时，“有一张好的可视化图表始终相当重要。”同样，P7 也指出，有向无环图中因果关系的可视化呈现，是与非数据科学家受众沟通的重要方式。数据科学家需要了解因果结构和因果推断过程的细节，而外部协作者或领域专家“只想知道”协变量之间的关系，并看到结果。”除了可视化本身的有效性之外，参与者还认为，DAG 模块改进了现有工具。例如，P9 谈到撰写论文时不得不在 PowerPoint 和 Photoshop 中手动创建有向无环图，最后表示“这样会好得多。”

在 DAG 模块中自动识别不同的变量类型，对参与者（T3）很有帮助。研究期间，许多参与者认为，选定处理变量和结果变量后，DAG 模块会动态突出显示重要变量类型，这一功能很实用。正如 P1 所说，“它会立刻给所有内容上色，这样也能一目了然地看出哪些节点具有什么背景。”总体而言，参与者认为，这项功能对那些不太熟悉因果推断、“不知道哪些是混杂因素、碰撞变量和中介变量”（P9）的用户很有帮助；对于数据分析师也同样有用，他们或许会发现“图中某处有我们之前没意识到的混杂因素或中介变量”（P4）。此外，参与者也喜欢在 DAG 中使用颜色编码来突出显示各类变量。通过提供视觉反馈，使变量类型一目了然且易于快速识别，DAG 模块更好地支持了选择需要调整的变量这一任务。正如 P10 所说，“用视觉方式读懂一张图需要花一些时间”，因此“这种自动标注非常有用”。不过，尽管许多参与者认可这项功能，也有人提醒不要过度依赖自动识别变量。P9 称之为“一把双刃剑”，并担心变量可能会被错误突出显示，因为有人“可能会创建一张将变量标注错误的 *DAG*。”如果调整了错误的变量，后续的因果推断分析就可能因此产生误差。

参与者喜欢能够快速、交互式地在 DAG 模块上进行迭代。与当前用于创建有向无环图的技术或工具相比，许多参与者表示，他们更偏好 DAG

模块所提供的交互式功能。例如，P1 表示，“它提供了手动编辑的选项，这一点很好，你不必把所有内容都用某种 *YAML* 之类的东西写出来。”同样，P2 之前有通过 R 学习有向无环图的经历，但他评论说，当时的可视化并不具备交互性，而这“本来会非常有用。”在讨论过程中，P9 进一步阐述了他对交互性的偏好。此前，他曾使用 PowerPoint 和 Photoshop 等工具创建有向无环图，但觉得“不断迭代某个东西才是真正让人恼火的地方。”相比之下，在使用 Causalvis 有向无环图模块时，他很欣赏“能够快速迭代，并添加或移除节点和关系，”而交互式界面使这一点成为可能。

参与者希望 DAG 模块能提供更多注释功能。参与者经常提出的一项建议是，为节点之间的边添加注释。对于数据分析人员而言，这项功能有助于他们更好地理解数据集。正如 P1 所说，在边上添加相关系数会很有帮助，因为“因果关系不仅仅是结构，还包括对事物之间关系的一些数值估计。”这些信息有助于她判断某个混杂因素是否与处理和结果都高度相关，因此在调整过程中是否应优先处理。相比之下，其他参与者希望添加注释来解释关系或与合作者沟通。例如，P9 希望“添加信息来描述这些路径，而不只是画一支箭头。”同样，P11 通常每隔一两周才与合作者会面一次，因此希望标出存疑的连线，并指出有向无环图中需要“由领域专家审查”的部分。总之，对于在因果推断过程中经常需要与领域专家合作的数据科学家而言，注释有助于促进协作。P10 的这番话对此作了很好的总结：*“[*注释*]* 并非因果推断的一部分。这些只是用于沟通。但在客户场景中，沟通很重要，所以我认为这很有用。”

5.3.2 队列评估器。 **CohortEvaluator 模块中的详细协变量视图，进一步揭示了协变量平衡和阳性假设违背（T4）情况下的数据分布**。研究中的许多参与者都很喜欢在详情视图中分别查看每个协变量的分布（图 6 ④），并认为这为因果推断中更常用的倾向得分图和 aSMD 图增添了实用功能。参与者从以下几个方面描述了这张附加图表的好处。正如 P1 所说，“例如，有了 *aSMD* 图，你只能看到平均值。我觉得查看整个分布更好，因为它能提供更多信息。”P2 进一步谈到了自己以往的经历，他评论道：“每当我做倾向评分匹配时，如果想查看探索性图表，我就会逐一查看 *[*协变量*]*。”相比之下，谈到使用 CohortEvaluator 模块时，P2 表示：“不用为获取这些图表分别编写代码，我觉得这很好。”

参与者从倾向评分图中选择实例后，希望从队列评估器模块获得反馈。在专家评估研究期间，参与者发现，可以在倾向评分图上进行框选以获取未得到良好平衡的样本。例如，P5 表示：“我真的很喜欢在 *[*倾向评分图*]*上滚动，看看这些样本具体是什么。真的非常神奇。”然而，许多参与者希望在完成选择后从该模块获得更多视觉反馈，例如单独呈现所选样本的协变量值。P5 表示：“如果能针对不平衡的选择提供一些可视化，我觉得会非常有意思。用一种直接、快速的可视化方式说明为什么你的数据在这些情况下没有被抽样。”同样，P3 希望将所选样本的协变量与整个队列进行比较，因为“看到分布呈现出的差异会相当有意思。”除了可视化反馈之外，研究中的一些专家还希望 Jupyter 笔记本能更详细地解释这些可视化内容以及如何与其交互。例如，P10 表示，即使在倾向评分图上进行框选之后，“仍然不清楚这些人是谁，”也不清楚哪些协变量使他们区别于队列中的其他人。在这种情况下，他希望笔记本提供下一步应采取哪些分析步骤的示例——“不必改变设计，也不必改变你的软件包。只需在这里添加一些使用案例，说明你可以采取哪些操作。”P11 也提出了类似建议，表示笔记本示例应包括“更多关于你推荐使用哪些软件包的指导”，以便在样本被选中后对其进行特征分析并排除样本。

5.3.3 Treatment Effect Explorer。 **TreatmentEffectExplorer 模块中的交互式可视化减少了参与者反复探索和比较数据中亚组所需的工作量（T5）**。我们研究中的许多参与者认为，这些可视化比他们当前采用的方法有所改进。例如，P1 表示：“将所有内容可视化，而不只是查看数字，这很不错。”同样，P8 “非常喜欢这个三变量可视化，*[*因为*]* 这是一件非常难做到的事情。”参与者还表示，交互式选择不同变量的方式“非常直观”（P4），而且< शैली id='17'> “非常有趣”（P6）。这种交互性有助于探索不同的亚组，并“获得平均处理效应之间的不同比较”（P4）。对于经常与领域专家合作、以识别数据中的异质亚组的参与者而言，这一点尤其有用。在探索过程中，领域专家可能会要求比较不同变量，或根据不同阈值对亚组进行分层。正如 P6 所描述的，使用当前方法时，往往会出现“这种无休止的反复更改”。相比之下，TreatmentEffectExplorer“使用起来很方便”。

TreatmentEffectExplorer 模块支持围绕因果推断进行沟通和讲述。除了分析任务外，不少参与者还提到，“数据科学家的工作就是构思故事”（P11）。尤其是顾问，需要向客户、顾客和合作者传达结果，并帮助他们理解结果。然而，在因果推断领域，这可能颇具挑战。正如 P10 所说，“大多数数据科学家都知道自己需要讲述一些故事，但在因果推断的情境下，我认为由于这个理念还很新，大多数数据科学家不知道如何讲述因果关系，因为这样的故事更难讲。”为此，参与者表示，TreatmentEffectExplorer 模块中的可视化非常适合这一用途。与下载数据并

手动创建可视化相比，P11 很喜欢“这个工具能提供一些现成的图来帮助我们讲故事.....这为我们节省了很多时间。”

应为 TreatmentEffectExplorer 提供更多可视化自定义选项和指导。尽管参与者普遍认可此模块中实现的可视化和交互功能，但并非所有参与者都熟悉雨云图。P5 发现“我花了一点时间才理解这个可视化”，P3 则评论说“不到一分钟，但也不是一下子就明白了。”为了更好地解读可视化，许多参与者希望能根据自己的熟悉程度和专业知识自定义图表。例如，P8 建议将“一些可视化的数据（例如箱线图）设为可开关的选项”；P4 则希望“能够启用或禁用不同的 *[小提琴图]*，因为我通常不会绘制这些图。”最终，参与者希望能获得更多关于如何解读 TreatmentEffectExplorer 模块中可视化内容的说明。P4 建议在可视化中加入“事先提供一本说明手册，让你知道自己看到的是什么”，而 P3 则希望有“一些指引，告诉你要找什么”。

5.3.4 版本历史。**版本历史模块有助于追踪数据溯源，但还应加入更多功能（T7）**。许多参与者提到，记录分析过程是他们希望或需要纳入工作流程的事情。在谈到自己的因果推断研究时，P2 表示：“我绝对需要 *[这样的工具]*。我确实需要记录队列中有多少患者”；P8 则评论道：*“[对于]* 我现在正在做的事情，*[版本控制]* 会非常有帮助，因为我正在用很多有向无环图处理大量数据。”将不同队列及其对应的平均处理效应可视化，还有一个好处：在采用交叉验证方法时，能帮助分析人员评估所估计的平均处理效应是否稳健。正如 P1 所解释的，“你会希望看到平均处理效应基本稳定，处于一个不太大的范围内。这样你就可以说，我们的结果确实稳健，而且模型具有真正的泛化能力，而不是对我们得到的任何数据都过拟合。”不过，也有一些参与者指出，版本历史模块还需要记录更多信息，才能真正发挥作用。在因果推断中使用机器学习模型时，该模块还应记录“机器学习参数”（P11）以及“模型的具体设定”（P6）。

参与者希望版本历史模块能帮助他们比较有向无环图。除了跟踪分析过程外，多位参与者还提到，他们希望版本历史模块能帮助他们比较有向无环图的不同版本。参与者认为，可视化能够减少进行此类比较所需的精力，“让人轻松看懂有向无环图 *1* 是什么、有向无环图 *2* 是什么，以及它们之间有何不同”（P1）。同样，P2 表示，“进行这样的比较，让比较更容易，会非常有帮助。”更具体地说，参与者希望通过可视化比较，找出因果图中的独特结构。例如，P6 表示，他想知道“哪些边只出现在一个图中，而不出现在另一个图中”；P8 则希望该模块“着重向我展示局部结构，也就是变量的实际邻居有哪些，这样我就能更确定它不是路径形式中的碰撞点。”P10 更进一步，建议直接在 DAG 模块中记录版本历史，因为“我编辑时，如果能以某种方式查看历史记录，会很有帮助。”他解释说，能够查看早期版本可以提醒自己与合作者之前讨论过的内容，进而帮助决定需要进行哪些编辑。最终，研究参与者一致认为，比较和对照有向无环图的不同版本将增强版本历史模块。P3 的表述最能体现这一点，他说：“展示出来会是第一个好处，但如果能进行比较就更好了。”

6 讨论

本节将反思 Causalvis 对现有因果推断工作流程的贡献，并讨论其对未来工作的启示。

通过交互式可视化支持快速迭代检验假设。因果推断的过程要求数据分析师多次迭代工作流程的各个步骤，以探索不同的因果结构、完善队列，并探索不同亚组中的异质性处理效应。在我们的评估研究中，参与者认为，与需要通过编程手动编辑静态可视化、过程既耗时又繁琐的现有工具相比，DAG 和 TreatmentEffectExplorer 模块的交互性有助于更快地进行迭代。因此，Causalvis 更好地支持了诸如协作探索不同因果结构假设（**T1**）或比较不同队列亚组中的异质性处理效应（**T5**）等任务。不过，参与者也认为 CohortEvaluator 模块仍有改进空间。尽管用户可以在倾向评分图上刷选处理组和对照组之间不平衡的样本（**T4**），但模块中的可视化并不会对此交互作出更新。用户必须自行从数据集中手动排除选中的样本，并重新运行模块，才能看到可视化更新。许多研究参与者认为这一过程不够直观，并希望可视化能够动态更新，显示所选样本与整个队列之间的差异，从而为调整选择、完善队列提供必要的指导。因此，未来开发因果推断可视化时，应更充分地考虑在整个工作流程中快速迭代的需求，尤其要重视队列构建和完善任务。

向领域专家和合作者解释并沟通因果推断。在因果推断中，数据科学家不仅要完成分析任务，还经常需要与领域专家沟通、发表研究结果，并帮助客户和合作者理解结果。为此，我们评估研究中的许多参与者都认为，Causalvis 中的可视化在分析和沟通方面都很有效。参与者认为，DAG 模块中的保存和共享功能优于现有的因果结构模型可视化工具。他们还喜欢 TreatmentEffectExplorer 模块能够帮助他们在分析过程中迅速识别异质性处理效应，同时向合作者解释结果，并“讲述*[一个]*关于因果关系的故事”

（P10）。综上所述，这凸显了因果推断可视化需要同时支持分析和沟通。我们的评估研究还揭示了使用注释来记录与领域专家讨论内容的潜力。通过形成性研究和评估研究，我们发现，因果推断工作流程中的协作可能并不连贯。数据科学家可能每隔一到两周才与领域专家会面一次，而在此期间，他们必须迅速验证更改并进行必要的完善。在对因果结构建模时，这一点尤为重要，许多参与者建议为 DAG 添加注释功能，以更好地支持这种双向协作。在未来的工作中，研究人员可以进一步探索如何将注释整合到各种可视化模块中，从而更好地满足因果推断过程中的沟通与协作需求。 [56] 和 [7] 等现有研究也可以为这些注释功能的设计提供参考。

通过敏感性分析评估估计稳健性。 因果推断的一个特点是缺乏真实数据，无法据此评估所估计处理效应的准确性。因此，因果推断分析人员通常会进行敏感性分析 [31]，以检验不同分析选择的稳健性和泛化能力。于是，分析人员可能会多次迭代因果推断流程，每次迭代采用不同的协变量组合，或测试不同的队列 [41]。正如 P1 在评估研究中所提到的，版本历史模块可能支持类似敏感性分析的评估（见 5.3.4）。例如，如果平均处理效应在不同协变量和队列下保持稳定（见 9 ②），分析人员就能更有信心地认为其估计可能具有稳健性和泛化能力。这种对合理分析决策进行组合测试的做法，与多宇宙分析 [54, 55], 有许多相似之处；多宇宙分析会自动穷举所有分析决策组合，以确保结果稳健。尽管 Causalvis 的设计具有高度交互性，并不天然适用于穷举式组合搜索，分析人员仍可指定最合理的分析情景并逐一迭代。因此，为评估和审查多宇宙分析结果而开发的许多可视化策略，同样可能适用于可视化所探索的因果推断备选方案子集。例如，版本历史模块可直接作为森林图，用于对比不同设计选择下的平均处理效应。未来，我们希望借鉴现有研究，例如 [42] 和 [35]，了解如何扩展该模块，以帮助因果推断分析人员进行敏感性分析并评估估计稳健性。

7 设计研究的经验教训

回顾与因果推断专家开展这项设计研究的过程，我们发现了一些遇到的挑战，并总结了由此获得的方法论经验。

设计可视化软件包，以补充现有工作流程和任务（T6）。 在设计研究过程中，我们做出的一个关键决定是明确 Causalvis模块应如何融入当前工作流程，并补充现有分析工具，而不是打断或取代它们。为此，我们需要了解当前的因果推断流程和软件包。我们

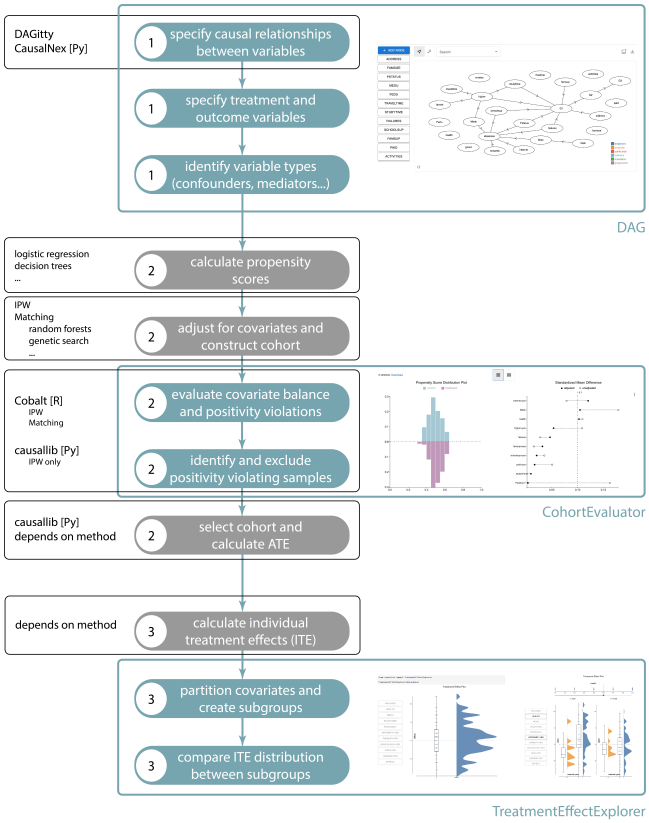

图10：因果推断流程概要。三步工作流程（1-3）中的每个步骤都细分为更具体的分析活动。Causalvis支持的分析活动以绿色突出显示。灰色方框表示现有软件包或算法支持的其他活动。此处列举的只是部分示例，并非现有工具的完整清单。为清晰起见，我们省略了表示步骤之间迭代关系的箭头。

在形成性访谈中，我们与因果推断专家密切合作，逐步完善了对因果推断的理解，并将其归纳为本文提出的三步工作流程（见图3）。我们还找出了当前工具未能充分支持的某些分析任务，并记录了每个步骤所产生的数据格式和分析输出。这项形成性研究为 Causalvis 模块的开发提供了指导。对于每个模块，我们都着重为现有工具支持不足的任务开发可视化，同时整合分析人员开发的外部软件包或算法（见图10）。例如，在 Cohort-Evaluator 模块的开发过程中，我们了解到分析人员使用各种不同的算法来计算倾向得分。因此，我们决定不在模块中重新实现倾向评分计算，而是接受计算好的数值作为输入，以兼容现有用户方法的多样性。本文展示了如何设计用于因果推断的可视化软件包，使其能够补充现有的分析工作流程和工具，而不是打断或取代它们。在因果推断以外的领域中，

也可能存在需要多个步骤且工作流程同样复杂的数据分析任务。因此，本文采用的工作流发现过程或许也适用于这些领域。未来的研究还可以探讨全面的设计决策标准，以便可视化工具能够最有效地补充现有的分析方法和用户任务。

融入计算环境（T8）。在通过 Causalvis 的早期设计进行构思时，我们最初考虑在交互式计算环境之外的独立可视分析系统中实现这些可视化模块。然而，我们最终放弃了这种做法，因为形成性研究中的因果推断专家提到，他们通常在 JupyterLab 计算环境中工作。在不同系统间切换很可能会打断用户的工作流程 [6, 10], ，而在计算环境之外引入独立应用程序，也可能干扰高度迭代的分析。因此，我们决定将 Causalvis 开发为可在 JupyterLab 计算环境中使用的可视化模块。为此，我们必须应对模块设计和评估方式上的挑战。在模块设计过程中，我们从最早的原型和线框图阶段就将 JupyterLab 纳入考量，并使用伪代码截图展示可视化模块将如何融入计算环境。这有助于进一步明确各模块的输入和输出要求。例如，在查看初始线框图后，一位形成性研究中的专家要求 DAG 模块也支持将 networkx 图作为输入，以便他能轻松地将 Causalnex 包的输出直接传递给 Causalvis。同样，在后续评估中，我们创建了笔记本来演示如何在 JupyterLab 中使用这些可视化，并要求参与者完成包含直接修改笔记本内容的任务（见 5.2）。通过这种研究设计，我们收集到了有助于各模块更好地融入 JupyterLab 的反馈。例如，参与者提出了有关数据类型的问题（“为什么要转换为字典？”，*P11*），并对预期输出提出了建议（“我希望这里的输出是数据集中的索引，而不是数据本身”，*P3*）。通过在预期使用环境中开展研究，我们得以评估各模块与计算环境本身的集成效果，以及如何改进应用程序编程接口。

与不同领域和角色的专家合作。在本研究中，我们很幸运，能在三个不同时间通过线上会议获得专家的反馈（详见第 3.2 节）。第一次会议旨在了解用户进行因果推断分析的工作流程；第二次会议则基于原型和草图集思广益并确认功能；最后一次评估旨在评估 Causalvis 的可用性。我们学到的一条重要经验是，招募背景多元的参与者，并从不同视角了解用户工作流程，十分重要。所有参与者都是因果推断方面的专家，但他们在日常数据分析活动中所涉的领域和担任的角色各不相同。向多元化的参与者群体学习，使我们能够针对广泛的使用场景进行设计。例如，顾问重视使用可视化向合作者解释流程，并围绕因果推断“讲述故事”；而医疗保健数据分析师最感兴趣的是比较不同亚组的处理效应。我们推测，如果没有接触到多元化的参与者群体，我们只能提出不完整的解决方案；我们还认识到，在为特定任务设计可视化软件包时，了解不同用户的应用需求至关重要。

8 结论

本文介绍了 Causalvis 的设计研究。Causalvis 是一个用于支持因果推断的可视化 Python 软件包。在为期三个月的时间里，我们与专家密切合作，首先梳理了因果推断工作流程，并确定了相关分析任务。随后，我们采用迭代式设计过程，开发了四个可视化模块，用于支持因果结构建模、队列构建与优化，以及处理效应探索等任务。评估研究的结果表明，设计可视化以支持工作流程各个步骤中的快速迭代十分重要；同时，因果推断工具的设计应兼顾分析与沟通。最后，我们还分享了开发组件化可视化以支持灵活工作流程的方法论经验，以及如何设计和评估可视化，使其能够集成到特定的计算环境中。未来，我们计划进一步探索如何设计用于因果推断的可视化，以支持对有向无环图的可视化比较，以及如何实现注释功能，促进数据科学家与合作者之间的沟通。

致谢

我们谨向所有参与者表示感谢，感谢他们在整个研究过程中慷慨地贡献宝贵时间并分享见解。我们还要感谢匿名审稿人对本文提出的周到而详尽的反馈，并感谢 Georgia Tech Visualization Lab 的成员以及 IBM Research 的研究人员在整个研究过程中提供的宝贵建议、深思熟虑的评论和建设性反馈。本项目部分得到了 NSF IIS-1750474 和 NSF 2247790 的支持。

参考文献

[1] Micah Allen、Davide Poggiali、Kirstie Whitaker、Tom Rhys Marshall 和 Rogier A Kievit。2019年。雨云图：用于稳健数据可视化的多平台工具。*Wellcome open research* 4（2019）。[2] Joshua D Angrist、Guido W Imbens 和 Donald B Rubin。1996年。使用工具变量识别因果效应。《美国统计学会杂志》91，434（1996），444–455。[3] Susan Athey 和 Guido W Imbens。2017年。应用计量经济学的现状：因果关系与政策评估。《经济展望杂志》31，2（2017），3–32。[4] Alex Bäuerle、Ángel Alexander Cabrera、Fred Hohman、Megan Maher、David Koski、Xavier Suau、Titus Barik 和 Dominik Moritz。2022年。Symphony：为机器学习编排交互式界面。载于*CHI*计算系统中的人为因素会议。1–14。[5] Paul Beaumont、Ben Horsburgh、Philip Pilgerstorfer、Angel Droth、Richard Oentaryo、Steven Ler、Hiep Nguyen、Gabriel Azevedo Ferreira、Zain Patel 和 Wesley Leong。2021年。*CausalNex*。https://github.com/quantumblacklabs/causalnex [6] Benjamin B Bederson。2004年。保持流畅状态的界面。*Ubiquity* 5，27（2004），1。[7] Arjun Choudhry、Mandar Sharma、Pramod Chundury、Thomas Kapler、Derek WS Gray、Naren Ramakrishnan 和 Niklas Elmqvist。2020年。可视化中的从前：理解文本叙事在因果关系中的运用。*IEEE*可视化与计算机图形学汇刊 27，2（2020），1332–1342。[8] Paulo Cortez 和 Alice Maria Gonçalves Silva。2008年。使用数据挖掘预测中学生学业表现。（2008）。

[9] Tuan Nhon Dang、Paul Murray、Jillian Aurisano 和 Angus Graeme Forbes。2015年。ReactionFlow：用于生物通路因果关系分析的交互式可视化工具。载于BMC会议论文集，第9卷。BioMed Central，1–18。[10] Niklas Elmqvist、Andrew Vande Moere、Hans-Christian Jetter、Daniel Cernea、Harald Reiterer 和 TJ Jankun-Kelly。2011年。用于信息可视化的流体交互。Information Visualization 10，4（2011），327–340。[11] Niklas Elmqvist 和 Philippas Tsigas。2003年。使用动画生长多边形进行因果关系可视化。载于2003年IEEE信息可视化研讨会（IEEE目录号：03TH8714）。IEEE，189–196。[12] Niklas Elmqvist 和 Philippas Tsigas。2003年。生长方块：因果关系的动画可视化。载于2003年ACM软件可视化研讨会论文集。17–ff。[13] Nicolas F Fernandez、Gregory W Gundersen、Adeeb Rahman、Mark L Grimes、Klarisa Rikova、Peter Hornbeck 和 Avi Ma'ayan。2017年。Clustergrammer：用于高维生物数据的基于Web的热图可视化与分析工具。Scientific data 4，1（2017），1–12。[14] Kiran Gadhave、Zach Tyler Cutler 和 Alexander Lex。2021年。复用交互式分析工作流。（2021）。[15] Andrew Gelman 和 Guido Imbens。2013年。为什么要追问原因？正向因果推断与逆向因果问题。技术报告。美国国家经济研究局。[16] John Gerring。2005年。因果关系：社会科学的统一框架。《理论政治学杂志》17，2（2005），163–198。[17] Bhavya Ghai 和 Klaus Mueller。2022年。D-BIAS：用于应对算法偏差的基于因果关系的人在回路系统。arXiv预印本 arXiv:2208.05126（2022）。[18] Thomas A Glass、Steven N Goodman、Miguel A Hernán 和 Jonathan M Samet。2013年。公共卫生中的因果推断。《公共卫生年度评论》34（2013），61。[19] Clark Glymour、Kun Zhang 和 Peter Spirtes。2019年。基于图模型的因果发现方法综述。《遗传学前沿》10（2019），524。[20] Emily Granger、Tim Watkins、Jamie C Sergeant 和 Mark Lunt。2020年。高影响力医学期刊发表论文中倾向评分诊断使用情况综述。BMC医学研究方法学20，1（2020），1–9。[21] Sander Greenland、Judea Pearl 和 James M Robins。1999年。流行病学研究中的因果图。Epidemiology（1999），37–48。[22] Noah Greifer。2020年。协变量平衡表和图：cobalt软件包指南。访问日期：2020年3月10日（2020）。[23] Grace Guo、Maria Glenski、ZhuanYi Shaw、Emily Saldanha、Alex Endert、Svitlana Volkova 和 Dustin Arendt。2021年。VAINE：用于自然实验的可视化与人工智能。载于2021年IEEE可视化大会（VIS）。IEEE，21–25。[24] Miguel A. Hernán。2021年。公共卫生研究方法——加强基于观察性数据的因果推断。《新英格兰医学杂志》385，15（2021），1345–1348。https://doi.org/10.1056/NEJMp2113319 arXiv:https://doi.org/10.1056/NEJMp2113319 PMID: 34596980。[25] Miguel A Hernan 和 James M Robins。2020年。《因果推断：如果》（第1版）。Chapman Hall/CRC。[26] Paul W Holland。1986年。统计学与因果推断。《美国统计学会杂志》81，396（1986），945–960。[27] Md Naimul Hoque 和 Klaus Mueller。2021年。Outcome-explorer：因果关系引导的交互式可视界面，用于可解释的算法决策。IEEE可视化与计算机图形学汇刊（2021）。[28] Mike Horn、Amartya Banerjee 和 Matthew Brucker。2022年。TunePad Playbooks：为创意音乐编程设计计算笔记本。载于CHI计算系统中的人为因素会议。1–12。[29] J. D. Hunter。2007年。matplotlib：二维图形环境。Computing in Science & Engineering 9, 3 2007 90 95 10 1109 2007 55 （），–。[30] 2010 https://doi.org/ 。/MCSE。 。Guido W Imbens 和 Donald B Rubin。 。Rubin 因果模型。载于Microeconometrics。Springer，229–241。[31] Bertrand Iooss 和 Andrea Saltelli。2017年。敏感性分析引言。载于不确定性量化手册。Springer，1103–1122。[32] Tobias Isenberg、Petra Isenberg、Jian Chen、Michael Sedlmair 和 Torsten Möller。2013年。评估可视化实践的系统综述。IEEE可视化与计算机图形学汇刊19，12（2013），2818–2827。[33] Zhuochen Jin、Shunan Guo、Nan Chen、Daniel Weiskopf、David Gotz 和 Nan Cao。2020年。事件序列数据的视觉因果关系分析。IEEE可视化与计算机图形学汇刊27，2（2020），1343–1352。[34] Nivedita R Kadaba、Pourang P Irani 和 Jason Leboe。2007年。使用动画可视化因果语义。IEEE可视化与计算机图形学汇刊13，6（2007），1254–1261。[35] Alex Kale、Matthew Kay 和 Jessica Hullman。2019年。研究综合中不确定性下的决策：为分叉路径花园进行设计。载于2019年CHI计算系统中的人为因素会议论文集。1–14。[36] Alex Kale、Yifan Wu 和 Jessica Hullman。2021年。因果支持：使用可视化对因果推断进行建模。IEEE可视化与计算机图形学汇刊28，1（2021），1150–1160。[37] DaYe Kang、Tony Ho、Nicolai Marquardt、Bilge Mutlu 和 Andrea Bianchi。2021。Toonnote：使用交互式数据漫画改善计算笔记本中的沟通。载于《2021年CHI计算系统中的人为因素会议论文集》。1–14。[38] Smiti Kaul、David Borland、Nan Cao 和 David Gotz。2021。使用反事实改进可视化解读。《IEEE可视化与计算机图形学汇刊》28，1（2021），998–1008。[39] Mary Beth Kery、Marissa Radensky、Mahima Arya、Bonnie E John 和 Brad A Myers。2018。笔记本中的故事：使用文学编程工具进行探索性数据科学。载于《2018年CHI计算系统中的人为因素会议论文集》。1–11。[40] Mary Beth Kery、Donghao Ren、Kanit Wongsuphasawat、Fred Hohman 和 Kayur Patel。2020。笔记本编程的未来是灵活流动的。载于《2020年CHI计算系统中的人为因素会议扩展摘要》。1–8。[41] Jing Lei、Max G'Sell、Alessandro Rinaldo、Ryan J Tibshirani 和 Larry Wasserman。2018。用于回归的分布无关预测推断。《美国统计学会杂志》113，523（2018），1094–1111。[42] Yang Liu、Alex Kale、Tim Althoff 和 Jeffrey Heer。2020。Boba：编写和可视化多宇宙分析。《IEEE可视化与计算机图形学汇刊》27，2（2020），1753–1763。[43] Jorge Piazentin Ono、Juliana Freire 和 Claudio T Silva。2021。Jupyter 笔记本中的交互式数据可视化。《计算科学与工程》23，2（2021），99–106。[44] João Palmeiro、Beatriz Malveiro、Rita Costa、David Polido、Ricardo Moreira 和 Pedro Bizarro。2022。Data+ Shift：支持数据科学家对数据分布偏移进行可视化调查。arXiv 预印本 arXiv:2204.14025（2022）。[45] Judea Pearl。1995。用于实证研究的因果图。《生物计量学》82，4（1995），669–688。[46] Judea Pearl。2012。重新审视 do 演算。arXiv 预印本 arXiv:1210.4852（2012）。[47] Judea Pearl 等。2000。模型、推理与推断。英国剑桥：剑桥大学出版社 19，2（2000）。[48] Julia M Rohrer。2018。清晰地思考相关性与因果关系：用于观察性数据的图形因果模型。《心理科学方法与实践进展》1，1（2018），27–42。[49] Paul R Rosenbaum 和 Donald B Rubin。1983。倾向得分在因果效应观察性研究中的核心作用。《生物计量学》70，1（1983），41–55。[50] Donald B Rubin。1973。通过匹配消除观察性研究中的偏差。《生物计量学》（1973），159–183。[51] Adam Rule、Ian Drosos、Aurélien Tabard 和 James D Hollan。2018。利用带注释的单元格折叠，助力计算笔记本的协作式复用。《ACM人机交互汇刊》2，CSCW（2018），1–12。[52] Michael Sedlmair、Miriah Meyer 和 Tamara Munzner。2012。设计研究方法论：来自一线实践与文献积累的反思。《IEEE可视化与计算机图形学汇刊》18，12（2012），2431–2440。[53] Yishai Shimoni、Ehud Karavani、Sivan Ravid、Peter Bak、Tan Hung Ng、Sharon Hensley Alford、Denise Meade 和 Yaara Goldschmidt。2019。用于指导因果推断中的模型选择与队列定义的评估工具包。arXiv 预印本 arXiv:1906.00442（2019）。[54] Uri Simonsohn、Joseph P Simmons 和 Leif D Nelson。2020。规格曲线分析。《自然·人类行为》4，11（2020），1208–1214。[55] Sara Steegen、Francis Tuerlinckx、Andrew Gelman 和 Wolf Vanpaemel。2016。通过多宇宙分析提高透明度。《心理科学观点》11，5（2016），702–712。[56] Chase Stokes、Vidya Setlur、Bridget Cogley、Arvind Satyanarayan 和 Marti Hearst。2022。取得平衡：整合文本与图表时读者的收获与偏好。arXiv 预印本 arXiv:2208.01780（2022）。[57] Krishna Subramanian、Nur Hamdan 和 J an Borchers。2020。随性的笔记本与僵化的脚本：理解数据科学编程。载于《2020年IEEE可视语言与以人为中心的计算研讨会（VL/HCC）》。IEEE，1–5。[58] Johannes Textor、Benito Van der Zander、Mark S Gilthorpe、Maciej Liśkiewicz 和 George TH Ellison。2016。使用有向无环图进行稳健的因果推断：R 软件包 dagitty。《国际流行病学杂志》45，6（2016），1887–1894。[59] April Yi Wang、Will Epperson、Robert A DeLine 和 Steven M Drucker。2022。循环中的差异：支持探索性数据分析中的数据比较。载于《CHI计算系统中的人为因素会议》。1–10。[60] April Yi Wang、Dakuo Wang、Jaimie Drozdal、Xuye Liu、Soya Park、Steve Oney 和 Christopher Brooks。2021。怎样的笔记本才算文档完善？对 Kaggle 上数据科学家文档实践的案例研究。载于《2021年CHI计算系统中的人为因素会议扩展摘要》。1–7。[61] Jun Wang 和 Klaus Mueller。2015。视觉因果关系分析师：用于因果推理的交互式界面。《IEEE可视化与计算机图形学汇刊》22，1（2015），230–239。[62] Jun Wang 和 Klaus Mueller。2017。让视觉因果关系分析切实可行。载于《2017年IEEE可视分析科学与技术会议（VAST）》。IEEE，151–161。

[63] Jun Wang 和 Klaus Mueller。2022。DOMINO：针对时间依赖现象的可视化因果推理。*IEEE*可视化与计算机图形学汇刊（2022）。[64] Zijie J Wang、Katie Dai 和 W Keith Edwards。2022。StickyLand：打破计算笔记本线性呈现方式。载于*CHI*计算系统中的人为因素会议扩展摘要。1–7。[65] Zijie J Wang、David Munechika、Seongmin Lee 和 Duen Horng Chau。2022。NOVA：创建适用于笔记本的可视分析的实用方法。*arXiv* 预印本 *arXiv:2205.03963*（2022）。[66] Michael L. Waskom。2021。seaborn：统计数据可视化。开源软件期刊 6, 60 2021 3021 10 21105 03021 （），。[67] https://doi.org/。/joss。Yifan Wu、Joseph M Hellerstein 和 Arvind 2020 2 Satyanarayan。。B：在计算笔记本中连接代码与交互式可视化。载于第*33*届*ACM*用户界面软件与技术研讨会论文集。152–165。[68] Cindy Xiong、Joel Shapiro、Jessica Hullman 和 Steven Franconeri。2019。可视化数据中的因果幻觉。*IEEE*可视化与计算机图形学汇刊26，1（2019），853–862。[69] Chi-Hsien Eric Yen、Aditya Parameswaran 和 Wai-Tat Fu。2019。视觉因果关系分析的探索性用户研究。载于计算机图形学论坛，第38卷。Wiley Online Library，173–184。[70] Amy X Zhang、Michael Muller 和 Dakuo Wang。2020。数据科学工作者如何协作？角色、工作流程与工具。*ACM*人机交互汇刊4，CSCW1（2020），1–23。[71] Chengbo Zheng、Dakuo Wang、April Yi Wang 和 Xiaojuan Ma。2022。讲述计算笔记本中的故事：借助人工智能辅助创建演示文稿，以展示数据科学工作。载于*CHI*计算系统中的人为因素会议。1–20。

因果推断工具和软件包

A.1 DAGitty

DAGitty（图 A1）是一款允许用户交互式指定因果结构图的应用程序。它是一款功能强大且全面的程序，但用户需要了解因果结构术语（例如混杂因素、条件独立性）及其表示法，并且善于运用这些概念。

A.2 Causalnex

Causalnex（图 A2）是一款 Python 软件包，可执行因果发现，并允许用户通过指定节点及节点间的连线来建模因果结构。Causalnex 中的有向无环图可视化是静态的，必须通过手动编写代码进行编辑。

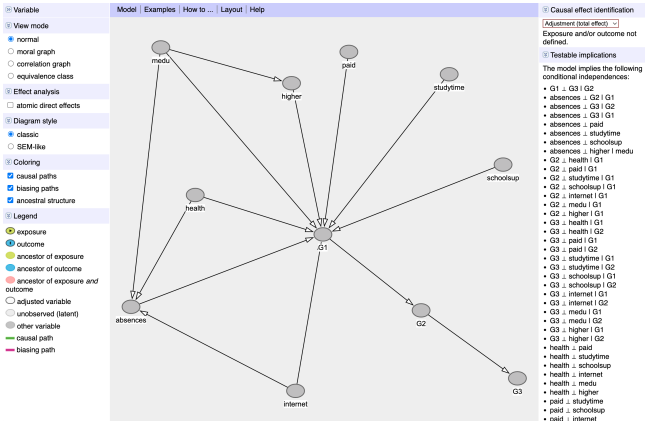

图 A1：DAGitty [58] 应用程序的屏幕截图，展示了 UCI 机器学习数据集中一部分变量构成的有向无环图。

A.3 Causallib

Causallib（图 A3）是一个 Python 工具包，实现了多种用于因果推断方法的机器学习模型，包括本文讨论的常用逆概率加权和匹配方法。该软件包还提供了可视化功能，帮助分析人员评估所使用的机器学习模型。与我们的工作密切相关的是逆概率加权方法中包含的 aSMD 图和倾向得分分布。所有 causallib 可视化均为静态可视化。

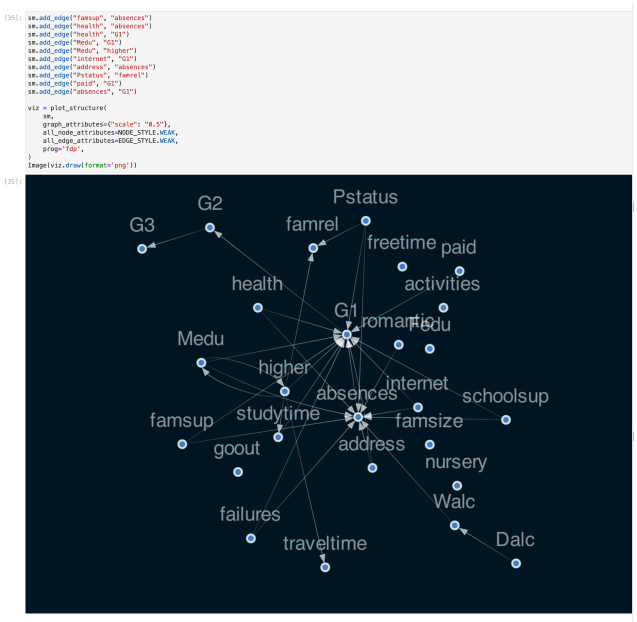

图 A2：Causalnex [5] 软件包的屏幕截图，展示了 UCI 机器学习数据集中各变量构成的有向无环图。上方的 Jupyter 笔记本单元格展示了分析人员如何手动添加有向箭头，以表示变量对之间的因果关系。

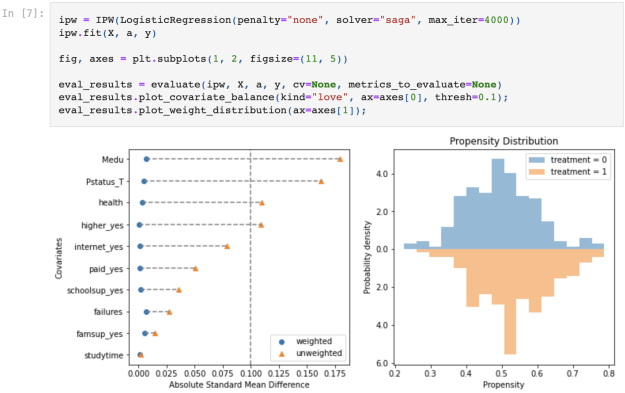

图 A3：causallib [53]软件包的屏幕截图，展示了队列在使用逆概率加权进行协变量调整后生成的 aSMD 图和倾向评分分布。

A.4 Cobalt

Cobalt（图 A4）是一个 R 软件包，可生成可视化结果，帮助评估通过匹配及其他方法生成的因果推断队列的结果。必须通过单独的函数调用来实例化可视化。所有 Cobalt 可视化均为静态可视化。

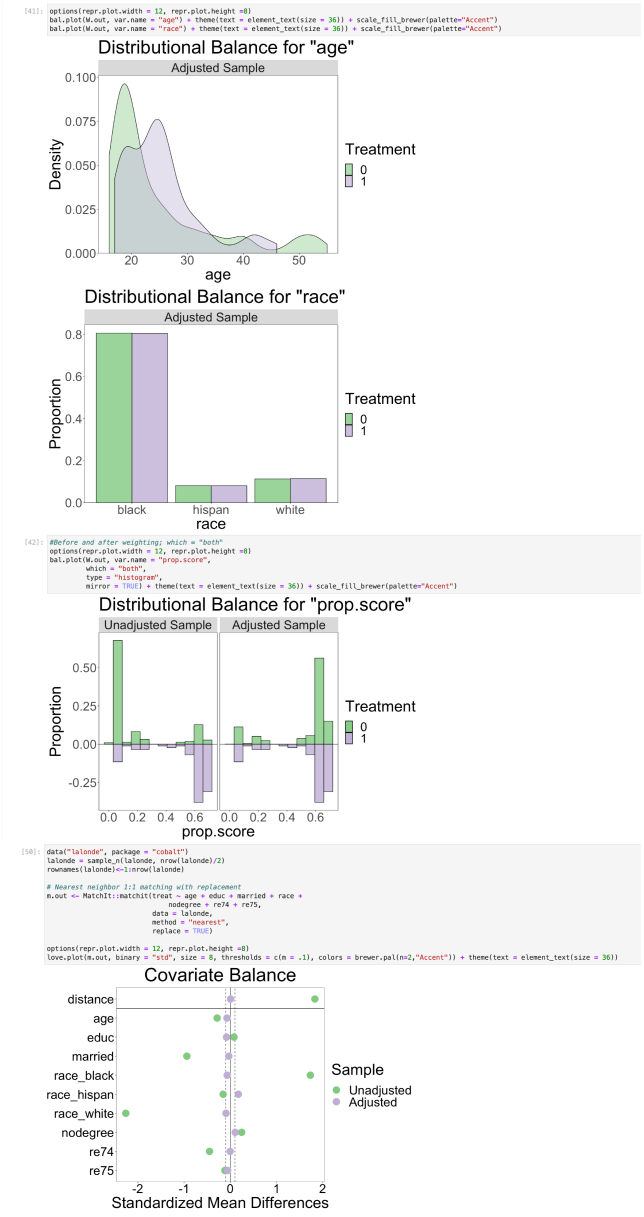

图 A4：Cobalt [22] 软件的屏幕截图，展示了使用匹配进行协变量调整后某个队列的各个协变量分布、aSMD 图和倾向评分分布。