因果结构学习中的开放问题：英国新冠肺炎的案例研究

Anthony Constantinou[a]、Neville K. Kitson[a]、Yang Liu[a]、Kiattikun Chobtham[a] Arian Hashemzadeh Amirkhizi[b, c]、Praharsh A. Nanavati[d]、Rendani Mbuvha[a]和Bruno Petrungaro[a]。

a. 贝叶斯人工智能研究实验室、机器智能与决策系统（MInDS）研究组，电子工程与计算机科学学院（EECS）、伦敦玛丽女王大学（QMUL），英国伦敦，E1 4NS。b. 谢里夫理工大学，伊朗德黑兰。c. 多伦多大学，加拿大多伦多。d. 印度科学教育与研究学院， 印度中央邦博帕尔。

电子邮箱：a.constantinou@qmul.ac.uk，n.k.kitson@qmul.ac.uk，yangliu@qmul.ac.uk，k.chobtham@qmul.ac.uk，arian.hashemzadeh@mail.utoronto.ca，praharsh19@iiserb.ac.in，r.mbuvha@qmul.ac.uk，b.petrungaro@qmul.ac.uk。

摘要：因果机器学习（ML）算法能够恢复图结构，揭示因果关系。此类算法提供的因果表征具有透明性和可解释性，这对于在关键的现实问题中进行决策是必要的。然而，与关联机器学习相比，因果机器学习在实践中的影响有限。本文探讨因果机器学习面临的挑战，并将其应用于英国新冠肺炎疫情数据。我们汇集了来自不同公开来源的数据，并研究各种结构学习算法从这些数据中学到了什么。我们探索不同数据格式对不同类别学习算法的影响，并从图结构、模型维度、敏感性分析、混杂变量、预测推断和干预推断等方面评估各个算法及算法组所产生的结果。我们利用这些结果指出因果结构学习中的开放问题及未来研究方向。为促进后续研究，我们在网上公开了所有图、模型、数据集和源代码。

关键词：贝叶斯网络结构学习、因果发现、因果机器学习、因果模型、基于知识的系统、概率图模型。

注意：这是以下开放获取出版物的预印本：

Constantinou, A., Kitson N. K., Liu, Y., Chobtham, K., Hashemzadeh, A., Nanavati, P. A., Mbuvha, R., 和

开放获取链接：https://www.sciencedirect.com/science/article/pii/S0957417423015713

1 引言

1.1. 因果机器学习

以往关于因果发现的研究通常聚焦于实验，例如临床研究和社会科学中广泛使用的随机对照试验（RCTs）。然而，开展受控实验往往既昂贵又耗时，而且有时甚至不合乎伦理或根本无法实施。此外，尽管随机对照试验有助于科学研究估计干预的效果，但它们并不

总能准确估计该效果（Deaton and Cartwright, 2018年）。因此，通常需要采用新方法来补充随机对照试验所得的结果。

评估和建模因果关系，对于识别并解释我们所经历的因果现象至关重要。要理解世界如何运作，从而更好地影响世界，就需要通过超越预测验证的措施，检验有关某些事件为何发生的假设。如今，对可验证性的要求促使公众重新认识到黑箱机器学习（ML）解决方案的局限性：这类方案往往可解释性有限。因此，它们在为需要解释依据的干预措施提供建议方面，能力有限（Guidotti等人，2018年；Schölkopf等人，2021）。

因果机器学习是一个研究领域，广义而言，它聚焦于旨在从数据中恢复因果结构的无监督学习算法。这些方法可以在完全可解释的情况下指导干预。尽管过去几十年间，因果机器学习领域的方法取得了巨大进展，但这些创新往往体现为理论上的突破，在实践中的影响有限；至少与关联机器学习的方法相比是如此。部分原因在于，发现准确的因果关系一直被视为一项极其困难的任务。当处理真实数据时，这一点尤其明显，因为真实数据往往存在各种缺陷，不同于合成实验。例如，真实数据可能包含缺失、有偏或错误的值，也可能因降维而造成信息丢失、其分布随时间发生变化，或未能涵盖所有关注变量。这些缺陷违背了许多算法所依据的假设，因而会对算法的性能产生负面影响（Constantinou等人，2021）。

1.2. 案例研究：新冠肺炎

2019冠状病毒病（新冠肺炎）大流行病凸显了确定前所未有的政策干预措施有效性的必要性，而因果模型能够为此提供支持。首例已知的新冠肺炎病例于2019年12月在中国武汉被发现。该病毒在数周内传播至全球，世界卫生组织（WHO）于2020年3月宣布新冠肺炎为大流行病。截至2022年10月11th 日，英国累计报告感染病例达2370万例，死亡人数达20.8万人；全球累计感染病例达6.22亿例，死亡人数达660万人（Our World in Data，2022a）。

为应对新冠肺炎的快速传播，各国都实施了量身定制但大体相似的政策干预措施，主要涉及限制人口流动和人员接触。在英国，政府的强制规定包括保持社交距离、佩戴口罩、限制参加各类活动及其他聚集性活动、鼓励接种疫苗，以及实施严格的封锁措施，仅允许必要的人员流动。这些政策通常以国家层面的科学和学术研究为依据（帝国理工学院COVID-19应对团队，2021），旨在控制住院人数。

在大数据时代，学术文献中已经有数量惊人的论文，几乎研究了新冠肺炎大流行病的各个方面。这些研究包括新冠肺炎与以下因素之间关系的研究：

- 维生素D浓度。Hastie 等人（2020年）指出，没有证据表明维生素D能够解释人们对 COVID-19感染的易感性；
- 肥胖或 BMI。Yates 等人（2020年）报告了早期证据，表明 BMI、腰围与新冠肺炎之间存在剂量－反应关联；
- 个人防护装备（PPE）的作用。Nguyen 等人（2020年）发现，与普通社区人群相比，一线医护人员感染新冠肺炎的风险更高，而这通常是由于个人防护装备不足；
- 精神障碍。Yang 等人（2020年）报告称，临床确诊患有大流行病前精神障碍的个体，住院治疗和死亡的风险较高；
- 临床、地区和遗传特征；研究指出，男性、受教育程度较低者以及非白人和黑人族裔感染的风险较高

新冠肺炎（Chadeau-Hyam 等，2020年），而这些因素又有助于预测新冠肺炎死亡情况（Elliot 等，2021），Kolin 等人（2020年）认为，劳动力市场中的歧视可能是黑人群体感染新冠肺炎相对风险较高的原因之一，并证实了 A 型血与新冠肺炎之间的关联；

- 能够预测长期新冠的特征；Sudre 等人（2021）报告称，疾病早期特征、年龄和性别基本可以解释长期新冠；
- 脑结构变化方面，Douaud 等人（2022）通过调查英国生物样本库中 785 名参与者在感染新冠肺炎前后的脑部变化，报告了脑部异常的有力证据。

其他相关研究包括：O’Connor 等人（2020）报告称，英国封锁似乎影响了英国成年人口的心理健康和福祉；Jarvis 等人（2020）得出结论，英国公众采取的保持身体距离措施减少了接触频率，并导致感染人数大幅下降；Menni 等人（2021）报告称，辉瑞-BioNTech（BNT162b2）和牛津-阿斯利康（ChAdOx1 nCoV-19）新冠肺炎疫苗被发现可在接种疫苗 12 天后降低感染新冠肺炎的风险。

尽管许多新冠肺炎研究报告了感兴趣因素之间可能存在的因果联系，但只有少数研究采用了某种形式的因果建模。这些研究包括：a）Mastakouri 和 Schölkopf（2020）的研究，他们利用因果时间序列，根据各联邦州实施的限制政策，研究德国各地区新冠肺炎传播过程中的因果关系；b）Horn 等人（2020）的研究，他们通过模拟混杂变量和中介回归变量，探讨新冠肺炎在免疫精神病学中的因果作用；c）Sahin 等人（2020）的研究，他们构建假设性的因果回路图，以探究 COVID-19 大流行的复杂性及相关政策干预；d）Friston 等人（2020）的研究，他们运用变分贝叶斯方法和假设性的条件依赖关系，针对流行病学人群构建动态因果模型，以预测伦敦的新冠肺炎病例；e）Fenton 等人（2020）的研究，他们阐释了为何有必要在数据分析中纳入因果解释，以避免新冠肺炎检测中的估计偏差；f）Chernozhukov 等人（2021）的研究，他们构建假设性因果模型，以探究口罩、政策和行为对美国早期 COVID-19 大流行的因果影响。我们仅发现一篇运用因果结构学习的论文，即 Gencoglu 和 Gencoglu（2020）关于发现并量化大流行病特征、Twitter 活动与公众情绪之间因果关系的研究。

1.3. 论文目的与结构

本文聚焦于新冠肺炎数据，这些数据记录了英国在两年半时间内大流行病的发展情况，重点涉及病毒检测、感染、住院、死亡、疫苗接种、政策干预和人口流动。从因果视角来看，这些因素颇具研究价值，因为人们认为它们能够反映彼此之间的因果相互作用，并可作为因果机器学习的良好试验平台。

与以往新冠肺炎研究的一项重要区别在于，本研究重点评估结构学习算法，而非利用这些算法来探索新冠肺炎的相关知识。具体而言，本研究旨在探讨结构学习算法的优势与局限，并将其应用于一项有关新冠肺炎、对大多数读者而言都易于理解且至关重要的现实案例研究，同时依据结构学习领域尚待解决的问题，为未来研究提出方向。

本文结构如下：第2节介绍因果模型和结构学习的背景信息；第3节描述我们收集并预处理新冠肺炎数据的过程；第4节描述我们如何依据人类知识构建因果图；第5节介绍所研究的结构学习算法；第6节呈现研究结果；最后，我们在第7节给出结论。

2 预备知识：因果结构学习与因果建模

本节不使用技术符号，简要介绍因果建模和结构学习的一些主要概念，以帮助读者理解本文所采用的方法和呈现的研究结果。有关结构学习和因果建模的详细信息，我们建议读者参阅以下资料：Koski 和 Noble（2012）关于贝叶斯网络的综述，Scanagatta 等人（2019）关于结构学习的综述，Kitson 等人（2023）对74种结构学习算法所作的全面评述，Koller 和 Friedman（2009）关于概率图模型的著作，Darwiche（2009）关于贝叶斯网络的著作，以及 Fenton 和 Neil（2018）聚焦于基于知识的贝叶斯网络的著作。

2.1. 因果建模

贝叶斯网络（BN）通常用有向无环图（DAG）表示，其中包含节点和有向边；每个节点代表一个数据变量，每条有向边代表一种依赖关系。在离散贝叶斯网络中，变量之间的关系由条件概率表（CPT）表示，而在连续型贝叶斯网络中，变量之间的关系则由条件分布或函数关系表示。因果贝叶斯网络（CBN）是一种贝叶斯网络模型，其有向边被假定为具有因果关系。由此可知，CBN 只能由单一的 DAG 结构表示，而贝叶斯网络则可以由其对应的马尔可夫等价类中的任意 DAG 结构表示；我们将在第 2.2 小节对此进行介绍。

CBN 模型可用于预测推断和诊断推断，也可用于干预推断和反事实推断。Pearl（2018年）将其称为因果阶梯：第一阶关注我们能从关联本身学到什么；第二阶关注如何模拟假设性干预，以衡量其效果，而无需进行实验；第三阶则关注如何回答有关过去本可能采取的其他行动的反事实问题。Pearl 的 do 算子框架用于表示和模拟因果模型中的物理干预（Pearl，2012年），可用于对超越关联推断的第二阶情形进行建模。在本研究中，我们将探讨因果阶梯的前两阶。

2.2. 因果结构学习

因果结构学习是一种无监督学习过程，可分为以下两类：a）组合优化，大多数算法通过搜索布尔值或离散值邻接矩阵来表示边的存在与否，也可能包含不同类型的边；b）较新的方法——连续优化，使用实值邻接矩阵，可借助现成优化器求解，并施加无环约束。组合优化包含不同类别的学习方法，包括：

i. 基于约束的算法会执行一系列边缘独立性检验和条件独立性检验，以识别变量之间的无向边，并确定其中一些边的方向。

ii. 基于评分的算法依靠启发式搜索技术探索可能图构成的搜索空间，并利用目标函数为访问到的每个图评分。这类学习方法还可进一步分为近似解和精确解。

近似学习算法通常效率较高，但可能会停在局部极大值解；而精确学习算法的计算成本更高，但能保证返回图搜索空间中得分最高的图。不过，这一保证通常要求将图搜索空间限制在指定的节点入度范围内[1]，，并且图结构往往通过组合优化进行局部探索，而非借助启发式搜索进行全局探索。

iii. 混合算法结合了上述两类结构学习方法。

[1] 每个节点要探索的父节点最大数量。

依赖连续优化的方法可分为线性和非线性得分方法。线性方法包括构建线性结构因果模型（SCM），或构建关系为线性的连续型或函数型贝叶斯网络（BN），并使用梯度下降优化权重矩阵中的系数。非线性方法通常是对线性方法的扩展，融入神经网络模型，并允许邻接矩阵中的关系呈非线性。

每一类结构学习算法都有其自身的一系列假设、优缺点，文献中对于哪种学习类别更可取或更合适尚无共识。无论算法属于哪种学习类别，不同算法往往会生成不同类型的图。这在一定程度上是因为许多算法对输入数据作出了不同的假设，并依赖不同的目标函数。大多数被考虑的目标函数都倾向于评分等价；这意味着，对于属于同一马尔可夫等价类的图结构，这些函数会给出相同的得分。这是因为评分等价函数假定，无法从观测数据中恢复所有有向关系。

或许最广为人知的一类图是有向无环图（DAG），其中包含节点和有向边，且不存在循环关系。生成有向无环图结构的算法所使用的目标函数并非评分等价，或者只是从得分最高的等价类中随机返回一个有向无环图。该等价类由完成的部分有向无环图（CPDAG）表示，其中包含无法根据观测数据确定方向的无向边（以及有向边）。换言之，CPDAG表示一组马尔可夫等价的有向无环图，其中等价类中的每个有向无环图都会产生相同的目标得分。

一种更复杂的图是最大祖先图（MAG），除了有向边和无向边外，还包含表示混杂的双向边。与有向无环图（DAG）或部分有向无环图（CPDAG）不同，MAG 能够捕捉有关潜在混杂因素和祖先关系的信息[2]。MAG 的等价类由部分祖先图（PAG）表示，这与有向无环图和 CPDAG 之间的关系类似。考虑潜在混杂因素可能性的算法会生成 MAG 或 PAG，通常属于基于约束的学习或混合学习类别。

3. 数据收集与预处理

数据收集自不同的公开来源，其中大部分信息来自英国政府发布的新冠肺炎官方数据网站。汇总后的数据集包含 18 列和 866 行，每列对应一个变量，每行对应一天的结果。这些数据记录了英国大流行病在 866 天内的发展情况，时间跨度为 2020年1月30th 至2022年6月13th 。

我们在第3.1小节中，根据数据变量所捕捉信息的类型对其进行描述和分类。原始数据包含连续型变量和分类变量，也存在缺失数据值。第3.2小节介绍了我们如何对数据进行预处理，使其适用于所考虑的不同结构学习算法。

3.1. 数据变量的分类与描述

表1列出了所有数据变量及其描述和数据来源。请注意，日期变量仅为提供信息而添加。由于算法假设数据行之间不存在时间关系，因此在结构学习过程中无法将日期纳入考虑；算法会将日期视为一个包含与样本数量相同状态数的变量，这不会对学习到的结构产生任何影响（我们将在第4节和结论中讨论这一局限性）。我们将数据变量分为以下八类：

[2] 最大祖先图可能包含其他类型的边，也可用于表示选择变量。由于这些内容超出了本研究的范围，我们不在预备知识中介绍这些细节。

病毒检测：此类别仅包含一个变量，即四大支柱中的检验。该变量记录了用于检测COVID-19感染的不同检验类型，涵盖英国政府GOV.UK（2022a）定义的支柱1、2、3和4下的检测。具体而言，支柱1下的检测由英国国家医疗服务体系（NHS）和英国卫生安全局（UKHSA）开展；支柱2下的检测由英国政府的COVID-19检测计划开展；支柱3下的检测包括抗体血清学检测；支柱4下的检测则用于国家监测。

感染：此类别包含三个变量，即阳性检测、新增感染和再感染。这些变量反映每日病例信息，以及这些病例是新增感染还是再感染。

住院：此类别包含住院入院人数、住院患者和MVB中的患者这三个变量。这些变量分别反映因新冠肺炎入院的患者人数、因新冠肺炎住院的患者人数，以及因新冠肺炎使用机械通气病床（MVB）的患者人数。

疫苗接种：此类别仅包含一个变量，称为第二剂接种率，用于反映2^{nd}剂疫苗的接种情况，并表示符合接种条件的人口中已接种疫苗的比例。在英国，接种两剂疫苗才被视为“完全接种”。后来建议接种加强剂次，但为求简明，本文不予讨论。

死亡：此类别包含两个变量，即死亡证明上注明COVID的死亡和超额死亡率，分别反映死亡证明上列有新冠肺炎的死亡人数，以及总体超额死亡率。

出行：三个变量以指数的形式反映英国的出行信息，分别是交通活动、工作和学校活动，以及休闲活动。这些变量对应一系列综合指数，涵盖航班、公共汽车、火车和交通枢纽站、对公园、零售场所、杂货店和餐厅的访问，以及步行、出行、居家办公和学校活动。如表1所示，尽管这些数据由英国政府提供，但数据来源于伦敦交通局（TfL）、Google、Apple、Citymapper 和 OpenTable 等第三方提供商。学校指数是构成工作和学校活动变量的指数之一；由于我们无法在网上轻易找到相关信息，因此按照表A.4中的说明手动构建了该指数。

政策：口罩和封锁这两个变量反映了英国有关封锁措施和口罩令的重要政策干预。由于我们无法在网上轻易找到这些数据，因此根据表1所列来源并按照表A.4和A.5中的说明，手动构建了这两个变量。

其他：变量“季节”和“主要COVID变体”可视为背景信息，无法归入之前定义的任何类别。这些变量记录了季节（例如夏季）以及占多数的COVID-19变异株相关信息。它们也是按照表A.4所述手动构建的。

表1. 用于构建新冠肺炎英国数据集的原始数据变量说明。所有数据变量的取值或状态均遵循序数分布。

编号	变量名	类别	Var type	说明和数据来源
0	Date	n/a	离散化	记录观测数据的日期。不用于结构学习。
1	所有四大支柱	病毒检测	续	支柱1、2、3和4的检验总数（GOV.UK，2022a）。
2	阳性检测	感染	续	按样本统计的病例数（即新冠肺炎检测呈阳性的人数）日期统计的病例数（即新冠肺炎检测呈阳性的人数）（GOV.UK，2022b）。
3	新增感染	感染	续	按标本日期统计的新增感染（GOV.UK，2022b）。
4	再感染	感染	续	按标本日期统计的新发再感染病例（GOV.UK，2022b）。
5	医院入院	住院治疗	续	因新冠肺炎入院的患者人数（GOV.UK，2022c）。
6	住院患者医院	住院治疗	续	患有新冠肺炎的住院患者人数（GOV.UK，2022c）。
7	MVB中的患者住			MVB中患有新冠肺炎的患者人数（GOV.UK，2022c）。
8	第二剂接种率	疫苗	续	报告的2nd 剂疫苗接种率（GOV.UK，2022d）。
9	死亡病例伴有新冠肺炎 on证书	死亡	续	按记录的死亡日期统计、死亡证明上注明新冠肺炎的每日死亡人数（GOV.UK，2022e）。
10		死亡	续	报告死亡人数与根据往年数据推算的同一时期死亡人数之间的百分比差异。(Our World in Data，2022c)
11	活动	出行	续	英国政府提供的 TfL 公交和地铁综合数据指数活动、Google 提供的交通枢纽站活动数据、Citymapper 提供的出行数据活动数据、Apple 的步行活动数据以及英国的航班活动数据（GOV.UK，2022年f；英国国家统计局网站，2022年；欧洲航空安全组织，2022年）。
12	活动	出行	续	Google提供的英国政府居家办公综合数据指数以及工作场所活动、零售场所和休闲活动、杂货和药房活动，以及新冠肺炎疫情期间英国的学校运营指南(GOV.UK，2022f；2022h；2022i；2022j；维基百科，2022b)。
13		出行	续	英国政府汇总了 Google 提供的公园 посещения 数据指数，以及 OpenTable 提供的餐厅预订数据（GOV.UK，2022年f）。
14		政策	分类：	维基百科，2022a）。
15		政策	分类：	英国政府，2022年)
16		其他	分类。	
17	变异株	其他	分类。	数据，2022b)

3.2. 连续型、离散和混合型、完整和不完整数据集

原始数据集包含连续型变量和分类变量，其中一些变量不完整。然而，许多算法只能处理一种数据格式，而且大多数算法假设输入数据集完整，不含缺失数据值。为了能够针对所有算法接受的各种可能输入数据测试这些算法，我们从原始数据派生出以下七个数据集：

i. 离散四分位数不完整：所有连续变量均按四分位数离散化，且未对缺失值进行插补。

ii. 离散四分位数完整：所有连续变量均按四分位数离散化，并根据离散化后的值对数据集中的所有缺失值进行插补。iii. 离散K均值不完整：所有连续变量均使用K均值聚类进行离散化，且未对缺失值进行插补。iv. 离散K均值完整：所有连续变量均使用K均值聚类进行离散化，并根据离散化后的值对数据集中的所有缺失值进行插补。v. 连续型不完整：所有分类变量均转换为连续变量，且未对缺失值进行插补。vi. 连续型完整：所有分类变量均转换为连续变量，并根据连续值对数据集中的所有缺失值进行插补。vii. 混合型完整：对含缺失值的原始数据集进行插补。

例如，我们没有构建混合不完整数据集，因为没有算法能够接受包含缺失值的混合数据集。所有七个数据集均已在线公开，可从 Bayesys 仓库下载（Constantinou等人，2020年）。

连续型数据集是通过将所有分类变量转换为 0 到 1 范围内的连续值构建的，这些值对应于各类别的顺序。我们采用两种常见的数据离散化方法。第一种方法是求出每个连续变量的四分位区间，并将其转换为四种状态 $\{Very\ Low,\ Low,\ Hig$和$Very\ Hig\}$，每种状态对应相应的四分位区间。第二种方法是对每个变量的取值应用无监督的 k-均值聚类，根据聚类而非四分位数来确定这四种状态。我们使用 sklearn 版本 1.1.2，并将超参数设为 $k = 4$，使其与前一种方法生成的状态数量一致；同时确保由这些聚类得出的状态按序排列，且其顺序与根据四分位数得出的状态顺序一致。这两种方法各有差异：基于四分位数的离散化会产生平衡分布，而聚类则不会。也就是说，平衡分布能够最大限度地增加可用于为条件概率表（CPT）中各参数进行参数化的样本量；而聚类方法则体现了经典的离散化思路，即力求最大化数据拟合。

最后，由于许多缺失数据值不太可能属于随机缺失，我们采用马尔可夫毯缺失森林（MBMF）插补算法，该算法在随机缺失和系统性缺失两种假设下执行插补。MBMF 算法旨在利用缺失性的图形表达形式——即 m 图，来恢复部分观测变量的马尔可夫毯[3]。m 图是一种将观测变量与缺失性的潜在成因（称为缺失指示变量）一并纳入的图。这种方法在随机缺失和系统性缺失两种情况下，均被证明能提升插补准确率，优于最先进水平（Liu 和 Constantinou，2023年）。

4 从知识构建因果图

我们首先基于人类知识构建因果图。知识图谱见图3，包含表1中描述的17个变量（不包括日期）。它代表了所有作者达成的共识。作者们通过小组讨论探讨不确定的关系并解决分歧。需要注意的是，尽管大多数作者此前有将因果模型应用于医疗保健领域的经验，但没有一位作者是流行病学专家。基于此，我们假设图3所示的知识图谱代表了关于COVID-19大流行的常识。该图谱基于以下五个关键假设构建：

i. 感染率受季节和COVID-19变体的影响 ;

[3] 给定变量的马尔可夫毯是能够完全解释其取值的变量子集。在有向无环图模型中，给定节点的马尔可夫毯包含该节点的父节点、子节点以及其子节点的父节点。

ii. 住院风险受COVID-19变体的影响；iii. 政府关于封锁措施和口罩的政策，以及疫苗接种率，受到新冠肺炎住院病例数和死亡人数的影响；iv. 出行受到政府封锁政策的影响，也受到季节的影响；v. 超额死亡率受到新冠肺炎死亡人数的影响，也受到封锁措施的影响，因为封锁导致出行发生变化，进而以难以衡量的方式影响了其他人的生活。

需要特别指出的是，在构建知识图谱时，我们遇到了关系循环的问题。例如，大流行病的状况（如感染率和住院治疗情况）会影响封锁决策，但封锁措施及其他政策行动预计也会影响大流行病未来的发展轨迹。同样，感染率会影响出行，但出行也会影响未来的感染率。我们认识到，某些变量存在反馈回路，但由于本研究不对时间关系进行建模，因此无法纳入其中一些关系。

知识图谱假设起点是大流行病状况。例如，我们必须先观察到较高的住院率，之后才会观察到政策干预（如封锁）。由于图1中绘制的边表示同一日期的关系，我们假设大流行病状况已经受到知识图谱中未显示的过往政策和出行的影响。例如，如果我们观察到感染率较低，且这可以用几周前实施的出行限制来解释，那么我们假设出行对感染的影响并非发生在同一天；相反，同日影响的方向是从感染到出行。这是因为政策干预（如封锁或限制出行）的效果会在之后才显现，无法在同一数据行中观察到；例如，在数据中观察到封锁，并不一定意味着当天的感染情况受到当天实施的封锁影响。

基于此，我们假设，在知识图谱中，政策和出行与感染和住院之间不应存在同日期的边。同样，我们假设政策会受到大流行病当前状况的影响。因此，知识图谱中描绘的同日期关系反映的是观测事件，而非干预事件；例如，对出行进行干预并不能准确模拟封锁干预中的同日期关系，因为这需要将此类干预与未来的观测结果相联系。

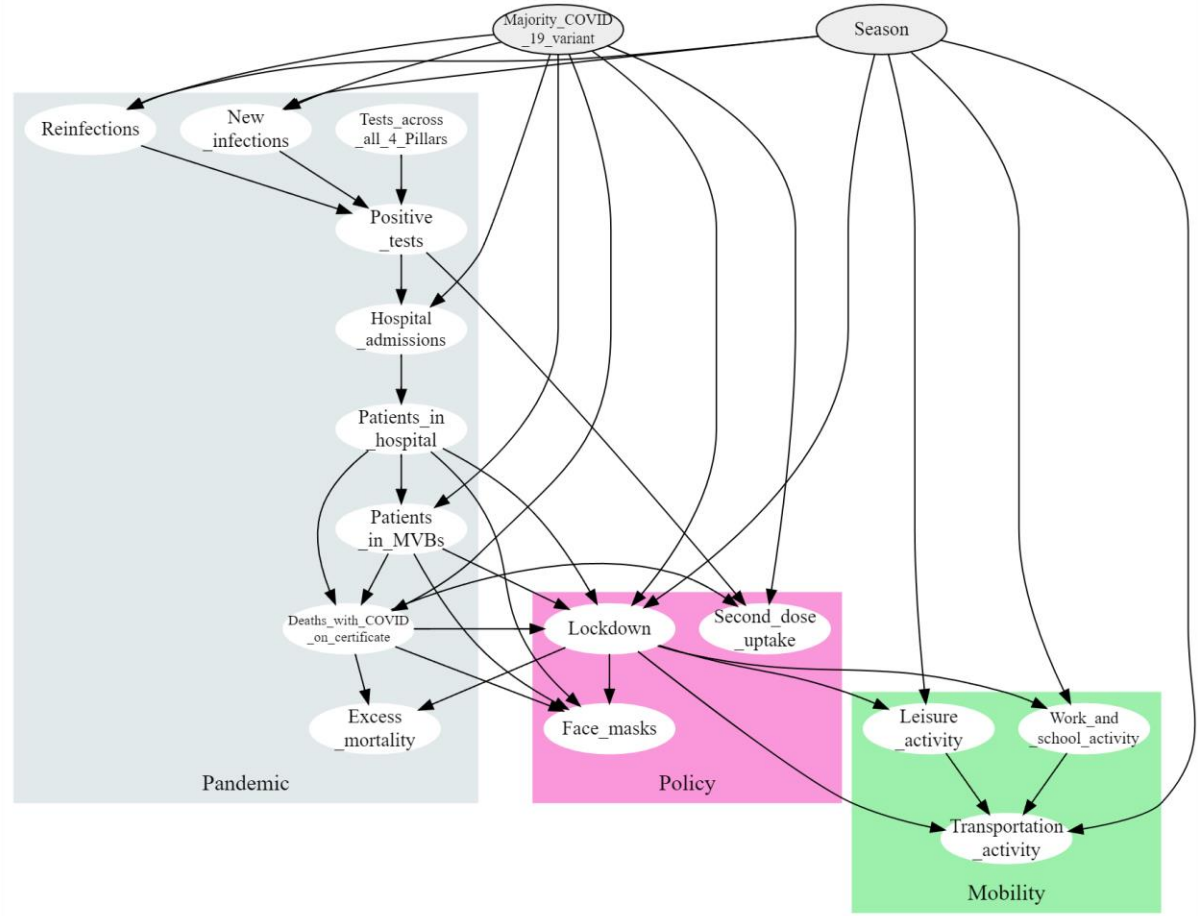


图1。知识图谱。它包含表1中描述的17个变量（不包括日期）、37条有向边，节点最大入度为5，出度为7，节点度为7。

5 从数据中学习因果结构

我们采用了涵盖不同学习类别的结构学习算法，这些算法对输入数据有不同的假设，并生成不同类型的图结构。具体而言，我们考虑基于约束的算法、近似和精确的基于评分算法、混合学习算法以及连续优化算法，它们的输出分别为有向无环图、CPDAG、最大祖先图或部分祖先图（定义请参见第2.1小节）。此外，其中一些算法假设数据完整或不完整，并进一步假设输入数据为分类的、连续型或混合型（定义请参见第3.2小节）。

表2列出了算法及其对输入数据所作的假设。每种算法都会应用于其能够处理的所有不同数据格式。此外，如果某种算法专为不完整数据而设计，则会将其应用于不完整数据集，而非完整数据集。我们还考虑采用模型平均方法，为属于特定组或结构学习类别的算法集合获得整体结构（详情见下文）。

总共，我们测试了29种不同算法，包括模型平均学习器，共进行了表2中列出的64项实验。选择这些算法是为了考察所有结构学习类别中常用的算法，以及一些不太常用但具有独特特点的算法。除非另有说明，以下所有算法都会从相应的等价类中返回一个随机的有向无环图或最大祖先图结构（如果返回的不是实际的等价类）。关于这些算法的详细技术信息不在本文讨论范围内，可参见下文各算法对应的参考文献，以及第2节讨论的综述论文。按字母顺序排列，我们测试的算法如下：

- CCHM：Chobtham和Constantinou（2020）提出的一种混合算法。该算法首先采用CFCI的方法（见下文）来学习骨架并确定v-结构。随后，它在最大祖先图空间中应用爬山搜索来定向部分无向边，并使用do-算子定向其余无向边。
- FCI：快速因果推断（FCI）是一种基于约束的算法，它通过考虑数据中可能存在潜变量的情况，对 PC 算法进行了扩展（Spirtes 等人，1999年）。该算法首先生成图的骨架，然后进行条件独立性检验，以确定部分边的方向；其中还会生成表示混杂的双向边，最终输出部分祖先图（PAG）。
- FGES：一种“快速”（高效）的 贪婪等价搜索（GES）算法的并行化版本
  由 Chickering（2002年）提出。Ramsey 等人（2016年）提出的快速 GES（FGES）包含两个学习阶段，分别称为前向搜索阶段和后向搜索阶段。前向阶段从空图开始，在每次迭代中，将 CPDAG 空间内使目标函数最大化的边添加到图中。当找不到能够进一步提高目标得分的边时，FGES 便进入后向阶段，并以相同方式删除边；当无法通过删除边进一步提高目标得分时，算法停止。
- GFCI：Ogarrio 等人（2016）提出的贪婪快速因果推断（GFCI）是一种混合算法
  该算法首先使用 FGES 生成 CPDAG，然后进行条件独立性检验，以移除该骨架中多余的邻接关系，接着应用修订后的 FCI 定向规则生成部分祖先图。
- GOBNILP：Cussens（2011）提出的一种基于整数线性规划（ILP）的评分算法
  该算法通过结合剪枝和为每个节点分配父节点的最优方案，实现精确学习。ILP 保证返回使评分函数最大化的图，但由于计算复杂度高，这一保证通常仅适用于节点最大入度较低的情况。例如，为解决效率问题，其默认超参数将学习范围限制为每个节点三个父节点。
- HC：爬山算法是最简单、也是最早出现的基于评分
  结构学习算法之一（Bouckaert，1994年；Heckerman 等人，1995年）。该算法从空图开始，每次迭代通过添加、反转和删除弧来考察所有可能的相邻有向无环图。随后，它应用使目标得分最大化的图修改操作，并在目标得分不再增加时停止。
- HC_aIPW：Liu 和 Constantinou（2022）提出的一种混合算法，它将逐检验
  删除和逆概率权重（IPW）应用于基于评分的HC算法，以便更有效地处理离散变量中的随机缺失和系统性缺失数据。
- HCLC-V：采用VBEM的混合 爬山式潜在混杂因素搜索（HCLC-V），由
  Chobtham 和 Constantinou（2022）提出，可学习包含可能潜在混杂因素密度估计的图。该算法在最大祖先图空间中进行爬山搜索，并使用p-证据下界（也称为证据下界或变分下界）得分作为目标函数，以确定潜在混杂因素并学习其分布。p-证据下界得分不具备评分等价性，因此该算法会返回唯一的有向无环图。HCLC-V返回的有向无环图结构可能将潜变量作为观测变量。为确保以与其他算法相同的方式评估HCLC-V生成的图，我们会删除所有潜变量，并为潜在混杂因素引入必要的双向边，从而得到最大祖先图结构。
- MAHC：Constantinou等人（2022年）提出的模型平均爬山算法开始
  通过剪枝图的搜索空间来限制每个节点的候选父节点。该剪枝策略代表了应用于组合优化结构学习问题的激进版本。随后，它在爬山搜索过程中执行模型平均，并在以下假设下，移动到平均目标得分最大的邻接图：对该邻接图及其所有有效邻接图的目标得分进行平均，能够降低模型平均过程对数据噪声的敏感性，并使其更适用于真实数据。由于模型平均方法为每个节点分配多个BIC分数并对其求平均，因此得到的目标函数不具备评分等价性，并会导出唯一的DAG结构。

- MMHC：Tsamardinos等人（2006年）提出的混合最大-最小爬山（MMHC）算法，首先使用基于约束的最大-最小父节点与子节点（MMPC）算法构建骨架，然后使用爬山法确定这些边的方向。

- NOTEARS：Zheng 等人（2018年）提出的一种连续优化算法，它将传统的基于评分的组合优化问题转化为满足无环性要求的等式约束问题，并返回唯一的有向无环图。

- NOTEARS-MLP：Zheng 等人（2020年）对 NOTEARS 的扩展，利用基于偏导数的非参数稀疏性进行连续优化，从而在关系为非线性的假设下恢复有向无环图。

- Partition-MCMC：Kuipers 和 Moffa（2017年）提出的分区马尔可夫链蒙特卡罗（Partition-MCMC），用于搜索分区空间；分区也称为有向无环图中节点的部分拓扑排序。该方法通过考察与每个访问到的分区一致的所有有向无环图的得分，为各个分区评分，并返回一个与这些分区一致的随机有向无环图（不一定是得分等价的有向无环图）。
- PC-Stable：经典 Peter-Clark（PC）基于约束的算法（Spirtes 和 Glymour，1991年）的稳定版本。该算法从一个全连接图开始，删除导致边缘独立或条件独立的边。随后，通过一组 v-结构检验对其中一些边进行定向。Colombo 和 Maathuis（2014年）提出的 PC-Stable 变异株通过校正边删除的顺序，使 PC 不受从数据中读取变量顺序的影响。该算法生成一个部分有向无环图（PDAG），它是 CPDAG 的简化版本，也包含有向边和无向边，其中有向边表示 v-结构[4]。

- SaiyanH：由 Constantinou（2020年）提出的一种混合算法，它从一个更稠密的最大生成树图开始，并通过依次执行条件独立性检验、BIC最大化检验，以及最大化假设性干预效果的定向操作来确定这些边的方向。随后，它在有向无环图空间中进行禁忌搜索，并限制其不能访问包含不相交子图的图。这种学习策略可确保学习到的模型能够实现证据的充分传播。
- SED：Liu 等人（2022年）提出的虚假边检测（SED）算法可视为一个额外的学习阶段，可应用于任何其他结构学习算法的输出。该算法以图结构为输入，检查其中是否存在可能因测量误差而产生的假阳性边，并将其删除。具体做法是检查那些可能由测量误差导致的三顶点团。如表2所示，SED 被应用于另外五种基于评分、基于约束和混合算法。

- 结构 EM：结构 EM 是一种较早的算法，由 Friedman 于 1997年提出，用于从含有缺失值的数据中学习结构。它包含两个步骤：期望（E）步，在此步骤中推断缺失值以补全数据；以及最大化（M）步，在此步骤中使用完整数据集进行结构学习。它会从其等价类中返回一个随机的有向无环图。
- TABU：这是爬山法（HC）的一种扩展，采用禁忌搜索。它允许探索会降低目标得分的有向无环图，并维护一个禁忌列表，其中记录最近访问过的图，以防算法返回到近期访问过的图（Bouckaert，1995年）。与爬山法相比，这种方法使 TABU 能够进入新的图区域，而这些区域可能包含更优的局部最大值。

如表2所示，我们使用 Scutari（2023）的 bnlearn v4.8.1 软件包测试爬山法、TABU、PC-Stable、结构 EM 和 MMHC；使用因果发现中心（2023）的 Tetrad 软件包及其 Python 扩展 causal-learn 测试 FCI、GFCI 和 FGES；使用 Constantinou（2019）的 Bayesys v3.5 软件包测试 MAHC 和 SaiyanH；使用 Cussens 和 Bartlett（2015）的 GOBNILP v1.6.3 软件包测试 GOBNILP；使用 Suter 等人（2023）的 BiDAG 软件包测试 Partition-MCMC；并使用 Zhang 等人（2021）的 gCastle

[4] A v-结构指的是共同效应的因果类别；也就是说，$B \rightarrow A \leftarrow C$，其中$A$ 是B 和C的共同效应。

软件包测试 NOTEARS-MLP。所有算法均采用各软件包中实现的超参数默认值进行测试。请注意，我们还使用 Tetrad 软件包测试了 LiNGAM（Shimizu 等人，2006年），该算法用于估计线性因果模型的结构；也使用 gCastle 软件包测试了 DAG-GNN（Yu 等人，2019年）和 MCSL（Ng 等人，2022年），这两种算法采用非线性连续优化。然而，本研究测试的其他算法均在数秒或几分钟内完成结构学习，而这三种算法在三小时的结构学习运行时间内仍未返回结果，因此未纳入结果。

此外，尽管本研究考察了范围广泛的结构学习算法，但还有一些我们未曾探讨、值得在未来研究中考虑的算法。其中包括基于评分的 OBS（Teyssier 和 Koller，2005年）及其变体 ASOBS（Scanagatta 等，2015年）、INOBS（Lee 和 van Beek，2017年）和 WINASOBS（Scanagatta 等，2017年）。这些算法通过不同的节点排序遍历有向无环图的搜索空间，每种排序都会针对边的定向提供一组独特的约束，从而能够利用规模更大的数据集进行结构学习。在基于约束的学习方面，其他算法还包括 Li 等人（2019年）提出的 PC 修改版本，该版本确保已删除边的分离集与最终图保持一致；以及 Giudice 等人（2022年）提出的 Dual-PC，它通过改变条件独立性检验的执行顺序并优先处理某些高阶偏相关，降低了 PC 的计算复杂度。

5.1. 模型平均

除了分别研究各算法学习得到的图之外，我们还使用模型平均程序，获取并评估由特定一组图形输出生成的图：

- All_score-based：所有基于评分的算法的平均图；
- All_constraint-based：所有基于约束的算法的平均图；
- All_hybrid：所有混合学习算法的平均图；
- All_quartiles：对使用四分位数离散化的数据应用所有算法所得到的平均图；
- All_k-means：对使用 k 均值聚类离散化的数据应用所有算法后得到的平均图；

- 全部连续型：对连续数据应用所有算法后得到的平均图；
- All_mixed：对混合数据应用所有算法后得到的平均图。

我们在上文提到的 Bayesys 软件包中提供了这种模型平均方法。输入是从多个图中获得的一组边。模型平均程序假定有向边比无向边具有更高的确定性，因此优先处理有向边。该程序旨在尽可能多地确定边的方向（在我们的实验中，它确定了所有边的方向）。由于双向边表示因混杂而导致的不正确依赖关系，因此会被忽略；这与无边表示独立性的情况类似。给定一组有向边和无向边（其中包括对应于多个图的重复边），模型平均程序按以下步骤运行：

1. 按出现次数从高到低，将有向边添加到平均图中；a. 如果反方向的边已添加，则跳过该边；b. 如果添加该边会形成环，则反转该边并将其添加到边集 C；2. 按出现次数从高到低添加无向边；a. 如果该边已作为有向边添加，则跳过；3. 按出现次数从高到低，添加在 C 中找到的有向边；a. 如果该边已作为无向边添加，则跳过。

一个可选的超参数允许用户指定边被纳入平均图所需的最少出现次数。 我们指定阈值 θ ，使得相关学习得到的图中至少有$_1$个（共$_3$个）包含该边。具体而言：

- 基于 All_score 的：基于 31 个独立图构建，因此 $\theta = 10$;
- 全部基于约束的：构建时使用了 4 个独立图，因此 $\theta = 2$;
- All_hybrid：构建时使用了 17 个独立图，因此 $\theta = 6$;
- All_quartiles：构建时使用了 19 个独立图，因此 $\theta = 7$;
- All_k-means：构建时使用了 19 个独立图，因此 $\theta = 7$;
- 全部连续型：构建时使用了 14 个独立图，因此 $\theta = 5$;
- All_mixed：基于 5 个独立图构建，因此 $\theta = 2$;

表2. 64项结构学习实验。有关独立 GitHub 库的信息请参见第5节。

实验	算法	软件包	学习类别	输出数据格式		缺失性
1	CCHM	独立包混合		MAG	连续型	插补
2	FCI	TETRAD	基于约束的	PAG	离散型（四分位数）	插补
3	FCI	TETRAD	基于约束的	PAG	离散型（K均值）	插补
4	FCI	TETRAD	基于约束的	PAG	连续型	插补
5	FGES	TETRAD	基于评分	CPDAG	离散型（四分位数）	插补
6	FGES	TETRAD	基于评分	CPDAG	离散型（k 均值）	插补
7	FGES	TETRAD	基于评分	CPDAG	连续型	插补
8	FGES	TETRAD	基于评分	CPDAG	混合型	插补
9	GFCI	TETRAD	混合	PAG	离散型（四分位数）	插补
10	GFCI	TETRAD	混合	PAG	离散型（k 均值）	插补
11	GFCI	TETRAD	混合	PAG	连续型	插补
12	GOBNILP	GOBNILP	基于评分（精确）	CPDAG	离散型（四分位数）	插补
13	GOBNILP	GOBNILP	基于评分（精确）	CPDAG	离散型（k 均值）	插补
14	GOBNILP	GOBNILP	基于评分（精确）	CPDAG	连续型	插补
15	HC	bnlearn	基于评分	CPDAG	离散型（四分位数）	插补
16	HC	bnlearn	基于评分	CPDAG	离散型（k 均值）	插补
17	HC	bnlearn	基于评分	CPDAG	连续型	插补
18	HC	bnlearn	基于评分	CPDAG	混合型	插补
19	HC-aIPW	独立软件包	基于评分	CPDAG	离散型（四分位数）	不完整
20	HC-aIPW	独立软件包	基于评分	CPDAG	离散型（K均值）	不完整
21	HC-aIPW	独立软件包	基于评分	CPDAG	连续型	不完整
22	HCLC-V	独立包混合		DAG	离散型（四分位数）	插补
23	HCLC-V	独立包混合		DAG	离散型（K均值）	插补
24	MAHC	Bayesys	基于评分	DAG	离散型（四分位数）	插补
25	MAHC	Bayesys	基于评分	DAG	离散型（k 均值）	插补
26	MMHC	bnlearn	混合	CPDAG	离散型（四分位数）	插补
27	MMHC	bnlearn	混合	CPDAG	离散型（k 均值）	插补
28	MMHC	bnlearn	混合	CPDAG	连续型	插补
29	MMHC	bnlearn	混合	CPDAG	混合型	插补
30	NOTEARS	独立软件包	基于评分（连续优化）的DAG		连续型	插补
31	NOTEARS-MLP	gCastle	基于评分（连续优化）的DAG		连续型	插补
32	Partition-MCMC	BiDAG	混合	DAG	离散型（四分位数）	插补
33	Partition-MCMC	BiDAG	混合	DAG	离散型（K 均值）	插补
34	Partition-MCMC	BiDAG	混合	DAG	连续型	插补
35	PC-Stable	bnlearn	基于约束的	PDAG	离散型（四分位数）	插补
36	PC-Stable	bnlearn	基于约束的	PDAG	离散型（K均值）	插补
37	PC-Stable	bnlearn	基于约束的	PDAG	连续型	插补
38	PC-Stable	bnlearn	基于约束的	PDAG	混合型	插补
39	SaiyanH	Bayesys	混合	CPDAG	离散型（四分位数）	插补
40	SaiyanH	Bayesys	混合	CPDAG	离散型（k 均值）	插补
41	SED（爬山法）	独立软件包	基于评分	CPDAG	离散型（四分位数）	插补
42	SED（爬山法）	独立软件包	基于评分	CPDAG	离散型（k 均值）	插补
43	SED（整数线性规划）	独立软件包	基于评分	CPDAG	离散型（四分位数）	插补
44	SED（整数线性规划）	独立软件包	基于评分	CPDAG	离散型（k 均值）	插补
45	SED（MMHC）	独立包混合		CPDAG	离散型（四分位数）	插补
46	SED（MMHC）	独立包混合		CPDAG	离散型（K均值）	插补
47	SED（PC）	独立包混合		CPDAG	离散型（四分位数）	插补
48	SED（PC）	独立包混合		CPDAG	离散型（k 均值）	插补

49	SED (TABU)	独立软件包	基于评分	CPDAG	离散型（四分位数）	插补
50	SED（TABU）	独立软件包	基于评分	CPDAG	离散型（k 均值）	插补
51	结构 EM	bnlearn	基于评分	CPDAG	离散型（四分位数）	不完整
52	结构 EM	bnlearn	基于评分	CPDAG	离散型（k 均值）	不完整
53	结构 EM	bnlearn	基于评分	CPDAG	连续型	不完整
54	TABU	bnlearn	基于评分	CPDAG	离散型（四分位数）	插补
55	TABU	bnlearn	基于评分	CPDAG	离散型（k 均值）	插补
56	TABU	bnlearn	基于评分	CPDAG	连续型	插补
57	TABU	bnlearn	基于评分	CPDAG	混合型	插补
58	All_基于约束的	Bayesys	模型平均	DAG	n/a	n/a
59	全部连续型_	Bayesys	模型平均	DAG	n/a	n/a
60	全部_混合	Bayesys	模型平均	DAG	n/a	n/a
61	所有K均值_	Bayesys	模型平均	DAG	n/a	n/a
62	所有混合型_	Bayesys	模型平均	DAG	n/a	n/a
63	所有_四分位数	Bayesys	模型平均	DAG	n/a	n/a
64	所有基于得分的_	Bayesys	模型平均	DAG	n/a	n/a

6 评估与结果

我们首先考察由结构学习算法生成的图、平均图以及依据人类知识构建的图所得到的模型的维度。随后，我们在第6.2小节从图结构差异的角度评估这些图，在第6.3小节从推断和预测验证的角度进行评估，在第6.4小节评估模拟假设性干预效果的能力，并在第6.5小节进行敏感性分析。在第6.6小节，我们探讨其中一些算法能告诉我们什么关于混杂的信息。最后，我们在第6.7小节对这些图结构进行定性评估，并参考 COVID-19 案例研究中现实世界里合理的关系。

需要澄清的是，我们的目的并非独立评估这些算法，而是考察结构学习的整体能力，包括探究第5节所述模型平均的有效性。因此，我们更侧重于了解从结构学习算法整体中能够获得什么，而非单独考察各个算法。如下文各小节所示，在所考察的不同评估标准下，表现通常优于其他方法的平均图是 All_score-based 和 All_k-means，分别见图2和图3。正如下文各小节所讨论的，每种算法学习得到的大多数独立图以及平均图往往彼此不同。其余五个平均图见附录B中的图B1、B2、B3、B4和B5。

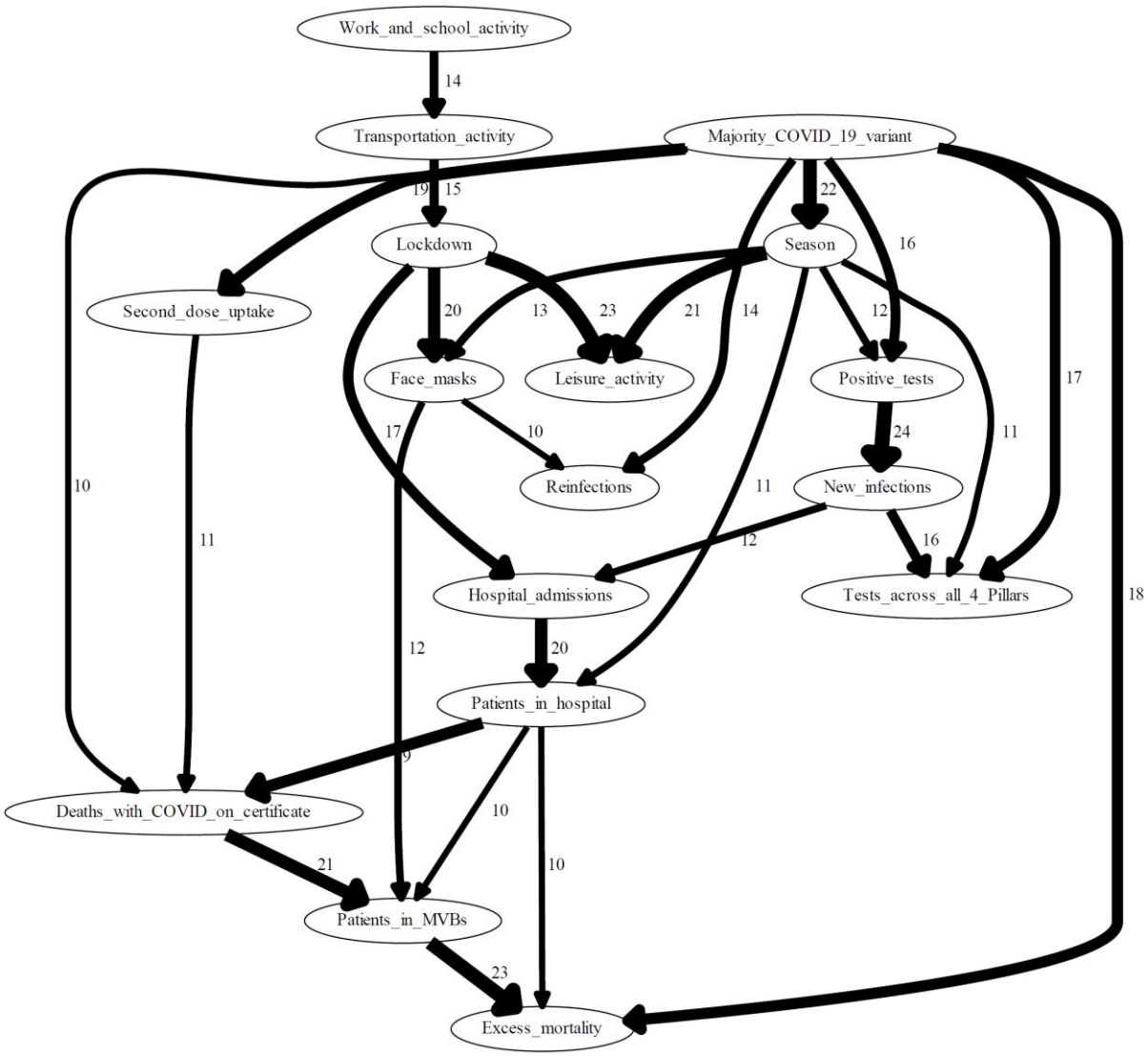


图2。通过31次基于评分的实验得到的 All_score-based 平均图。该图共包含29条边，边标签表示相应边在所考察的31个输出中出现的次数，边的宽度随出现次数增加。未显示在31个输入图中出现次数少于10次的边。

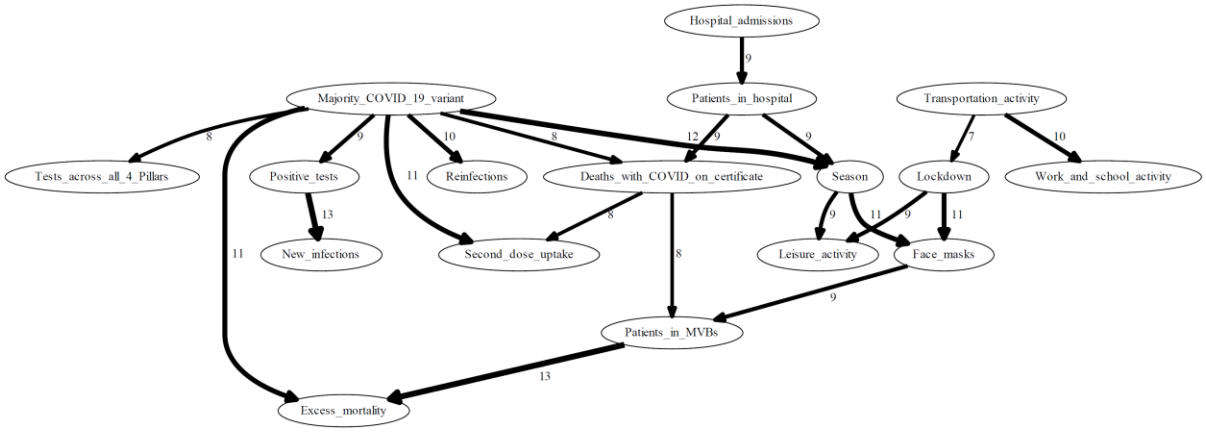


图3。通过19次实验得到的 All_k-means 平均图，这些实验使用 K均值聚类对输入数据集进行离散化。该图共包含21条边，边标签表示相应边在所考察的19个输出中出现的次数，边的宽度随出现次数增加。未显示在19个输入图中出现次数少于7次的边。

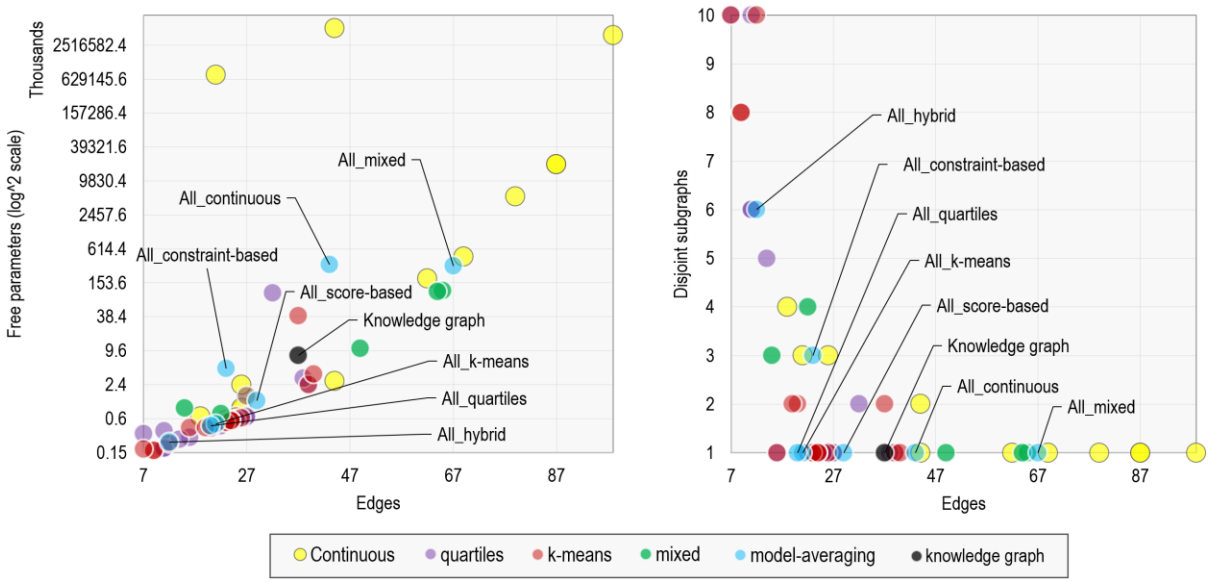


图4. 按输入数据格式分类的独立算法、平均图和知识图谱学习到的离散模型的二维散点图。

6.1. 维度

我们首先研究从不同图结构中获得的模型的维度。虽然最佳模型维度应达到何种水平并不总是明确，但本小节报告的发现有助于我们理解后续小节中报告的图性能和推断方面的差异。

图4展示了将所有图（无论输入数据格式如何）转换为离散贝叶斯网络模型后，边的数量与不相交子图数量及自由参数数量之间的关系。自由参数也称为独立参数，是衡量离散模型维度的一项指标。虽然使用自由参数来评判从连续数据中恢复的结构可能并不恰当，因为这一指标反映的是分类分布或离散分布的维度，但我们在此仅将其用于比较，以突出图复杂度方面的重要差异。此外，不相交子图的数量代表一张图中彼此独立的图形片段的数量。例如，附录B中图B2所示的平均图由六个不相交子图组成。理论上，生成包含不相交子图的图的算法并没有任何问题，但在实践中，我们可能需要能够实现完整证据传播的模型（Constantinou, 2020年）。

图4中的散点图很快揭示出，从数据中学习得到的图之间存在显著差异。具体而言，这些图中的边数从低至7条到高达98条不等。不同算法和模型平均方法所生成的边数差异，对自由参数的数量有显著影响，其范围从仅162个到超过50亿个。请注意，这种极端差异主要是由使用连续数据训练的算法造成的；根据图4所示结果，与使用分类数据学习得到的图相比，这些算法往往会生成更稠密的图。同一算法在不同类型的数据上训练时，也会出现这种差异。

需要说明的是，从推断的角度来看，某些学习到的结构会导致自由参数数量多到难以处理，这主要是因为它们为某些节点生成了大量父节点，而非因为所包含的边总数过多。例如，自由参数数量最多的 NOTEARS 仅学习到 44 条边，远少于许多其他算法，而这些算法得到的模型维度都更低。具体而言，NOTEARS 学习到的 44 条边中，有 15 条边从 15 个节点指向同一个节点。考虑到数据集仅包含 17 个变量，这种情况相当不现实；也就是说，在上述案例中，算法从可能的 16 个父节点中识别出了 15 个

父节点。结构 EM 产生了近 40 亿个自由参数，也生成了 15 条边指向同一个节点；但在这种情况下，学习得到的图中共有 98 条边。NOTEARS-MLP 产生了约 10 亿个自由参数，在发现的 21 条边中，共有 14 条边指向同一个节点。接下来是这三个极端案例之后的 TABU 和爬山法，它们生成的网络产生了近 2,000 万个自由参数；也就是说，比 NOTEARS 产生的自由参数少 258 倍，但就模型维度而言仍然非常高。TABU 和爬山法都学习到了 87 条边，节点最大入度为 11。上述五个极端案例都涉及连续数据的结构学习，这解释了为什么学习过程高估了分类数据的维度。具体而言，爬山法、TABU 和结构 EM 在分类数据上训练时产生的自由参数介于 552 至 653 之间，而在连续数据上训练时则从 2,000 万到 40 亿不等。这表明数据格式会对学习得到的输出产生显著影响。

总而言之，图4中的结果揭示了一个关键差异：在包含连续变量的数据上训练的算法（即图4中归类为连续型和混合型的算法）与在离散变量上训练的算法（即归类为四分位数和K均值的算法）之间存在显著差异。具体而言，前者会产生更稠密的图；当学习得到的图被转换为离散贝叶斯网络时，其维度可能变得难以用于推断，而后者则会产生更稀疏的图，并对应较低的模型维度。需要注意的是，这一观察同样适用于在不同数据格式上训练的相同算法。不出所料，知识图谱位于这两个极端的中点附近。就不相交子图而言，结果也呈现出类似模式：混合型和连续型数据集会产生包含较少不相交子图的图，而一些从离散数据中学习得到的稀疏图则包含多达10个不相交子图。

平均图能够有效缓和这些极端差异。例如，All_continuous 图包含的边少于大多数在连续数据上训练的算法所发现的边，而 All_k-means 包含的边多于大多数在经过K均值离散化的分类数据上训练的算法所发现的边。就模型维度和不相交子图数量而言，模型平均方法似乎也能产生更符合实际的图。

6.2. 图结构评估

图结构评估是指仅考察图结构差异，而不考虑推断或模型拟合的过程。本小节将考察从数据中学习得到的图与基于人类知识构建的图之间的图结构差异。

我们考虑三个相关指标。第一，结构汉明距离（SHD），即将学习得到的图转换为知识图谱（或已知时的真实值）所需的变更次数。第二，F1 分数，即召回率（R）与精确率（P）的调和平均数；也就是 $F1 = 2\frac{RP}{\_\_\_}$。第三，平衡评分函数（BSF），它会对$_{R+P}$得分进行平衡，从而考虑发现存在一条边的难度与发现不存在一条边的难度：

$$BSF = 0.5\left(\frac{TP}{a} + \frac{TN}{i} - \frac{FP}{i} - \frac{FN}{a}\right)$$

其中，TP、TN、FP 和 FN 是混淆矩阵中的各项，a 是边的数量，i 是真实值中的独立关系数量；其中 $i = {}^{|V|}{}_{(|V|}{}_{2}{}^{-1} - a$，而 $|V|$ 是节点数量。~~BSF 分~~数的取值范围为 -1 到 1，其中 -1 对应最差的图，即在假定的真实值中，若一对节点之间存在边，则图中不存在边；若不存在边，则图中存在边。得分为 1 对应与真实值相符的图，得分为 0 则对应信息量与空图或全连接图相当的图。最后，这三项指标应用于相应的 CPDAG 和部分祖先图；也就是说，它们表示马尔可夫等价类之间的差异。我们用于计算得分的规则见表 3。

表 3。图形评估指标在将学习得到的 CPDAG 或 PAG 与知识图谱进行比较时所使用的惩罚权重（即假阳性或假阴性）。为简化起见，对于包含边 o→和 o—o 的 PAG 输出，我们分别假设它们对应有向边和无向边。这些规则取自（Constantinou 等人，2022）。

Rule	True 图	学习得到的图	惩罚	推理
1	$A \rightarrow B$	$A \rightarrow B$	0	完全匹配
2	$A \rightarrow B$	$A \leftrightarrow B$, $A - B$, $A \leftarrow B$	0.5	部分匹配
3	任意边	无边	1	不匹配
4	$A \leftrightarrow B$	$A \leftrightarrow B$	0	完全匹配
5	$A \leftrightarrow B$	$A - B$, $A \leftarrow B$, $A \rightarrow B$	0.5	部分匹配
6	$A - B$	$A - B$	0	完全匹配
7	$A - B$	$A \leftrightarrow B$, $A \leftarrow B$, $A \rightarrow B$	0.5	部分匹配
8	无边	无边	0	完全匹配
9	无边	任意边/弧	1	不匹配

图5展示了一个散点图，将 SHD 分数与 F1 和 BSF 分数的平均值进行对比。阴影象限根据中位数区分结果。具体而言，落在绿色象限中的结果，其 SHD 和 F1+BSF 分数都优于中位数；而落在红色象限中的结果[2]则两项都低于中位数。黄色象限表示一项指标高于中位数、另一项低于中位数的情况。值得关注的一些结果包括：

i. 在连续型和混合型[5] 数据上训练的算法往往落在红色象限；ii. 在分类数据上训练的算法往往分布于所有象限，不受离散化方法的影响；iii. 精确学习落在黄色和红色象限；iv. 通过模型平均获得的图往往能同时提高两项得分，因为它们通常会朝绿色象限移动。

可以认为，知识图谱与从数据中学习得到的图之间存在显著分歧，因为没有任何算法或平均图产生的整体F1+平衡评分函数得分高于[2]035。。从SHD分数也可以得出相同~~结论，~~ 其范围在30到90之间。尽管如此，图1所示的知识图谱仅包含37条边；这意味着空图的SHD分数为37，反而优于大多数结构学习算法学习得到的图所产生的得分。

[5] 请注意，应用于混合数据的算法会固有地施加一项约束，禁止连续变量成为离散变量的父节点。因此，在图B1中，四个分类变量 Season、Face_masks、Majority_COVID_19_variant 和 Lockdown 均没有连续变量作为父节点。

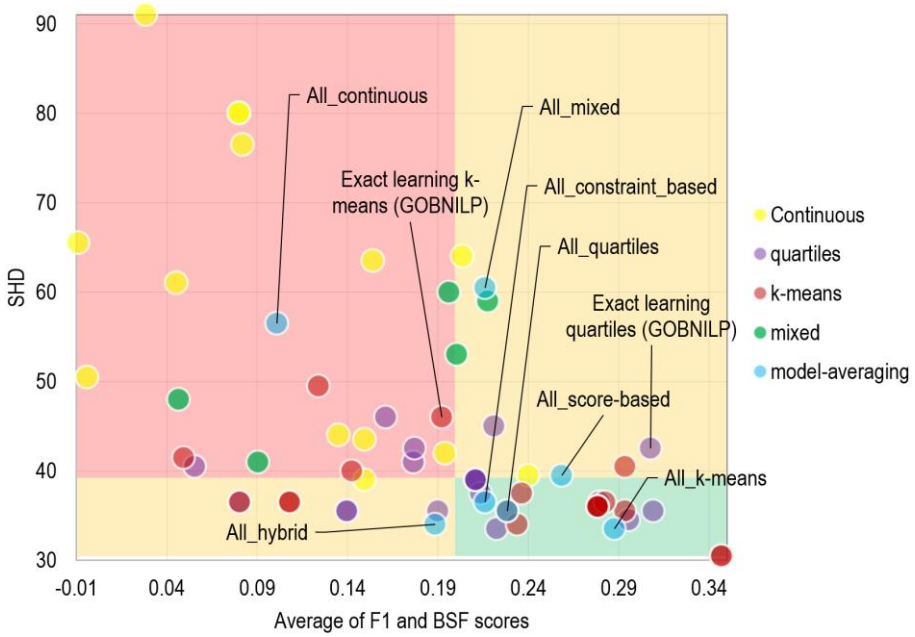


图5。二维散点图展示了知识图谱与以下方法生成的图之间的结构差异：a）按输入数据格式分类的独立算法；以及 b）应用于不同算法组的模型平均方法。阴影象限表示中位数边界。较低的 SHD 得分代表更好的性能，而较高的 F1 和 BSF 分数代表更好的性能。

我们通过衡量平均图与知识图谱之间的相似度，进一步考察这些结果。结果见表4，但此处省略了BSF指标，因为它具有非对称性；也就是说，该指标会受到我们指定为真实值的图的影响。总体而言，结果表明，平均图之间的大多数（但并非全部）得分都比平均图与知识图谱之间的得分更接近。这意味着，大多数算法生成的图与其他某种算法生成的图更接近，而非与知识图谱更接近；不过，有些从数据中生成的图彼此之间的差异，反而大于它们与知识图谱之间的差异。例如，最相似的平均图只需进行10次图结构修改（即SHD为10）即可变得完全相同，而最不相似的平均图则需要进行62次图结构修改才能变得完全相同。

表 4. 不同平均图对之间以及平均图与知识图谱之间的 F1（对角线下方）和 SHD（对角线上方）得分。绿色和红色分别表示相似度较高和较低。

学习方法	egde_wonK	;n-arisnoc_A	desab	suouninoc_A	drbyh_A	snaem-k_A	dex-m_A	seirauq_A	desab-erocs_A
知识		36.5	56.5	33	32.5	59	34.5	39.5
All_基于约束的	0.283		34	17.5	25	57.5	23	30.5
All_连续型	0.238	0.424		33.5	38	64	35	38
全部_混合	0.286	0.429	0.382		10.5	58	11	18
All_K均值	0.397	0.364	0.375	0.636		56.5	10.5	10
All_混合型	0.385	0.3	0.345	0.253	0.33		51.5	50
全部_四分位数	0.333	0.419	0.413	0.625	0.707	0.379		15.5
All_基于评分	0.348	0.327	0.417	0.537	0.76	0.438	0.633	

6.3. 基于推断的验证与预测验证

请回顾第6.2小节所介绍的图结构评估：它仅考察图结构，并未考虑模型的推断能力。本小节将探讨把学习到的结构参数化为离散贝叶斯网络后所具备的推断能力。我们通过模型拟合、模型选择以及针对分类数据进行的交叉验证检验来评估学习到的贝叶斯网络；换言之，评估对象是采用K均值和四分位数离散化方法处理后的数据集。

我们首先考察对数似然（LL）和贝叶斯信息准则（BIC）得分：前者表示模型对数据的拟合程度，后者是一种模型选择函数，用于平衡模型维度与模型拟合。图6以二维散点图绘出了这些得分，每种数据离散化方式各对应一幅图。左图基于经四分位数离散化的数据，右图则基于经K均值聚类离散化的数据。

图6中的结果表明，总体而言，从使用K均值离散化的数据所导出的图中学习到的模型，其拟合得分高于从使用四分位数离散化的数据所导出的模型。这是预期之中的，因为四分位数旨在实现平衡分布，而K均值这类聚类方法则会发现能够最大化拟合效果的分布。尽管LL 拟合效果存在差异，但两个图中的结果似乎是一致的。

我们发现，All_continuous图和All_mixed图产生的模型选择分数要差得多；这一观察结果与6.1节中的结果一致，后者表明，连续变量会导致更稠密的图，而这些图在转换为分类模型后，往往包含数量庞大的参数，因此在模型选择分数中会产生较大的复杂度惩罚。我们还注意到，模型维度的增加远远无法通过模型拟合方面的任何改进来证明其合理性。也就是说，All_mixed产生的BIC 得分尽管具有最高的LL 得分，却在图6中排名倒数第二，这表明LL 拟合的提升很可能是由过拟合造成的。此外，All_continuous模型的LL 拟合得分低于All_score-based，尽管前者是一个明显更简单的模型（根据BIC 得分判断）。

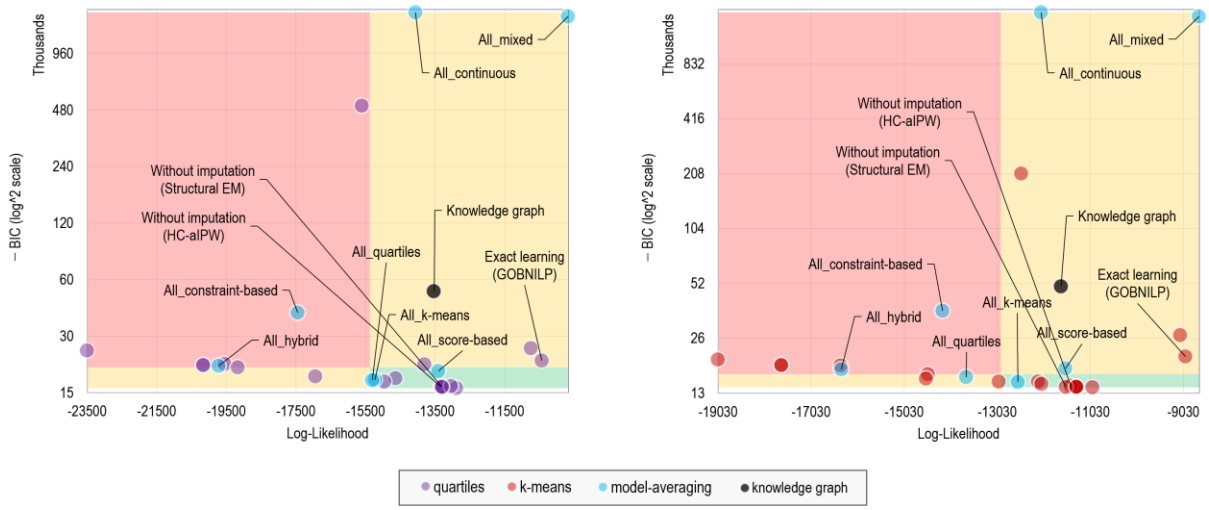


图6. 对数似然与 BIC 分数之间关系的二维散点图。左图假设输入数据按四分位数离散化，右图则假设输入数据通过 k均值聚类进行离散化。阴影象限表示中位数边界。排除了返回最大祖先图或部分祖先图输出的算法。

其他值得关注的结果包括从知识图谱和All_constraint-based图中获得的模型，它们产生的模型选择分数相对较低。对于知识图谱而言，这一结果并不令人意外，因为我们不会期望知识在其设计目标之外胜过算法；也就是说，算法通常旨在最大化某个模型选择分数或目标函数。对于All_constraint-based而言，这一观察结果部分可以由其学习类别（即基于约束的）解释，因为该类别侧重于局部条件独立性检验，往往会忽略诸如由LL 和BIC导出的全局分数。这也解释了为什么All_score-based在LL 和BIC 得分方面都表现得好得多。最后，颇有些出人意料的是，精确学习GOBNILP产生的模型选择分数低于中位数，尽管其设计目标是返回全局最大值。一种可能的解释是，精确学习在节点最大入度方面存在超参数限制；为处理计算效率问题，该限制默认为“3”，这意味着包含入度超过3的节点、且本可能带来更高模型选择分数的图并未得到探索。不过，这通常无法完全解释精确学习GOBNILP在模型选择分数上的相对较差表现，这种表现可能受到真实数据中自然存在的数据噪声影响。例如，Constantinou等人（2021）表明，与精确学习器或更复杂的近似学习器相比，简单学习器对数据噪声具有更强的抵抗力。

除了模型拟合和模型选择分数外，我们还根据交叉验证性能考察学习所得模型的推断能力。我们将学习所得模型载入 GeNIe BN软件（BayesFusion，2022年），并在两个离散数据集上进行10折交叉验证。图7展示了在全部17个变量上取得的平均交叉验证分类准确率；也就是说，对17个变量分别进行预测时，17次10折交叉验证结果的平均值。请注意，交叉验证准确率相差1%是一个相当显著的差异，因为这意味着在全部17个节点上的平均分类准确率相差1%。

图7中的结果与之前的结果存在一些一致性。具体而言，连续变量的结构学习再次被发现会导致较差的性能。这进一步支持了我们的假设，即从连续变量中学习到的图中观察到的图密度增加可能是由模型过拟合造成的。此外，由于维度极高，从连续数据中学习到的一些图无法进行参数化（参见图4），这些病例未被纳入图7。这些病例包括无法进行参数化的 HC、TABU、结构 EM、NOTEARS 和 NOTEARS-MLP 的连续数据版本，以及其 PDAG输出无法始终扩展为有向无环图结构的 PC（所有数据版本）。另一方面，从混合数据中学习到的许多结构确实表现出了较强的性能，这在一定程度上与之前关于连续数据的结果相矛盾。有趣的是，平均图 All_k-means 和 All_score-based 在交叉验证方面也表现良好，尤其是在 K均值病例下。

6.4. 干预分析

因果建模能够模拟干预。我们使用 GeNIe 软件（BayesFusion，2022年），该软件实现了 Pearl 的 *do* 算子，这是一种干预的数学表示（Pearl，2012年），用于衡量假设性干预的效果，并比较这些效果在不同学习到的结构之间的差异。鉴于这项新冠肺炎案例研究，我们对一组在现实生活中适合进行干预的关注变量开展干预分析。不过请注意，由于数据所反映的政策干预体现的是没有时间偏移的观察性关系（例如，当前政策与大流行病当前的状况相关），我们无法模拟政策干预的效果，因为这需要检验分布偏移（例如，探索政策的未来影响），而这些算法并不探索这一点；我们将在结论中详细讨论这一局限性。因此，我们衡量假设性干预对以下内容的影响：

1. COVID-19 检测的数量；即 do(四大支柱的检测)，如表6所示。例如，我们可能想要衡量增加 COVID-19 检测数量是否有益。2. COVID-19 感染的数量；即 do(New_infections)。例如，我们可能想要衡量控制感染率的效果。3. 住院患者中的 COVID-19 患者数量；即 do(住院患者)。例如，探究降低住院治疗风险或让住院患者提前出院的益处。4. MVB中的 COVID-19 患者数量；即 do(MVB患者)。例如，探究可能降低需要使用 MVB 风险的药物治疗或其他治疗的益处。

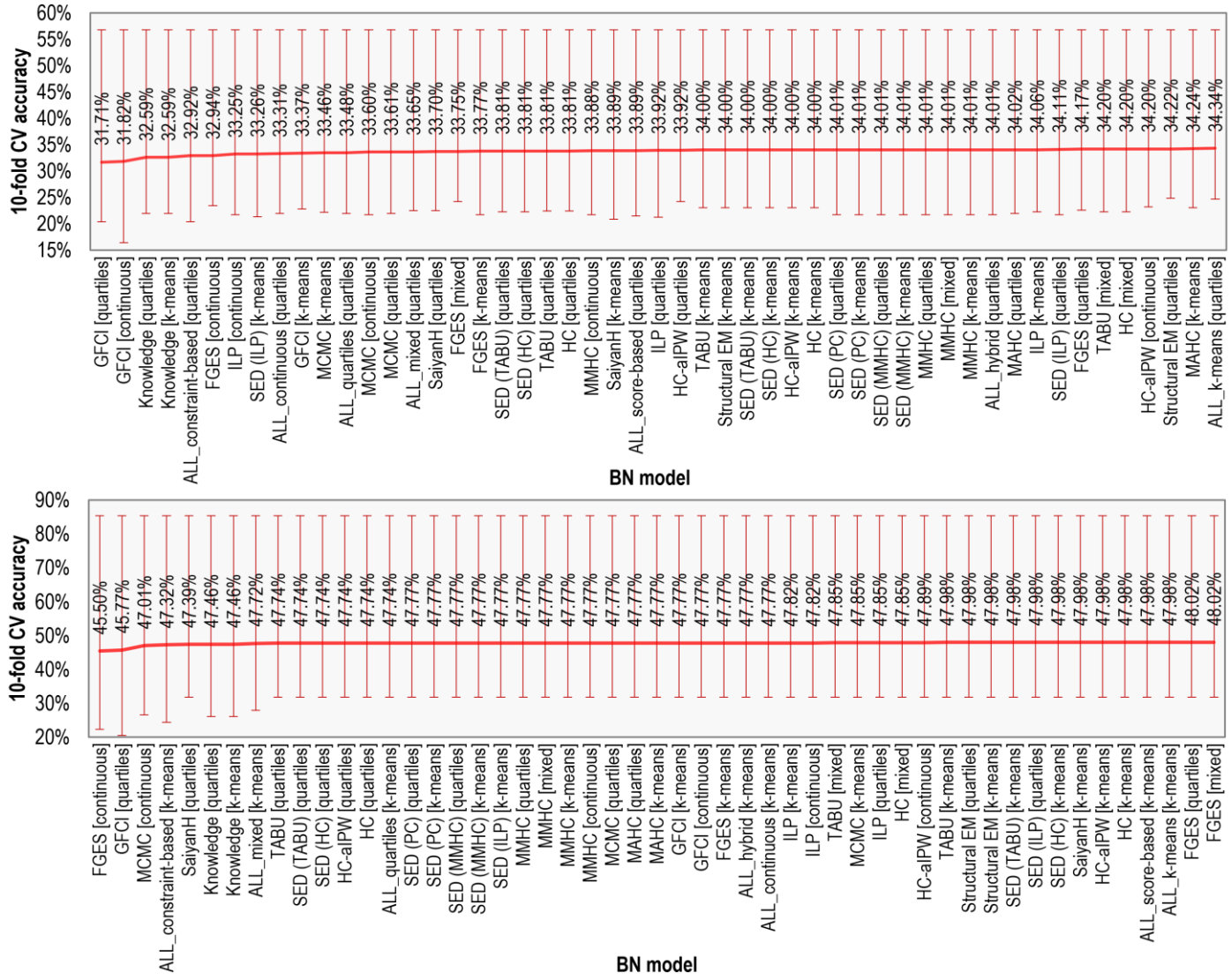


图7. 对全部17个节点的10折交叉验证分类准确率取平均值，并按性能从差到好排列。上方图表基于使用四分位数离散化的分类数据所参数化的模型，下方图表则基于使用K均值聚类离散化的分类数据。返回包含双向边的最大祖先图或部分祖先图输出的算法，以及爬山法、TABU、结构 EM、NOTEARS、NOTEARS-MLP 和 PC 学得的部分图，均不纳入此评估（说明请参阅正文）。误差线的宽度表示全部17个节点中最低与最高交叉验证得分之间的距离。

表5描述了我们用于衡量干预效应的线性尺度，表6则展示了上述干预作用于四个关注变量集合时的效应，分别基于七个平均图和知识图谱。由于所有变量均由序数分布表示，我们以分布偏移来衡量干预效应，并在线性尺度上将该偏移归一化到0至1之间。假设我们对变量A 进行干预，随后衡量该干预对变量B的效应，则（为简单起见，假设为确定性分布）：

表5。假设干预效应的衡量。效应得分根据分配给每个状态的概率进行加权[6] 。

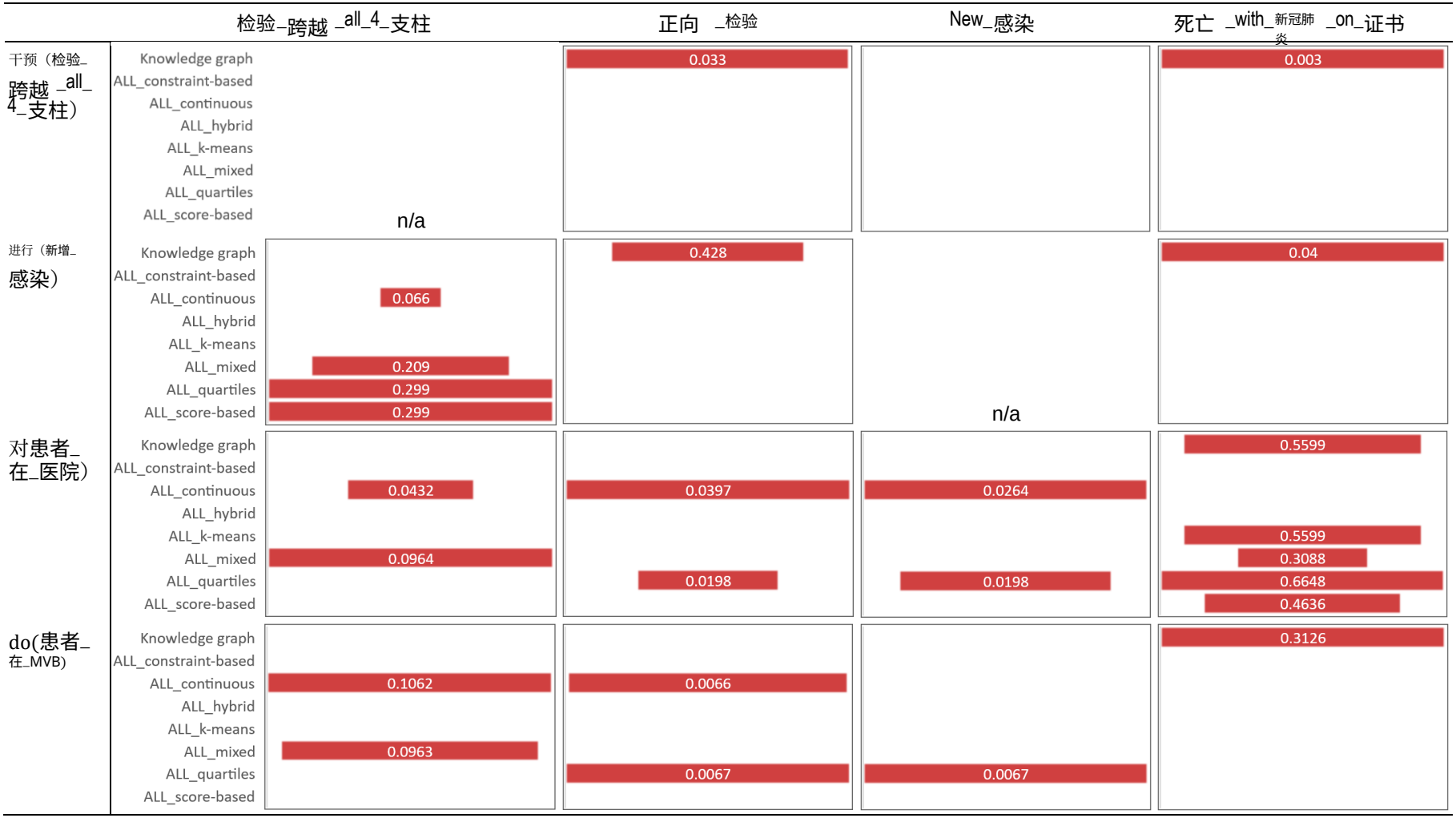
	do (A=极低_)	do (A=极高_)	**效应得分**
B	p(非常_低) = 1	p(非常_高) = 1	1
	p(非常_高) = 1	p(非常_低) = 1	1
	p(非常_低) = 1	p(高) = 1	0.66
	p(低) = 1	p(很_高) = 1	0.66
	p(非常_高) = 1	p(低) = 1	0.66
	p(高) = 1	p(非常_低) = 1	0.66
	p(非常_低) = 1	p(低) = 1	0.33
	p(低) = 1	p(高) = 1	0.33
	p(高) = 1	p(非常_高) = 1	0.33
	p(低) = 1	p(非常_低) = 1	0.33
	p(高) = 1	p(低) = 1	0.33
	p(非常_高) = 1	p(高) = 1	0.33
	p(非常_低) = 1	p(非常_低) = 1	0
	p(低) = 1	p(低) = 1	0
	p(高) = 1	p(高) = 1	0
	p(非常_高) = 1	p(非常_高) = 1	0

表6所呈现的结果表明，不同的图结构可能产生截然不同的干预效应。这些结果与第6.3小节中的预测验证结果形成对比，后者表明，图结构的巨大差异只会转化为预测性能上的微小差异。值得强调的结果包括：

i. 对检测的影响：大约一半的平均图表明，对感染进行干预会对检测产生中等影响，而对住院和MVB进行干预会对检测产生较弱影响。另一方面，知识图谱以及另一半平均图表明，在假设检验数量决定阳性病例数量的情况下，干预对检测没有影响。ii. 对阳性检测的影响：知识图谱表明，对感染数量进行干预会极大地影响观测到的阳性检测数量。所有其他干预和平均图均表明，对阳性检测数量的影响为零或可以忽略不计。iii. 对感染的影响：所有图结构都认为，所研究的干预措施均不会对感染数量产生有意义的影响。iv. 对死亡的影响：所有图结构都认为，对检验数量和感染率进行干预，对新冠肺炎死亡人数不会产生影响或仅产生可以忽略不计的影响（这相当令人意外）。另一方面，知识图谱（包括大多数平均图）都认为，对住院进行干预会极大地影响死亡人数。有趣的是，只有知识图谱表明，对患者需要收治至MVB的概率进行干预会影响死亡人数。

6 例如，如果 $do(A$ =非常_低)产生的效应分布为 { 0.5, 0.25, 0.15, 0.1 }，得分为 0.5×0+0.25×0.33+0.15×0.66+0.1×1=0.2815，而 $do(A$=非常_高)产生的效应分布为 {0.6, 0.3, 0.1, 0}，得分为 0.6×0+0.3×0.33+0.1×0.66+0×1=0.165，那么效应得分就是 |0.2815-0.165|=0.1165。

表6。指定的假设性干预对四个关注变量的影响。干预的影响以分布偏移表示，并按照表5所述归一化至0至1。结果分布在七个平均图和知识图谱中。没有结果／红色条表示没有影响。



6.5. 敏感性分析

与干预分析通过考察干预对子节点和后代节点的影响来探索假设性干预的效果不同，敏感性分析关注的是一个节点对其父节点和祖先节点的敏感程度。具体而言，高敏感性表示对某个节点的条件概率表参数进行微小更改，就会对其后验分布产生显著影响，这表明在贝叶斯网络结构给定的情况下，彼此敏感的变量之间存在依赖关系；而低敏感性则意味着即使大幅更改条件概率表参数，对其后验分布的影响也很小，这表明彼此不敏感的节点之间不存在依赖关系。Castillo等人（1997年）最初提出了敏感性分析，将其作为评估贝叶斯网络概率参数的一种方法。

我们使用 GeNIe BN软件（BayesFusion，2022年）进行敏感性分析。为简化起见，我们对单个变量进行敏感性分析，以展示不同学习到的结构在敏感性结论上的一些差异。我们重点关注新冠肺炎导致的死亡人数，并假设这是关注变量中的关键变量之一。

图8展示了根据由知识图谱和平均图得到的结构，对节点 Deaths_with_COVID_on_certificate 进行的敏感性分析。我们仅展示这些图中相关的局部结构，其中包含应用敏感性分析的目标节点的父节点和祖先节点。不属于目标变量的父节点和祖先节点集合的节点以灰色标示，其余节点以红色标示；红色越深，表示敏感性越强。请注意，不应比较不同图中红色的深浅。

不同学习到的结构所得结果之间存在的不一致也体现在敏感性分析中。具体而言，All_constraint-based 和 All_hybrid 图表明，死亡与其他所有节点均不敏感，因为这些图中没有目标节点 Deaths_with_COVID_on_certificate 的任何父节点。另一方面，All_continuous、All_k-means 和 All_quartiles 图表明，死亡对住院最为敏感；而 All_score-based、All_k-means 和知识图谱则表明，死亡对 COVID-19 变体最为敏感。相反，All_mixed 图表明，COVID-19 导致的死亡对封锁措施最为敏感。

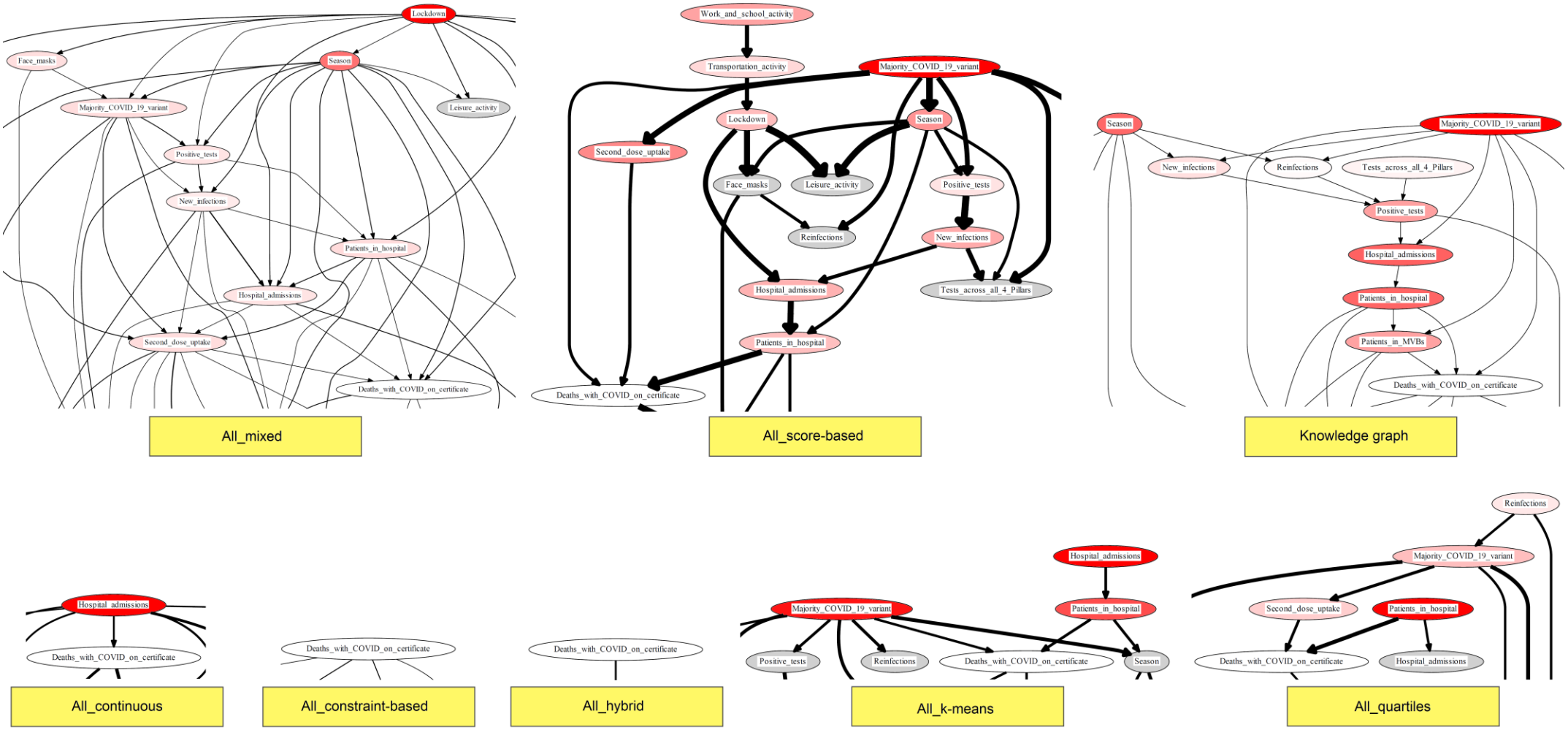


图8。基于知识图谱和平均图，对目标节点 Deaths_with_COVID_on_certificate 进行敏感性分析；这些图使用离散 K均值数据训练。图中仅显示包含相关父节点和祖先节点的图形片段（请注意，这些片段保留了平均图中的边宽）。白色节点表示目标节点，灰色节点不属于父节点和祖先节点集合，红色节点表示对目标节点敏感，其中更深的红色表示敏感性更强。

6.6. 混杂评估

本小节考察那些假设因果不足的算法。因果不足是指输入数据并未包含所有关注变量这一假设，其中一些关注变量可能缺少共同原因（即潜在混杂因素），从而导致学习到的结构存在混杂。如表1所示，九项结构学习实验基于能够生成包含双向边的图、以表示混杂的算法。这些实验涉及 FCI、GFCI、CCHM 和 HCLC-V 算法，它们生成的输出为最大祖先图或部分祖先图。

图9展示了发现的双向边所构成的平均图（不包含任何有向边或无向边），以此表示变量对之间可能存在的混杂。该平均图采用第5.1小节所述的同一模型平均方法生成，但重点改为双向边。不过，由于只有两条双向边满足第5节所定义的阈值θ（在图9中以黑色显示；本例中为$\theta = 3$），我们也选择展示那些仅差一次计数、未达到该阈值的边，并以红色显示。

不同算法和所考虑的数据格式在混杂方面得出的结果也存在不一致。例如，在九项实验中，只有三项实验包含图9所示的相同两条双向边，而对于其他所有边，一致程度都更低。此外，鉴于双向边表示混杂，而混杂又会导致伪边，这意味着假设因果充分性的算法（即本文研究的其他所有算法）会将这些伪边识别为真实边。换言之，我们预计前几小节中测试的算法会在图9中由双向边连接的节点对之间发现一条边。然而，参照图9中的黑色双向边，七个平均图中只有四个在Excess_mortality与Patients_in_hospital之间包含一条边，只有两个在Tests_across_all_4_Pillars与Work_and_school_activity之间包含一条边；并且这些边中的大多数都出现在All_continuous图和All_mixed图中，而这两类图都是稠密图。这一观察结果进一步凸显了所研究的不同结构学习方法之间存在的分歧。

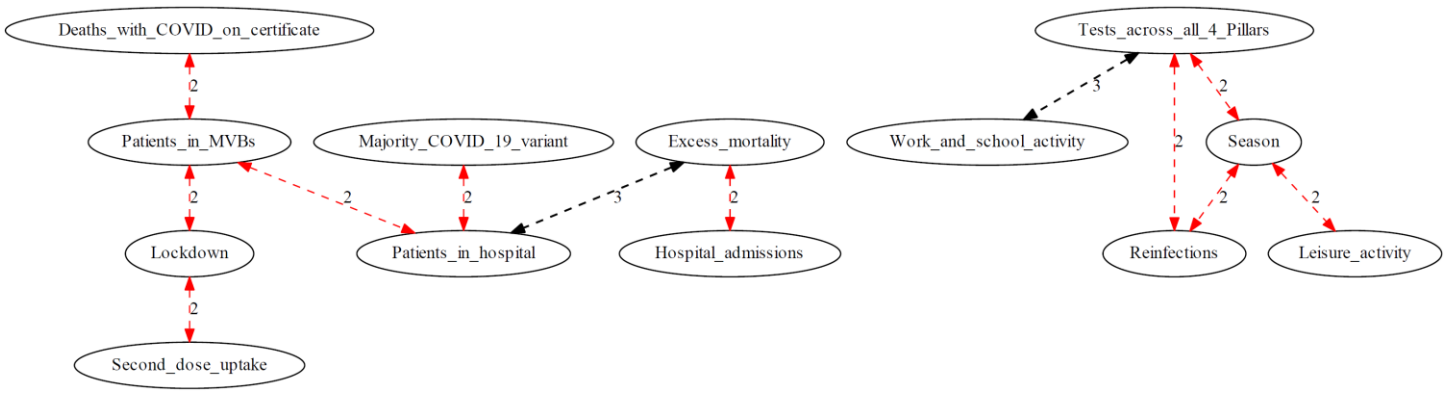


图9. 仅包含双向边（不含有向边和无向边）的平均图，表明存在混杂。黑色边表示按照第5节中模型平均过程所定义的阈值 θ，满足该阈值的边。由于只有两条边满足此标准（即，$\theta = 3$ 在九个参与平均的单独图中），我们还用红色标出了未达到该阈值、但仅差一次计数的双向边。

6.7. 案例研究评估

本小节重点对结构学习算法恢复出的因果关系进行定性评估。我们关注那些已得到充分理解，且我们相对确信其存在于真实值中的关系。因此，这些定性评估基于一系列基于知识的假设，即哪些关系可能是正确的，哪些可能不是正确的。这一主观评估见表7，重点考察七个平均图。

表7。对七个平均图进行定性评估，并参照所列的新冠肺炎关系；这些关系被假定为常识。如果结构学习所产生的学习结果中，超过半数与因果知识一致，我们就认为这些结果符合因果知识。

定性评估	预期	观察	与……一致因果知识?
感染	任意感染与任意检测之间的关系，不对因果方向作具体要求。	存在于……中的关系 共 7 个图中的第 7 个。	Yes
疫苗接种	第二_剂_接种率之间的关系和 死亡证明上注明*COVID*的死亡 或_ _ _ _超额_死亡率，无特定要求关于因果关系的方向。	存在这种关系的情况：7 个图中有 4 个。	Yes
检测	任意检测与任意感染之间的关系，不要求因果方向。	关系存在于 7 个图中有 7 个。	Yes
死亡	与住院治疗（任何类型）的关系以及新冠肺炎导致的死亡。因果关系必须从住院治疗指向死亡。	存在关联 7 张图中有 7 张，并且在 0 张图中方向正确图。	部分
住院治疗	与医院_入院之间的关系和 患者_在_医院或患者_在_*MVB*。因果关系必须从住院入院人数传递至住院患者/MVB。	关系存在于 7个图中的7个，且在3个 图中方向正确。	部分
政策A中封锁	与出行之间的关系（任意）。因果关系必须从封锁措施指向出行。	存在于 7 个图中的 7 个，且在 2 个图中方向正确个图。	部分
政策A中封锁	与住院治疗（任意）。因果关系必须从住院治疗指向封锁措施。	关系存在于 7个图中的2个，且在 0 中正确定向图。	No

尽管前几小节指出学习到的结构之间存在相当明显的不一致，但表7中的结果表明，结构学习在识别或部分识别我们认为代表常识的大多数关系方面表现良好。学习到的结构与常识之间唯一相对明显的分歧涉及住院与封锁措施之间的关系。然而，这一特定关系包含一个时间维度，而假设条件分布随时间保持静态的结构学习算法无法识别这一点。更具体地说，我们知道，住院率较高时会实施封锁措施，而住院率较低时会放松封锁措施。然而，与大多数机器学习算法一样，由于所研究的结构学习算法并不预期变量之间的关系会随时间变化，这些算法从数据中读取到的是：无论是否存在封锁措施，住院率都曾处于低位和高位，因此无法识别触发封锁的时间因素，也无法识别封锁的时间效应；也就是说，根据封锁的严格程度，住院率往往会在封锁后的数周内下降。因此，未发现住院与封锁之间存在关系，凸显的是这些算法无法恢复此类随时间变化的关系，而不是表明这是一个重要发现。

7 结论

本研究探讨了因果结构学习中的一些未决问题，以及这些问题在新冠肺炎大流行建模中的应用。我们设计了64项实验，采用29种不同算法，并将其应用于我们收集的有关英国新冠肺炎大流行的数据。由于因果模型能够让我们模拟假设性干预的效果，我们认为，新冠肺炎这一问题要求人们针对此前未曾观察到的事件迅速做出前所未有的决策，因此为因果结构学习提供了绝佳的试验场。本节将基于这一案例研究总结因果结构学习中的未决问题，并提出未来研究方向。为促进未来研究，我们通过 Bayesys 仓库[7] (Constantinou等人，2020年)在线公开所有学习得到的图、平均图、贝叶斯网络模型、数据集和源代码。

我们采用的流程用于全面考察这组结构学习算法，并且可应用于任何数据集。具体而言，

> i. 第3.2小节介绍了我们将原始混合数据集转换为离散数据集和连续数据集所采用的流程。此外，本小节还介绍了在假设缺失值并非随机缺失的情况下，我们如何填补缺失数据值。ii. 第5节介绍了所考虑的一系列算法，涵盖结构学习的不同类别和不同类型，以及用于将这些算法应用于数据的软件包或软件。iii. 第5.1小节介绍了我们采用的模型平均方法，用于根据不同组结构学习算法生成的输出获得平均图。iv. 第6.1小节介绍了我们用于衡量每个学习到的结构维度的函数和指标。v. 第6.2小节介绍了用于衡量不同学习到的结构之间图结构差异的图结构指标。vi. 第6.3小节介绍了我们评估每个学习到的结构的推断能力所采用的流程。该流程包括模型选择和拟合优度函数，以及通过交叉验证评估学习到的结构中所有节点的预测能力。vii. 第6.4小节介绍了我们如何对关注变量进行假设性干预，以估计不同学习到的结构所确定的干预效果。viii. 第6.5小节介绍了在给定不同学习到的结构的情况下，我们如何对给定节点与其父节点和祖先节点之间的关系进行 敏感性分析。ix. 第6.6小节介绍了我们用于识别可能对输入数据中的观测变量造成混杂的潜在混杂因素的结构学习算法。

本研究的主要发现包括：a）学习到的结构对算法的选择高度敏感；b）模型平均能够有效且高效地降低基于算法的敏感性；c）与从离散或分类数据中学习到的结构相比，从连续数据中学习到的结构更稠密，且更容易出现过拟合和虚假因果关系；d）结构学习未考虑分布偏移是一个关键局限。下文将详细讨论这些发现。

最明显的结果是，在所考察的各种结构学习算法之间，发现学习到的结构高度不一致。也就是说，这些算法生成的图在边的数量、发现的具体边以及这些边的方向上都大相径庭。比较不同学习类别的算法（例如基于评分或基于约束的算法）时，这种不一致性会加剧；输入数据格式不同（例如分类的或连续型）时，情况也是如此。有人或许会认为，依赖不同学习类别的算法表现不同在一定程度上是合理的，但结果表明，这种不一致性在同一学习类别且输入数据格式相同的算法之间也普遍存在。还需强调的是，这种

[7] http://constantinou.info/downloads/bayesys/bayesys_repository.pdf

算法之间的分歧程度只会导致预测验证结果出现微不足道的差异（见图7），但当评估扩展至干预分析（见表6）或敏感性分析（见图8）时，则会造成显著差异。这些实证发现进一步表明，预测验证无法为因果推理相关问题提供有意义的答案。

许多结构差异无法用算法之间的差异来解释，因为同一种算法往往会根据输入数据格式的不同而生成截然不同的图。为减少学习得到的图之间的不一致性，一种常见方法是对一组图进行模型平均，从而获得能够代表这组学习得到的图的平均图。我们也对此进行了研究，按照学习类别或数据格式对算法分组。虽然模型平均确实能够降低变异性，但我们还发现，每个组的平均图彼此都不相同。这项工作的后续研究可以着重开发更复杂的模型平均方法。例如，本研究假设每种结构学习算法对平均图的贡献相同。然而，采用加权平均可能更有益，并优先考虑那些由已知准确度高于其他算法的算法学习得到的边。

这种程度的不一致性也存在于支持潜变量的算法中（也称为假设因果不足的算法）。由于这些算法旨在恢复能够揭示潜在混杂因素所导致的虚假关系的结构，我们会预期，不考虑潜在混杂因素的算法会在其学习得到的图中发现这些虚假边。然而，我们的研究结果不仅表明，假设因果不足的算法恢复出的虚假边各不相同，还表明，许多预测出的虚假边并未出现在不支持潜变量的算法所学习得到的大多数结构中。混杂效应方面的这些不一致性，引发了人们对支持潜变量的结构学习算法有效性的质疑。未来一个可能的研究方向，是更加注重改进并恰当评估旨在恢复潜在混杂因素的算法。例如，通过同时考察真正例率和假正例率，了解这些算法正确识别潜在混杂因素的成功率，将会很有帮助。此外，还应研究数据噪声或输入数据中的缺陷可能如何影响这些比率；这些问题在真实世界的数据中往往存在。

另一个值得强调的结果涉及精确学习，它保证从所探索的图中返回得分最高的图。结果表明，在实践中，精确学习的表现不如许多近似学习算法，而且这一结果还包括恢复出得分低于大多数近似学习算法所恢复结构的结构（参见图6）。这些结果与Constantinou等人（2021）报告的结果一致，我们提出了两种可能的解释。首先，对精确学习的节点最大入度超参数施加通常必要的限制（本文将其设为3，即默认设置），不可避免地限制了精确学习探索更稠密图的能力，而这些图本可能具有更高的目标得分。其次，精确学习通常依赖剪枝策略，剪去不包含得分最高图的搜索空间部分。然而，这一保证假设输入数据是干净的。在实践中，真实数据并不满足这一保证。例如，Constantinou等人（2021）表明，与精确学习器或更复杂的近似学习器相比，简单学习器对数据噪声具有更强的韧性。尽管精确学习代表了一项重要的理论探索，但精确解在实践中是否有用仍是一个开放问题。未来的研究可以探讨，在输入数据包含噪声或不完善的真实环境中，精确解是否比近似解更加准确。

总体结果还揭示了连续分布中的一些有趣模式。具体而言，研究发现，支持从连续数据中学习的算法所学得的图，比它们从相应的离散化数据中学得的图稠密得多。研究还发现，这些更稠密的图与基于知识的因果图相距更远。这一结果也适用于连续优化。连续优化最初因 NOTEARS 算法而被视为一种新的学习类别，并受到积极评价（Zheng 等人，2018年），但实践表明其表现并不理想（Constantinou 等人，2021；Kaiser 和 Sipos，2022年）。据此，从连续数据中进行结构学习不仅很可能面临较高的模型过拟合风险，而且恢复出的许多边也很可能反映的是关联而非因果关系。未来有关从连续数据中进行结构学习和连续优化的研究方向之一，是考察这些算法恢复出的伪边比例与真实边比例，尤其是在

真实图稀疏的情况下。理解为什么从连续数据中学习往往会产生比从相应离散化数据中学习所得图稠密得多的图，这一点十分重要。

最后，我们注意到，本文研究的传统因果结构学习算法无法察觉随时间可能发生的分布偏移。在处理新冠肺炎等时间序列数据时，这一点尤为重要，因为我们知道，其中捕捉到的一些效应存在时间滞后。例如，Dehning 等人（2020）表明，只有通过变点检测（即识别模型从一个模型变为新模型的时间点），才能揭示新冠肺炎政策干预的有效性；这也解释了为何本文测试的大多数结构学习算法都无法确定封锁与新冠肺炎住院病例之间的关系。很少有研究探讨因果结构学习如何处理数据中随时间变化的分布偏移（Kummerfeld 和 Danks，2013；Kocacoban 和 Cussens，2019；Huang 等人，2020；Bregoli 等人（2021）），尽管这些算法均未被结构学习领域采用。Glymour 等人（2019）承认，“从时间序列中估计因果生成过程这一一般问题，距离解决尚远”。时变因果结构的恢复问题仍缺乏充分研究。其他领域的相关研究，尤其是聚焦于变点检测的研究（Fearnhead 和 Liu，2007；Adams 和 MacKay，2007；Saatçi 等人，2010；Knoblauch 等人，2018），或许能为今后扩展现有方法以解决这一问题提供宝贵见解。

我们得出结论：准确的因果结构学习仍是一个尚待解决的问题，任何研究在确定因果关系时都不应依赖单一的结构学习算法。我们还指出，恢复因果结构时出现的不一致并非仅限于因果机器学习算法，而是同样存在于因果知识中；也就是说，每位作者都根据输入数据变量集合构建了不同的知识图谱，之后才达成共识。这些结果凸显了确定因果结构的难度、考虑多种信息来源的必要性，以及采用模型平均程序的益处。

作者贡献声明

	Conceptualization	Methodology	Software	Validation	Formal analysis	Investigation	Resources	Data Curation	Writing – Original Draft Preparation	Writing – Review & Editing	Visualization	Supervision
Constantinou:	X	X	X	X	X	X	X	X	X	X	X	X
Kitson:	X	X		X	X					X		
Liu:	X	X		X	X					X		
Kiattikun:	X	X		X	X					X		
Hashemzadeh:	X			X	X					X		
纳纳瓦蒂:	X			X	X					X		
姆布瓦氏:	X									X		
Petrungaro:	X									X		

附录A：数据整理补充信息

表A.4。构成工作和学校活动变量第12项，以及表1所述的第14、16和17项的学校指数手动构建方法。

No.	**Name**	详细信息
n/a	学校	引入了与以下内容一致的状态{开放、部分开放、关闭} 英国在COVID-19大流行期间针对学校的运营指南 （GOV.UK，2022h；2022i；2022j；维基百科，2022b）。
14	口罩	引入与口罩令相符的状态 {No, Optional, Yes} 在新冠肺炎大流行期间的英国（GOV.UK，2022g； 维基百科，2022a）。
16	季节	引入与以下内容相对应的状态 {冬季，春季，夏季，秋季} 四季。
17	主要COVID变体	引入 状态 {初始变体，阿尔法变体，德尔塔变体，奥密克戎变体，奥密克戎BA.2变体} 与英国政府关于变异株的数据一致（GOV.UK，2022k） 以及 Our World in Data 的数据（Our World in Data，2022b）。

表A.5. 根据政府研究所（2022）提供的英国COVID-19大流行期间封锁强制令信息，手动构建封锁（第15号）变量。

	政府措施					
变量状态	全面 全国范围的 封锁	允许 参加 工作场所 或学校	**Non-** 必要的 场所 **open**	限制措施 对社交 聚会的限制	社交 社交距离 限制措施	口罩 可选 没有 强制规定
无或有限措施						✓
社交距离					✓	
弱封锁			✓	✓		
中度封锁		✓				
严格封锁	✓					

附录B：平均图

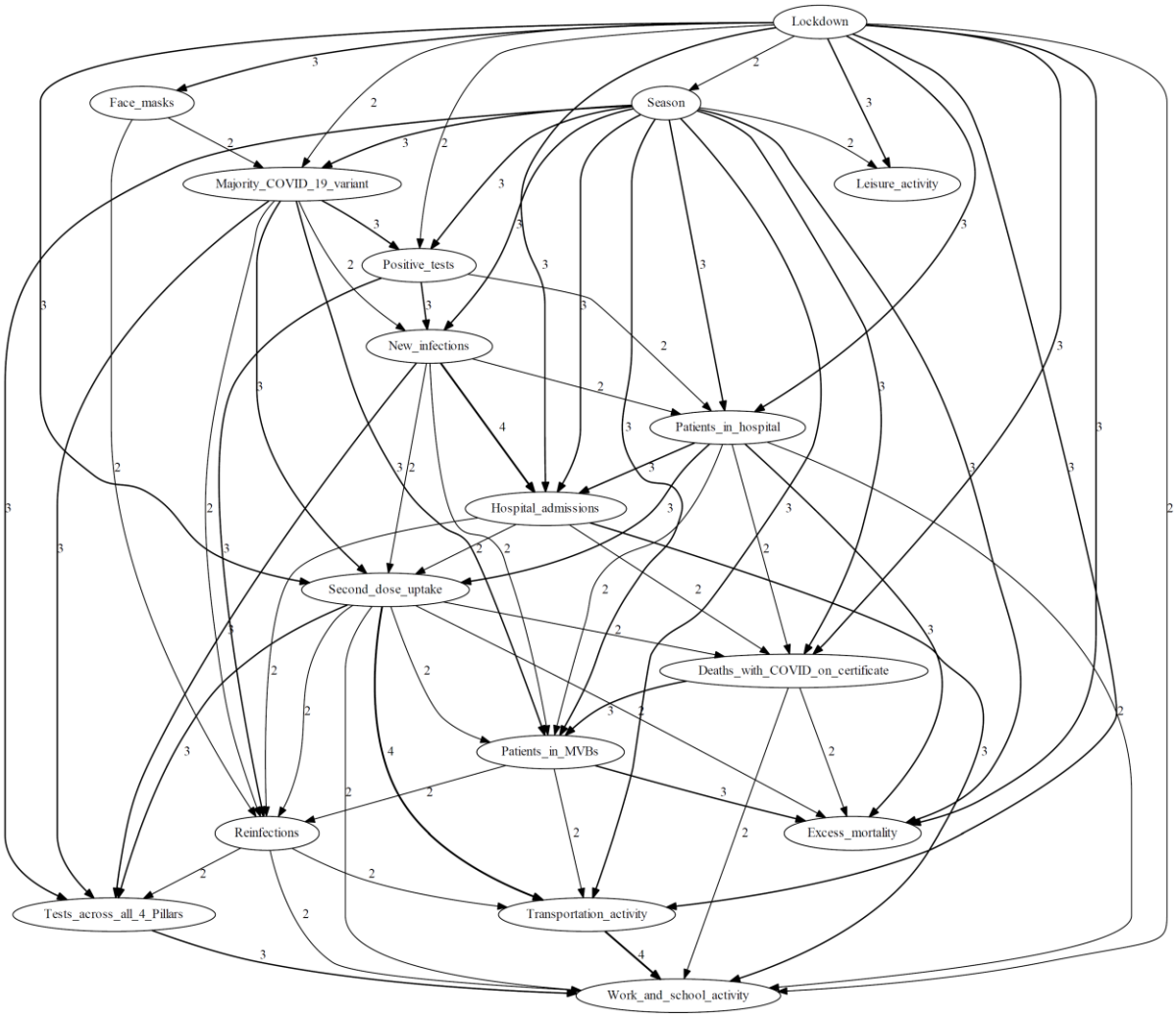


图 B.1。在输入数据集包含混合数据的情况下，通过 5 次实验获得的 All_mixed 平均图。该图共包含 67 条边，其中边标签表示给定边在所考虑的 5 个输出中出现的次数，且边的宽度随该次数增加而增大。在 5 个输入图中出现次数少于 2 次的边不包括在内。

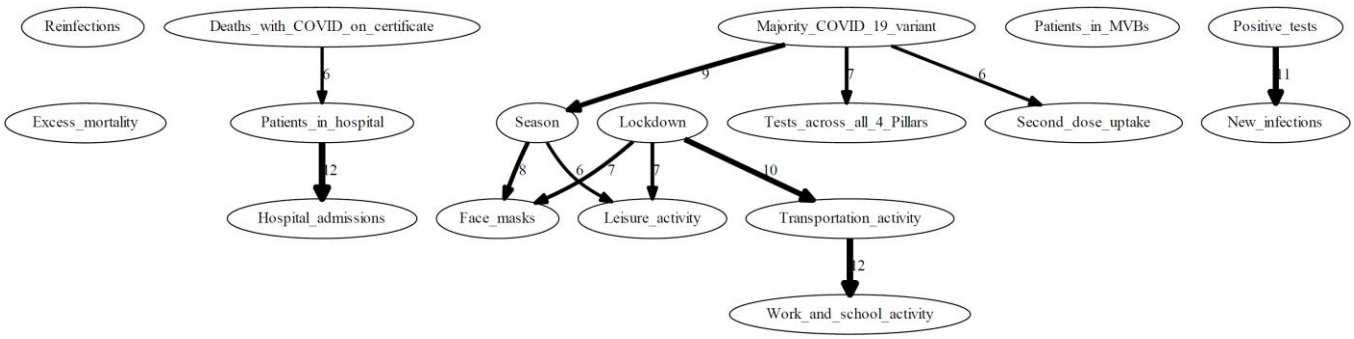


图 B.2. 通过 17 次混合学习实验得到的 All_hybrid 平均图。该图共包含 12 条边，其中边标签表示相应边在所考虑的 17 个输出中出现的次数，边的宽度随出现次数的增加而增大。在 17 个输入图中出现次数少于 6 次的边未予纳入。

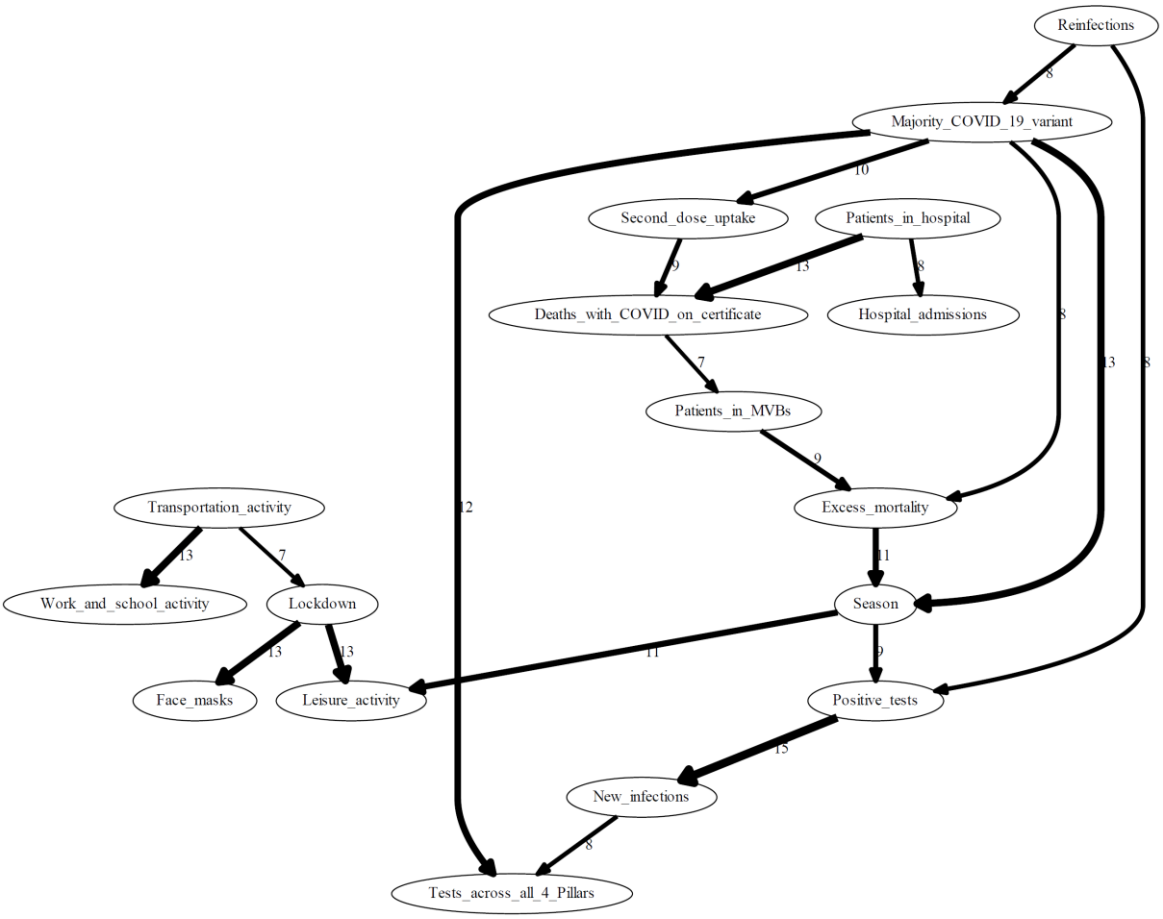


图 B.3。对 19 次实验所得的 All_quartiles 平均图，这些实验的输入数据集包含混合数据。该图共包含 20 条边，边标签表示所考虑的 19 个输出中相应边出现的次数，边的宽度随出现次数增加。未包含在 19 个输入图中出现次数少于 7 次的边。

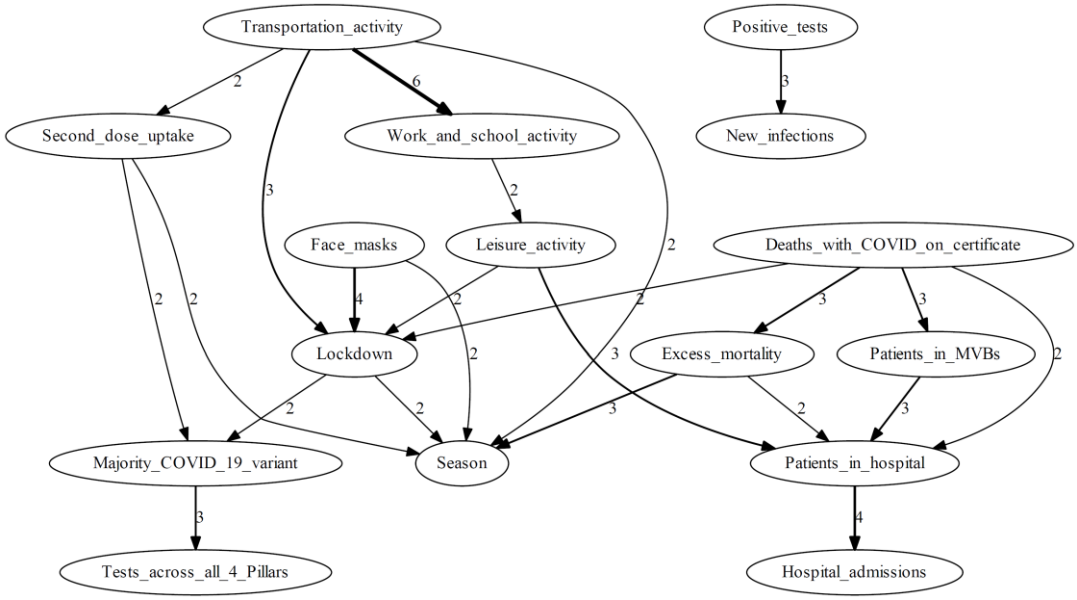


图 B.4。对 4 次基于约束的实验所得的 All_constraint-based 平均图。该图共包含 23 条边，边标签表示所考虑的 4 个输出中相应边出现的次数，边的宽度随出现次数增加。未包含在 4 个输入图中出现次数少于 2 次的边。

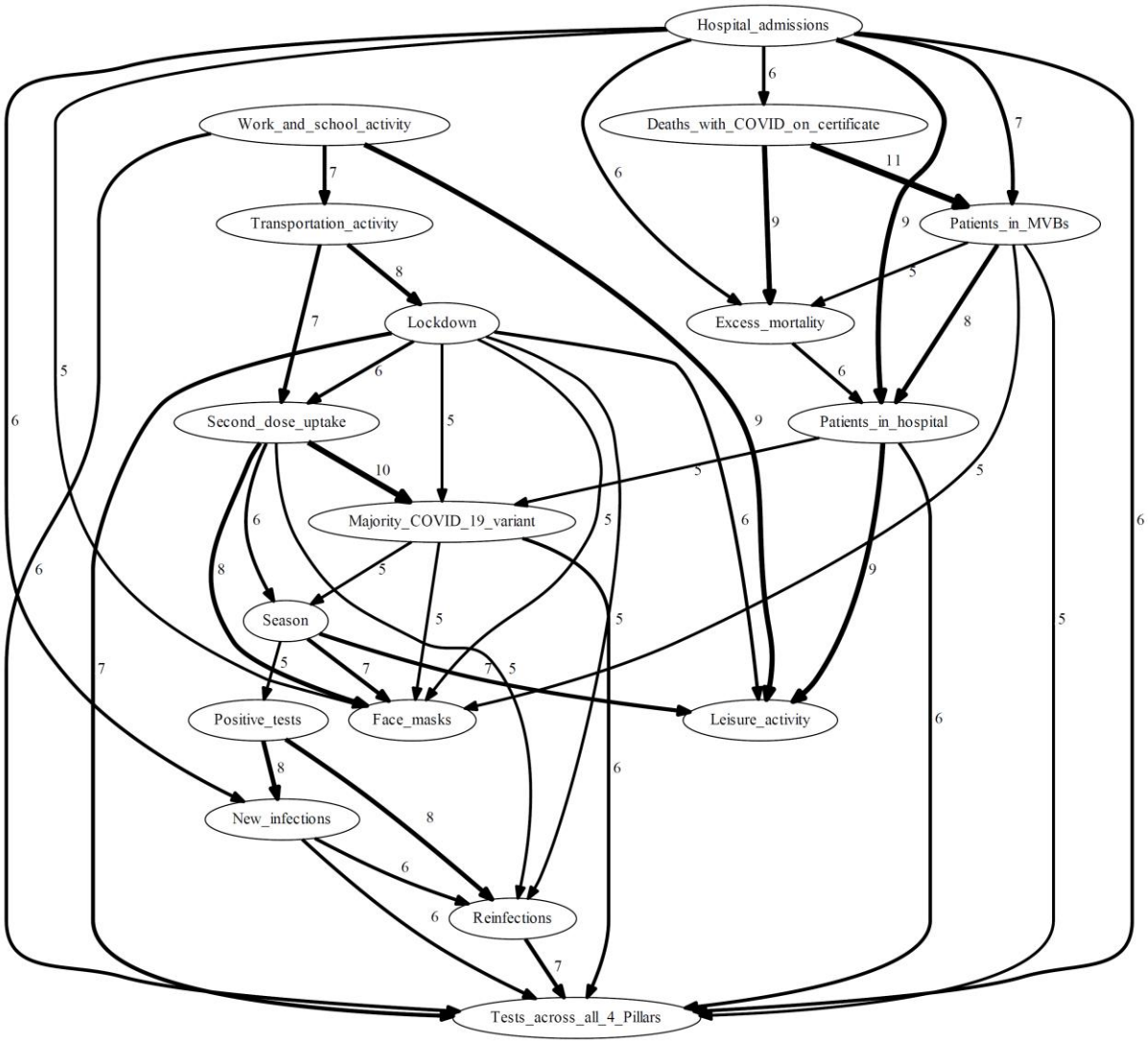


图 B.5。在 14 次实验中获得的全部连续型平均图，其中输入数据集包含连续数据。该图共包含 43 条边，边标签表示给定边在所考察的 14 个输出中出现的次数，且边的宽度随该次数增加而增大。在 14 个输入图中出现次数少于 5 次的边不予显示。

参考文献

Adams, R. P. 和 MacKay, D. J. C. (2007). 贝叶斯在线变点检测。arXiv:0710.3742 [在线]。可访问：https://support.bayesfusion.com/docs/GeNIe.pdf [访问日期：2023年2月9日2023]。

Bouckaert, R. (1994年). 贝叶斯信念网络学习算法的性质。载于 *10th*不确定性与智能会议论文集，第 102–109 页。

Bouchaert, R. (1995年). 贝叶斯信念网络：从构建到推断。博士论文，乌得勒支大学。

Bregoli, A., Scutari, M., 和 Stella, F. (2021). 一种用于结构学习的基于约束的算法连续时间贝叶斯网络。《国际近似推理杂志》，第138卷，第105–122页。

Castillo, E.F.、Gutierrez, J. M. 和 Hadi, A. S.（1997年）。离散贝叶斯网络中的敏感性分析。《IEEE系统、人及控制论汇刊——A辑：系统与人》，第27卷，第4期，第412–423页。
Chadeau-Hyam, M.、Bodinier, B.、Elliott, J.、Whitaker, M. D.、Tzoulaki, I. 等（2020年）。COVID-19 检测呈阳性和阴性的风险因素：对英国生物样本库数据的谨慎深入分析。《国际流行病学杂志》，第49卷，第5期，第1454–1467页。Chernozhukov, V.、Kasahara, H. 和 Schrimpf, P.（2021）。口罩、政策和行为对美国早期COVID-19大流行的因果影响。《计量经济学杂志》，第220卷，第1期，第23–62页。Chickering, D.M. 和 Meek, C.（2002年）。寻找最优贝叶斯网络。载于《*18th*人工智能中的不确定性会议》（UAI-2002）论文集，第94–102页。Chobtham, K. 和 Constantinou, A.（2020年）。存在潜变量时具有因果效应的贝叶斯网络结构学习。载于《*10th* 国际概率图模型会议》（PGM-2020）论文集，PMLR，第138卷，第101–112页。Chobtham, K. 和 Constantinou, A.（2022年）。贝叶斯网络中潜在混杂因素的发现与密度估计

具有证据下界的网络。载于 *11th* 《国际概率图模型会议》（PGM-2022）论文集，西班牙阿尔梅里亚，2022年10月。

Colombo, D. 和 Maathuis, M. H.（2014）。与顺序无关的基于约束的因果结构学习。《机器学习研究杂志》，第15卷，第1期，第3741–3782页。

Constantinou, A.（2019）。Bayesys 用户手册。贝叶斯人工智能研究实验室、机器智能与决策系统（MInDS）研究组，伦敦玛丽女王大学，英国伦敦。[在线]。网址：http://bayesianai.eecs.qmul.ac.uk/bayesys/ [访问日期：2月2日 2023]。

Constantinou, A. C.、Liu, Y.、Chobtham, K.、Guo, Z. 和 Kitson, N. K.（2020）。Bayesys 数据与贝叶斯网络资料库。贝叶斯人工智能研究实验室，伦敦玛丽女王大学，英国伦敦。[在线]。网址：http://bayesian-ai.eecs.qmul.ac.uk/bayesys/ [访问日期：2月2日 2023]。

Constantinou, A.（2020年）。实现证据完全传播的贝叶斯网络学习。IEEE Access，8，第124845–124856页。

Constantinou, A. C.、Liu, Y.、Chobtham, K.、Guo, Z. 和 Kitson, N. K.（2021）。大规模实证验证噪声数据下的贝叶斯网络结构学习算法。《国际近似推理杂志》，第131卷，第151–188页。

Constantinou, A.、Liu, Y.、Kitson, N. K.、Chobtham, K. 和 Guo, Z.（2022）。通过剪枝和模型平均策略实现有效且高效的结构学习。《国际近似推理杂志》，151，第292–321页。

因果发现中心。（2023）。Tetrad 手册。因果发现中心。 [在线]。可用：https://htmlpreview.github.io/?https:///github.com/cmu-phil/tetrad/blob/development/docs/manual/index.html[访问日期：2月2023]。

Cussens, J.（2011年）。使用割平面进行贝叶斯网络学习。载于 *27th* 会议论文集，主题为人工智能中的不确定性（*UAI–2011*），第153–160页，AUAI 出版社。

Cussens, J. 和 Bartlett, M.（2015年）。GOBNILP 1.6.2 用户/开发者手册。约克大学。[在线]。获取地址：~~https://www.cs.york.ac.uk/aig/sw/gobnilp/manual.pdf [~~访问日期：2月 2023]。Darwiche, A.（2009年）。贝叶斯网络建模与推理。剑桥大学出版社。de Campos, C.P.、Zeng, Z. 和 Ji, Q.（2009年）。利用约束进行贝叶斯网络结构学习。载于《*26th* 国际机器学习年度会议（ICML–09），第113–120页。Deaton, A. 和 Cartwright, N.（2018年）。理解和误解随机对照试验。《社会科学与医学》，第210卷，第2–21页，ISSN 0277-9536。

arXiv 预印本，2023年。

Dehning, J.、Zierenberg, J.、Spitzner, F.P.、Wibral, M.、Neto, J.P.、Wilczek, M. 和 Priesemann, V.（2020年）。推断新冠肺炎传播中的变化点揭示了干预措施的有效性。《科学》，第 369 卷，第 6500 期，p.eabb9789。

Douaud, G., Lee, S., Alfaro-Almagro, F., Arthofer, C., Wang, C., 等人（2022年）。SARS-CoV-2 与……相关。英国生物样本库中大脑结构的变化。《自然》，第604卷，第697–707页（2022年）。

Elliott, J.、Bodinier, B.、Whitaker, M.、Delpierre, C.、Vermeulen, R. 等（2021）。英国生物样本库队列中的新冠肺炎死亡率：重新审视并评估风险因素。《欧洲流行病学杂志》，第36卷，第299–309页。

欧洲航空安全组织（2022年）。新冠肺炎对欧洲航空交通网络的影响。欧洲航空安全组织。航空导航安全（欧洲航空安全组织）。[在线] 可访问：https://www.eurocontrol.int/covid19 [访问日期：6月26日2022]。

Fearnhead, P. 和 Liu, Z.（2007）。多重变点问题的在线推断。《皇家统计学会杂志：B辑》，第69卷，第589–605页。

Fenton, N. E. 和 Neil, M.（2018年）。使用贝叶斯网络进行风险评估与决策分析（2nd 版）。CRC出版社。

Fenton, N. E.、Neil, M.、Osman, M. 和 McLachlan, S.（2020年）。COVID-19感染与死亡率：有必要将因果解释纳入数据分析，并避免检测偏差。《风险研究杂志》，第23卷，第7–8期：新冠肺炎特刊，第862–865页。

Friedman, N.（1997年）。存在缺失值和隐变量时的信念网络学习。载于《国际机器学习会议论文集》，第97卷，第125–133页。

Friston, K. J、Parr, T.、Zeidman, P.、Razi, A.、Flandin, G. 等（2020年）。新冠肺炎的动态因果建模。Wellcome Open Research，第 5 卷，89 合乐

Gencoglu, O. 和 Gruber, M.（2020年）。新冠肺炎期间 Twitter 活动的因果建模。Computation，第 8 卷，第 4 期，文章 85。

Giudice, E.、Kuipers, J. 和 Moffa, G.（2022年）。用于结构学习的双重PC算法。载于会议论文集第 *11th* 届国际概率图模型会议（PGM-2022），阿尔梅里亚，西班牙，PMLR，2022年。

Glymour, C., Zhang, K., 和 Spirtes, P.（2019年）。基于图的因果发现方法综述。模型。Frontiers in Genetics，文章编号 524。

GOV.UK（2022a）。冠状病毒病（COVID-19）：英国检测。英国政府关于冠状病毒病（COVID-19）数据与洞察的官方网站。 [在线] 网址：https://coronavirus.data.gov.uk/details/testing [访问日期：6月26日2022]。

GOV.UK（2022b）。冠状病毒病（COVID-19）：英国病例。英国政府关于冠状病毒病（COVID-19）数据和见解的官方网站。[在线] 可查阅：https://coronavirus.data.gov.uk/details/cases[访问日期：6月26日2022]。GOV.UK（2022c）。冠状病毒病（COVID-19）：英国医疗保健。英国政府关于冠状病毒病（COVID-19）数据和见解的官方网站。[在线] 可查阅：https://coronavirus.data.gov.uk/details/healthcare [访问日期：6月26日2022]。GOV.UK（2022d）。冠状病毒病（COVID-19）：英国医疗保健。英国政府关于冠状病毒病（COVID-19）数据和见解的官方网站。[在线] 可查阅：https://coronavirus.data.gov.uk/details/vaccinations [访问日期：6月26日2022]。GOV.UK（2022e）。冠状病毒病（COVID-19）：英国死亡情况。英国政府关于冠状病毒病（COVID-19）数据和见解的官方网站。[在线] 可查阅：https://coronavirus.data.gov.uk/details/deaths [访问日期：6月26日2022]。

arXiv 预印本，2023年。

GOV.UK（2022f）。冠状病毒病（COVID-19）：流动性报告。大伦敦政府。[在线] 可查阅：https://data.gov.uk/dataset/4a475119-698e-49f3-afcc-8739558e1696/coronavirus-covid-19-mobility-report[访问日期：6月26日2022]。

GOV.UK（2022g）。冠状病毒病（COVID-19）：面部遮盖物：何时佩戴、豁免情形及如何选择合适的面部遮盖物。卫生与社会保健部。[在线] 可查阅：https://www.gov.uk/government/publications/face-coverings-when-to-wear-one-and-how-to-make-your-own/face-coverings-when-to-wear-one-and-how-to-make-your-own [访问日期：7月6日2022]。

GOV.UK（2022h）。远程教育良好实践。教育部。[在线] 可查阅：https://www.gov.uk/government/publications/remote-education-good-practice/remote-education-good-practice[访问日期：7月6日2022]。

GOV.UK（2022i）。可进入学校或教育机构的关键工作人员的子女和弱势儿童环境。教育部。[在线] 可查阅：https://www.gov.uk/government/publications/coronavirus-covid-19-maintaining-educational-provision/guidance-for-schools-colleges-and-local-authorities-on-maintaining-educational-provision [访问日期：7月6日2022]。

GOV.UK（2022j）。冠状病毒疫情期间学校应对措施。教育部。[在线]网址：https://www.gov.uk/government/publications/actions-for-schools-during-the-coronavirus-outbreak[访问日期：7月6日2022]。GOV.UK（2022k）。冠状病毒病（COVID-19）：研究与分析：COVID-19变异株：经基因组确认的病例数。英国卫生安全局。[在线] 网址：https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers[访问日期：7月6日2022]。Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D.（2018）。解释黑箱模型的方法综述。《ACM计算调查》，第51卷，第5期，文章93。Hastie, C. E., Mackay, D. F., Ho, F., Celis-Morales, C. A., Katikireddi, S. V., 等（2020）。英国生物样本库中的维生素D浓度与COVID-19感染。《糖尿病与代谢综合征：临床研究与评论》，第14卷，第4期，第561–565页。Heckerman, D., Geiger, D., and Chickering, D.（1995）。贝叶斯网络学习：知识与统计数据的结合。《机器学习》，20，第197–243页。Horna, S. R., Westona, S. J., and Fisher P. A.（2020）。识别COVID-19在免疫精神病学模型中的因果作用。《大脑、行为与免疫》，第88卷，第6–8页。Huang, B., Zhang, K, Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Scholkopf, B.（2020）。从异质/非平稳数据中进行因果发现。《机器学习研究杂志》，第21卷，第1–53页。帝国理工学院COVID-19应对团队。（2021）。2020—2021年报告（2021年3月22日）。伦敦帝国理工学院，DOI：https://doi.org/10.25561/87192。政府研究所。（2022）。英国政府冠状病毒封锁与措施时间线：2020年3月至2021年12月。政府研究所分析。[在线] 网址：https://www.instituteforgovernment.org.uk/charts/uk-government-coronavirus-lockdowns Jarvis, C. I., Van Zandvoort, K., Gimma, A., Prem, K., CMMID COVID-19工作组，等（2020）。量化保持身体距离措施对英国COVID-19传播的影响。《BMC医学》，第18卷，文章124。Kaiser, M., and Sipos, M.（2022）。处理有量纲量时NOTEARS不适用于因果图发现。《神经处理快报》，第54卷，第1587–1595页。Kolin, D. A, Kulm, S., Christos, P. J., and Elemento, O.（2020）。英国生物样本库COVID-19患者的临床、区域和遗传特征。《PLoS ONE》，第11卷，e0241264。

arXiv 预印本，2023年。

Kitson, N. K., Constantinou, A., Guo, Z., Liu, Y. 和 Chobtham, K.（2023年）。贝叶斯网络综述结构学习。《人工智能评论》， https://doi.org/10.1007/s10462-022-10351-w

Knoblauch, J.、Jewson, J. 和 Damoulas, T.（2018年）。基于 β-散度的非平稳流式数据双重稳健贝叶斯推断。载于*32nd* 神经信息处理系统会议（NeurIPS-2018）论文集，蒙特利尔，加拿大。

Kocacoban, D. 和 Cussens, J.（2019年）。存在潜变量时的在线因果结构学习。载于*18th* IEEE 国际机器学习与应用会议（ICMLA）论文集，佛罗里达，美国。

Koller, D. 和 Friedman, N.（2009年）。概率图模型：原理与技术。麻省理工学院出版社。Koski, T. J. T. 和 Noble, J.（2012年）。贝叶斯网络与结构学习综述。《Mathematica》

《Applicanda》，第40卷，第1期，第55–103页。

Kummerfeld, E. 和 Danks, D.（2013）。追踪随时间变化的图结构。载于《神经信息处理系统进展》第26届会议论文集（NIPS-2013），美国内华达州

Lee, C. 和 van Beek, P.（2017）。用于贝叶斯网络结构学习的得分与搜索元启发式方法。载于《人工智能进展：计算机科学讲义》，加拿大人工智能会议，第129–141页

Li, H., Cabeli, V., Sella, N. 和 Isambert, H.（2019）。基于约束的因果结构学习与一致的分离集。载于*33rd* 神经信息处理系统会议论文集（NeurIPS- 2019），加拿大温哥华。

Liu, Y. 和 Constantinou, A.（2022年）。从包含系统性缺失值的数据中进行贪心结构学习。《机器学习》，第111卷，第3867–3896页。

Liu, Y.、Constantinou, A. 和 Guo, Z.（2022年）。在存在测量误差的情况下改进贝叶斯网络结构学习。《机器学习研究杂志》，第23卷，第324期，第1–28页。

Liu, Y. 和 Constantinou, A.（2023年）。利用马尔可夫毯发现改进缺失数据的插补。在 *11th* 国际学习表征会议（ICLR-2023）论文集，卢旺达基加利。

Mastakouri A. A. 和 Schölkopf, B.（2020年）。德国新冠肺炎传播的因果分析。在 *34th* 神经信息处理系统会议（NeurIPS–2020）论文集，加拿大温哥华。

Menni, C., Klaser, K., May, A., Polidori, L., Capdevila, J., 等（2021）。疫苗接种副作用与 SARS-CoV-2 感染：基于英国 COVID 症状研究应用用户的前瞻性观察研究。《柳叶刀·传染病》，第 21 卷，第 7 期，第 939–949 页。

Ng, I., Zhu, S., Fang, Z., Li, H., Chen, Z. 和 Wang, J.（2022）。基于掩码梯度的因果结构学习。载于《2022 年 SIAM 国际数据挖掘会议（SDM）论文集》。 工业与应用数学学会。

Nguyen, L. H., Drew, D. A., Graham, M. S., Joshi, A. D., Guo, C., 等（2020）。一线医护人员和普通社区人群感染新冠肺炎的风险：一项前瞻性队列研究。《柳叶刀·公共卫生》，第 5 卷，第 9 期，第 e475–e483 页。

Ogarrio, J. M., Spirtes, P. 和 Ramsey, J.（2016年）。一种用于潜变量模型的混合因果搜索算法。载于*8th* 国际概率图模型会议（PGM-2016）论文集，第 368–379 页

O'Connor、R.、Wetherall、K.、Cleare、S.、McClelland、H.、Melson、A.等（2021）。COVID-19大流行期间的心理健康与福祉：英国成年人在英国COVID-19心理健康与福祉研究中的纵向分析。《英国精神病学杂志》，第218卷第6期，第326–333页。

arXiv 预印本，2023年。

英国国家统计局网站（2022）。冠状病毒病（COVID-19）：最新洞察：生活方式：英国航班七日平均数量仍低于冠状病毒大流行前2019年水平（84%）。英国国家统计局。[在线] 可访问：https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/lifestyle#:~:text=约十分之三的成年人，在家中的时间%20有所增加。[访问日期：7月6日 2022]。

Our World in Data（2022a）。超额死亡率：探索全球情况：COVID-19 数据探索器。[在线]可访问：https://ourworldindata.org/coronavirus#explore-the-global-situation [访问日期：10月11日2022]。

Our World in Data（2022b）。按变异株划分的 SARS-CoV-2 序列：在此前两周内分析的序列中，各变异株组所占比例（指标：变异株，国家：英国）。全球共享禽流感数据倡议（GISAID）。[在线] 可访问：https://ourworldindata.org/explorers/coronavirus-data-explorer [访问日期：7月6日2022]。

Our World in Data（2022c）。超额死亡率：与基于此前 s
年份的预测值相比的全因死亡人数（指标：超额死亡率（%），国家：英国）。人类死亡率数据库（HMD）和 th *e*
世界死亡率数据集（世界死亡率数据集）。[在线] 可访问：https://ourworldindata.org/explorers/coronavirus-data-explorer [访问于7月6日2022]。

Pearl, J.（2012年）。再论因果演算。主题演讲，人工智能不确定性会议 *(UAI–2012)*，加利福尼亚州卡特琳娜岛。

Pearl, J. 和 Mackenzie, D.（2018年）。《为什么之书》：因果关系的新科学。Basic Books。

Ramsey, J.、Glymour, M.、Sanchez-Romero, R. 和 Glymour, C.（2016年）。百万个及更多变量：用于学习高维图形因果模型的快速贪婪等价搜索算法及其在功能性磁共振图像中的应用。《国际数据科学与分析杂志》，第3卷，第121–129页。

Reisach, A. G.、Seiler, C. 和 Weichwald, S.（2021）。警惕模拟有向无环图！因果发现基准可能很容易被

Saatçi, Y.、Turner, R. D. 和 Rasmussen, C. E.（2010）。高斯过程变点模型。载于《*27th* 国际机器学习会议》（ICML-10）论文集，第927–934页，海法，以色列。

Sahin, O.、Salim, H.、Suprun, E.、Richards, R.、MacAskill, S. 等（2020）。构建初步因果回路图，以理解新冠肺炎大流行的复杂难题。《系统》，第8卷，第2期，文章编号20。

Scanagatta, M.、de Campos, C. P.、Corani, G. 和 Zaffalon, M.（2015）。学习包含数千个变量的贝叶斯网络。载于《神经信息处理系统进展》第28卷（NIPS-2015），第1864–1872页。

Scanagatta, M.、Corani, G. 和 Zafalon, M.（2017）。改进贝叶斯网络结构学习中的局部搜索。载于《贝叶斯网络高级方法》（AMBN-2017）论文集，第45–56页。

Scanagatta, M., Salmerón, A. 和 Stella, F.（2019年）。基于数据的贝叶斯网络结构学习综述。《人工智能进展》第8卷，第425–439页。

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A. 和 Bengio, Y.（2021）。迈向因果表征学习。《IEEE会刊》第109卷，第5期，第612–634页。

Scutari, M.（2023年）。软件包“bnlearn”。[在线] 可访问：https://www.bnlearn.com/documentation/bnlearn-manual.pdf [访问日期：2月2日2023]。

Seng, J., Zecevic, M., Dhami, D. S. 和 Kersting, K.（2022年）。拆解 NOTEARS：通过方差操控控制图预测。arXiv:2206.*07195* [cs.LG]

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J.（2006年）。一种线性非高斯无环模型，适用于因果发现。《机器学习研究杂志》，第 7 卷，第 2003–2030 页。

arXiv 预印本，2023。

Spirtes, P. 和 Glymour, C.（1991年）。一种快速恢复稀疏因果图的算法。《社会科学计算评论》，第 9 卷，第 1 期，第 62–72 页。

Spirtes, P.、Meek, C. 和 Richardson, T.（1999）。存在潜变量时进行因果推断的算法和选择偏差，载于：Clark Glymour、Gregory Cooper（编），《计算、因果与发现》，麻省理工学院出版社，马萨诸塞州剑桥，1999年，第211–252页

Sudre, C.H.、Murray, B.、Varsavsky, T.、Graham, M. S.、Penfoldet, R. S. 等（2021）。长期新冠的特征与预测因素《自然·医学》，第27卷，第626–631页。

Suter, P.、Kuipers, J.、Moffa, G. 和 Beerenwinkel, N.（2023）。使用R软件包BiDAG进行贝叶斯结构学习与贝叶斯网络采样。《统计软件杂志》，第105卷，第9期，第1–31页。

Teyssier, M. 和 Koller, D.（2005年）。基于排序的搜索：一种用于学习贝叶斯网络的简单有效算法网络。在第*21st* 届人工智能中的不确定性会议（UAI-2005）论文集中，第 584–590 页。

Tsamardinos, I.、Brown, L. E. 和 Aliferis, C. F.（2006年）。最大-最小爬山贝叶斯网络结构学习算法。《机器学习》，第 65 卷，第 31–78 页。

维基百科（2022a）。COVID-19疫情期间的口罩：按国家和地区比较：英国：英格兰。[在线] 可获取：https://en.wikipedia.org/wiki/Face_masks_during_the_COVID-19_pandemic#cite_note-495 [访问日期：7月6日2022]。

维基百科（2022b）。COVID-19疫情对英国教育的影响。[在线] 可用：https://zh.wikipedia.org/wiki/新冠肺炎大流行病对英国教育的影响[访问于7月6日2022]。

Yang, H., Chen, W., Hu, Y., Chen, Y., Zeng, Y., 等人（2020年）。大流行病前精神障碍与……的风险。新冠肺炎：一项基于英国生物样本库的队列分析。《柳叶刀·健康长寿》，第 1 卷，第 2 期，第 e69–e79 页。

Yates, T., Razieh, C., Zaccardi, F., Davies, M. J. 和 Khunti, K.（2020年）。肥胖与新冠肺炎风险：英国生物样本库分析。《初级保健糖尿病》，第14卷，第566–567页。

Yu, Y., Chen, J., Gao, T., Yu, M.（2019年）。DAG-GNN：使用图神经网络进行DAG结构学习。载于*36th* 国际机器学习会议（ICLM-2019）论文集，加利福尼亚州长滩，PMLR 97。

Zhang, K., Zhu, S., Kalander, M., Ng, I., Ye, J., Chen, Z., Pan. L.（2021）。gCastle：用于因果发现的 Python 工具箱。arXiv:2111.15155 [cs.LG]

Zheng, X., Aragam, B., Ravikumar, P., 和 Xing, E.P.（2018）。无撕裂DAG：用于结构学习的连续优化。载于*32nd* 神经信息处理系统国际会议（NeurIPS–2018）论文集，第 9492–9503 页，加拿大魁北克省蒙特利尔。

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., 和 Xing, E. P.（2020）。稀疏非参数DAG的学习。载于*23rd* 人工智能与统计学会议（AISTATS-2020）论文集，意大利巴勒莫。PMLR：第 108 卷。