arXiv:2303.06423v1 [q-bio.QM] 11 Mar 2023

从超大型数据集中学习可解释因果网络：应用于 400,000 份乳腺癌患者医疗记录

Marcel da Câmara Ribeiro-Dantas[a], 1、**Honghao Li**[a], 1、**Vincent Cabeli**[a], 1、**Louise Dupuis**[a], 1、**Franck Si mon**[a]、**LizaHettal**[a]、**Anne-Sophie Hamy**[b], c, [d]和 **Hervé Isambert**[2], a

[a]法国国家科学研究中心 UMR168、居里研究所、巴黎文理研究大学、索邦大学，法国巴黎；[b]法国国家健康与医学研究院 U932、居里研究所，法国巴黎；[c]医学肿瘤科、居里研究所，法国圣克卢；[d]外科、居里研究所、巴黎大学，法国巴黎

本稿件编制于2023年3月14日

发现因果效应是科学研究的核心，但仅有观察性数据时，这仍是一项挑战。在实践中，因果网络难以学习和解释，且研究范围仅限于相对较小的数据集。我们提出了一种更可靠、可扩展的因果发现方法（iMIIC），该方法基于一般互信息上确界原理，大幅提升了推断所得因果关系的精确率，同时能够区分真实原因与推定的、潜在的因果效应。我们在合成数据以及来自美国监测、流行病学和最终结果计划的396,179名乳腺癌患者的真实医疗保健数据上展示了 iMIIC 的应用。超过90%的预测因果效应似乎是正确的，其余出乎意料的直接和间接因果效应则可从诊断程序、治疗时机、患者偏好或社会经济差异等方面加以解释。iMIIC 的独特能力为在多个研究领域发现可靠且可解释的因果网络开辟了新途径。

因果发现 | 可解释因果网络 | 医疗保健数据 |乳腺癌

Nationwide 医疗记录包含大量关于人类健康的真实世界数据，其中包括一些个人、家庭和社会经济信息。这些信息往往不仅会影响健康状况，还会影响诊断时间、医疗治疗，并最终影响患者的生存。此外，在临床试验中，通常会对这类非医疗因素进行控制；临床试验会依据严格的入组标准筛选特定患者群体。然而，由于缺乏能够在不预设假设的情况下分析这些数据的无监督方法和工具，真实世界医疗记录中丰富的信息仍未得到充分利用。这凸显了开发新的机器学习策略来分析医疗保健数据的必要性，以揭示患者病史中记录的所有可用信息之间未被察觉的关联和可能的因果关系，图1a。

早期关于因果发现方法的开创性研究（1，2）早已指出，原则上，仅凭观察性数据学习因果关系是可行的。从本质上说，因果发现根据涉及至少三个变量的特定相关模式推断因果关系，这超越了广为流行的观点——两两相关并不意味着因果关系。然而，尽管观察性数据占据了众多领域中绝大多数可用数据集，但在缺乏系统性干预的情况下，揭示因果关系仍然极具挑战性；而涉及人类健康时，系统性干预可能不切实际、成本过高或有违伦理。

尽管因果发现与旨在学习图模型的方法密切相关（1–4），但大多数结构学习方法实际上并非为揭示因果关系而设计。特别是，最大似然方法，例如搜索与评分（5）或图形套索（6）方法，都受限于特定的模型类别，假设图要么完全有向、要么完全无向，因此无法学习图边的因果或非因果性质。相比之下，基于约束的因果发现方法假设更广泛的图类别，并且仅凭观察性数据就能确定某些边的方向（1，2），图1b。为此，这些方法首先以条件独立关系的形式学习结构约束。这些关系提供了有关观测变量和未观测变量之间可能存在的因果关系的间接且较为隐晦的信息，如框 1 所述。然而，尽管在数据量无限时这些方法在理论上是可靠的（7），但基于约束的方法在实际应用中能够处理的数据集相对较小，因此仍不可靠且难以解释。

我们在此介绍先进的因果发现方法 iMIIC（可解释 MIIC）。该方法能够学习更可靠、易于解释的因果图模型，并处理规模大得多的数据集（例如，包含数十万份样本的数据集）。新颖的 iMIIC 方法大幅提升了近期结构学习方法 MIIC（基于多变量信息的归纳因果推断）的因果发现性能，该方法结合了基于约束和基于信息论的框架（8–10）。iMIIC 的性能依托于三个主要条件——

作者概要

仅有观察性数据时，揭示因果关系向来极具挑战性；对于绝大多数数据集而言，系统性扰动实验并不可行，情况更是如此。我们介绍一种可靠的因果发现方法 iMIIC，它能够从生物学、生物医学或其他复杂观察性数据中的多变量相关性，揭示真正的因果关系。iMIIC 基于信息论原理，能够分析整合自不同来源或实验技术的超大规模多模态数据集。我们利用来自 396,179 名乳腺癌患者的医疗保健数据，展示了 iMIIC 的因果发现性能。iMIIC 独特的图形化输出能够全局呈现变量之间的直接和间接关系，包括一些出乎意料的因果效应发现；这些发现可以结合临床实践、患者选择或社会经济背景加以解读。

M.C.R-D.、H.L.、V.C.、F.S. 和 H.I. 设计并实现了机器学习工具；M.C.R-D.、H.L.、L.D. 参与了数据分析；M.C.R-D.、H.L.、L.H.、A-S.H. 和 H.I. 参与了数据解读。H.I.、L.D.、M.C.R-D.、H.L. 撰写了稿件。作者声明不存在利益冲突。[1]M.C.R-D.、H.L.、V.C. 和 L.D. 对本研究贡献相同。[2]通讯请联系：电子邮箱 herve.isambertcurie.fr

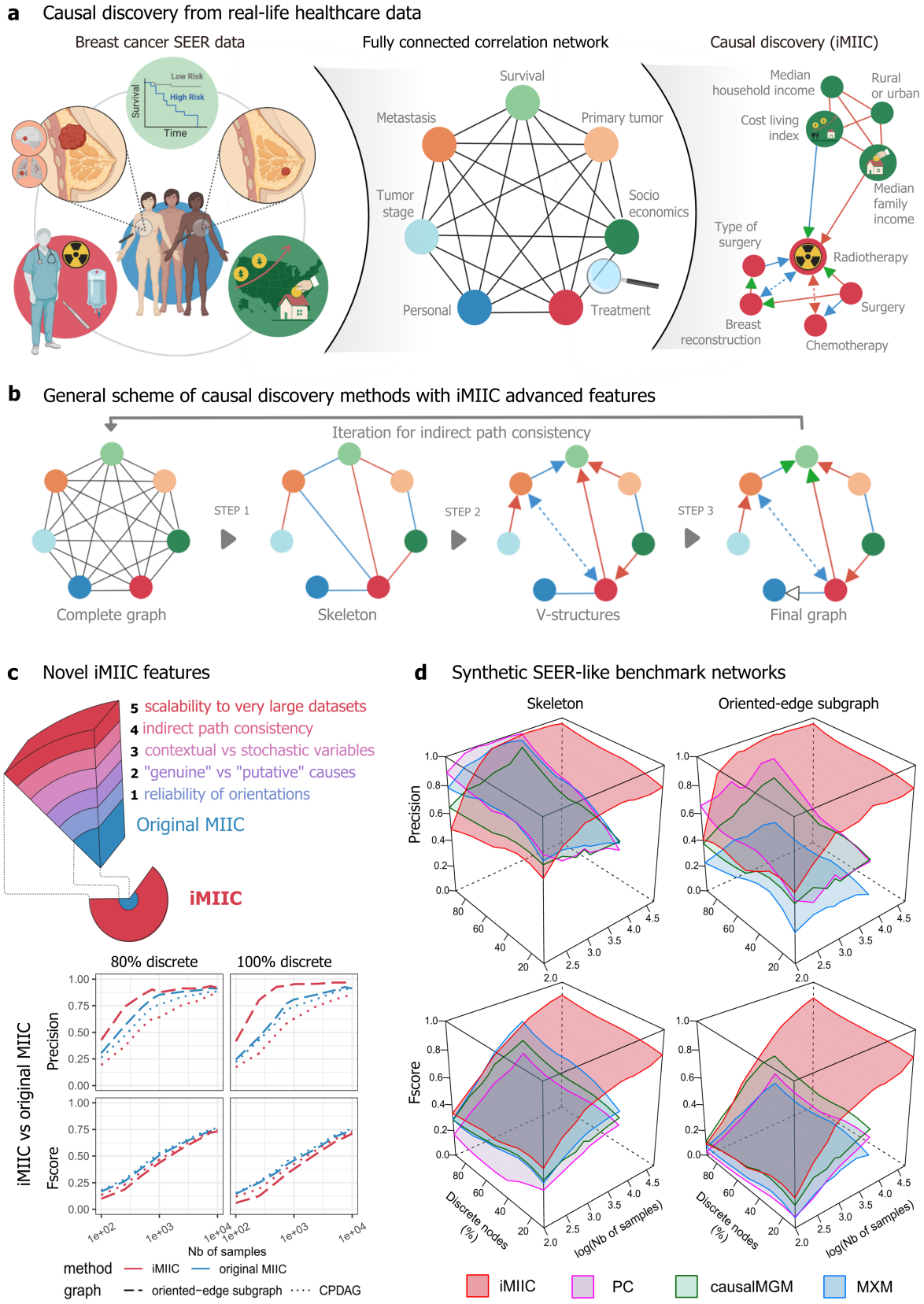


图1. 使用基于约束的方法从真实医疗保健数据中进行因果发现。(a) SEER数据库包含407,791份2010年到2016年确诊的乳腺癌患者医疗记录。因果发现旨在揭示此类具有全局相关性的数据集中的因果关系。 **(b)** 基于约束的方法的一般方案（包括 iMIIC 的新型高级功能，见正文和 补充信息附录）：步骤1，移除可省略边（确保间接路径一致性）；步骤2，定向“V结构”（以双向边表示可靠定向和潜在共同原因）；步骤3，传播定向，以白色箭头表示（并区分“推定的”与“真正的”原因，绿色箭头表示）。 **(c)** iMIIC 的新型高级功能，以及与原始 MIIC 的基准比较。 **(d)** 具有不同比例离散变量的合成类 SEER 基准网络，见正文、材料与方法以及图 S4-S6。使用 BioRender.com 创建

念上的进步及相关方法的发展。首先，iMIIC 基于普遍的信息论原理，定量提升了推断网络中定向的可靠性。在从真实医疗保健数据改编而来的、具有挑战性的基准数据集上，假阳性定向仅占少数百分比。其次，iMIIC 独特地能够区分“真正的”原因与“推定的”和“潜在的”因果效应。这一区分对于消除推断网络中定向边因果解释的歧义至关重要，框 1 中以直观示例说明了这一点。第三，iMIIC 能够量化间接效应，同时确保其与全局网络结构一致。这对于将间接贡献解释为通过推断网络中相应贡献者节点的间接路径十分重要，而其他因果发现方法通常无法做到这一点。此外，iMIIC 能够区分情境变量与随机变量，从而将外部控制变量纳入因果网络；最后，iMIIC 还可扩展至超大型数据集。这些独特能力为在多个研究领域中发现可靠且可解释的因果网络开辟了新途径。我们在合成数据以及来自监测、流行病学与最终结果（SEER）计划的乳腺癌患者超过 40万份医疗记录的真实医疗保健数据上，展示了 iMIIC 的因果发现性能（11），图 S1。

结果

因果发现方法概述及其局限性。基于约束的因果发现方法分多个连续步骤进行，见图1b（补充信息附录，第 1 节）。第一步是从初始的全连接网络中迭代移除所有可省略边，只要两个变量相互独立，或在给定一组所谓“分离集”的条件变量后条件独立即可。第二步则是对无向图（称为骨架）中的部分边进行定向，以形成所谓的“V结构”，$X \to Z \leftarrow Y$，它们是观察性数据中因果关系的标志，见框 1。最后，第三步旨在将 V 结构的定向传播至下游边，见图1b。然而，传统的基于约束的方法在有限数据集上缺乏稳健性，因为一长串不确定的判断会导致错误不断累积，从而降低最终网络的可靠性。尤其是，由条件变量的偶然组合所导致的伪条件独立会产生许多假阴性边，并最终降低推断出的定向的准确性。

最新的因果发现方法 MIIC（8、10）会先迭代收集显著的信息贡献者，再评估条件独立性，从而学习更稳健的因果图模型（补充信息附录，第 2 节）。实际应用中，与传统的基于约束的方法相比，MIIC 的策略减少了伪条件独立，并显著提高了灵敏度或召回率（即，正确恢复的边所占比例），见图 S2和 S3。然而，原始 MIIC 以及所有其他现有因果发现方法仍存在一些局限性。具体而言，（*i*）它们预测边定向的可靠性低于预测边保留的可靠性；（*ii*）它们揭示的是“推定的”而非“真实的”因果关系（框 1）；（*iii*）它们无法保证间接路径与全局网络结构一致；（*iv*）它们无法区分情境变量和随机变量；（*v*）它们的可扩展性有限。新颖的 iMIIC 方法有效克服了所有这些局限性，大幅提升了在大规模合成数据和真实观察性数据上进行因果发现的可靠性、可解释性和可扩展性。

iMIIC 提高了推断出的定向的可靠性。虽然原始 MIIC 在推断可靠定向方面显著优于传统的基于约束的方法，但 MIIC 骨架图与定向图预测之间通常仍存在明显的精确率损失，见图 S3。这是因为定向错误主要源于不一致的 V结构，$X \to Z \leftarrow Y$，其中心节点 Z 也可能被纳入不相连变量对的分离集$\{X, Y\}$，这与 V结构中头对头相连的关系相矛盾。对于包含水平数（过）多的离散变量的数据集，不一致的 V结构尤其常见。为避免此类不一致的定向，iMIIC 采用了更保守的定向规则，该规则基于一种通用的互信息上确界原理（15，16），并针对有限数据集进行了正则化（补充信息附录，第 3 节）。这一原理意味着，在评估（条件）独立性时，应以相同方式聚合分类变量或连续变量的各个水平。因此，定理 1（补充信息附录，第 3 节）要求校正独立变量之间的所有（条件）互信息。结合多元信息和定向评分的可扩展计算方法（补充信息附录，第 4 节），这一信息论原理大幅提升了预测方向的可靠性；与 MIIC 原始定向规则相比，灵敏度仅略有下降，见图 1c。尤其是在下文所述、从真实异质数据改编而来的高难度基准数据集上，iMIIC 的定向精确率超过 90%；而其他因果发现方法的定向精确率在样本量较大时通常停留在 50%–60% 以下，见图 1d（定向边子图精确率图）及图 S4-S6 （精确率图中的虚线）。

iMIIC 区分“真实”与“推定的”因果边。传统的基于约束的方法，乃至原始 MIIC 方法，都只能发现“推定的”因果关系，因为 V结构的定向先验地 既符合真实因果关系，也符合未观测共同原因所产生的效应，框 1 给出了一个直观示例。与之相反，iMIIC 通过排除未观测共同原因（或未测量混杂因素）对每条预测为真实因果边的影响，区分“真实”与“推定的”因果边。iMIIC 的这一独特特性是通过分别评估所有定向边的箭头头部和尾部概率来实现的（补充信息附录，第 4 节）。如果箭头头部和尾部概率在统计上都显著，则预测为真实因果边（以绿色箭头头部表示）；如果尾部概率在统计上不显著，或无法仅凭观察性数据确定，则因果边仍为“推定的”。同样，双向边表示未观测共同原因的效应，对应于两个显著的箭头头部概率；其他所有情况则以无向边表示（补充信息附录，第 5 节）。

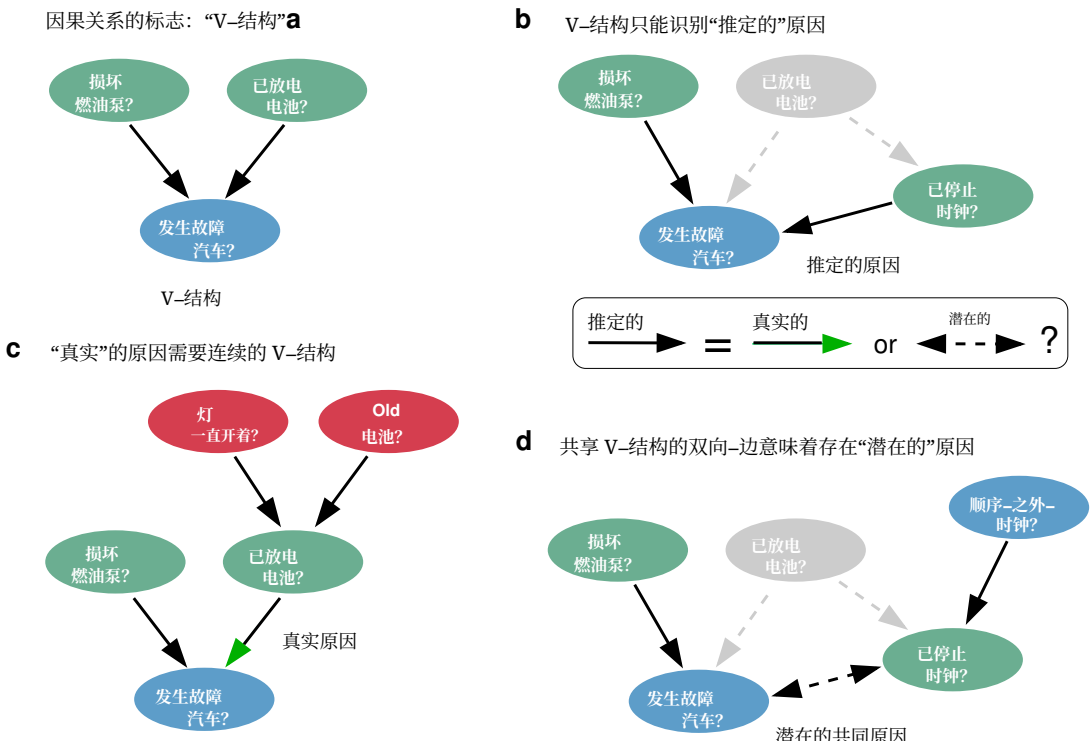
iMIIC 支持因果网络中的背景变量。iMIIC 中实现的箭头头部与尾部定向的独立概率框架，也允许纳入有关特定箭头头部或尾部定向的先验知识。例如，

框 1. 因果圆盘　　从观察性数据中发现原理：区分“真正的”原因与“推定的”和“潜在的”因果效应。

我们在此概述如何在纯观察性数据集中发现因果关系，并区分“真正的”原因与“推定的”及“潜在的”原因。我们通过一个符合直觉的虚构旧汽车数据集玩具示例来说明其原理。（**a**）在此类观察性数据集中，因果关系的特征是涉及三个变量的“V结构”子图，其中两个可能原因“燃油泵损坏？”和“电池没电？”彼此独立，因而不相连，并共同导致效应“汽车抛锚？”。该 V 结构的两条边都指向其中间变量“汽车抛锚？”，这是因为这两条边既不能是无向边，也不能指向“燃油泵损坏？”或“电池没电？”；否则，这些替代图模型将意味着两者相关，从而与“燃油泵损坏？”和“电池没电？”之间的独立性相矛盾。另一种情况是，在连续加性噪声模型这一特定假设下，有时仅凭两个变量也能发现因果关系（12）。然而，一般而言，因果发现至少需要三个变量，而且往往需要更多变量，因为 V 结构中可能原因之间的独立性通常以其他未在此讨论的变量为条件，这些变量构成一个分离集，见补充信息附录第 1 节。反过来，以 V 结构的顶端变量（此处为“汽车抛锚？”）为条件，会在彼此独立的可能原因之间引入虚假关联（1、2）。同样，若选取该顶端变量取特定值的数据集，也会因数据集中的选择偏倚而产生虚假关联（13、14）；例如，只选取“汽车抛锚？是”的数据时，“燃油泵损坏？”和“电池没电？”这两种可能原因之间会显现出某种反相关。（**b**）然而，V 结构实际上仍存在因果歧义（2），因为它们只能识别出“推定的”原因，这些原因可能是以绿色箭头头部表示的“真正的”原因，也可能表明存在未测量混杂因素，即数据集中未观察到、并以双向边表示的潜在共同原因。例如，常被用作“电池没电？”代理变量的“时钟停了？”，也会与“燃油泵损坏？”形成类似的 V 结构；然而，众所周知

已知“时钟停了？”不可能是“汽车抛锚？”的真实原因，因为篡改汽车的时钟实际上不可能导致汽车抛锚。（**c**）在缺乏背景知识且未对变量进行直接干预的情况下，要证明“电池没电？”确实是“汽车抛锚？”的真实原因，就必须排除“电池已放电”和“汽车发生故障”之间存在未观测共同原因（即未测量混杂因素）的可能性。为此，需要找到“电池没电？”上游的另一个 V 结构（例如“灯还开着？”→“电池没电？”←“旧电池？”），或者掌握关于某个上游（推定）原因的先验知识，并证明至少一个上游变量对下游变量“汽车抛锚？”的效应完全是间接的，且至少部分由中间变量“电池已放电”所中介。这要求找到一个条件独立关系：给定一个包含中间变量“电池已放电”的分离集后，上游变量与“汽车抛锚？”相互独立。（**d**）反过来，要排除某个推定原因是真实原因的可能性，就必须通过找到第四个变量（例如“时钟故障？”）来证明这种关系实际上源自未观测共同原因；该变量定义了另一个 V 结构，并与（b）中的 V 结构共同在“汽车抛锚？”和“时钟停了？”之间形成一条双向边。

先进的 iMIIC 方法通过评估每条边两端各自的箭头头部或尾部方向概率，区分真正的因果边与推定因果边，以及无向边和双向边（参见结果与补充信息附录第 4 和第 5 节）。因此，iMIIC 能发现四种具有不同因果解释的边。iMIIC 还可以估计任意一对变量之间的直接和间接信息贡献，补充信息附录第 6 节。然而，与其他因果发现方法一样，iMIIC 不会量化因果效应；量化因果效应需要额外的假设（可识别性），而这些假设通常无法在观察性研究中检验（2）。特别是，推定原因的因果效应不可识别，这意味着仅凭观察性数据无法量化对推定原因进行干预的结果。



在图模型中纳入少数几个背景变量，有助于根据数据集的性质指定控制参数或实验条件，或描述患者的个人档案（例如性别、出生年份）。与数据集中的大多数其他变量不同，此类背景变量不会随机变化；根据假设，它们的所有边都不应有入箭头，即 $p_{\text{tail}} = 1$。这体现了我们的先验知识：背景变量不可能是数据集中其他已观测或未观测变量的结果。

iMIIC 强制实现间接路径一致性，并量化其信息贡献。基于约束的因果发现方法在第一步中移除可省略边，其依据是：所有不相连变量之间的统计关联，都应能通过最终网络中的间接路径以图形方式解释。然而，实际情况往往并非如此（17）。尤其是，无法保证在迭代移除边的过程中识别出的分离集，在最终网络的间接路径方面仍保持一致。为此，iMIIC 采用了一种新颖的

算法方案（17），以确保所有用于移除可省略边的分离集都与最终推断图一致。具体做法是迭代重复基于约束的结构学习方案，同时仅选取与前一次迭代所得骨架或部分定向图一致的分离集，如图 1b 所示。这种间接路径一致性提升了 iMIIC 推断网络在间接效应方面的可解释性；间接效应也通过间接信息贡献进行量化（补充信息附录，第 6 节）。

iMIIC 在合成基准数据集上的表现优于现有方法。我们在多个类 SEER 基准数据集上，将 iMIIC 与原始 MIIC 以及其他先进的基于约束的方法进行了基准测试；这些数据集包含不同比例的离散变量，见材料与方法。图 1c 表明，对于离散变量占比较高的类 SEER 基准数据集，iMIIC 显著提高了定向精确率，代价是定向灵敏度和 F 分数略有下降。例如，对于 $N = 500$ 个样本，在完全离散的类 SEER 数据集中，iMIIC 的定向精确率（及 F 分数）已达到 93%（及 25%），而原始 MIIC 为 64%（及 35%）；对于离散变量占比为 80%（与实际 SEER 数据集相同）的数据集，iMIIC 的定向精确率（及 F 分数）超过 85%（及 32%），而原始 MIIC 为 73%（及 39%），见图 1c。此外，在推断出的定向的可靠性和灵敏度方面，iMIIC 大幅优于其他先进的基于约束的方法，见图 1d 和图 S4-S6。尤其是，在这些类 SEER 数据集的所有样本量和离散变量占比下，iMIIC 的定向 F 分数约为 PC算法的两倍（18、19）。例如，对于离散变量占比为 80%（与实际 SEER 数据集相同）的基准数据集，iMIIC 在 $N = 10^3$时的精确率（及 F 分数）已达到 88%（及 44%）；相比之下，保守 PC（19、20）约为 60%（18%），causalMGM（21）为 50%（36%），MXM（22）为 24%（18%）。对于$N = 10^4$，iMIIC 的精确率（F 分数）达到 92%（73%）；相比之下，保守 PC 约为 75%（40%），causalMGM 为 62%（55%），MXM 为 30%（30%）。最后，对于$N = 10^5$，iMIIC 的定向精确率和 F 分数均超过 90%，这一表现超出了其他方法可达到的样本量范围。关于连续变量占比更高时的比较，见材料与方法。

应用于全国乳腺癌医疗记录。我们在来自美国国家癌症研究所监测、流行病学与最终结果（SEER）计划的大型乳腺癌数据集（11）上应用了 iMIIC。该计划收集了占美国人口 % 的人群中癌症诊断、治疗和生存数据，见图 1a。乳腺癌（23）是女性中最常见的浸润性癌，患者治愈率仅为 70–80%；在肿瘤亚型、诊断时的分期、初始及后续治疗，以及患者年龄、族裔、遗传易感性、生活方式或社会经济状况方面，均存在显著差异。已有大量针对 SEER 癌症数据的回顾性关联研究（24–27）和少数因果推断研究（28–31），使其成为评估因果发现方法在真实医疗保健数据上实际表现的独特基准资源。

此处我们展示 iMIIC 对 2010—2016年 SEER 乳腺癌数据进行的因果发现分析。共有 407,791 份医疗记录，但由于部分患者有多个乳腺原发肿瘤，因此仅涉及 396,179 名不同患者。我们根据 51 个临床、社会经济和结局变量与乳腺癌的相关性，以及其冗余或缺失信息较少这一特点，选取了这些变量，见图 S1。

由此得到的乳腺癌网络（图 2a）提供了一个包含 280 条边、可解释的图模型；其中大多数因果关系要么已知，要么可依据常识、专家知识及临床实践予以排除。这项评估表明，iMIIC 推断出的真实或推定因果效应中，约 90% 是正确的；另有 8% 的因果关系根据临床和流行病学知识来看似乎合理（见 数据集 S1）。因此，iMIIC 的新型定向规则仅导致 2% 的因果边有误；相比之下，对同一 ∼400,000 名患者的 SEER 队列进行分析时，采用 MIIC 原有定向规则会产生约 15% 的错误因果边。此外，预测出的真实因果边中，没有一条连接非癌症特异性变量，例如容易受到潜在选择偏倚影响的个人或社会经济信息（13, 14）（框 1）。此外，对于真实因果边，可排除未测量的（潜在）混杂因素（框 1）；而已测量混杂因素的贡献则估计为间接路径贡献（见 补充信息附录，第 6 节）。然而，SEER 数据库的数据收集和分析还存在其他偏倚来源（32, 33）（下文将对此进行讨论）。这个包含 ∼400,000 名患者的临床网络也能抵御子抽样影响，因为它包含从三个各有 100,000 名患者的随机子集中学习得到的三个网络中 90% 的边，见图 2b。此外，三个 100,000 名患者子集网络中，88% 的边定向概率彼此相容；其中 92% 也与完整网络的边定向概率相容（见 数据集S1）。

iMIIC 乳腺癌网络的因果解释。现在，我们将探讨 iMIIC 推断出的 SEER 乳腺癌网络在临床和社会经济层面的解释，见图。2a。我们将重点关注 iMIIC 揭示的既符合预期又更令人意外的真实因果关系，并尝试从诊疗路径、治疗决策、患者偏好或医疗保健的社会经济决定因素等角度，解释那些违反直觉的因果预测。我们将从不同类别的变量及其相关子网络的角度呈现这些结果，首先讨论生存子网络，然后是原发肿瘤子网络、手术及后续治疗子网络，最后是社会经济子网络。

生存子网络。完整网络，图2a，包含四个与患者在2016年末生存状态相关的节点，构成一个生存子网络，其中包括所有与患者生存状态直接关联的变量，见图3a。除了每位患者的生存状态（死亡或存活）外，另外两个节点分别指定死因，即乳腺癌或其他原因导致的死亡；第三个连续变量表示生存或随访时长（月），并受研究的2010—2016年删失期限制。图3a显示，iMIIC正确识别出了已知的乳腺癌致死因素，例如诊断时转移（总死亡率为49.2%）；诊断时更严重的远处转移（脑和肝脏）也保留了直接

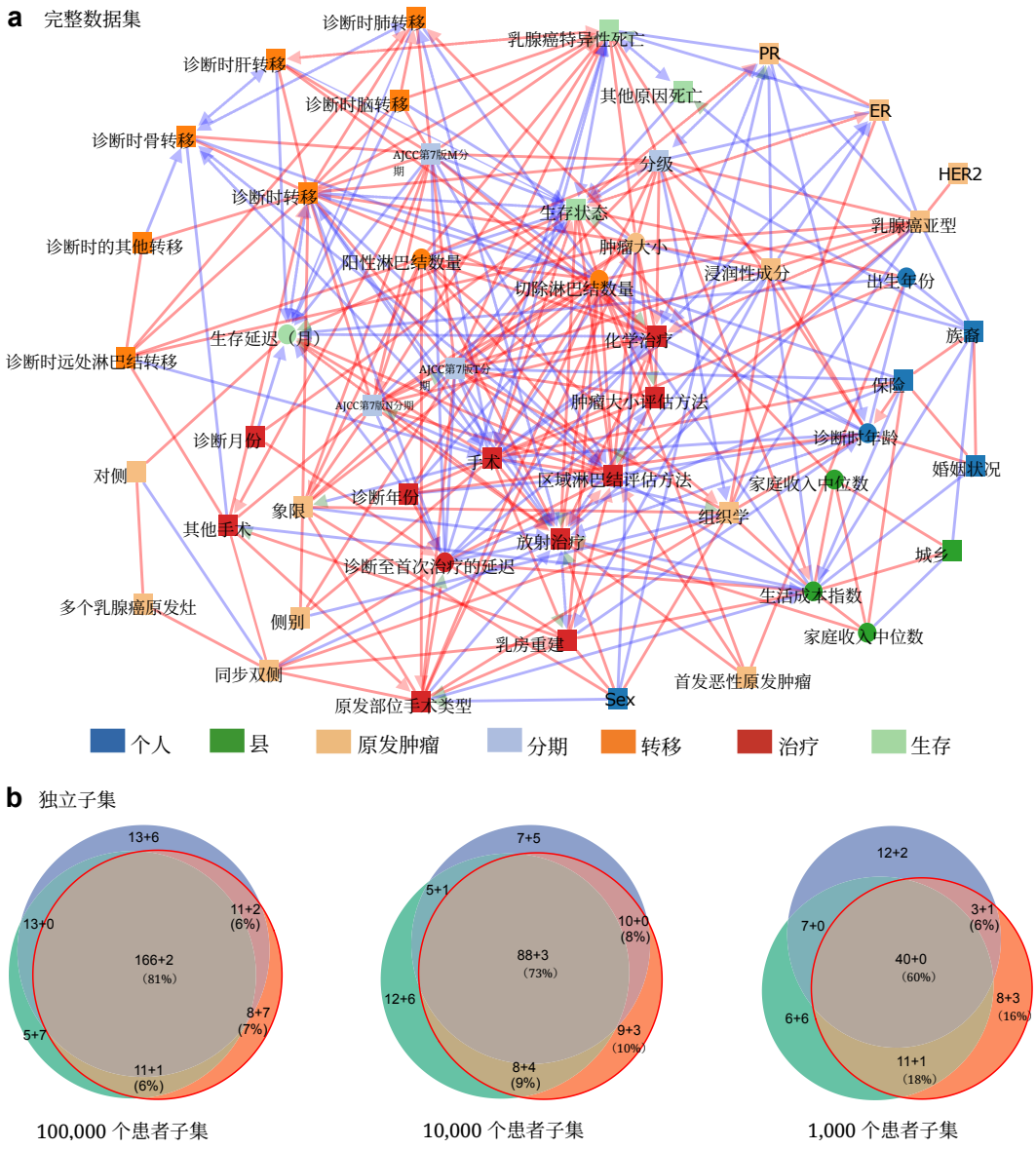


图2. 由 iMIIC 推断的 SEER 乳腺癌网络。（**a**）由 iMIIC 根据 SEER数据集推断的 51节点网络，包含 396,179 名于 2010年 至 2016年间确诊的乳腺癌患者。该骨架一致网络包含 280 条边，并包括 2 个背景变量：性别和出生年份。相应的方向一致网络包含 340 条边，见图 S7。有关 iMIIC 预测的每条边的列表及其因果性质，请参见 数据集S1。（**b**）比较从三个独立子抽样中推断出的网络，这些子抽样的患者子集大小分别为 100,000、10,000 或 1,000（从左至右）。欧拉图中共享边的数量（不论定向如何）以总和形式给出 $a+b$ 其中 a （分别为 b）对应于（a）中完整数据集网络包含（分别为不包含）的边数。百分比对应于总边数为中位数的子集网络（红色圆圈）。

关联，分别指向乳腺癌特异性死亡和生存状态，这解释了它们较高的死亡率，即脑转移（70.5%）和肝转移（59.5%）。类似地，转移阳性淋巴结数量以及分期变量（AJCC第7版T、N、M分期）均正确地与乳腺癌特异性死亡和生存状态相连，而未与任何其他死因相连。相比之下，iMIIC推断出生年份与其他原因导致的死亡之间，以及出生年份与生存状态之间存在因果关系，这符合预期。我们还可以看到，无论死因如何，患者死亡都被正确地预测为会导致生存时间缩短。然而，图3a也包含一些不那么直观的发现。具体而言，在完整数据集和全部三个各含100,000名患者的子集中，均稳健地推断出生存状态会“导致”放射治疗；在存活患者中，51%曾接受

接受放射治疗的患者中，仅有 27% 的患者死亡，图 3b。这表明，诊断后最初几个月内过早死亡，可能会使部分患者无法接受放射治疗；如果他们存活时间更长，本来可能会接受这种治疗。生存状态与放射治疗之间这种短期因果效应，与未接受放射治疗时前 3–6 个月生存延迟分布迅速下降的现象一致，图 3c；这也符合诊断后接受放射治疗的典型时间范围，具体取决于放射治疗是在手术后作为第二种治疗实施，还是在手术和化学治疗后作为第三种治疗实施（34）。总而言之，生存状态对放射治疗的这种短期因果效应，超过了方向相反的因果效应，即放射治疗对患者长期生存的有益作用。这表明存在显著的“不死时间偏倚”（32）

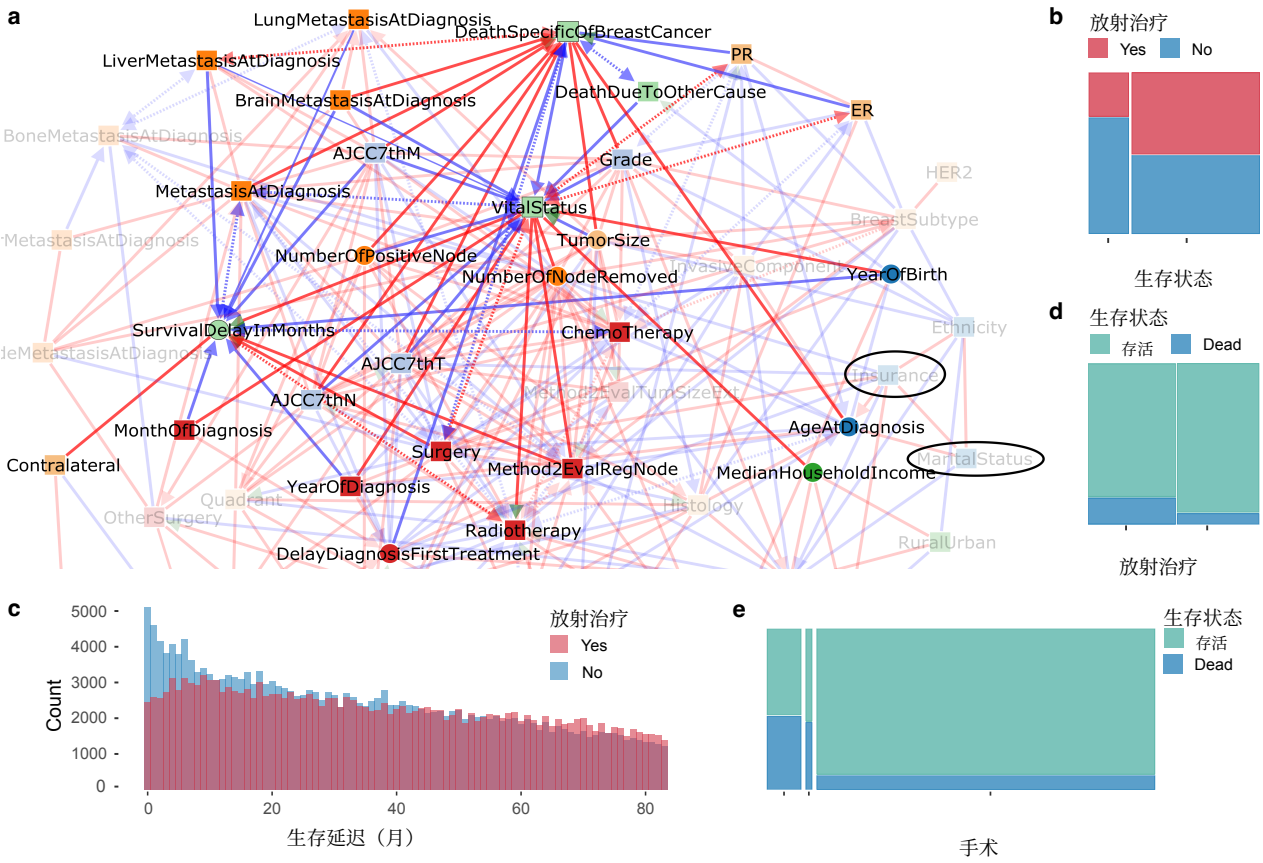


图 3. iMIIC 根据 SEER 乳腺癌数据集推断出的生存子网络。 **(a)** 突出显示与生存变量（生存状态、乳腺癌特异性死亡、其他原因死亡、生存延迟（月））直接关系的子网络。网络中未显示与其他变量（如保险和婚姻状况）存在直接连接，这可根据符合网络骨架的间接路径贡献来解释，详见正文和 补充信息附录第 6 节。 **(b)** 放射治疗与生存状态的联合分布，突出显示二者之间违反直觉的因果关系，详见正文。 **(c)** 接受或未接受放射治疗的患者生存延迟（月）的直方图。每个区间代表一个月。蓝色的早期峰值表明，许多患者在诊断后 3 至 6 个月内死亡，因此尚未接受放射治疗，这与 (a) 中预测的因果方向一致。这导致 (d) 中放射治疗表面获益被高估，详见正文。 **(d)** 生存状态与放射治疗的联合分布。 **(e)** 生存状态与手术的联合分布。

放射治疗表面上的获益，图 3d，需要通过“不死时间偏倚的标志时间法”（32， 35）予以校正，排除手术后特定时间内死亡的患者；或者通过利用观察性数据模拟目标试验（36）。相比之下，手术通常在诊断后 5 至 8 周内进行，被发现是导致患者生存延迟延长的主要原因，详见下文图 3e 和图 4a。

最后，我们注意到，一些被报告与生存变量相关的变量，实际上是间接而非直接与这些变量相连。保险（37、38）和婚姻状况（39、40）尤其如此。保险（类别为未参保 / 医疗补助 / 非医疗补助）对乳腺癌死亡的间接效应，已被证明可通过手术（50%）、化学治疗（14%）、婚姻状况（20%）、放射治疗（9%）和乳房重建（7%）间接解释，见《补充信息附录》第 6 节中的公式 10。同样，婚姻状况（类别为单身 / 已婚 / 分居 / 离婚 / 丧偶）对乳腺癌死亡的间接效应，已被证明可通过手术（58%）、出生年份（40%）和族裔（2%）间接解释。

*原发肿瘤子网络。*除诊断时转移外，激素受体（ER/PR）状态和原发肿瘤大小也被发现会直接影响患者的生存预后，见图S8a。具体而言，iMIIC推断，ER状态会使乳腺癌导致的死亡风险从17.7%（ER−）降至5.4%（ER+），其中PR状态的间接贡献很大（82%）。这与ER对PR的转录调控（41）相符；此外，ER+/PR− 患者的死亡率（11.8%）显著高于ER+/PR+ 患者（4.4%）。同样，iMIIC推断，原发肿瘤的组织学与其他多个变量存在直接关联，见图S8a，例如诊断时年龄（与早期报告一致（42）），以及同步双侧原发肿瘤（在首次诊断后6个月内发现）；存在小叶癌时，此类肿瘤发生的可能性几乎高出一倍，见图S8b。相比之下，与诊断后超过6个月才发现的对侧原发肿瘤之间未发现显著关联，见图S8c。

*手术及后续治疗子网络。*有趣的是，iMIIC 还揭示了手术在精确表征原发肿瘤方面的核心作用，见图 4a。例如，iMIIC 揭示了一个多少有些出乎意料的真实因果联系：手术会影响组织学类型。这反映出，手术后病理学家常会根据手术标本进一步明确组织学类型。这与手术后特定组织类型（例如

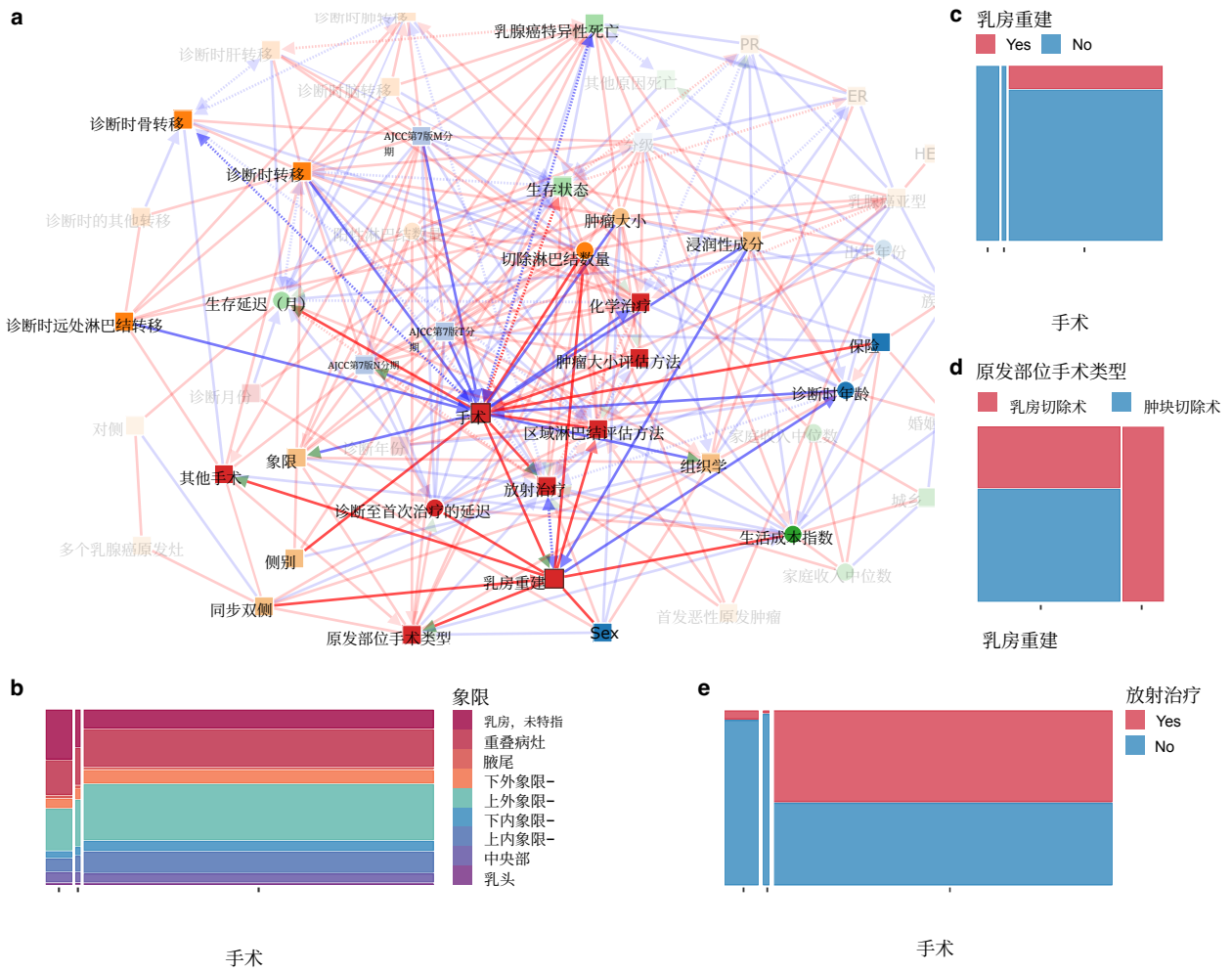


图4. iMIIC根据SEER 乳腺癌数据集推断出的手术与后续治疗子网络。 **(a)** 突出显示与手术和乳房重建直接相关的子网络。 **(b)** 象限与手术的联合分布。 **(c)** 乳房重建与手术的联合分布。 **(d)** 原发部位手术类型与乳房重建的联合分布。 **(e)** 放射治疗与手术的联合分布。关于手术在细化原发肿瘤特征及后续治疗决策（包括患者的个人选择）中所起作用的因果解释，请参阅正文。

手术后，浸润性导管癌与其他类型癌混合（+77%）和浸润性导管癌和小叶癌（+48%）、浸润性导管癌，未特指（+7.6%））的比例显著上升，而较笼统的组织学类型，如小叶癌，未特指（-11%）、癌，未特指（-91%）和腺癌，未特指（-95%）的比例则相应下降，这一结果一致。同样，iMIIC 正确推断出，分期变量 AJCC第7版N分期通常依据手术后的病理报告；而未进行手术（因诊断时存在远处转移或患者年事已高）则会导致原发肿瘤的乳房象限定位更常未予具体说明，见图 4a，也就是说，未建议手术时，30.4% 的病例为“乳房，未特指”；而建议手术时这一比例为 11.1%，见图 4b。

同样，iMIIC 揭示了手术在后续治疗决策中的核心作用，例如乳房重建和放射治疗，见图 4a。虽然乳房重建确实需要乳房手术，见图 4c，iMIIC 也正确推断出，原发部位手术类型（肿块切除术或乳房切除术）在很大程度上取决于早期乳腺癌患者在保乳和乳房重建方案之间的个人选择，见图 4a、d。同样，iMIIC 正确推断出放射治疗是乳房手术的一种常见“后果”，见图 4a，也就是说，53% 相比于未进行手术时的 4% 放射治疗率，见图4e；尤其是肿块切除术后（75%），以降低保乳手术后的复发风险。

社会经济子网络。图中完整的乳腺癌网络。2a 包含四个与每位患者居住县相关的社会经济变量：家庭收入中位数、家庭收入中位数、生活成本指数，以及各县的城乡人口规模。这四个社会经济变量实际上构成了一个完全连接的子图（即一个团），表明它们之间存在强烈的相互依赖关系，并且与许多其他变量直接相连，见图。5a。有趣的是，生存状态仅通过家庭收入中位数与这个县级变量团相连，这与先前有关预期寿命和收入之间关联的研究结果一致（43）。相比之下，所有其他与县级变量团相连的患者特定变量（如肿瘤分级、放射治疗、乳房重建、保险）实际上都至少与生活成本指数有一条连线，这凸显了医疗卫生系统与全球经济的融合。尤其是，较高生活成本与更有利的乳腺癌

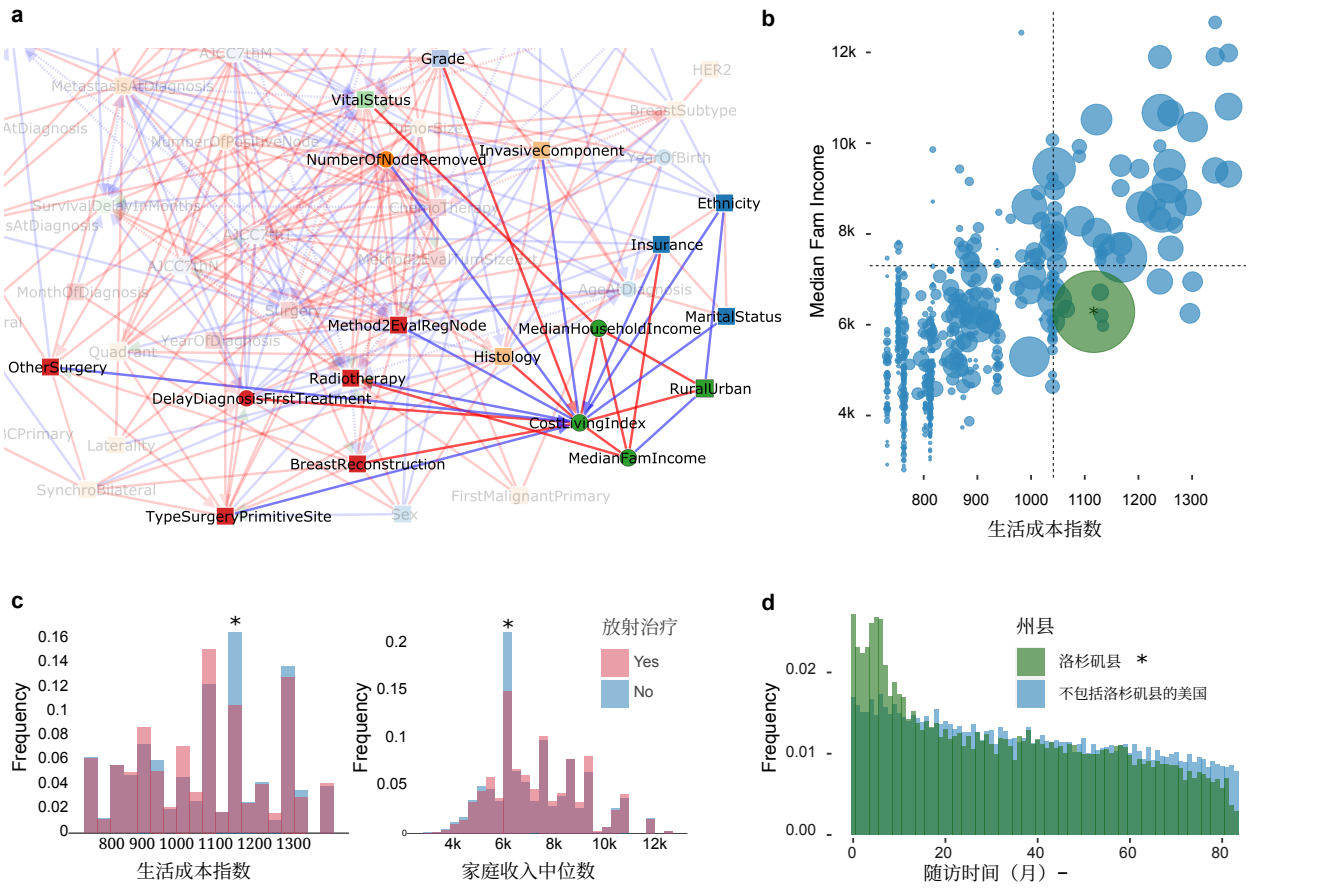


图 5. 根据 SEER 乳腺癌数据集通过 iMIIC 推断的社会经济子网络。（**a**）突出显示与县级社会经济变量（生活成本指数、家庭收入中位数、家庭收入中位数和城乡）直接关系的子网络。（**b**）家庭收入中位数与生活成本指数联合分布的气泡图。气泡面积表示该县的患者人数。虚线分别表示生活成本指数均值和家庭收入中位数均值。带星号的绿色气泡表示洛杉矶县，该县占完整数据集的 10%。（**c**）按放射治疗分组的生活成本指数和家庭收入中位数直方图。带星号的区间对应洛杉矶县。（**d**）洛杉矶患者与 SEER 中纳入的其他美国各县患者的随访时间（月）直方图。

预后之间存在直接关联（例如诊断时浸润性成分较少）。这可能是因为预防性医疗条件更好，包括更容易获得乳腺癌筛查中心服务以及更全面的保险保障。然而，各县之间也存在显著差异，表现为保险和放射治疗与家庭收入中位数的关联方向，与其和生活成本指数的关联方向相反，见图。5a。这些耐人寻味的发现可追溯至洛杉矶县，该县约占整个数据集的10%；尽管其生活成本指数高于平均水平（处于第58至67百分位范围），家庭收入中位数却低于平均水平（处于第29至38百分位范围），见图。5b。这必然加重了许多患者的经济负担：2010年至2016年间，洛杉矶县共有39,089名乳腺癌患者确诊。尽管这些患者中有18%享有医疗补助保险（而整个数据集中的比例为10%），许多人仍不得不选择价格较低但保障有限的私人保险，其中共付政策的负担相当显著；尤其是在2014年1月《平价医疗法案》实施之前，还有一些人甚至没有保险（2013年未参保比例为3.4%，而2014年为1.5%）。因此，许多洛杉矶患者似乎放弃了接受昂贵治疗。具体而言，洛杉矶仅有32.6%的患者接受放射治疗，而全国范围内这一比例为50%——不包括洛杉矶县的范围更广；图5c 所示的差异只能部分归因于各县对门诊患者放射治疗漏报程度的不同（32、33）。此外，根据洛杉矶的随访时间分布相比数据集其余部分迅速下降这一情况，估计有 7% 的洛杉矶患者似乎中断了治疗，或迁往 SEER 数据库未涵盖的其他县（全国这一比例为 1.5%，不包括洛杉矶县）。见图5d。这一比例指的是在诊断后不到一年内、且在 2016年12月本研究结束前一年以上进行最后一次医疗接触的患者所占比例。

讨论

全国性医疗保健数据（例如本文分析的 SEER 乳腺癌数据集）从方法论角度来看尤其值得关注；它们提供了真实场景中的基准数据集，有助于评估因果发现方法在真实数据上的可靠性，因为大多数因果关系预测都可以依据专家知识、临床实践，或潜在的数据收集与选择偏倚加以验证或否定。此外，机器学习方法的可解释性对于临床数据应用尤为重要，因为人工智能辅助的建议很难仅依赖黑箱分类器，还需要用医疗从业者能够理解的理由加以解释。然而，

除临床数据之外，因果发现方法有望成为用于解释广泛领域中各类观察性数据的关键机器学习方法；由于实际、成本或伦理方面的原因，这些领域无法开展系统性扰动实验。具体而言，因果发现可以通过预测特定干预的因果效应来指导生物学研究（44），例如基因表达或基因沉默；随后可通过靶向小干扰RNA、基因敲除或基于CRISPR的编辑实验对其进行检验。

在 SEER 的乳腺癌数据集背景下，iMIIC 揭示了许多符合预期的因果关系，例如转移的不良后果，以及 ER+状态对乳腺癌所致死亡的保护效应；又如，研究结束时，出生年份是其他原因导致死亡的主要原因。另一方面，研究发现，保险保障或婚姻状况虽被报道可降低乳腺癌所致死亡风险，但其效应完全是间接的，且主要通过治疗（60-80%）介导，尤其是手术（$>$50%）。事实上，手术似乎是乳腺癌治疗的基石：它首先有助于明确组织学类型，继而指导放射治疗和乳房重建的治疗决策，最终延长患者的生存时间。然而，iMIIC 也正确推断出，原发部位的手术类型（肿块切除术或乳房切除术）很大程度上取决于早期乳腺癌患者在保乳或重建方案之间的个人选择。相比之下，放射治疗和化学治疗等其他治疗似乎对乳腺癌结局的影响较不显著，这或许部分是因为 SEER数据库中有些治疗信息报告不全（32，33）。放射治疗甚至似乎是生存状态的结果，而非原因，这表明，诊断后的头几个月内早逝，可能会使一些患者无法接受放射治疗；如果他们存活得更久，本来可能会接受这种治疗。最后，iMIIC 发现县级社会经济变量（例如家庭收入中位数和生活成本指数）与患者特定变量（例如肿瘤分级、放射治疗、乳房重建和保险）之间存在直接关联，凸显了医疗卫生系统与全球经济的融合。平均而言，较高的生活成本与更有利的癌症预后相关，这可能是由于预防性医疗服务更完善、保险保障更全面所致；与此同时，iMIIC 还揭示了各县家庭收入与生活成本指数之间存在巨大差异（例如洛杉矶县），导致经济负担加重，患者不得不放弃昂贵的治疗，甚至中断治疗。

总之，iMIIC 是一种通用的因果发现方法，可从广泛的生物学和临床数据中揭示直接且可能具有因果性的关系，以及网络一致的间接效应。iMIIC 能够处理异质数据（即结合分类变量和连续变量的数据）、包含未观测潜变量的数据（即存在未测量混杂因素的数据），以及含有缺失数据的变量；这类情况在许多现实应用中普遍存在。重要的是，iMIIC 完全无监督，不需要预设假设或专家知识。具体而言，iMIIC 会自动调整已测量混杂因素（以间接贡献的形式），并通过排除或突出显示每条因果边上未测量混杂因素的效应（框 1），从而区分真实原因与推定的和潜在的因果效应。尽管 iMIIC 无法避免可能影响观察性数据的数据收集偏倚和选择偏倚，但它基于稳健的信息论框架，因此在解释具有挑战性的数据类型时尤为可靠，例如整合自不同来源的多模态数据（例如临床、个人和社会经济数据，如本文及文献（10、45）所示），或采用不同实验技术获取的数据（例如单细胞转录组学（8、44、46）和成像数据（10））。原则上，只要只有观察性数据可用，iMIIC 就可应用于广泛的其他领域，以揭示因果关系并量化间接贡献。随着许多数据科学领域几乎无限的数据集不断涌现，我们亟需可扩展的因果发现方法；我们相信，iMIIC 能够基于因果解释，为多个研究领域带来独特见解。

材料与方法

类 SEER 数据集生成。为便于与其他因果发现方法进行比较，我们使用从完整的乳腺癌患者 SEER数据集中抽取的 10,000 个患者子集中推断出的网络结构生成类 SEER 合成数据集，具体如下。首先，使用蒙特卡洛图生成算法（47）生成具有类似类 SEER 度分布、且各节点的连接存在额外变异性的随机网络骨架。随后，按照节点的随机排序为这些骨架定向，以获得有向无环图，并设定离散变量与连续变量的不同占比。无父节点变量的边缘分布选取为类似典型的 SEER 边缘分布（图 S1）；其他变量则使用混合类型结构方程模型（结构方程模型）进行模拟（10），例如见图 S2。对于每种离散节点占比（以十分位为步长），生成 25 个基准网络，并分别用于生成 100,000 个样本。

因果发现评分。为进行评估，我们将网络重构视为一项二元分类任务，并计算精确率、召回率和 F 分数等经典性能指标，以评估（i）骨架、（ii）完全部分定向无环图（CPDAG）以及（iii）定向边子图重构。CPDAG 评分采用与骨架评分相同的指标，但将 CPDAG 中非定向边的错误定向，以及定向边的未定向或反向定向评为“假阳性”。然而，从因果发现的角度来看，这些错误并不等同。因此，我们引入了定向边子图得分，其评估范围仅限于理论 CPDAG 与推断图中仅包含定向边的子图。这些定向边评分在基准比较中予以重点展示，旨在专门评估方法在因果发现方面的表现，即评估其对原则上可从观察性数据中学习的定向边的识别能力，以及因果结构学习方法实际预测这些边的表现。

基准评估的因果发现方法。五种能够分析混合类型数据集的因果发现方法，在类 SEER 生成数据集上进行了比较：

- 可解释 *MIIC*（*iMIIC*）在所有设置下均使用默认参数运行。
- 原始 *MIIC*(9, 10) 在所有设置下均使用默认参数运行（图 1c 和图 S3）。
- *PC*(18) 使用 *pcalg* 包（19）运行，并采用稳定选项（48）以及多数规则（48）（图 S3）或保守规则（20）（图 1d 和图 S4）进行定向。使用 *bnlearn* 包（49）中的“ci.test”函数对混合类型数据进行独立性检验（离散变量与连续变量之间使用“mi-cg”选项，离散变量之间使用“mi”选项，连续变量之间使用“mi-g”选项），显著性检验的阈值设为默认值 $\alpha = 0.01$。

- *causalMGM* (21) 使用 *rCausalMGM* R 包运行。初始图使用 mgm() 函数计算，三个 lambda 参数均设为 0.05；随后使用 pcMax() 函数并采用默认的 $\alpha = 0.01$ 参数获得定向。

- *MXM* (22)，一种混合型 PC 约束方法，使用 *MXM* R 包运行。通过 pc.skel() 函数获取图的骨架，采用“comb.mm”独立性检验和默认的 $\alpha = 0.01$ 显著性检验阈值，并通过 pc.or() 函数进行定向。

计算时间。当某方法在连续变量比例较高的情况下，单个网络的平均计算时间达到 1 小时（连续变量比例较低时，单个网络约为 10 分钟）时，即停止基准测试；对于每个样本量下生成的 250 个网络，这相当于最长运行时间约为 115 小时。

基准测试结果。在具有不同比例离散变量的类 SEER 基准数据集上，将 iMIIC 与最先进的基于约束的方法 PC、causalMGM 和 MXM 进行了基准测试，见图 1d 和图 S4-S6。正文讨论了离散变量占 80% 的数据集结果，这一比例与真实 SEER 乳腺癌数据集中实际的比例相符。同样，对于连续变量比例较高的情况，图 1d 和图 S4-S6 表明，与最先进的基于约束的方法相比，iMIIC 在预测方向的可靠性和灵敏度方面均显著更优。例如，在离散变量仅占 20% 的类 SEER 基准数据集中，iMIIC 的精确率和 F 分数分别已达到 81% 和 64%；对于 $N = 10^3$，其精确率和 F 分数分别为 53% 和 29%（保守 PC）、50% 和 40%（causalMGM）以及 29% 和 25%（MXM）。对于 $N = 10^4$，iMIIC 的精确率和 F 分数达到 88% 和 78%，而保守 PC 约为 60% 和 45%，causalMGM 为 52% 和 50%，MXM 为 22% 和 28%。最后，iMIIC 在 $N = 10^5$上的结果达到 86%（81%），这一样本量是其他方法所能达到的样本量之外的。

数据、材料和软件可用性。乳腺癌患者数据集来自监测、流行病学和最终结果计划，可通过 https://seer.cancer.gov/seertrack/data/request/ 获取。使用 iMIIC 进行因果发现，可通过开放访问服务器 https://miic.curie.fr 完成，也可运行 https://github.com/miicTeam/miic_R_package 上提供的 R 包。用于基准比较的其他 R 包可在以下网址获取：https://r-forge.r-project.org/projects/pcalg、https://cran.r-project.org/web/packages/bnlearn、https://github.com/tyler-lovelace1/rCausalMGM 和 https://cran.r-project.org/web/packages/MXM。

致谢。我们谨向 Irène Buvat、Laura Cantini、Michèle Sebag、Jean-Christophe Thalabard 和 Nathalie Vialaneix 致谢，感谢他们就稿件初稿提出的宝贵讨论意见。

1. P Spirtes, CN Glymour, R Scheines, D Heckerman, 因果关系、预测与搜索。（麻省理工学院出版社），（2000年）。2. J Pearl，因果关系。（剑桥大学出版社），（2009年）。3. J Runge, P Nowack, M Kretschmer, S Flaxman, D Sejdinovic，在大型非线性时间序列数据集中检测并量化因果关联。*Sci. Adv*。**5**（2019年）。4. J Runge 等，推断地球系统科学时间序列中的因果关系。*Nat. Commun.***10**，2553（2019年）。5. D Heckerman, D Geiger, DM Chickering，学习贝叶斯网络：知识与统计数据的结合。*Mach. Learn*。**20**，197–243（1995年）。6. J Friedman, T Hastie, R Tibshirani，使用图形套索进行稀疏逆协方差估计。*Biostatistics* **9**，432–441（2008年）。7. J Zhang，存在潜在混杂因素和选择偏倚时因果发现定向规则的完备性。*Artif. Intell.***172**，1873–1896（2008年）。8. L Verny, N Sella, S Affeldt, PP Singh, H Isambert，利用基因组数据中的多元信息学习含潜变量的因果网络。*PLoS Comput. Biol.* **13**，e1005662（2017年）。9. N Sella, L Verny, G Uguzzoni, S Affeldt, H Isambert，Miic online：一个从非扰动数据重建因果或非因果网络的网络服务器。*Bioinformatics* **34**，2311–2313（2018年）。10. V Cabeli 等，基于混合类型数据的信息估计量从医疗记录中学习临床网络。*PLOS Comput. Biol.*。**16**，e1007866（2020年）。

11. N Howlader 等，SEER癌症统计回顾，载于《SEER癌症统计回顾》。（美国国家癌症研究所），第1975–2016页（2018年）。12. J Peters, JM Mooij, D Janzing, B Schölkopf，使用连续加性噪声模型进行因果发现。《机器学习研究杂志》15，2009–2053（2014年）。13. DL Sackett，分析性研究中的偏倚。《慢性病杂志》32，51–63（1979年）。14. MA Hernán, S Hernández-Díaz, JM Robins，选择偏倚的结构化方法。《流行病学》15，615–625（2004年）。15. TM Cover, JA Thomas，《信息论要素》。（Wiley），第2版，（2006年）。16. V Cabeli, H Li, M da Câmara Ribeiro-Dantas, F Simon, H Isambert，基于有限数据集互信息上确界原理的可靠因果发现，发表于WHY21研讨会上的论文，神经信息处理系统大会第35届会议。（NeurIPS），（2021年）。17. H Li, V Cabeli, N Sella, H Isambert，使用一致分离集进行基于约束的因果结构学习。《神经信息处理系统进展》（NeurIPS）32（2019年）。18. P Spirtes, C Glymour，一种快速恢复稀疏因果图的算法。《社会科学计算评论》9，62–72（1991年）。19. M Kalisch, M Mächler, D Colombo, MH Maathuis, P Bühlmann，使用R软件包pcalg通过图模型进行因果推断。《统计软件杂志》47，1–26（2012年）。20. J Ramsey, P Spirtes, J Zhang，邻接忠实性与保守因果推断，载于第22届人工智能不确定性大会论文集，UAI。（AUAI出版社，美国俄勒冈州），第401–408页（2006年）。21. AJ Sedgewick 等，用于整合性因果分析的混合图模型：在慢性肺病诊断和预后中的应用。《生物信息学》35，1204–1212（2018年）。22. M Tsagris, G Borboudakis, V Lagani, I Tsamardinos，基于混合数据的因果发现。《国际数据科学与分析杂志》6，19–30（2018年）。23. N Harbeck 等，乳腺癌。《自然综述：疾病概述》5，6（2019年）。24. AM Alaa, D Gurdasani, AL Harris, J Rashbass, M van der Schaar，利用机器学习指导乳腺癌辅助治疗的使用。《自然·机器智能》3，716–726（2021年）。25. C Lee 等，应用新型机器学习框架，利用监测、流行病学和最终结果数据库（SEER）预测男性非转移性前列腺癌特异性死亡率。《柳叶刀数字健康》3，e158–e165（2021年）。26. G Mendiratta 等，美国人口的癌症基因突变频率。《自然通讯》12，5961（2021年）。27. HG Welch, PC Prorok, AJ O’Malley, BS Kramer，乳腺癌肿瘤大小、过度诊断与乳腺X线筛查的有效性。《新英格兰医学杂志》375，1438–1447（2016年）。28. MS Leapman 等，前列腺癌患者中前列腺磁共振成像使用方面种族差异的中介因素。《JAMA肿瘤学》，2022年3月3日在线发表。29. LC Petito 等，接受不同治疗方案的癌症患者总生存期估计。《JAMA网络开放》3，e200452（2020年）。30. RC Nethery, Y Yang, AJ Brown, F Dominici，利用公开数据开展癌症聚集调查的因果推断框架。《英国皇家统计学会杂志A辑（统计学会）》183，1253–1272（2020年）。31. L Wang，基于贝叶斯分析挖掘临床变量之间的因果关系以进行癌症诊断。《生物数据挖掘》8，13（2015年）。32. HS Park, S Lloyd, RH Decker, LD Wilson, JB Yu，监测、流行病学和最终结果数据库的局限性与偏倚。《癌症当前问题》36，216–224（2012年）。33. R Jagsi 等，监测、流行病学和最终结果注册数据中放射治疗接受情况的漏报。《癌症》118，333–341（2011年）。34. SY Chen 等，早期乳腺癌保乳手术后化学治疗和放射治疗的时机：一项回顾性分析。《肿瘤学前沿》10，571390（2020年）。35. JR Anderson, KC Cain, RD Gelber，根据肿瘤反应分析生存。《临床肿瘤学杂志》1，710–719（1983年）。36. MA Hernán, JM Robins，随机试验不可用时利用大数据模拟目标试验：表1。《美国流行病学杂志》183，758–764（2016年）。37. X Han, KR Yabroff, E Ward, OW Brawley, A Jemal，《患者保护与平价医疗法案》实施前后新诊断癌症患者保险状态和诊断分期的比较。《JAMA肿瘤学》4，1713（2018年）。38. T Ermer 等，了解Medicaid扩展对美国癌症医疗的影响。《JAMA肿瘤学》8，139（2022年）。39. L Hinyard, LS Wirth, JM Clancy, T Schwartz，婚姻状况对65岁以下女性乳腺癌相关结局的影响：一项SEER数据库分析。《乳腺》32，13–17（2017年）。40. Z Zhai 等，婚姻状况对不同年龄、种族和激素受体状态患者乳腺癌生存的影响：一项基于人群的研究。《癌症医学》8，4906–4917（2019年）。41. J Bonéy-Montoya, YS Ziegler, CD Curtis, JA Montoya, AM Nardulli，孕酮受体基因表达的远距离转录调控。《分子内分泌学》24，346–358（2010年）。42. C Fisher 等，乳腺癌组织病理学与年龄的关系。《英国癌症杂志》75，593–596（1997年）。43. R Chetty 等，2001—2014年美国收入与预期寿命之间的关联。《JAMA》315，1750（2016年）。44. C Desterke 等，推断骨髓造血干细胞支持性基质生态位细胞群中的基因网络。《iScience》23，101222（2020年）。45. N Sella 等，从大型乳腺癌队列交互式探索全球临床网络。《NPJ数字医学》5，113（2022年）。46. S Affeldt, L Verny, H Isambert，3off2：一种基于二点和三点信息统计量的网络重建算法。《BMC生物信息学》17，12（2016年）。47. F Viger, M Latapy，根据给定度序列高效、简单地生成随机简单连通图，载于《计算机科学讲义》。（施普林格柏林海德堡），第440–449页（2005年）。48. D Colombo, MH Maathuis，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》15，3741–3782（2014年）。49. M Scutari，使用bnlearn R软件包学习贝叶斯网络。《统计软件杂志》35，1–22（2010年）。

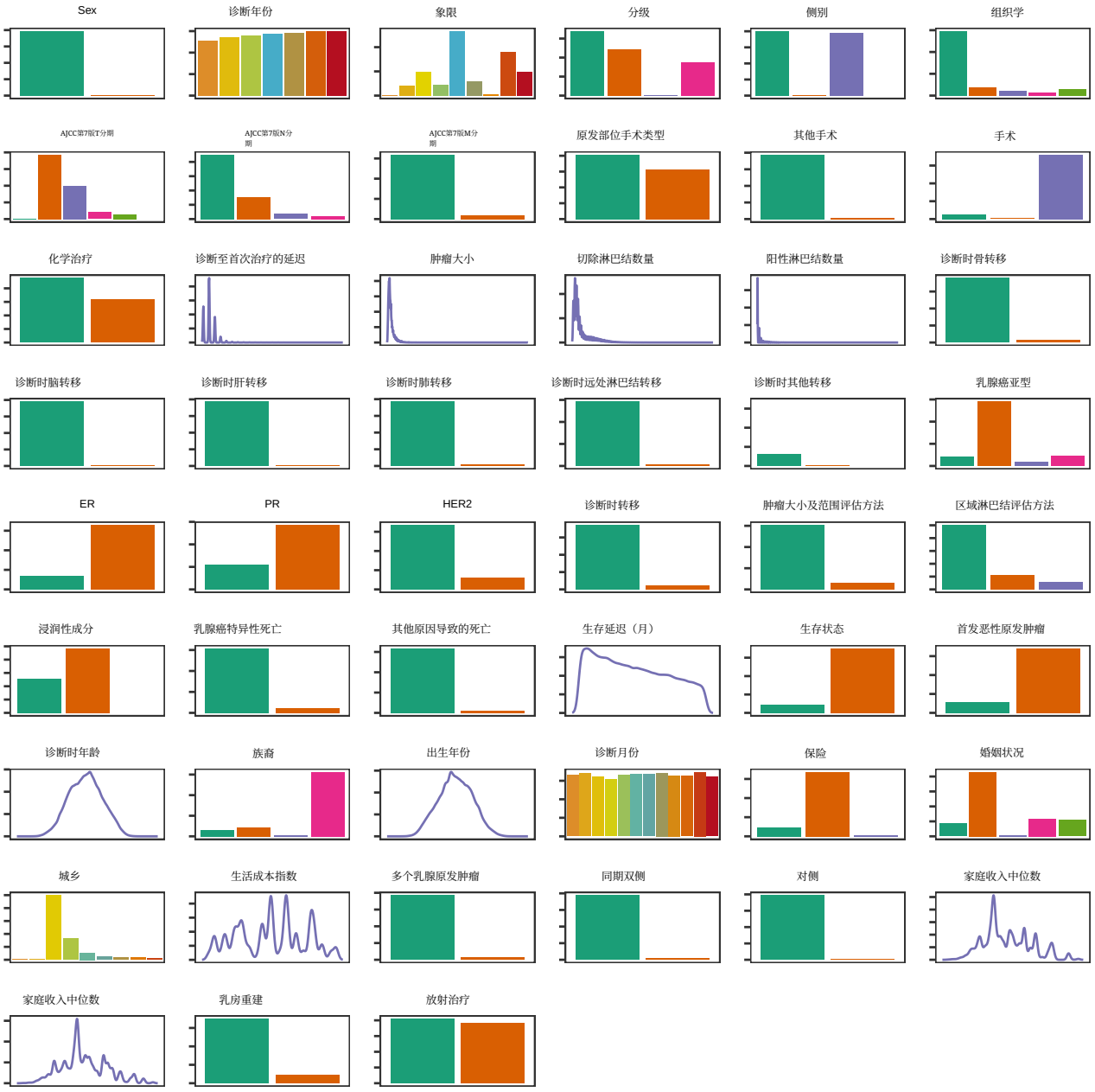


图 S1. 为乳腺癌选取的 51 个 SEER 变量的分布。2010—2016年期间，SEER 中共有 407,791 条乳腺癌记录，但由于部分患者有多个乳腺原发肿瘤，因此仅涉及 396,179 名不同患者。对于每位患者，我们选取 SEER 中记录的首个乳腺原发肿瘤，并在变量 多个乳腺癌原发灶中标明其在 2010—2016年期间乳腺癌原发灶的总数。我们还构建了 同步双侧 变量，用于识别两侧乳房均有肿瘤且诊断时间间隔不足 180天的患者；而 对侧 则用于识别首个乳腺原发肿瘤确诊超过 180天后，对侧乳房出现后续肿瘤的患者。部分类别变量合并了某些类别，原因是这些类别含义大致相同，或在患者中出现频率过低（即 <0.1% 的患者，不包括该变量存在缺失数据的患者）。这些变量包括 族裔、原发部位手术类型、手术、其他手术、诊断时的其他转移、保险和组织学。因此，患者占比低于 0 49, 531 43 59, 63 69, 73 75 . % 的类别被合并，并重新命名为“其他”。乳房重建 是根据 原发部位手术类型 构建的（即 SEER 手术编码范围 - - - 和 - 被归为 “ 是 ”，而其他手术编码则被归为 “ 否 ” ）。放射治疗 由 放射治疗与手术的顺序生成，该变量的缺失数据（0.05%）远少于原始 放射治疗 变量（49%）。肿瘤大小 合并了两个不同的变量，分别包含 2004—2015年和 2016+ 的肿瘤大小。同样，诊断时转移 变量中 2016年的大量缺失信息，根据特定转移变量所含信息进行了补全（即诊断时骨转移、诊断时肺转移、诊断时肝转移、诊断时脑转移、诊断时的其他转移）。最后，家庭收入中位数 和 家庭收入中位数 是这些连续变量在 2007—2011年、2008—2012年、2009—2013年、2010—2014年、2011—2015年和 2012—2016年各期间的平均值。

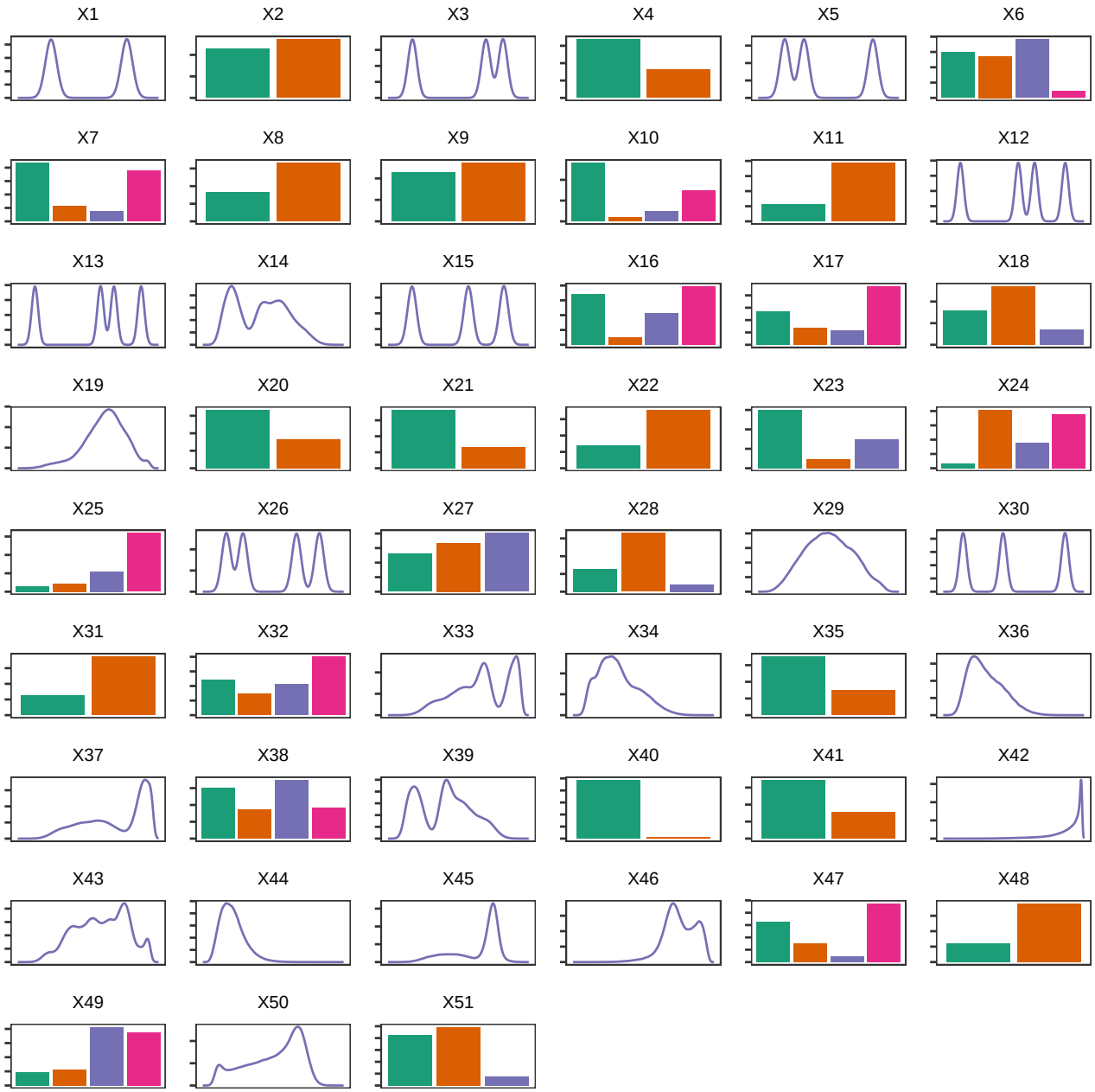


图 S2. 模拟的类 SEER 数据集示例。 使用混合类型结构方程模型（结构方程模型）获得的模拟类 SEER 数据集的边缘分布示例（此处离散变量约占 60%）（10），参见材料与方法中的数据生成与基准测试。

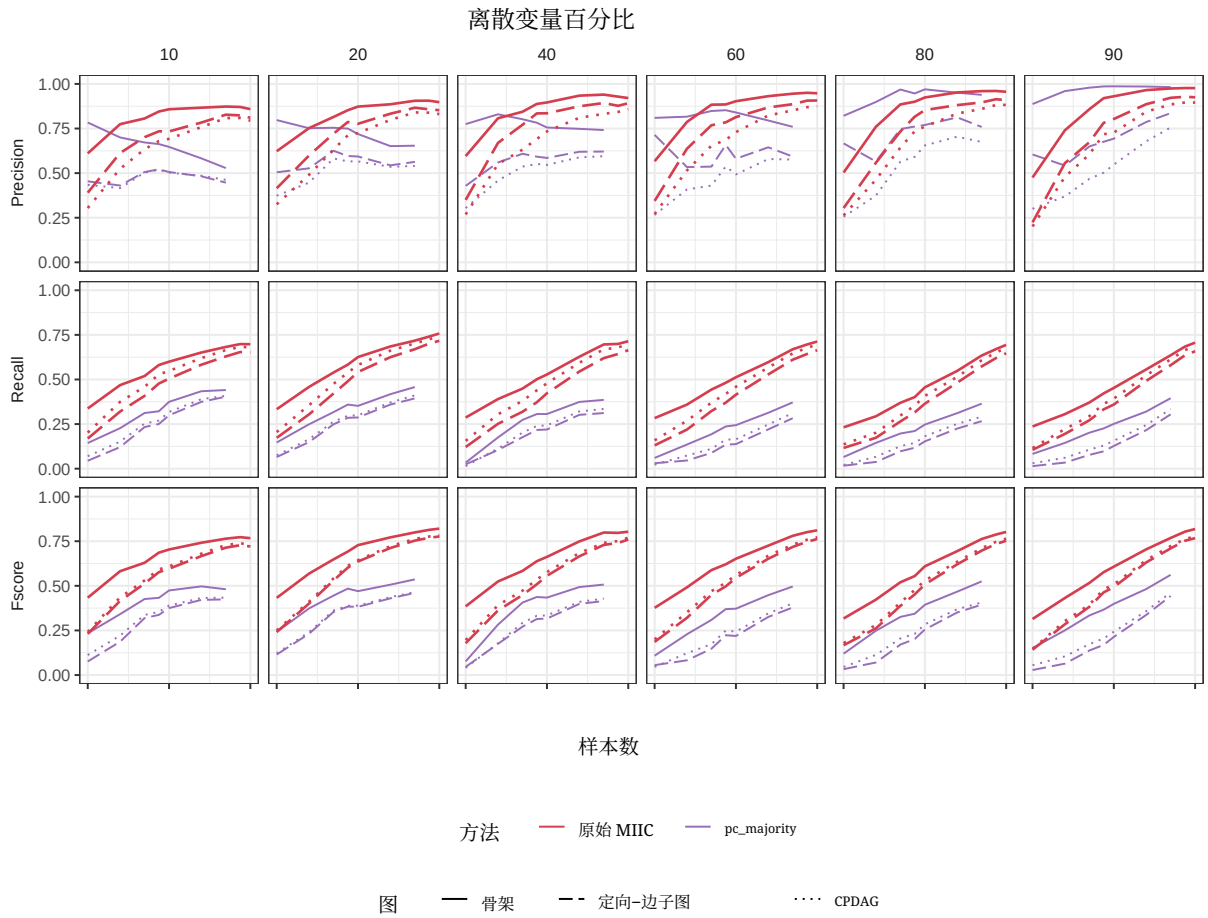


图 S3. 原始 MIIC 与 PC 在类似 SEER 的基准测试上的比较。 参数设置请参见《材料与方法》中的“数据生成与基准测试”。定向边子图得分（虚线）仅针对理论 CPDAG 与 推断图中仅包含定向边的子图进行计算。这些定向边得分旨在专门评估该方法在因果发现方面的表现，即评估那些原则上可从观察性数据中学习到的定向边与 因果结构学习方法实际预测出的定向边之间的情况。

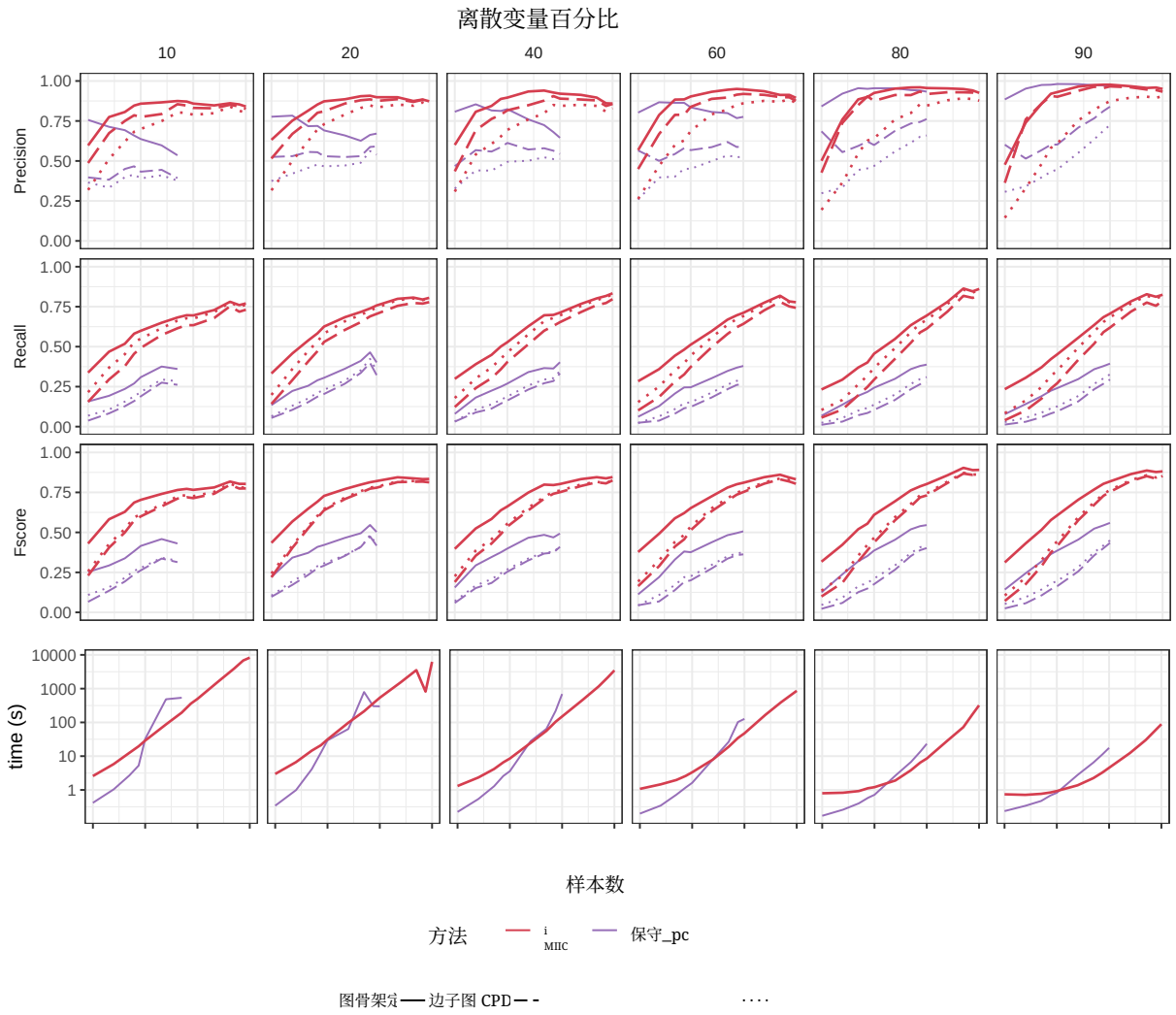


图 S4. iMIIC 与 PC 在类似 SEER 的基准测试中的比较。 参数设置请参见材料与方法中的数据生成与基准测试。定向边子图得分（虚线）仅针对理论 CPDAG 与推断图中仅包含定向边的子图进行计算。这些定向边评分旨在专门评估该方法在因果发现方面的表现，即评估原则上可从观察性数据中学习到的定向边与因果结构学习方法实际预测出的定向边之间的差异。

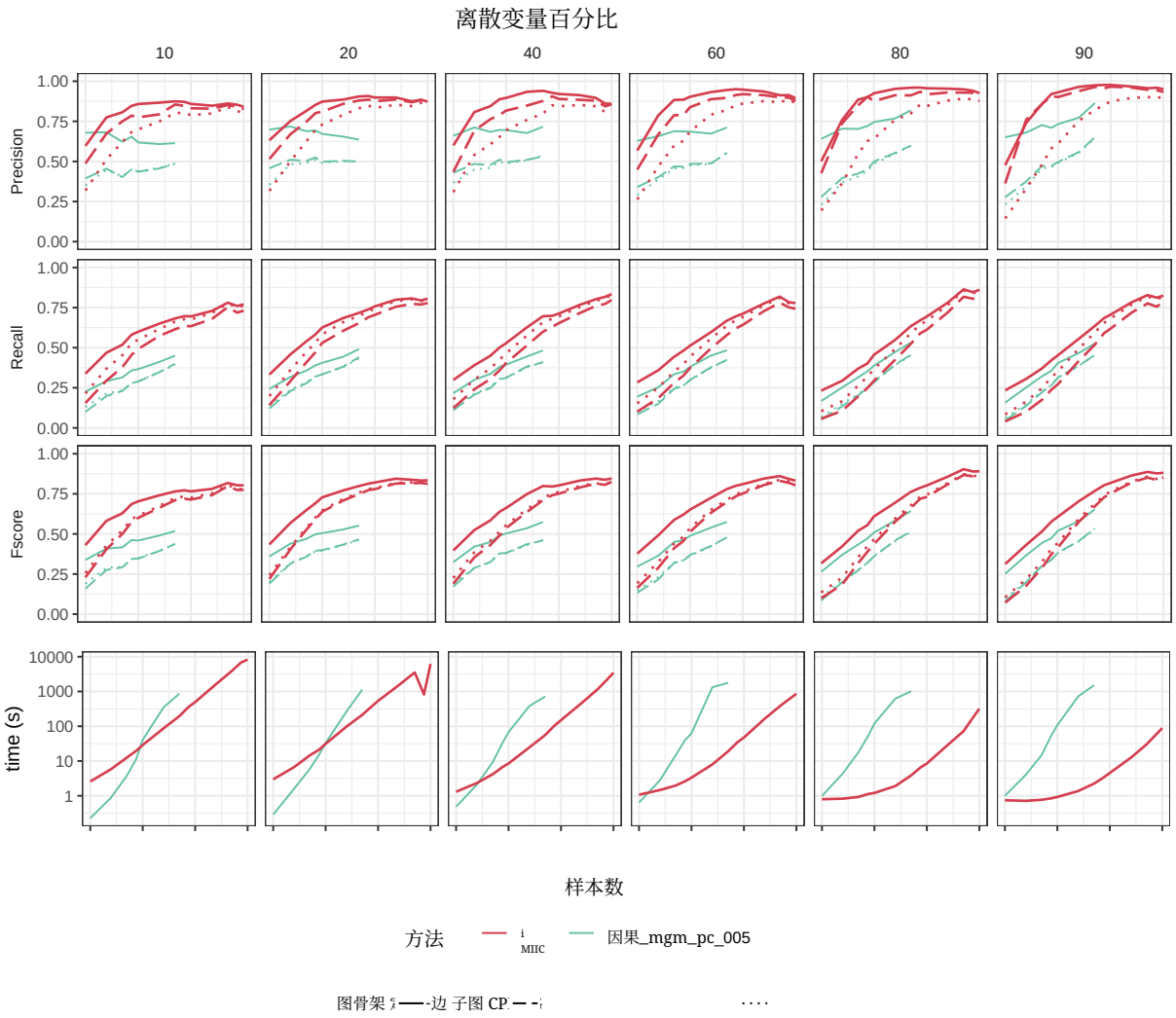


图 S5. iMIIC 与 causalMGM 在类似 SEER 的基准测试中的比较。 参数设置见《材料与方法》中的“数据生成与基准测试”。定向边子图得分（虚线）仅针对理论 CPDAG 与推断图中仅包含定向边的子图进行计算。这些定向边评分旨在专门评估该方法在因果发现方面的表现，即评估原则上可从观察性数据中学习的定向边，以及因果结构学习方法实际预测出的定向边。</output>

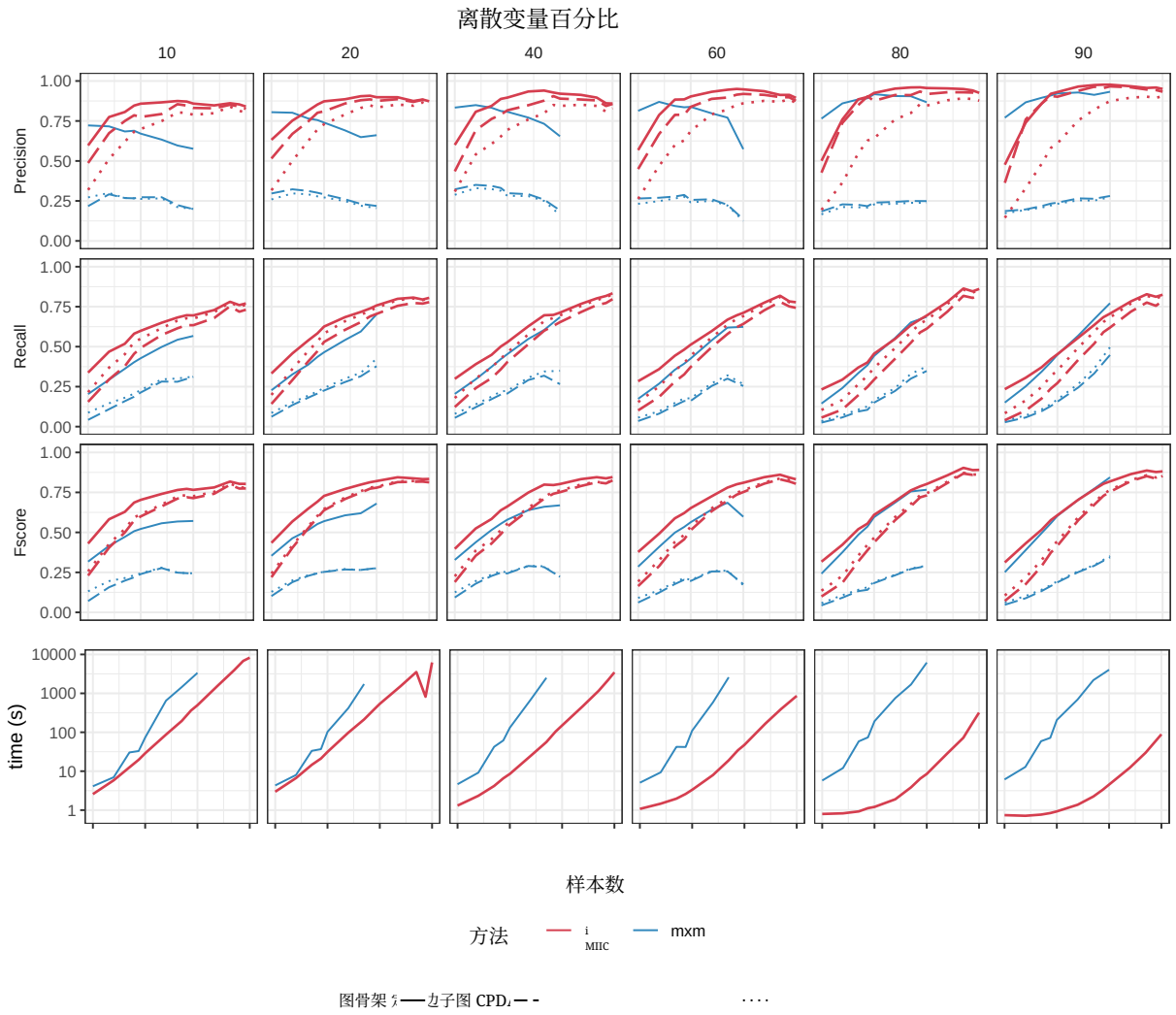


图 S6. iMIIC 与 MXM 在类似 SEER 的基准测试中的比较。 参数设置请参见《材料与方法》中的"数据生成与基准测试"。定向边子图得分（虚线）仅针对理论 CPDAG 与推断图中仅包含定向边的子图进行计算。这些定向边得分旨在专门评估该方法在因果发现方面的表现，也就是评估其对原则上可从观察性数据中学习到的定向边的预测表现，与因果结构学习方法实际预测出的定向边进行比较。

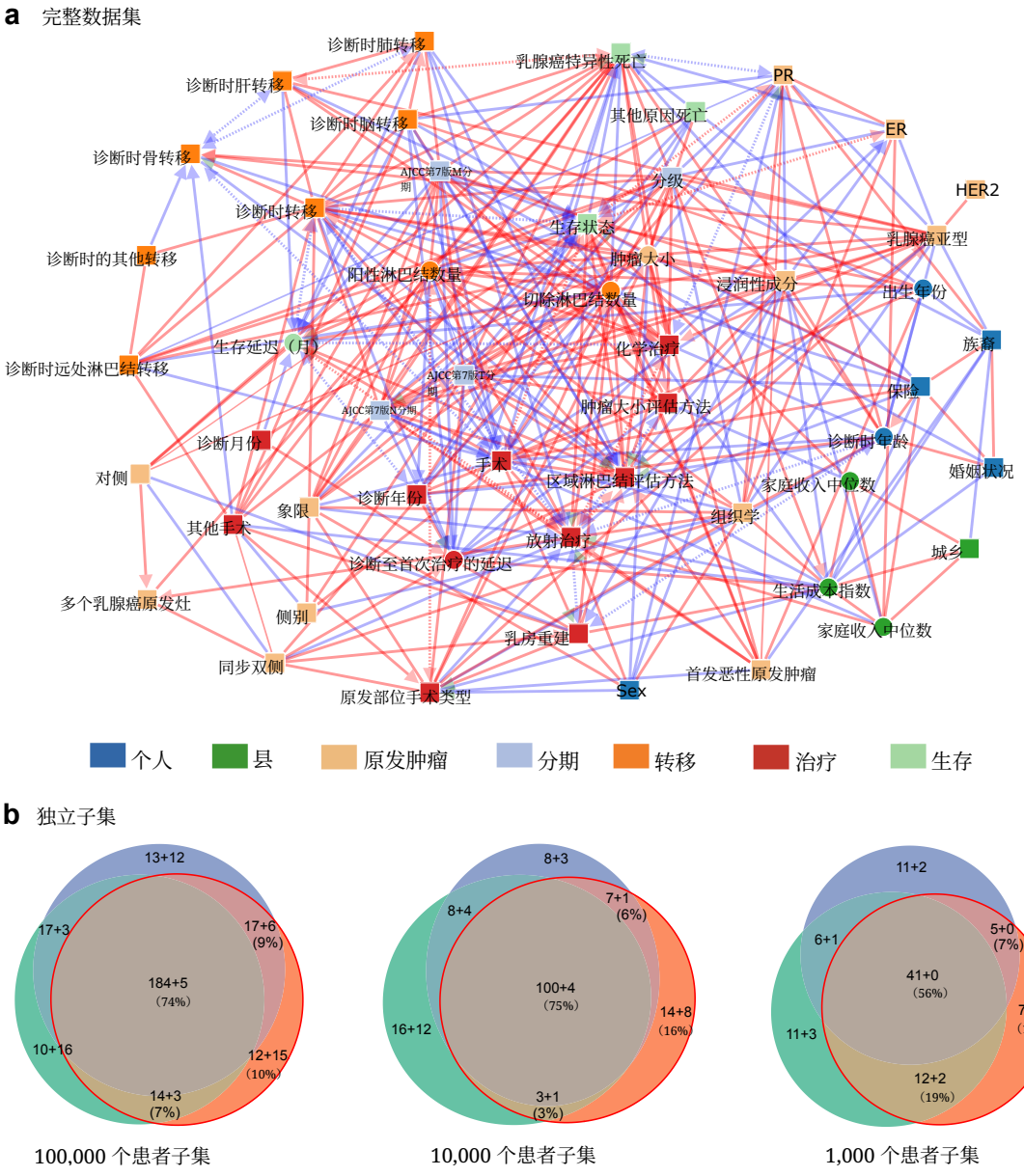


图 S7. 通过 iMIIC 推断的 SEER 乳腺癌方向一致网络。(a) iMIIC 根据 SEER 数据集推断出的 51节点网络，该数据集包含 396,179 名在 2010年 至 2016年 期间确诊的乳腺癌患者。这个方向一致网络包含 340 条边和 2 个背景变量：性别和出生年份。iMIIC 预测的每条边的列表及其因果性质见 数据集S1 。 **(b)** 比较从三个独立子抽样中推断出的网络；每个子抽样的患者子集大小相同，分别为 100,000、10,000 或 1,000（从左到右）。欧拉图中共享边的数量（无论定向如何）以总和表示 $a + b$ 其中 a （分别 b）表示完整数据集网络 (a) 中包含（分别为不包含）的边数。百分比对应于总边数居中的子集网络（红色圆圈）。

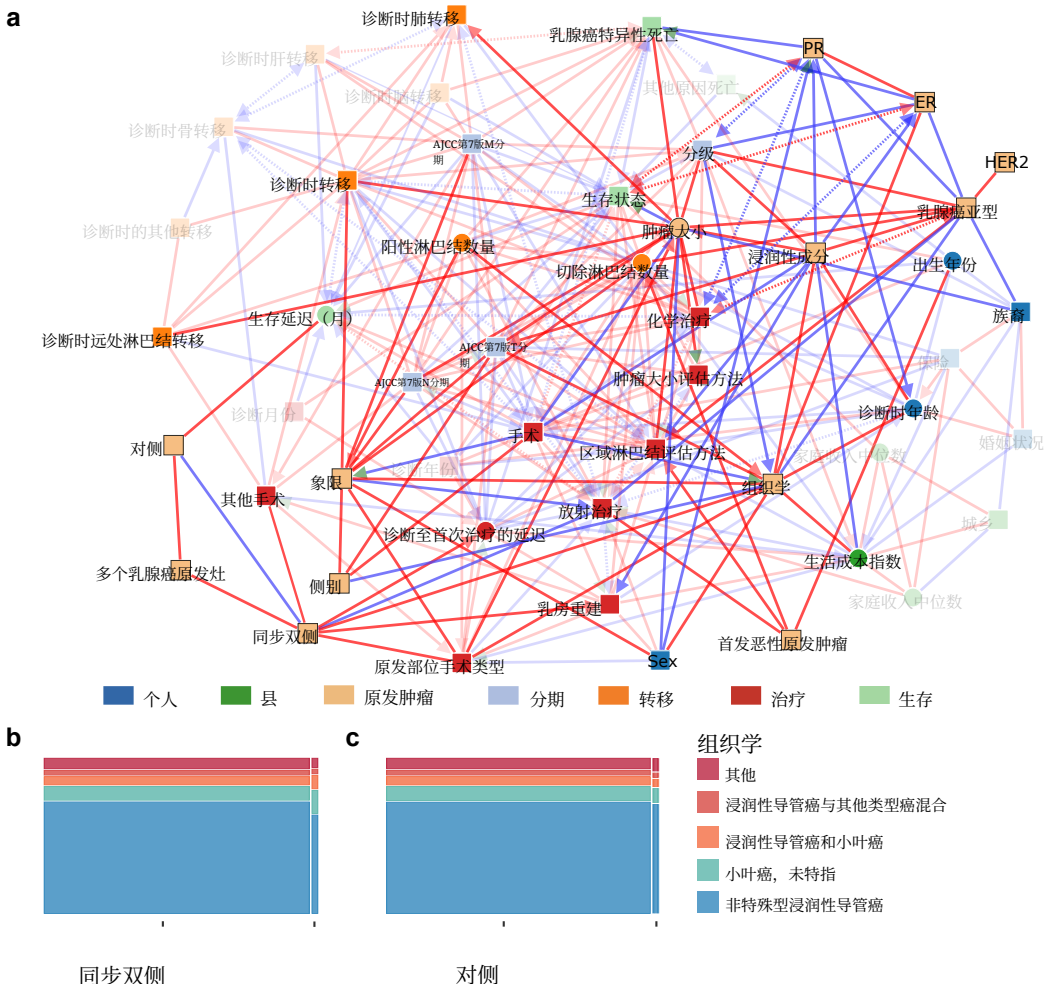


图 S8. iMIIC 根据 SEER 乳腺癌数据集推断的原发肿瘤子网络。（**a**）突出显示与原发肿瘤变量（对侧肿瘤、多个乳腺癌原发灶、同步双侧、侧别、象限、组织学、首发恶性原发肿瘤、肿瘤大小、浸润性成分、孕激素受体、雌激素受体、人表皮生长因子受体2和乳腺癌亚型）直接关联的子网络。（**b**）组织学与同期双侧肿瘤的联合分布。（**c**）组织学与对侧肿瘤的联合分布，参见正文。