arXiv:2303.07342v1 [stat.AP] 13 Mar 2023

新发疾病中的观察性因果推断：新冠肺炎案例研究

Alexander Peysakhovich MetaAI Research

Yin Aphinyanaphongs纽约大学朗格尼医学中心

摘要

所有观察性因果推断都面临一个关键问题：它依赖于一个无法验证的假设，即已观察到的特征足以作为治疗混杂因素的替代指标。本文认为，在医学案例中，尚未制定标准化治疗指南的情况下，这一假设更有可能成立。新发疾病的出现就是此类情况的一个例子。我们研究了纽约市医院早期新冠肺炎病例，并表明，对两种重要疗法——抗凝治疗和类固醇治疗——进行观察性分析所得的结果，与后来通过随机试验及其他证据相结合而制定的指南一致。我们还通过一个警示案例指出，观察性因果推断不能机械地应用，分析人员需要具备领域专业知识：数据中有一种疗法看似极有前景，但这一结果实际上是由医院政策中的一个特殊情况造成的。

1 引言

评估疗法的有效性是医学中的一项首要问题。黄金标准是采用随机试验（Concato 等人，2000）。然而，随机试验可能耗资高昂且耗时较长。第二种证据来源是利用非随机数据（例如，在医院正常运营期间收集的数据）进行观察性因果推断。

观察性因果推断面临的最大问题是，患者治疗分配并非随机的——具体而言，治疗分配与结局之间存在相关性，也就是受到混杂影响（Angrist 和 Pischke，2008；Imbens 和 Rubin，2015）。因此，与随机试验不同，治疗组和对照组之间出现正向（或负向）差异，并不一定意味着这种差异是由治疗造成的。

大量观察性因果推断方法（例如，基于倾向评分的方法（Rosenbaum 和 Rubin，1983）、基于模型调整的方法（Van der Laan 和 Rose，2011），或双重稳健方法（Bang 和 Robins，2005））试图利用患者的可观测特征消除混杂，以解决这一问题。本质上，这会构造出一个合成治疗组和对照组，此时治疗分配便“近似随机”。
所有这些方法背后的关键假设是，观测到的特征足以作为混杂因素的替代指标（Imbens and Rubin，2015）。然而，按定义，这一假设无法验证，因此，一般而言，通过观察性因果推断得出的效应是否可信，取决于“消除混杂”是否可能充分（或不充分）。

在医学领域，有些病症有根深蒂固的治疗方案。在这种情况下，我们往往无法将接受治疗的患者与“相似”的对照明确匹配起来。这是因为，符合治疗条件的患者如果没有接受治疗，通常是出于某种特定原因，使其成为“特殊”患者（或许是由于无法观测到的原因），因此未必能与接受治疗的患者良好匹配。

预印本。

相比之下，有时治疗方案并没有根深蒂固。例如，对于信息日新月异的新发疾病，医生对这类疾病经验甚少，而且不同医生可能会对病情相似的患者采取不同的措施。在这种情况下，我们认为观察性因果推断可能极为有用，因为接受治疗组与对照组的分布会有相当大的重叠。

我们研究了纽约市新冠肺炎早期暴发的案例。在这一阶段，患者病情恶化的两大主要原因是血栓事件（凝血）（Cantador 等，2020）和炎症（Shang 等，2020）。我们运用观察性因果推断，考察两种疗法的疗效——强化抗凝治疗和使用类固醇。我们发现，与预防性抗凝相比，普遍采用治疗性抗凝似乎没有明显益处。相比之下，类固醇治疗确实具有积极作用。这正是随后一年新冠肺炎治疗中，随机试验所确立的指导意见（RECOVERYGroup，2021年；Sadeghipour 等，2021年；REMAP-CAP 及研究者，2021年）。

对于去混杂假设是否成立，目前没有统计检验方法（Angrist 和 Pischke，2008）。因此，观察性因果推断要求分析人员深入了解数据生成过程。我们展示了一个警示性案例：一种疗法——Xa 因子抑制剂。即使在调整了其他已观测因素后，接受这种疗法的患者似乎仍有更好的结局。然而，这种正相关不太可能是因果关系；相反，医生往往会给那些被认为状况良好、可以出院的患者开具 Xa 因子抑制剂处方，而这些患者通常是在住院的最后几天才开始接受治疗。

综上所述，我们认为，在新发疾病的情境中，观察性因果推断尤其有用，因为缺乏治疗方案会使治疗组和对照组的分布产生必要的重叠。尽管在没有额外假设的情况下，观察性研究结果并不能保证具有因果性，但分析人员对数据生成机制的了解可以使这些假设更为可信。观察性因果推断的结果应作为医生知识的补充，用于帮助制定治疗方案，并提出有待通过成本更高的随机试验验证的潜在假设。

2 数据集

我们使用了一份去标识化的数据集，其中包含2020年3月至5月期间因新冠肺炎入住纽约地区某医院系统的患者。
对于所有患者，我们构建了一组协变量，包括人口统计学特征（年龄、性别、种族/族裔、吸烟状况）、已知合并症（既往心肌梗死、心力衰竭、血管疾病、痴呆、肺部疾病、风湿性疾病、消化性溃疡病、糖尿病、癌症、肝病、艾滋病毒/艾滋病以及综合 Charlson 评分）、门诊用药（患者是否服用镇痛药、抗组胺药、抗关节炎药、抗真菌药、抗生素、抗寄生虫药、抗凝药、降糖药、肿瘤坏死抑制剂、抗肿瘤药、抗帕金森病药、抗血小板药、抗病毒药、心血管药物、避孕药、止咳制剂、中枢神经系统药物、自主神经系统药物、利尿剂、胃肠道药物、激素、免疫抑制剂、肌肉松弛剂、产前维生素、精神治疗药物、镇静剂、戒烟药、甲状腺药物或维生素），以及住院前36小时内采集的首项实验室检查指标（我们采用36小时，是因为能够获取实验室检查结果的时间，而非医嘱时间；关于所用实验室检查指标的更多说明见下文），并汇总住院前24小时内生命体征（心率、血氧饱和度、血压、体温、呼吸频率）的统计数据（均值、最小值、最大值）。

2.1 实验室指标

并非所有患者都有全部实验室指标。为处理这一问题，我们首先剔除缺失值超过 20% 的所有实验室检查指标。作为规则的例外，我们在抗凝分析中保留 D-二聚体（凝血的衡量指标）作为协变量。25 % 的患者缺少这项指标，但我们已有相关知识，临床医生会在抗凝治疗分配时使用该指标。

其中许多实验室检查指标呈现出严重的右偏分布，因此我们采用对数变换使其正态化。此外，由于数据质量问题，我们发现了一些超出生物学合理性的数值。为处理这些问题，我们对每项实验室检查指标在第99百分位数处进行缩尾处理。

我们最终使用的实验室检查指标包括：白蛋白、碱性磷酸酶、丙氨酸氨基转移酶、天冬氨酸氨基转移酶、胆红素、血尿素氮、C反应蛋白、钙、氯化物、肌酐、D-二聚体、嗜酸性粒细胞、铁蛋白、血细胞比容、乳酸脱氢酶、淋巴细胞、血小板体积、单核细胞、中性粒细胞、钾、蛋白质、凝血酶原时间、钠、白细胞。

2.2 结局指标构建

我们将21天无器官支持天数（21OSFD）作为主要关注的结局指标。咨询了在我们医院工作的医生后，我们选择了这一指标，因为它已用于现有的随机试验（例如）或（Abdelhady等人，2021年）。

21天无器官支持天数是指患者无需使用升压药、肾脏替代治疗、高流量氧疗或有创机械通气的天数。住院期间死亡的患者，其该项指标记为 -1 。

我们考察的是在观察期结束前至少21天入院的患者，因此我们的数据中不存在删失。

3 种研究治疗

我们研究了两种治疗：强化抗凝治疗和类固醇。下面我们将介绍每种治疗的医学细节，这些细节会影响我们的一些分析选择。

3.1 抗凝

新冠肺炎患者出现不良结局的一个主要诱因是血栓形成（Bilaloglu等人，2020）。处理凝血问题的标准方案是使用抗凝治疗（AC）。抗凝治疗通常分为两种剂量：较低的预防剂量，用于预防血栓形成；较高的治疗剂量，通常在检测到血栓后使用（Lloyd等人，2008）。

收集这些数据时，关于新冠肺炎患者血栓事件的发生率是否足以支持采取更积极的抗凝治疗策略，存在很大争议。我们在此评估采用这种强化策略的效果。请注意，我们医院几乎所有患者住院期间都会接受某种形式的抗凝治疗——在我们的数据中，92%的患者接受了某种抗凝治疗，首次接受抗凝治疗的中位时间为8小时，因此我们并未评估“完全不进行抗凝治疗”的组别。

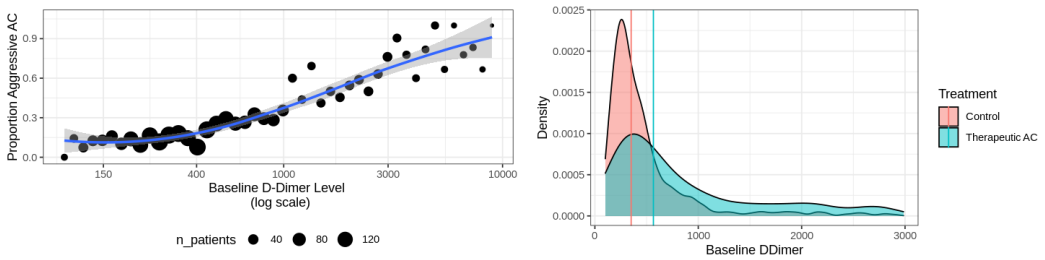
我们重点考察两种可能的策略：尽早强化抗凝治疗，或采取更保守的策略，即先给予预防剂量，之后如有需要再提高剂量。我们将治疗定义为患者住院前72小时内接受治疗剂量（治疗组）或仅接受预防剂量（对照组）的抗凝治疗。我们剔除了在此期间接受不止一种抗凝剂量的患者。[1]

治疗性抗凝治疗的定义包括所有静脉给药的肝素、利伐沙班、华法林、达比加群或大剂量（$\geq$60mg）依诺肝素。预防性抗凝治疗包括皮下注射肝素，每24小时用量为$\leq$ 15,000单位，或小剂量（低于60mg）依诺肝素。我们的样本中，绝大多数抗凝治疗使用的是肝素。

通过与临床医生交流，我们了解到，在此期间，D-二聚体被用作可能出现严重凝血的指标。我们的数据表明，几乎所有D-二聚体水平极高（¿3000 ng/mL）的患者都接受了强化抗凝治疗。因此，我们将分析范围限定为D-二聚体水平低于3000的患者。

总体而言，我们的样本中有23%接受了强化治疗。下方的图1显示了按基线D-二聚体水平划分的强化治疗观察概率，以及治疗组内D-二聚体的分布。可以看出，在“升高”类别中存在大量重叠，而该类别也涵盖了大多数患者。

[1]经与医生讨论后，我们选择了72小时窗口。将窗口改为24、48、72或96小时，并不会改变主要结果。需要权衡的是，使用更长的窗口会减少我们的样本量，因为我们会剔除在研究窗口内接受多种抗凝治疗方案的患者。



(a) D-二聚体预测治疗　　(b) 不同治疗下的 D-二聚体分布

图1：正如预期，D-二聚体水平较高的患者接受强化抗凝治疗的可能性要大得多（图1a）。去除极高水平的数据（≥ 3000）后，我们可以看到不同治疗组的 D-二聚体水平有很大重叠。

3.2 类固醇治疗

新冠肺炎患者出现不良结局的另一项主要驱动因素是炎症介导的肺损伤。治疗炎症的一种方法是使用糖皮质激素（类固醇）。在我们的分析中，如果患者在入院后的前72小时内接受了类固醇（地塞米松、氢化可的松或泼尼松）治疗，则判定其接受了类固醇治疗。

在此，我们的完整病例分析纳入了 2282 名患者，其中 190 名（约 8%）接受了类固醇治疗。与抗凝治疗分析不同，我们没有根据实验室检查结果设置排除标准，因为我们并未发现任何一项实验室检查被所有临床医生采用。相反，多种因素综合作用，使得部分（但非所有）医生决定使用类固醇治疗。

2020年新冠肺炎疫情在纽约市暴发时，主流观点并不一致（Shang 等，2020；Russell 等人，2020），随机试验（RECOVERYGroup，2021年）则直到很久以后才出现。尽管随机试验之后类固醇已被广泛使用，但在我们的样本中，只有 $\sim 8\%$ 的患者在住院最初 72 小时内接受了类固醇治疗。

4 结果

对于这两种处理，我们进行了 3 项分析：第一，考察处理与结局之间的总体（未调整）相关性。第二，通过线性回归调整观测到的协变量，并报告处理项的系数。第三，进行倾向评分匹配。

为了进行倾向评分匹配，我们对全部协变量进行逻辑回归，以构建倾向评分。由于大多数单位未接受处理，我们按每 1 个处理单位匹配 3 个对照单位，卡钳设为。05。所有分析均使用 R 包 Matching（Sekhon，2008）完成。

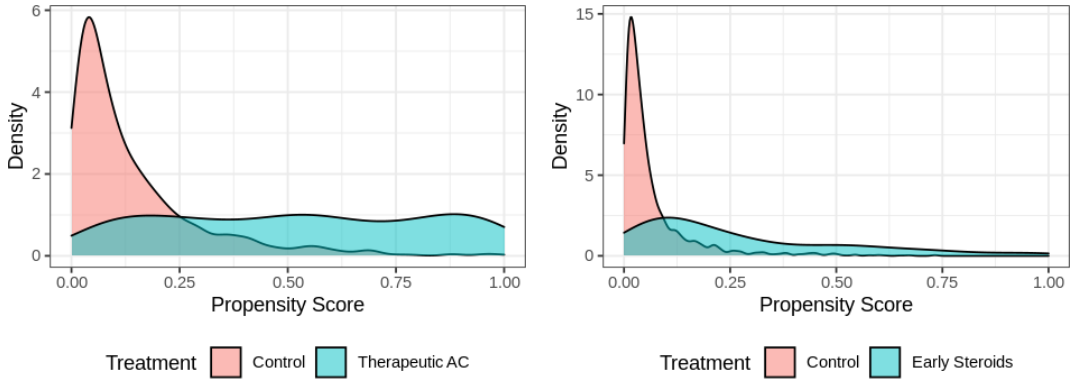
我们在倾向评分匹配中估计处理组平均处理效应（ATT）。近期研究表明，当接受处理的个体比例较低时（如我们的两个案例），观察性数据中的回归所估计的结果接近 ATT（Słoczy´nski，2020），因此我们的估计具有可比性。

4.1 重叠分析

我们通过考察倾向评分的重叠情况，以及匹配前后协变量的平衡情况，来评估匹配分析的质量。由于协变量数量众多，我们只在平衡图中展示“重要”的协变量。我们通过使用 R 软件包 $glmnet$（Hastie 等人，2016）运行两次交叉验证的 L1 正则化回归，分别预测结局和治疗，从而筛选出这些重要协变量。随后，我们取这两个模型所选协变量的并集（请注意，从理论上讲，只有同时影响结局和治疗的变量才是真正的混杂因素，因此这种筛选方式非常保守）。

图2展示了两种治疗的倾向得分分布。我们发现有一些无法匹配的样本（对数据的检查表明，这些样本似乎是病情非常严重的

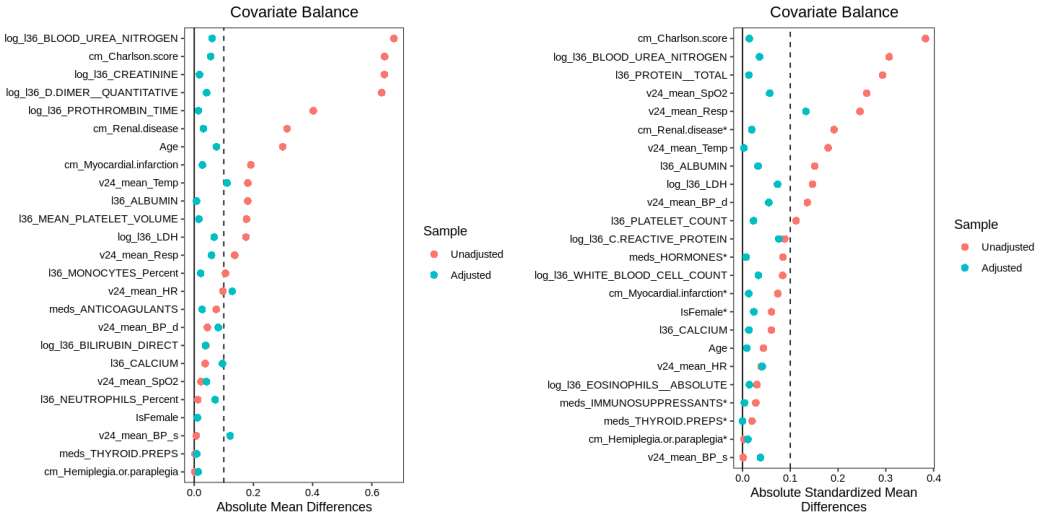
接受强化治疗的患者）。不过，两种分布有很大重叠，因此可以进行匹配。在抗凝治疗分析中，我们的匹配排除了84名（约33%）接受治疗的患者；在类固醇分析中，则排除了44名（~ 23%）接受治疗的患者。



(a) 强化抗凝治疗　　(b) 类固醇使用

图2：数据集中倾向得分分布。使用大小为.05.的卡钳时，我们能够匹配66%的接受强化抗凝治疗的患者和77%的接受类固醇治疗的患者。

图3展示了上述定义的“重要”变量在两种治疗下的平衡图[2] 。我们看到，倾向评分匹配使两种治疗组的均值分布相对可比。



(a) 强化抗凝治疗　　(b) 类固醇使用

图3：分析的平衡图。倾向评分使用所有实验室检查指标和生命体征进行训练。为节省篇幅，平衡图仅显示重要特征，这些特征由用于预测治疗和结局的 Lasso线性模型筛选得出。以 $v24$ 开头的特征为生命体征，其中 BP 表示血压，HR 表示心率。以 $l36$ 开头的特征为实验室指标。以cm 开头的特征为合并症。以 $meds$ 开头的特征为门诊用药。

[2]我们使用 R cobalt 软件包（https://cran.r-project.org/web/packages/cobalt/ index.html）生成了这些图。

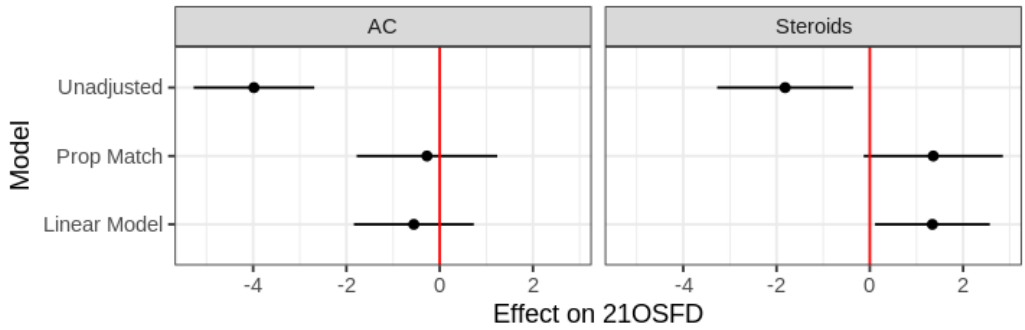


图 4：在抗凝治疗和类固醇治疗中，病情较重的患者都更有可能接受相应治疗。然而，经过调整或匹配后，我们发现抗凝治疗与结局之间没有明显关联，而类固醇治疗与结局之间存在正相关。误差线表示 95% 置信区间。

4.2 因果效应估计

最后，我们在图4中展示了因果效应估计。可以看到，在未进行平衡或调整的情况下，两种治疗与结局之间都存在基线负相关。因此，基线病情较重的患者更有可能接受这两种治疗。

然而，经过调整或匹配后，我们发现类固醇治疗与结局之间存在正相关（回归点估计为 1.34，95% 置信区间为 $= [.11, .257]$ $p < .05$；匹配点估计为 1.36，95% 置信区间为 $= [-.13, 2.85]$ $p = .07$）。相比之下，强化抗凝治疗的效果很小，点估计接近 0（回归点估计为 $-.55$，95% 置信区间为 $=[-1.82, .75]$；匹配点估计为 $-.27$，95% 置信区间为 $= [-1.8.8])$ 。

4.3 外部验证

我们的观察性研究结果基于大流行初期的数据，与随着大流行发展而演变的治疗政策 www.covid19treatmentguidelines.nih.gov/therapies/ antithrombotic-therapy/ 相一致；随机试验数据也显示，与标准护理相比，开始积极的治疗剂量抗凝治疗并无获益（Sadeghipour 等，2021）。其他观察性研究提示治疗效果存在异质性（Paranjpe 等人，2020）。我们的统计功效不足，无法在研究人群中检验主效应以外的任何效应（请记住一条经验法则：与检验主效应相比，检验一个异质性切分需要约 $16\times$ 倍的数据（Gelman，2018））。

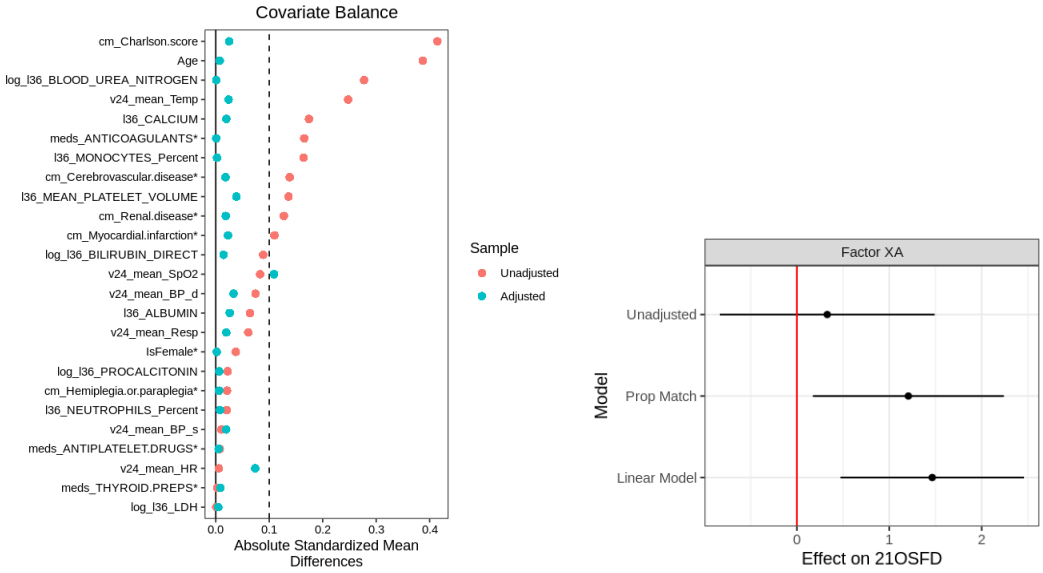
相比之下，随机试验已显示，类固醇（尤其是地塞米松）可有效降低新冠肺炎患者的不良事件（RECOVERYGroup，2021）；另有观察性研究考察了效果的异质性（Lengerich 等人，2021）。类固醇治疗是临床医生治疗新冠肺炎患者的标准手段 https://www.covid19treatmentguidelines.nih.gov/management/clinical-management/hospitalized-adults--therapeutic-management/。

5 结论与注意事项

虽然前文似乎暗示，在治疗方案尚未固定的情况下，观察性因果推断相对直接，但事实远非如此。本文将论证，因果推断并非单纯的统计练习，而是要求分析者了解底层的数据生成过程。

我们还考虑另一类抗凝药物，即 Xa 因子抑制剂（FXI）。这些药物与我们前面考虑的主要基于肝素的抗凝治疗有两点不同：首先，它们影响

凝血通路中不同于肝素作用的部分（Ansell，2007年）；其次，它们有口服剂型，而我们此前考虑的大多数抗凝治疗药物要么通过皮下注射给药，要么通过静脉给药。
我们采用一种简单的治疗定义：如果患者在医院住院期间接受过任何 Xa 因子抑制剂（$N_{control} = 1963$），就将其定义为接受了治疗（$N_{treated} = 318$）。
我们按照主要分析中的做法进行了相同的分析（调整后的线性模型、倾向评分匹配）。图5展示了平衡图（与上文一样，我们仅展示通过 Lasso回归选出的“重要”变量）以及效应估计值。



（a）倾向评分平衡　　（b）Xa因子分析结果

图5：模型结果似乎表明，Xa因子对患者有很强的保护作用（作用大小与类固醇相当）。然而，这很可能是数据生成过程造成的假象——我们知道，在收集这些数据时，即将出院的患者会被开具 Xa 因子抑制剂作为门诊治疗，并在出院前开始服用。

模型结果似乎表明，FXI 对患者有很强的保护作用（作用大小与类固醇相当）。然而，这很可能是数据生成过程造成的假象——我们知道，在收集这些数据时，即将出院的患者会被开具 Xa 因子抑制剂作为门诊治疗，并在出院前开始服用。请注意，我们得出错误结论的部分原因在于治疗变量的定义方式：我们考察的是曾经 接受过治疗的患者，而不是在住院早期接受治疗的患者（如抗凝治疗/类固醇分析中所做的那样）。

我们确实注意到，结果中似乎有一些迹象表明存在问题。首先，我们看到治疗与结局之间的基线相关性呈弱正相关。这意味着，总体而言，我们没有看到类固醇/抗凝治疗中那种病情越重的患者越可能接受治疗的关系。虽然这或许有合理的解释，但这也提醒我们应保持一定谨慎。其次，我们看到，即使经过匹配，重要协变量仍然存在不平衡。这进一步表明，接受治疗的个体与对照组中的个体存在很大差异。
关键在于，观察性因果推断在医学中仍是一种有用的工具，尤其是在治疗指南尚未确立之时。对于抗凝治疗和类固醇治疗，我们都清晰地得到了随机对照试验所证实的结果。然而，正如我们的 Xa 因子示例所示，要从观察性研究中获取知识，必须谨慎定义治疗变量，并理解数据生成过程。

参考文献

Hesham Abdelhady、Marwa Abdelrazik、Zakee Abdi、David Abdo、Amina Abdulle、Lynn Abel、Shrefee Abouzeenni、Gail Abrahamson、Yousuf Abusamra、Lisa Adams 等。2021年。恢复期血浆对新冠肺炎危重患者无器官支持天数的影响：一项随机临床试验。*JAMA* 326卷，第17期（2021年），1690–1702页。

Joshua D Angrist 和 J¨orn-Steffen Pischke。2008。《大体无害的计量经济学》。普林斯顿大学出版社。

J Ansell。2007。Xa因子还是凝血酶：Xa因子是更好的靶点吗？《血栓与止血杂志》5（2007年），60–64。

Heejung Bang 和 James M Robins。2005。缺失数据与因果推断模型中的双重稳健估计。《生物统计学》61，4（2005），962–973。

Seda Bilaloglu、Yin Aphinyanaphongs、Simon Jones、Eduardo Iturrate、Judith Hochman 和 Jeffrey S Berger。2020。在纽约市医疗卫生系统中新冠肺炎住院患者的血栓形成。《JAMA》324，8（2020），799–801。

Estefan´ía Cantador、Alberto N´ú˜nez、Pilar Sobrino、Victoria Espejo、Luc´ía Fabia、Lydia Vela、Luis de Benito 和 Javier Botas。2020。新冠肺炎患者全身动脉血栓事件的发生率及其后果。《血栓形成与溶栓杂志》50（2020），543–547。

John Concato、Nirav Shah 和 Ralph I Horwitz。2000。随机对照试验、观察性研究与研究设计的层级。《新英格兰医学杂志》342，25（2000），1887–1892。

Andrew Gelman。2018。估计交互作用所需的样本量是估计主效应的16倍。https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Trevor Hastie、Junyang Qian 和 Kenneth Tay。2016。《glmnet简介》。

Guido W Imbens 和 Donald B Rubin。2015。《统计学、社会科学与生物医学科学中的因果推断》。剑桥大学出版社。

Benjamin J Lengerich、Rich Caruana、Alex Peysakhovich、Leora Horwitz 和 Yin Aphinyanaphongs。2021年。中性粒细胞淋巴细胞比值作为新冠肺炎治疗中糖皮质激素疗效的预测因子。*medRxiv*（2021年）。

NS Lloyd, JD Douketis, I Moinuddin,W Lim, 和 MA Crowther。2008年。抗凝预防可预防住院内科患者无症状深静脉血栓形成：一项系统综述和荟萃分析。《血栓与止血杂志》6, 3 (2008), 405–414。

Ishan Paranjpe、Valentin Fuster、Anuradha Lala、Adam J Russak、Benjamin S Glicksberg、Matthew A Levin、Alexander W Charney、Jagat Narula、Zahi A Fayad、Emilia Bagiella 等。2020年。治疗剂量抗凝与新冠肺炎住院患者住院生存率的关联。《美国心脏病学会杂志》76, 1 (2020), 122–124。

RECOVERYGroup。2021年。住院新冠肺炎患者使用地塞米松。《新英格兰医学杂志》384, 8 (2021), 693–704。

ACTIV-4a、REMAP-CAP和ATTACC研究者。2021年。对新冠肺炎危重患者采用肝素进行治疗性抗凝。《新英格兰医学杂志》385, 9 (2021), 777–789。Paul R Rosenbaum和Donald B Rubin。1983年。倾向评分在观察性研究因果效应分析中的核心作用。《生物计量学》70, 1 (1983), 41–55。Clark D Russell、Jonathan E Millar和J Kenneth Baillie。2020年。临床证据不支持使用皮质类固醇治疗2019新型冠状病毒肺损伤。《柳叶刀》395, 10223 (2020), 473–475。

Parham Sadeghipour、Azita H Talasaz、Farid Rashidi、Babak Sharif-Kashani、Mohammad Taghi Beigmohammadi、Mohsen Farrokhpour、Seyed Hashem Sezavar、Pooya Payandemehr、Ali Dabbagh、Keivan Gohari Moghadam等。2021年。中等剂量与标准剂量预防性抗凝对入住重症监护病房的新冠肺炎患者血栓事件、体外膜肺氧合治疗或死亡率的影响：INSPIRATION随机临床试验。*JAMA*325, 16 (2021), 1620–1630。

Jasjeet S Sekhon. 2008. 具有自动平衡优化功能的多变量和倾向评分匹配软件：R匹配程序包。《统计软件杂志》，即将出版(2008)。

Lianhan Shang、JianpingZhao、YiHu、Ronghui Du 和 Bin Cao. 2020. 关于皮质类固醇在2019新型冠状病毒肺炎中的应用。《柳叶刀》395, 10225 (2020), 683–684。

Tymon Słoczy´nski. 2020. 治疗效应存在异质性时如何解读OLS估计目标：较小群体获得更大的权重。《经济学与统计学评论》(2020), 1–27。

Mark J Van der Laan 和 Sherri Rose. 2011. 目标化学习：观察性研究的因果推断 *l**a*和实验数据。Springer Science & Business Media。