arXiv:2304.03243v1 [cs.AI] 6 Apr 2023

医疗保健中的合成数据

Daniel McDuff,[1*] **Theodore Curran,**[1] **Achuta Kadambi**[2]

[1]美国西雅图华盛顿大学[2]美国洛杉矶加利福尼亚大学洛杉矶分校（UCLA）

*联系作者；电子邮箱：dmcduff@uw.edu。

摘要

合成数据正成为构建人工智能系统的重要工具。模拟器能够系统地大规模生成数据。这些数据既可单独用于系统训练和测试，也可与真实数据结合使用。在“真实”训练样本可能成为瓶颈的情况下，合成数据尤其具有吸引力。尽管医疗保健领域的数据量正呈指数级增长，但要为新任务创建数据集，或创建能够反映多样病症和因果关系的数据集，并非易事。此外，这些数据高度敏感，且往往与特定患者相关。近期研究开始展现合成数据在医学诸多领域的潜力，但目前尚无相关文献的系统综述。本文阐述了利用物理仿真和统计模拟创建数据的理由，以及这些数据在医疗保健和医学中的拟议应用。我们指出，尽管合成数据可以 促进隐私保护、公平、安全以及持续学习和因果学习，但也可能引入缺陷和盲点，并传播或加剧偏见。

***K*关键词** 医疗保健、机器学习、合成数据

1 引言

医疗保健中生成的数据量正在呈指数级增长[1]，每年产生艾字节级的患者记录、医学成像、实验室检查、生理测量以及大量其他数据。数据驱动学习方法很可能在未来许多年里成为科学研究和临床研究中最重要的工具之一 [1, 2]。人工智能（AI）算法正日益得到采用，以自动化数据分析，并帮助提高决策的效率和有效性。在医疗保健领域，人工智能已被探索用于各种应用，包括医学影像中的病理检测（例如放射摄影、计算机断层扫描 [计算机断层扫描], 和磁共振成像 [磁共振成像]） [3],心血管疾病的诊断（例如心电图 [心电图]） [4], 使用电子健康记录（EHR） [5, 6], 分析和预测健康结局，以及从医学文献中挖掘信息 [7]。

新的神经网络模型（例如卷积神经网络、长短期模型和 Transformer）显著提升了许多此类任务的表现。这些大参数模型旨在从数据中“端到端”地学习复杂的非线性关系。因此，它们是用于训练的数据观测结果的产物，而数据或许是构建模型时最重要的考量，这一点已被广泛认可。一个例证是新兴的以数据为中心的 人工智能领域，该领域关注通过反复迭代和协作来改进用于构建人工智能系统的数据 [8], ，并强调整理数据集是提升模型效果最有效的方式。

在模型开发流程中，获取、整理和标注这些数据既成本高昂，耗费人力，也面临技术上的挑战。确保这些数据具备适当的多样性，并能反映实践中观察到的真实关系和分布，并非易事。在医疗保健领域，受隐私顾虑、技术限制、成本和缺乏激励等因素影响，数据的收集、管理和共享十分复杂。根据欧盟《通用数据保护条例》（GDPR），患者隐私和适当的数据使用同意是重要的伦理和法律考量。

[1]https://www.statista.co m/statistics/1037970/global-医疗保健-d ata-volume/

以及美国《健康保险流通与责任法案》（HIPAA）。随着法规日益倾向于保护患者及其隐私，要收集足够数量、能够发挥神经网络模型可扩展性优势的带标签样本，也变得越来越困难。

如果数据可以被“生成”，而不是被“收集”，会怎样？合成数据生成是指由计算机模型按需生成样本，如今正成为人工智能模型开发中的一种非常强大的工具。合成数据生成可以采取物理模型（例如计算机模拟器、计算机图形处理管线）、统计模型（例如生成网络），或二者的混合形式（融合模型） [9]。实践者越能精确控制其所创建示例的属性，这些模型就越强大。用于训练人工智能模型的合成数据是一个价值数十亿美元的市场，据估计，最迟到 2030 [10] 用于训练模型的数据中大多数将是合成数据[2]。在包括计算机视觉 [11, 12, 13] 和自然语言处理 [14, 15] 在内的许多机器学习领域，利用仿真扩充现有数据集已经得到了广泛探索，并取得了令人鼓舞的结果。如今，生成模型正越来越多地在医学和医疗保健中的特定应用内接受测试 [16, 17]。在这些应用中，合成数据可以用于扩充真实数据集，以提供补充示例，也可以作为训练系统的唯一数据。支持医疗保健中采用合成数据的理由是：使用合成数据训练的人工智能模型可以与使用真实数据训练的模型相媲美 [18]；更进一步的研究表明，使用合成数据和真实数据组合训练的人工智能模型，其表现优于仅使用真实数据训练的模型 [19, 20]。一些公司已经开始利用合成数据开展商业医疗保健应用[21, 22]；然而，这样做仍然存在许多挑战与风险。

2 合成数据生成

2.1 物理模型

物理模型（见图1A）需要了解如何对数据生成过程进行参数化。计算机图形引擎模拟现实世界的物理学，是创建物理环境、结构和事件的渲染图像的一种方式。例如，要设计一个用于模拟心脏血容量脉搏的模型，就需要对心血管系统的流体力学进行参数化。包含大量参数的精细计算机仿真通常创建成本高、耗时长，而且可能需要大量计算资源。由于需要手动对模型进行参数化，这些仿真必须针对特定任务设计，未必能够泛化；这意味着不同领域的研究人员需要设计各自独立的处理管线。物理模型的优点在于可解释，且其能够和不能够生成的样本具有已知约束；在这方面，它们为创建数据提供了一种潜在的更安全的选择。然而，模型失配可能发生，因为现实世界往往极其复杂，物理模型无法反映可能观测到的全部方差。如果已知存在未知因素，并可通过为模型的使用设置防护措施加以缓解，那么简单模型或许可以接受；然而，未知的未知会对这些模型的实际效用构成重大得多的风险。

2.2 统计模型

统计生成模型能够捕捉数据集的概率表示，并可从中抽取样本。这类模型通常被称为“生成式人工智能”（ChatGPT 就是统计生成模型的一个示例[3]），经过训练后可模仿数据集的分布。生成对抗网络（GANs） [25], 变分自编码器（VAEs） [26], 流模型 [27], 和扩散模型 [28, 29, 30] 都是生成架构的示例。有关用于合成数据的深度生成模型综述，请参阅 Eigenschink 等人 [31]。这些模型都是经过训练的模型，尽管通常采用无监督训练，因此其表现取决于用于创建它们的数据。与物理模型相比，统计模型通常能更忠实地捕捉观测到的数据分布；不过，只有当训练数据具有代表性时，这才是优势。使用这些模型可以模拟具有极高分辨率且内容十分逼真的图像 [32]。如果有可用数据，创建统计模型的复杂程度远低于某些物理模拟器。例如，渲染高度逼真的人体视频需要精心构建、以物理原理为基础的图形模型，而训练生成模型只需少得多的工作，就能生成逼真的视频。统计模型通常远不如物理模型“可控”或“易于理解”，因为每个参数或参数组（例如网络中的一层）并不会根据物理学或任何其他理论基础以合乎逻辑的方式相互作用。例如，对于心脏血容量脉搏，生成模型可能会生成具有十分合理生理特性的波形，但可能很难确定哪些参数对应于前向波的速度或心脏的.....

[2] https://blogs.gartner.com/andrew-white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/ [3] https://openai.com/blog/chatgpt

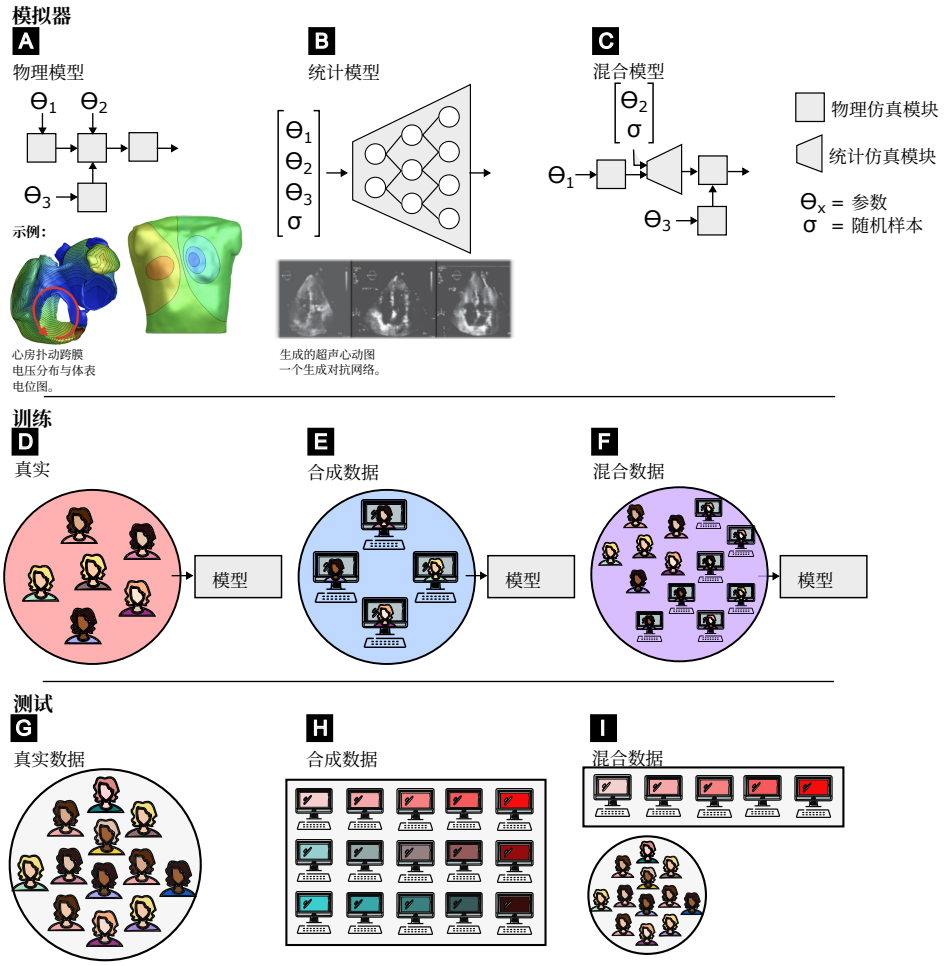


图 1：仿真可以是 A）确定性物理模拟、B）统计生成式模拟，或 C）物理模拟与统计模拟相结合的混合模拟。机器学习模型可以使用来自这些仿真的 D）真实数据、E）合成数据，以及 F）真实数据与合成数据的混合数据进行训练。合成数据可用于测试/性能评估，与 G）使用真实数据进行测试、H）系统生成的合成数据，以及 I）两者结合进行对比。心房扑动电压分布图改编自 Dossel 等人 [23]，超声心动图改编自 Madani 等人。 [24]

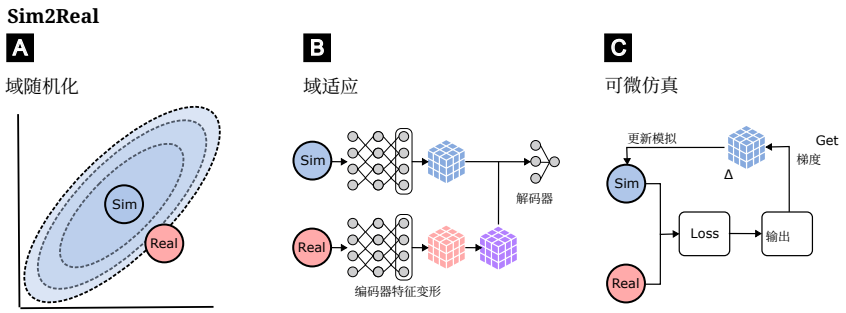


图2：实现 Sim2Real 的方法包括迁移学习方法，例如（A）域随机化或（B）域适应。由于有时很难将模型从仿真迁移到真实数据，另一种策略是通过（C）可微仿真等方法，以物理学与数据相结合的方式提高模拟器的逼真度。

心率变异性。监督学习可用于创建能够沿某些维度进行控制（条件化）的模型；然而，仍需开展更多工作，以确保此类模型生成的样本在医学等应用中的可靠性。

2.3 Sim2Real

理想情况下，使用合成数据训练的模型应能泛化到真实世界条件。当模型在真实领域中的表现与在合成领域中的表现不同时，就存在 Sim2Real 领域差距。为了尽可能缩小这一差距，可以构建更逼真的模拟器，但模拟我们的物理世界并非易事。图2展示了三种缩小领域差距的框架。域随机化会扰动仿真数据的生成过程，从而改变生成数据的分布。如图2a所示，当分布偏移足够大时，现实世界就会落入仿真数据所涵盖的范围内。例如，使用某一型号数码相机的模型生成 RGB图像，可能无法生成与另一型号相机拍摄的图像相匹配的图像；不过，改变像素分布或许可以解决这种差异。图2b展示了另一种称为域适应的技术。在仿真世界中训练的模型，通常会在使用从仿真数据中学习到的特征时表现得更好。域适应旨在对真实世界数据的特征进行变换，使其与模拟器中的特征相匹配。域适应和域随机化都属于迁移学习，即在一个领域的数据上训练模型，再用另一个领域的数据对其进行测试。第三种称为可微仿真的技术采取更直接的方法，试图让模拟器更加逼真。可微仿真旨在构建一个参数可优化的模拟器，进而利用学习来设定模拟器的最优参数。设想一个机器人学习拾取具有复杂反射特性的闪亮物体。初始模拟器或许无法准确模拟光泽外观，但在收集一些图像并观察模拟器与真实数据之间的差异后，就可以学习如何优化模拟器。例如，可以扩展模拟器，使其更有效地模拟反射。在医疗保健领域，任何形式的合成患者都必须能够泛化到真实数据，而 Sim2Real 技术很可能会发挥作用。例如，利用域适应提升基于合成数据训练的 MRI肿瘤分割 [33]和相干断层成像（计算机断层扫描）材料分解模型的性能 [34]。

3 医疗保健中合成数据的理由

完整、易于获取且具有代表性的患者数据，对于临床和科学研究、内部运营评估与质量改进，以及临床诊断和决策支持等潜在的人工智能驱动工具都至关重要。然而，这类数据难以获取；下文将探讨支持在医疗保健领域使用合成数据的主要论据。

数据共享与隐私。 医疗保健领域的数据共享面临诸多障碍，有时甚至令人费解。患者数据主要因隐私顾虑而受到限制，相关限制由 HIPAA 和 GDPR 等法规执行。即使在同一家医疗保健机构内部，由于各项政策理所当然地将隐私放在首位，患者数据的共享也十分繁琐。此外，医疗保健系统历来缺乏共享数据的激励机制，因为获取数据成本高昂，因此数据一直被视为专有资产。近期的州级监管举措为医院和诊所之间增加数据共享提供了经济激励，希望以此降低成本、改善患者护理，但其适用范围似乎仅限于患者护理用途，最终影响尚不明确[35]。同样，医学研究界也缺乏数据共享文化，部分原因是专业激励机制不协调[36], 数据格式针对临床护理而非分析进行了优化；更重要的是，患者隐私顾虑，以及由此产生的以合乎伦理和法律的方式共享数据所需的成本、时间与精力。少数临床患者数据库例外地向研究界开放，但由于在人口统计特征、疾病和数据类型方面存在局限，这些数据库仍远远不够（例如，MIMIC-IV [37], eICU [38],AmsterdamUMCdb [39], HiRID [40], PIC [41]）。

尽管合成数据无法解决共享数据的所有障碍，但它可以克服其中或许最大的障碍——患者隐私。事实证明，当前去标识化患者数据的标准容易受到恶意方法的攻击，例如与其他已知记录进行匹配，从而暴露原始标识符[42, 43]。合成数据提供了一种独特的身份保护方法，因为样本并非从某个个体生成，而是来自多个个体特征的融合 [44]。在其他行业中，保护隐私的数据集使数据更易于由商业和学术实体共享和使用，从而促进协作、扩大影响并实现二次分析 [45]。最近发布的一种用于保护数据隐私的通用（即“与任务无关”）人工智能方法，可能会对医疗保健数据产生影响 [46]。无需直接使用真实数据，即可构建计算机图形模型和其他物理仿真。还可以创建具有可微分隐私保证的统计模型 [47]。即使在个体层面，也可以在不暴露真实样本的情况下创建合成数据，从而构建一个“数字孪生”。最终，由使用真实数据训练的模型创建的合成数据是否属于受保护的健康信息，将留待

向监管机构和机构审查委员会。虽然并非专门针对健康数据，但在 Liu 等人 [48] 和Zhang等人 [49]的研究中可以找到一些被认为能够保护隐私的、用于合成机器学习数据的通用工具。

公平性与公平。 在人工智能领域，“公平性”虽未有明确的定义 [50], ，通常是指在法律和社会规范的背景下，某个模型或数据分布能够准确反映相关群体，且不带有偏见；同样，在医疗保健领域，“公平”通常是指在人类受试者研究或患者护理中，人口统计特征，即年龄、性别、族裔和种族，得到均等分布。必须确保医疗干预不会使部分人群处于不利地位。对于医疗器械而言，确保公平尤其是一个复杂的问题。正如 Kadambi [51]所解释的，医疗器械的公平性不仅仅是人工智能问题，还必须从物理层（传感器如何感知数据）、人工智能层（算法如何处理数据）以及解释层（人类将如何解读输出）进行研究。虽然合成数据并非“快速解决方案”，但它们可以帮助减轻偏见，并在多个层面改善公平性与公平。例如，近期证据表明，脉搏血氧仪在较深肤色上表现不佳 [52]。如果数据集包含在光学上具有挑战性的受试者，人工智能模型就能更加公平。不幸的是，很难收集大量具有极端肤色的真实受试者。在这种情况下，可以生成具有多样表型的合成数据。已有研究表明，在多样化合成数据集上训练的人工智能模型能够改善公平性 [53, 54]。有些违反直觉的是，一个公平的算法通常对每一位利益相关者而言都是更好的算法。不仅纳入少数群体表型能够改善模型对少数群体和整体人群的表现，也能改善其对多数群体的表现；Chari 等人将这一现象称为“少数群体纳入促进多数群体提升”（MIME） [55]。在合成世界中，或许可以创建融合多种难处理表型的合成人类，从而让人工智能学到更好的模型。这样的数据集在现实世界中可能很难甚至不可能获得，尤其是在罕见病方面。

罕见事件。 在医疗保健领域，疾病、综合征、极端值或正态分布中的其他异常值等罕见事件不胜枚举。因此，机器学习方法往往专门用于检测这类样本 [56, 57]。然而，收集此类事件的样本来训练算法颇具挑战，这通常意味着数据集存在严重的类别不平衡 [56]。可以使用模拟器生成罕见事件的合成示例，从而弥补在类别不平衡数据上训练的模型灵敏度不足的问题。示例包括根据光电容积描记或心电图检测心律失常 [58] ，或从医学影像中检测癌症 [59]。罕见事件的合成数据可用于补充现有真实数据，以改进人工智能算法。不过，以这种方式合成罕见事件时，确保仿真可靠且不包含混杂伪影就更加重要；否则，基于这些数据训练的模型可能只会学会识别合成样本。

数据不完整或不足。 除了罕见情况之外，鉴于训练人工智能算法所需的数据量极其庞大，大多数情况都可以从将合成数据与真实数据结合使用中获益，这种做法称为数据增强 [60]。许多研究论文介绍的方法只是使用生成式人工智能模型生成补充样本，而这些样本本质上只是对真实训练数据的排列组合 [18, 61, 62]。

安全与测试。 已有大量文献使用合成数据对机器学习系统进行性能测试 [13, 63]。合成数据能够实现参数的系统性变化，其优势在于只需对生成器进行一次训练和性能表征，随后便可以高效且系统地评估许多分类器，且变化可能比用于训练生成器的变化多得多——本质上是在流形中连续采样，而不仅仅是训练数据集中的离散点。然而，如果模拟器存在偏见，这可能会传播真实世界中存在的系统性不平等 [64]，并导致已经处于不利地位的群体面临更高风险。模型在训练集中的少数群体上可能表现不佳，测试性能也可能并不能真实地代表实际情况。

对于人体受试者研究而言，某些干预、实验和数据收集是不道德的和/或不安全的。合成数据能够在不会对患者或其他研究对象造成风险的情况下，对安全关键环境中的方法进行评估。例如，脓毒症预警系统可以先在合成数据上进行测试和微调，然后再部署。虽然这些方法适用于了解某个模型在医疗保健领域中的一些局限性，但合成数据能否用于适当的性能测试，仍有许多问题有待解答。这些问题涉及合成数据在相关任务上的表现、合成数据的可信度，甚至还涉及基于合成数据训练的模型提供和接受医疗服务时所遵循的社会规范。

持续学习。 由于模拟器能够按需生成数据，因此可用于持续学习。如果发现模型存在弱点或偏见，就可以创建该领域或类别的更多样本。如果模型部署所处的环境或病症发生变化，则可以更新模型。重要的是，合成数据通常可以非常快速地生成，从而使模型能够迅速更新，尽量减少病症变化对模型性能的影响。相比之下，如果需要真实世界的临床数据，则可能需要经过机构审查委员会审批、招募患者、数据收集和模型再训练等漫长流程。再者，数字孪生的概念意味着可以创建与特定个体相似的数据，并在模型部署到该患者身上之前，对个性化模型进行微调。

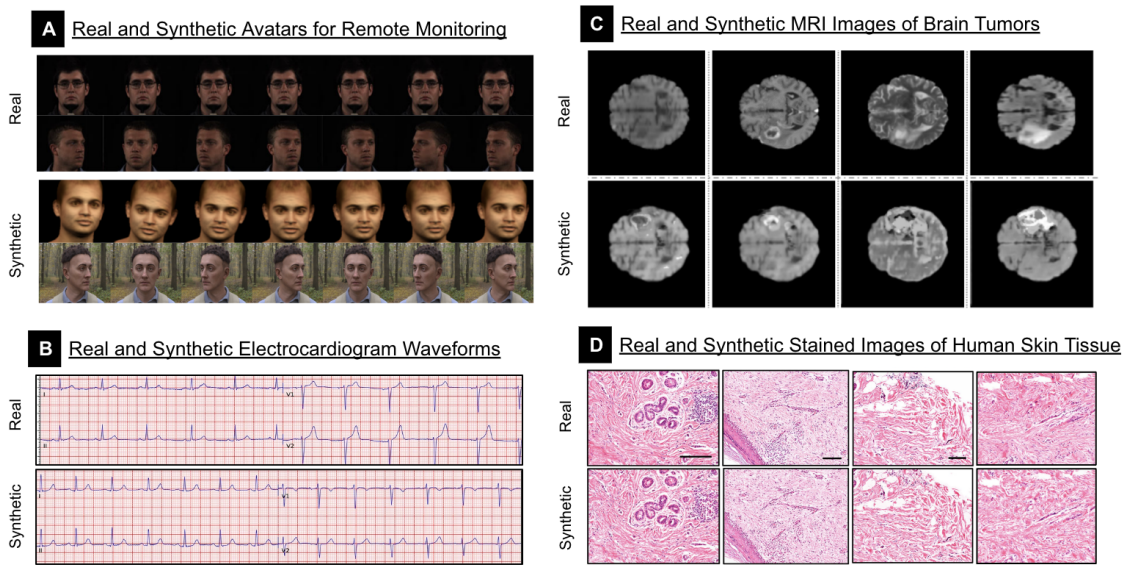


图3：A)用于远程监测的真实和合成患者视频 [69, 54, 70]。B)真实和合成心电图 [71] C) 脑肿瘤的真实和合成磁共振成像 [18]。D)来自 TA-PARS 非辐射吸收对比图像的真实和合成苏木精-伊红染色 [72]。

可解释人工智能。 医疗保健领域对人工智能的采用将在很大程度上取决于临床医生，他们实际上是临床应用的终端用户。临床医生基于一种心智模型做出决策，而这种模型建立在公认的理论基础之上，包括病理生理学、药物的作用机制、实践经验，以及对当前文献和临床试验的了解。相比之下，人工智能往往会得出无法通过传统方式充分理解或解释的推论。尽管人工智能的一般流程是可以理解的，但其中起关键作用的确切参数和关联往往未知；若纳入许多参数，这些因素几乎无法知晓。医疗保健领域的可解释人工智能问题十分复杂，因为 Ghassemi 等人对现有方法能否满足临床医生和监管机构的要求提出了质疑 [65]。合成数据可用于检验我们对模型的理解：通过改变或省略被认为重要的参数，观察输出是否发生变化。基于物理模型生成的合成数据可用于帮助解释模型，例如生成临床医生熟悉且可理解的数据，包括可查看的表格数据、图像或心电图。人类天生会信任自己理解的事物，而不信任自己不理解的事物。因此，要让人工智能不仅经过验证，还能以临床医生熟悉的方式得到解释，这对其成功应用至关重要。

因果模型。 人工智能的可解释性包括考察模型中的因果关系，而这需要基于既有理论建立合理的机制。在医学和医疗保健中的安全关键场景中，因果与反事实推理和推断可能产生重大影响；在这些场景中，数据驱动预测模型可能会被误用于进行因果推断 [66] ，而理解相关性以外的内容是一个重要组成部分。研究人员使用视频和合成数据集进行因果推理 [67, 68], ，因为仿真，尤其是物理仿真，具有已知的生成机制。创建用于训练和测试因果模型的训练数据时，了解底层数据的因果结构是一项巨大优势。

合成数据在医疗保健中的 4 个应用

在现代医疗保健系统中，患者数据存储在电子健康记录（EHR）中。电子健康记录包含种类繁多的数据，从非结构化数据（包括医学成像和自然语言记录），到结构化数据（包括连续变量、类别变量和有序变量），这些数据都表现出线性和非线性交互。大多数数据都是时间序列，其中数据点的顺序和时间间隔对于因果关系和效应量至关重要。数据以表格、图像或自然语言记录的形式存储，其设计历来优先考虑临床护理和计费，而非数据分析。得益于上述优势，合成数据已应用于心脏病学、皮肤病学、组织学和临床心理学等多种领域（示例见表1）。统计生成式人工智能模型是最常用的工具，但可能不适用于某些应用 [59]。现在，我们将探讨合成数据在医疗保健中的应用示例，以及这些研究得出的共同结论。

表1：为医疗人工智能系统创建数据的示例仿真。

应用	模态	模拟器	示例
心脏病学	心脏波形	动力学模型	心电图 [73],光电容积描记 [74],心音图 [75]
		统计	心电图 [19]
	心脏成像	物理模型	超声心动图 [76, 77]
		统计	超声心动图 [78, 79], 心脏磁共振成像 [80]
皮肤病学	皮肤病变	统计	皮肤状况图像 [81]
胃肠病学	内窥镜检查	物理模型	单目内窥镜深度估计 [82]
神经病学	脑部医学成像	物理与统计模型	脑肿瘤图像分割 [83]
眼科学	视网膜病变	统计模型	早产儿视网膜病变图像性 [84]
传染病	检测	统计模型	新冠肺炎胸部X线片检测。 [3, 85]
组织学	检测	统计模型	肾细胞癌检测 [59], Col-皮肤组织图像的着色 [72]
临床心理学	患者记录	统计模型	对精神健康诊断进行分类 [86]

心电图 = 心电图，光电容积描记 =光电容积脉搏波，心音图 =心音图

结构化数据。电子健康记录中的结构化数据（或称表格数据）包括患者人口统计信息、既往病史、药物、诊断、过敏史、生命体征、液体平衡、实验室检查结果、微生物学数据、病理学数据，以及部分与医疗操作和医学成像相关的数据（即不包括图像本身）。Benaim等人展示了如何利用合成结构化数据开展临床观察性研究。他们旨在验证商业人工智能平台MDClone能否在保留已知关联的同时，从真实电子健康记录数据生成合成结构化数据 [21]。他们研究了出院后高危患者接受胃保护用质子泵抑制剂处方的比例、ST段抬高型心肌梗死病例的入院至球囊扩张时间作为180天时心力衰竭和死亡率预测指标的情况、心力衰竭患者入院期间的血尿素氮水平作为3年死亡率预测指标的情况、接受磁共振成像与计算机断层扫描后发生急性肾损伤的风险，以及接受甘精胰岛素与地特胰岛素治疗的住院患者发生低血糖的风险。随后，Foraker等人在另一个医疗卫生系统中研究了三个情境，以验证同一人工智能平台 [22]。他们根据儿科重症监护病房中儿童死亡风险评分（PRISM）的真实数据、住院患者脓毒症预测模型的真实数据，以及用于追踪衣原体感染率的公共卫生工具的真实数据，生成了合成数据。研究表明，该平台能够生成具有与真实数据相似特征的合成数据；重要的是，基于合成数据得出的结果在统计学上与真实数据结果相似。这些结果依托的数据集具有如下特点：患者数量远多于所用变量数量，且样本量较大，从而降低删失数据的影响 [21]。这些研究属于观察性研究，似乎仅旨在共享数据以供回顾性分析。

基于来自超过10亿条重症患者记录的数据集（Amsterdam UMCdb [39]）建模生成的合成数据，被用于比较合成数据生成和患者重新识别的方法。他们的分析强调，不仅需要高质量的合成数据生成方法，还需要用于评估此类数据的良好且稳健的指标 [87]。

自然语言记录。电子健康记录（EHR）中包含大量医疗保健提供者在记录中撰写的自然语言文本——其质量和准确性具有主观性，但在理想情况下，这些文本能提供经过提炼的叙述，突出患者病程中最关键的部分以及预期的发展方向。这对于心理健康记录尤为重要，因为心理健康记录在很大程度上比结构化数据更多地依赖非结构化自然语言文本。Ive等人表明，使用合成出院摘要训练的模型可以准确预测心理健康患者的诊断和表型，因此，将合成自然语言数据作为训练数据可以产生与原始结果相当的分类结果 [86]。

其他行业在自然语言处理方面取得的进展为医疗保健带来了希望，合成自然语言文本可能会从中受益。ChatGPT 近期的流行展现了现代语言模型强大的能力、影响力和规模。这类语言模型依赖于不受限制地访问从互联网上搜集的大量自然语言文本。鉴于隐私保护目前施加的限制，要利用电子健康记录中的自然语言记录，开发类似的临床决策支持语言模型并非易事。由 MedGPT 驱动的 CogStack Foresight 当前测试版展现了作为一种基于原始文本的概率诊断工具的潜力，但模型训练所用的数据仅略多于一百万份患者记录，且大多来自同一家医院。

system.[4] 根据真实记录生成的合成自然语言文本，或许能够用于训练模型，或至少提供增强训练数据，以提升此类模型服务更广泛患者群体的能力。

生理测量。合成患者数据应用最广泛的领域之一，是生理信号处理和下游分类。生理测量指任何非结构化的连续数据流，例如12导联心电图、心电遥测，或与血流动力学相关的波形。研究人员尤其关注心脏信号（即心电图）的合成，以及心音图和光电容积脉搏波（PPG）的合成 [73, 19, 75]。例如，参数化动力学模型能够有效捕捉心电图的特性 [73, 88], ，从而生成窦性心动过缓、窦性心动过速、心室扑动、心房颤动和室性心动过速的合成心电图。值得注意的是，使用真实心电图训练并辅以合成心电图增强的模型，在心律诊断方面表现出更高的准确性 [19, 89]。将常微分方程（ODE）系统纳入生成式人工智能优化过程的动力学统计模型，既能控制波形的动态特性，又能利用生成式人工智能模型的数据驱动特性 [19]。随着人们日益关注生理信号之间的关系（例如，根据心电图和光电容积描记计算的脉搏到达时间可作为血压预测指标），未来的研究应考虑能够模拟不同测量值之间关系的合成数据，而不应仅关注单一测量值。

利用相机测量生理信号的优势在于相机相对普及，而且无需身体接触即可采集测量数据 [90]。然而，合成视频比合成一维波形复杂得多。基于图形的模型可用于创建多种物理模型 [70, 91]。不过，这些模型生成的数据目前还不像“照片级真实”的数据。统计模型提供了一种更易实现的方法，无需复杂的合成流程 [54]。半合成 方法也已被提出，即仅合成部分输入数据。这意味着，对于数据空间中难以模拟的部分，可以依赖真实样本；对于更容易模拟的样本，则可补充合成训练数据。例如，在光电容积描记中，皮肤的光学特性会使肤色影响波形的准确性。在 [92], 中，研究人员对真实患者的数据进行风格迁移，以模拟不同肤色，从而创建半合成数据，促进更公平的生命体征检测。

医学影像。 医学影像领域已开始采用人工智能方法，自动化临床评估流程中的部分环节。人们曾乐观地声称，这些模型很快就能达到人类水平，甚至取代人类专家（例如放射科医生） [93] ，但事实并非如此。尽管如此，仍有许多令人振奋的示例表明，机器学习有助于从这些数据中获取更深入的见解。合成图像生成已成功用于改进多种模型，例如对组织病理学切片中的恶性细胞进行分类 [59], 、对照片中的皮肤病变进行分类 [81],、根据胸部X线片检测新冠肺炎 [85, 3], 、根据CT扫描对肝脏病变进行分类 [61] ，以及分割MRI数据中的脑肿瘤 [83]。大多数研究采用了统计生成式人工智能模型（最常见的是生成对抗网络） [85, 81] （见表1），但在物理模型特征已得到充分描述的情况下，也有研究使用物理模型和混合模型 [94, 95]。具体而言，混合方法通过物理信息驱动的数据增强来扩充合成数据。例如，Zhao等人 [94] 对合成数据进行增强，使其符合由解剖结构和MRI图像采集所产生的空间形变场。Amirrajab等人 [80] 将心脏模拟器与生成模型结合，生成合成心脏MR图像。Zhang 等人 [96] 大胆尝试使用生成式人工智能模型将三维心脏MR图像转换为二维CT格式，同时从理论上避免辐射和放射性造影剂暴露。随着新型成像设备不断普及，对高效创建训练数据集的方法的需求也将持续增长 [97]。合成数据很可能会发挥日益重要的作用。

5 挑战与风险

虽然这些示例凸显了合成数据在医学和医疗保健人工智能领域的广泛应用，但也引发了人们对软件漏洞以及当前政策挑战的担忧 [59]。合成数据极具吸引力，但也伴随着真实风险；由于合成数据属于新兴事物且应用迅速，当前文献和法规对其中许多风险的讨论仍十分有限。

仿真引擎的缺陷与局限。 与模拟器相关的最重大技术挑战，是创建能够反映现实世界真实复杂性的样本，并确定能否理解和描述仿真的局限性。对于物理模型而言，这些局限性通常更加明确，因为这类模型建立在已知或部分已知的确定性结构和参数集之上。能够模拟传感器（例如内窥镜或相机）以及解剖结构逼真的身体部位（例如结肠或面部）的精准物理模型 [82, 70] ，提供了一种有原则的方法，或许更符合现行临床法规 [59]。然而，对于统计生成式人工智能模型，局限性可能更难界定。由于统计模型需要经过训练，目前尚不清楚它们能否有效生成超出训练数据分布范围的样本

[4] https://cogstack.org/cogstack-foresight-beta-launched/

。这意味着它们本质上受限于用于训练的数据，而它们如何补充现有数据这一问题仍有待充分解答。例如，语言模型可以拼接短语，却未必能产生任何新见解 [98]。针对电子健康记录（EHR）中结构化数据的统计合成数据生成方法，容易生成违背预期关系的数据，例如性别与疾病之间的关系 [99]。一项关于姑息治疗中结构化表格数据合成生成的综述指出，尽管相关论文数量众多，方法也不断创新，但对于如何评估合成数据生成方法的性能，仍缺乏共识 [17]。这些问题将阻碍其应用，并使合成数据难以赢得信任。

未知的未知。 在对一种现象建模时，存在已知的已知（例如，可参数化关系或属性）、已知的未知（例如，已知存在但因缺乏经验数据或证明而无法量化的关系），以及未知的未知（例如，可能存在也可能不存在、但被模型开发者忽略的关系）。其中最后一种最成问题。要确信模拟器具有足够的保真度，使未知的未知不会产生影响，并非易事，甚至可能无法证明。迄今为止，研究人员通常依赖经验证据，证明在合成数据上训练的模型能够泛化到真实测试数据。然而，仿真在真实数据上的表现只是整个问题的一部分。未来，人们或许会使用仿真，不仅提升性能，还能帮助提炼或发现此前未知的潜在关系 [100, 101]。在这些情况下，模拟器可能需要预言机（即“人在回路”）提供输入，以识别故障 [102, 59]。

缺乏用于评估通过合成数据训练的模型的标准与法规。 美国食品药品监督管理局的基于人工智能的医疗器械软件（人工智能医疗器械软件）行动计划[5] 没有讨论仿真数据、合成数据或增强数据。Chen 等人详细讨论了该行动计划在合成数据方面的局限性 [103], ，并得出结论：需要制定新的法规，并针对不同的应用场景进行专门调整。尽管人们对合成数据的兴趣迅速增长，且其具有潜在的积极应用场景，但目前尚无明确的法律法规规范其在医疗保健领域的使用 [10]。导致这一问题的原因之一是，新方法往往在缺乏充分透明度和文档的情况下被披露。其后果是无法开展客观评估 [104]。

代表性与偏见。基于物理模型的方法和统计方法都可能无法体现患者和病症的真实多样性。由于学习到的生成模型会生成相对于底层数据分布的统计结果，代表性不足通常会导致模型产生偏见 [105] ，而且已有研究发现，大参数模型会放大偏见 [98]。已有大量文献可帮助解决这些偏见 [106, 107], ，但面向模型创建者的工具仍然有限。模型卡有助于记录用于创建生成模型的数据，使采用这些模型的从业者能够了解其局限性和缺陷。 [108]。

数据泄漏。 尽管合成数据在保护患者隐私方面具有巨大潜力，但合成数据生成处理管线存在泄露个人信息的风险 [109, 87]。目前没有任何方法能够严格判定合成数据是否与真实样本存在足够差异，从而被视为已匿名化 [10]。对于任何模拟器而言，都存在隐私与效用之间的权衡：必须在隐私保护的益处与拟议任务效用的降低之间进行权衡 [88]。这对于统计模型尤其如此，但对于依赖可识别人类数据要素（如面部或身体图像）的参数模型同样成立 [90]。过拟合的模拟器，或使用少量真实患者数据集训练的模拟器，尤其容易受到逆向工程方法的攻击，而这些方法可能暴露原始数据中的可识别信息 [55]。对于特别敏感的情况，只要样本量足够大，就可以从模型中排除个体记录，而不会损害模型保真度。通过恶意攻击，或因无意生成了与特定个体相匹配的数据而泄露私人健康信息（PHI），都会带来问题，因此 Chen 等人 [59] 认为，统计生成模型并不适用于医疗保健领域的某些合成数据生成应用。正是这种观点促使 Jordan 等人针对一个受保护的合成数据集开展了一项新颖的挑战赛：该数据集由生成模型基于真实 ICU 数据集创建；参赛者分为“隐藏者”团队和“寻找者”团队，前者试图创建受保护的合成数据集，后者则试图根据合成数据重新识别患者 [87]。

6 结论

模拟器颇具吸引力，因为它们能够按需生成样本，用于训练和/或测试数据密集型机器学习方法。合成数据可以让人们共享数据，而这些数据若直接共享，可能会违反患者隐私保护规定。在某些情况下，合成数据甚至可以替代那些收集起来不合乎伦理或不安全的样本，或者替代因许多疾病罕见而难以实际收集的样本。然而，在临床应用中使用合成数据也伴随着若干挑战与风险。数据生成过程中的缺陷可能导致

[5] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

模型出现缺陷。具体而言，模拟器的局限意味着无法生成某些类型的真实样本；偏见可能导致生成的数据无法充分代表实际情况；未知的未知意味着合成数据无法得到完全信任；此外，个人数据泄漏仍是风险，因为模拟器可能暴露敏感数据。要应对这些挑战，还需要开展更多工作；我们也观察到许多合成数据被相当草率地应用的示例。在许多方面，仿真会放大当前的人工智能模型：它们能够扩大数据集的规模和多样性；但我们也必须承认，如果使用不当，它们同样会放大人工智能处理管线中的错误和偏见。确保基于仿真训练的模型在真实数据上接受测试，并验证其确实有用，是我们确保仿真发挥价值所能采取的最有效措施。这进一步凸显了生成合成数据的从业者与负责收集真实数据并据此进行推断的临床专家之间开展闭环协作的必要性。这也让我们重新关注以数据为中心的人工智能与仿真工作之间的密切联系。若要将仿真应用于实际医疗场景，就必须提供更多有助于评估和提升仿真质量的工具与标准。

致谢

作者谨此感谢 Aditya Gupta（UCLA）、PraneethVepakomma（MIT）和 Celsode Melo（美国陆军研究实验室）对本手稿早期版本提出的意见。Achuta Kadambi 部分获得了 NSF CAREER 奖（IIS-2046737）、陆军青年研究员奖和 DARPA 青年教师奖的资助。

参考文献

[1] Geoffrey Hinton. 深度学习——一种有望变革医疗保健的技术。*Jama*，320(11):1101– 1102，2018年。[2] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. 医疗健康领域机器学习的挑战与机遇综述。*AMIA Summits onTranslational ScienceProceedings*，2020:191，2020年。[3] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net：一种量身定制的深度卷积神经网络设计，用于从胸部X光图像中检测新冠肺炎病例。*Scientific Reports*，10(1):1–12，2020年。[4] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. 一种由人工智能驱动的心电图算法，可在窦性心律期间识别心房颤动患者：对结局预测的回顾性分析。*The Lancet*，394(10201):861–867，2019年。[5] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai：通过循环神经网络预测临床事件。载于医疗保健机器学习会议，第301–318页。PMLR，2016年。[6] Huiying Liang, Brian Y Tsui, Hao Ni, Carolina Valentim, Sally L Baxter, Guangjian Liu, Wenjia Cai, Daniel S Kermany, Xin Sun, Jiancong Chen, et al. 利用人工智能评估并准确诊断儿科疾病。*Naturemedicine*，25(3):433–438，2019年。[7] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 面向生物医学自然语言处理的领域特定语言模型预训练。*ACM Transactions on Computing for Healthcare (HEALTH)*，3(1):1–23，2021年。[8] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, et al. Dataperf：以数据为中心的人工智能开发基准。*arXiv*预印本 *arXiv:2207.10062*，2022年。[9] Celso M de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. 基于模拟器和合成数据的新一代深度学习。*Trends in cognitive sciences*，2021年。[10] Anmol Arora and Ananya Arora. 医疗保健领域的合成患者数据：不断扩大的法律漏洞。*The Lancet*，399(10335):1601–1602，2022年。[11] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 基于单幅深度图像的实时人体姿势分部识别。载于*CVPR 2011*，第1297–1304页。Ieee，2011年。[12] VSR Veeravasarapu, Rudra Narayan Hota, Constantin Rothkopf, and Ramesh Visvanathan. 通过图形模拟验证视觉系统模型。*arXiv*预印本 *arXiv:1512.01401*，2015年。[13] Robert M Haralick. 计算机视觉中的性能表征。载于*BMVC92*，第1–8页。施普林格，1992年。

[14] Alana Marzoev, Samuel Madden, M Frans Kaashoek, Michael Cafarella, and Jacob Andreas. 非自然语言处理：弥合合成数据与自然语言数据之间的差距。*arXiv*预印本 *arXiv:2004.13645*，2020年。[15] Dingquan Wang and Jason Eisner. 按需生成合成数据：句法分析案例。载于*2018*年自然语言处理实证方法会议论文集，第1325–1337页，2018年。[16] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 表格化健康记录的合成数据生成：系统综述。*Neurocomputing*，2022年。[17] Waldemar Hahn, Katharina Schütte, Kristian Schultz, Olaf Wolkenhauer, Martin Sedlmayr, Ulrich Schuler, Martin Eichler, Saptarshi Bej, and Markus Wolfien. 合成数据生成对改善姑息治疗中的患者分层所作的贡献。*Journal of Personalized Medicine*，12(8):1278，2022年。[18] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. 利用生成对抗网络进行医学图像合成，以实现数据增强和匿名化。载于医学影像模拟与合成：第三届国际研讨会 *SASHIMI 2018*，与 *MICCAI 2018* 联合举办，西班牙格拉纳达，*2018*年*9*月*16*日，论文集*3*，第1–11页。施普林格，2018年。[19] Tomer Golany, Kira Radinsky, and Daniel Freedman. Simgans：基于模拟器的生成对抗网络，用于心电图合成以改进深度心电图分类。载于国际机器学习会议，第3597–3606页。PMLR，2020年。[20] Daniel McDuff, Javier Hernandez, Erroll Wood, Xin Liu, and Tadas Baltrusaitis. 利用合成虚拟形象推进非接触式生命体征测量。*arXiv*预印本 *arXiv:2010.12949*，2020年。[21] Anat Reiner Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S Azzam, Johad Khoury, et al. 基于合成数据分析医学研究结果及其与真实数据结果的关系：对五项观察性研究的系统比较。*JMIRmedicalinformatics*，8(2):e16492，2020年。[22] Randi E Foraker, Sean C Yu, Aditi Gupta, Andrew P Michelson, Jose A Pineda Soto, Ryan Colvin, Francis Loh, Marin H Kollef, Thomas Maddox, Bradley Evanoff, et al. 找出差异：比较基于真实患者数据和合成衍生数据的分析结果。*JAMIA open*，3(4):557–566，2020年。[23] Olaf Dössel, Giorgio Luongo, Claudia Nagel, and Axel Loewe. 用于心电图解读的心脏计算机建模——综述。*Hearts*，2(3):350–368，2021年。[24] Ali Madani, Jia Rui Ong, Anshul Tibrewal, and Mohammad RK Mofrad. 深度超声心动图：面向心脏疾病自动诊断的数据高效监督式与半监督式深度学习。*NPJ digital medicine*，1(1):59，2018年。[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 生成对抗网络。*Communications of theACM*，63(11):139–144，2020年。[26] Diederik P Kingma and Max Welling. 自编码变分贝叶斯。*arXiv*预印本 *arXiv:1312.6114*，2013年。[27] Durk P Kingma and Prafulla Dhariwal. Glow：具有可逆1×1卷积的生成流。神经信息处理系统进展，31，2018年。[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 利用非平衡热力学进行深度无监督学习。载于国际机器学习会议，第2256–2265页。PMLR，2015年。[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020年。[30] Prafulla Dhariwal and Alexander Nichol. 扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展，34:8780–8794，2021年。[31] Peter Eigenschink, Stefan Vamosi, Ralf Vamosi, Chang Sun, Thomas Reutterer, and Klaudius Kalcher. 用于合成数据的深度生成模型。*ACM Computing Surveys*，2021年。[32] Tero Karras, Samuli Laine, and Timo Aila. 面向生成对抗网络的基于风格的生成器架构。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4401–4410页，2019年。[33] Manish Sahu, Ronja Strömsdörfer, Anirban Mukhopadhyay, and Stefan Zachow. Endo-sim2real：用于器械分割的基于一致性学习的域适应。载于医学图像计算与计算机辅助介入——*MICCAI 2020*：第*23*届国际会议，秘鲁利马，*2020*年*10*月*4*日至*8*日，论文集，第三部分*23*，第784–794页。施普林格，2020年。

[34] Juan FPJ Abascal、Nicolas Ducros、Valeriya Pronina、Simon Rit、Pierre-Antoine Rodesch、Thomas Broussaud、Suzanne Bussod、Philippe C Douek、Andreas Hauptmann、SimonArridge等。使用深度学习进行光谱计算机断层扫描中的材料分解：一种Sim2Real迁移方法。*IEEE Access*，9：25632–25647，2021年。[35] Edward Juhn和Erica Galvez。激励医疗保险计划、医院和医疗服务提供者之间的数据共享，以提高质量。*The American Journal of Managed Care*，28(12)：e426–e427，2022年。[36] Barbara E Bierer、Mercè Crosas和HeatherHPierce。将数据署名作为数据共享的激励，2017年。[37] Alistair EW Johnson、LucasBulgarelli、LuShen、Alvin Gayles、Ayad Shammout、Steven Horng、Tom J Pollard、BenjaminMoody、Brian Gow、Li-wei H Lehman等。MIMIC-IV：一个可自由访问的电子健康记录数据集。*Scientificdata*，10(1)：1，2023年。[38] Tom J Pollard、Alistair EW Johnson、JesseD Raffa、LeoA Celi、Roger GMark和OmarBadawi。eICU协作研究数据库：一个可自由获取的多中心重症监护研究数据库。*Scientific data*，5(1)：1–13，2018年。[39] Patrick J Thoral、Jan M Peppink、Ronald H Driessen、Eric JG Sijbrands、Erwin JO Kompanje、Lewis Kaplan、Heatherlee Bailey、Jozef Kesecioglu、Maurizio Cecconi、Matthew Churpek等。在重症监护医学会/欧洲重症监护医学会联合数据科学合作框架下负责任地共享重症监护病房患者数据：阿姆斯特丹大学医学中心数据库（AmsterdamUMCdb）示例。*Criticalcaremedicine*，49(6)：e563，2021年。[40] Martin Faltys、Marc Zimmermann、Xinrui Lyu、Matthias Hüser、StephanieHyland、GunnarRätsch和TobiasMerz。HiRID：高时间分辨率的重症监护病房数据集（版本1.1.1）。*PhysioNet*，2021年。[41] XianZeng、Gang Yu、Yang Lu、Linhua Tan、Xiujing Wu、Shanshan Shi、HuilongDuan、Qiang Shu和Haomin Li。PIC：儿科专用重症监护数据库。*Scientificdata*，7(1)：14，2020年。[42] Victor Janmey和Peter L Elkin。HIPAA去标识化数据集中的重新识别风险：MVA攻击。载于*AMIAAnnual Symposium Proceedings*，第2018卷，第1329页。美国医学信息学协会，2018年。[43] Ji Su Yoo、Alexandra Thaler、LatanyaSweeney和Jinyan Zang。患者隐私风险：对缅因州和佛蒙特州全州医院数据中患者的重新识别。*J Technol Sci*，2018：2018100901，2018年。[44] Debbie Rankin、Michaela Black、Raymond Bond、Jonathan Wallace、MauriceMulvenna、GorkaEpelde等。医疗保健领域使用合成数据进行监督机器学习的可靠性：一种保护隐私以实现数据共享的模型。*JMIRMedical Informatics*，8(7)：e18910，2020年。[45] Zahra Azizi、Chaoyi Zheng、Lucy Mosquera、Louise Pilote和Khaled El Emam。合成数据能否作为真实临床试验数据的代理？一项验证性研究。*BMJopen*，11(4)：e043497，2021年。[46] Abhishek Singh、Ethan Garza、Ayush Chopra、Praneeth Vepakomma、VivekSharma和RameshRaskar。解耦与采样：在任务无关的数据发布中保护敏感信息。*ECCV*，2022年。[47] James Jordon、Jinsung Yoon和MihaelaVanDerSchaar。PATE-GAN：生成具有差分隐私保证的合成数据。载于*International conference on learning representations*，2018年。[48] TerranceLiu、Giuseppe Vietri和Steven Z Wu。私有合成数据的迭代方法：统一框架与新方法。神经信息处理系统进展，34：690–702，2021年。[49] ZhikunZhang、Tianhao Wang、Jean Honorio、Ninghui Li、Michael Backes、Shibo He、JimingChen和Yang Zhang。Privsyn：差分隐私数据合成。2021年。[50] Sahil Verma和Julia Rubin。公平性定义释义。载于*Proceedings ofthe internationalworkshop onsoftwarefairness*，第1–7页，2018年。[51] Achuta Kadambi。实现医疗器械中的公平性。*Science*，372(6537)：30–31，2021年。[52] Olubunmi E Okunlola、Michael S Lipnick、Paul B Batchelder、Michael Bernstein、John RFeiner和PhilipE Bickler。脉搏血氧仪性能、种族不公平以及未来的工作。*Respiratory care*，67(2)：252–257，2022年。[53] DanielMcduff、J avier Hernandez、Xin Liu、ErrollWood和TadasBaltrusaitis。使用高保真度虚拟化身推进基于相机的心脏脉搏测量。*IEEE Transactions on Biomedical Engineering*，2022年。[54] Zhen Wang、Yunhao Ba、Pradyumna Chari、Oyku Deniz Bozkurt、Gianna Brown、Parth Patwa、Niranjan Vaddi、LalehJalilian和Achuta Kadambi。具有光电容积描记生理特征的人脸视频合成。载于*Proceedings of the IEEE/CVF Conference on ComputerVisionandPatternRecognition*，第20587–20596页，2022年。

[55] Pradyumna Chari、Yunhao Ba、Shreeram Athreya 和 Achuta Kadambi。Mime：通过少数群体纳入来提升多数群体的人工智能性能。载于欧洲计算机视觉会议，第 326–343 页。施普林格，2022年。[56] Der-Chiang Li、Chiao-Wen Liu 和 Susan C Hu。针对医学数据集类别不平衡问题的学习方法。《计算机与生物医学》，40(5)：509–518，2010年。[57] Tharindu Fernando、Harshala Gammulle、Simon Denman、Sridha Sridharan 和 Clinton Fookes。用于医学异常检测的深度学习——综述。《ACM Computing Surveys (CSUR)》，54(7)：1–37，2021年。[58] Serkan Kiranyaz、Turker Ince 和 Moncef Gabbouj。用于心脏心律失常的个性化监测与预警系统。《科学报告》，7(1)：1–8，2017年。[59] Richard J Chen、Ming Y Lu、Tiffany Y Chen、Drew FK Williamson 和 Faisal Mahmood。医学与医疗保健机器学习中的合成数据。《自然生物医学工程》，5(6)：493–497，2021年。[60] Zhengping Che、Yu Cheng、Shuangfei Zhai、Zhaonan Sun 和 Yan Liu。利用生成对抗网络提升电子健康记录深度学习风险预测能力。载于2017年 IEEE 国际数据挖掘会议（ICDM），第 787–792 页。IEEE，2017年。[61] Maayan Frid-Adar、Idit Diamant、Eyal Klang、Michal Amitai、Jacob Goldberger 和 Hayit Greenspan。基于生成对抗网络的合成医学图像增强，可提高卷积神经网络在肝脏病变分类中的性能。《Neurocomputing》，321：321–331，2018年。[62] Amjad Rehman Khan、Siraj Khan、Majid Harouni、Rashid Abbasi、Sajid Iqbal 和 Zahid Mehmood。利用 K 均值聚类、深度学习和合成数据增强进行脑肿瘤分割与分类。《显微镜研究与技术》，84(7)：1389–1399，2021年。[63] Daniel McDuff、Shuang Ma、Yale Song 和 Ashish Kapoor。利用生成模型表征分类器中的偏见。《神经信息处理系统进展》，32：5403–5414，2019年。[64] Aylin Caliskan、Joanna J Bryson 和 Arvind Narayanan。从语言语料自动推导出的语义包含类似人类的偏见。《科学》，356(6334)：183–186，2017年。[65] Marzyeh Ghassemi、Luke Oakden-Rayner 和 Andrew L Beam。当前医疗保健领域可解释人工智能方法所带来的虚假希望。《柳叶刀数字健康》，3(11)：e745–e750，2021年。[66] Mattia Prosperi、Yi Guo、Matt Sperrin、James S Koopman、Jae S Min、Xing He、Shannan Rich、Mo Wang、Iain E Buchan 和 Jiang Bian。面向可执行医疗保健的机器学习因果推断与反事实预测。《自然机器智能》，2(7)：369–375，2020年。[67] Li-Min Fu。Causim：基于规则的因果仿真系统。《仿真》，56(4)：251–256，1991年。[68] Daniel McDuff、Yale Song、Jiyoung Lee、Vibhav Vineet、Sai Vemprala、Nicholas Alexander Gyde、Hadi Salman、Shuang Ma、Kwanghoon Sohn 和 Ashish Kapoor。Causalcity：用于因果发现与推理的具备能动性的复杂仿真。载于因果学习与推理会议，第 559–575 页。PMLR，2022年。[69] Justin R Estepp、Ethan B Blackford 和 Christopher M Meier。利用多成像仪阵列，在非接触式医学成像光电容积描记中恢复运动伪影期间的脉率。载于2014年 IEEE 国际系统、人类与控制论会议（SMC），第 1462–1469 页。IEEE，2014年。[70] Daniel McDuff、Miah Wander、Xin Liu、Brian L Hill、Javier Hernandez、Jonathan Lester 和 Tadas Baltrusaitis。Scamps：用于生理信号摄像头测量的合成数据。arXiv 预印本 arXiv:2206.04197，2022年。[71] Vajira Thambawita、Jonas L Isaksen、Steven A Hicks、Jonas Ghouse、Gustav Ahlberg、Allan Linneberg、Niels Grarup、Christina Ellervik、Morten Salling Olesen、Torben Hansen 等。利用生成对抗网络生成的深度伪造心电图，标志着医学隐私问题终结的开端。《科学报告》，11(1)：21896，2021年。[72] Marian Boktor、Benjamin R Ecclestone、Vlad Pekar、Deepak Dinakaran、John R Mackey、Paul Fieguth 和 Parsin Haji Reza。无标记全吸收光声遥感（TA-PARS）的虚拟组织学染色。《科学报告》，12(1)：1–12，2022年。[73] Patrick E McSharry、Gari D Clifford、Lionel Tarassenko 和 Leonard A Smith。用于生成合成心电图信号的动力学模型。《IEEE 生物医学工程汇刊》，50(3)：289–294，2003年。[74] Diego Martin-Martinez、Pablo Casaseca-de-la Higuera、Marcos Martin-Fernandez 和 Carlos Alberola-López。光电容积描记信号的随机建模：一种分析合成方法及其应用。《IEEE 生物医学工程汇刊》，60(9)：2432–2441，2013年。[75] Ali Almasi、Mohammad B Shamsollahi 和 Lotfi Senhadji。用于生成合成心音图信号的动力学模型。载于2011年 IEEE 医学与生物学工程学会年度国际会议，第 5686–5689 页。IEEE，2011年。

[76] Martino Alessandrini, Mathieu De Craene, Olivier Bernard, Sophie Giffard-Roisin, Pascal Allain, Irina Waechter- Stehle, Jürgen Weese, Eric Saloux, Hervé Delingette, Maxime Sermesant 等。用于生成逼真三维合成超声心动图序列的流程：方法与开放获取数据库。*IEEE*医学影像学汇刊，34(7):1436–1451，2015年。[77] Adityo Prakosa, Maxime Sermesant, Hervé Delingette, Stéphanie Marchesseau, Eric Saloux, Pascal Allain, Nicolas Villain 和 Nicholas Ayache。结合生物物理模型与临床影像，生成合成但视觉逼真的心脏影像时间序列。*IEEE*医学影像学汇刊，32(1):99–109，2012年。[78] Cristiana Tiago, Andrew Gilbert, Ahmed Salem Beela, Svein Arne Aase, Sten Roar Snare, Jurica Šprem 和 Kristin McLeod。使用生成对抗网络生成带标注三维超声心动图影像数据集的数据增强流程。*IEEE Access*，10:98803–98815，2022年。[79] Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset 和 Kristin Mcleod。基于现有解剖模型生成带标注的合成数据：以超声心动图分割为例。*IEEE*医学影像学汇刊，40(10):2783–2794，2021年。[80] Sina Amirrajab, Samaneh Abbasi-Sureshjani, Yasmina Al Khalil, Cristian Lorenz, Juergen Weese, Josien Pluim 和 Marcel Breeuwer。XCAT-GAN：在解剖结构可变的XCAT仿体上合成三维一致的带标注心脏磁共振影像。载于医学图像计算与计算机辅助介入–*MICCAI 2020*：第*23*届国际会议，利马，秘鲁，*2020*年*10*月*4*日至*8*日，论文集，第*IV*部分 *23*，第128–137页。施普林格，2020年。[81] Amirata Ghorbani, Vivek Natarajan, David Coz 和 Yuan Liu。DermGAN：合成生成带有病理特征的临床皮肤影像。载于健康领域机器学习研讨会，第155–170页。PMLR，2020年。[82] Faisal Mahmood, Richard Chen 和 Nicholas J Durr。通过对抗训练对合成医学影像进行无监督反向域适应。*IEEE*医学影像学汇刊，37(12):2572–2581，2018年。[83] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest 等。多模态脑肿瘤影像分割基准（BRATS）。*IEEE*医学影像学汇刊，34(10):1993–2024，2014年。[84] Aaron S Coyner, Jimmy S Chen, Ken Chang, Praveer Singh, Susan Ostmo, RV Paul Chan, Michael F Chiang, Jayashree Kalpathy-Cramer, J Peter Campbell，早产儿视网膜病变影像与信息学联盟等。用于稳健且保护隐私的人工智能训练的合成医学影像：应用于早产儿视网膜病变诊断。眼科学，2(2):100126，2022年。[85] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman 和 Plácido Rogerio Pinheiro。CovidGAN：使用辅助分类器生成对抗网络进行数据增强，以提升新冠肺炎检测效果。*IEEE Access*，8:91916–91923，2020年。[86] Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart 和 Sumithra Velupillai。用于自然语言处理的人工心理健康记录生成与评估。*NPJ digital medicine*，3(1):1–9，2020年。[87] James Jordon, Daniel Jarrett, Evgeny Saveliev, Jinsung Yoon, Paul Elbers, Patrick Thoral, Ari Ercole, Cheng Zhang, Danielle Belgrave 和 Mihaela van der Schaar。捉迷藏隐私挑战：合成数据生成与患者身份再识别。载于*NeurIPS 2020*竞赛与演示赛道，第206–215页。PMLR，2021年。[88] Omid Sayadi, Mohammad B Shamsollahi 和 Gari D Clifford。利用高斯波动态模型进行合成心电图生成与贝叶斯滤波。生理测量，31(10):1309，2010年。[89] Youssef Skandarani, Alain Lalande, Jonathan Afilalo 和 Pierre-Marc Jodoin。生成对抗网络在心脏病学中的应用。加拿大心脏病学杂志，2021年。[90] Daniel McDuff。生命体征的摄像头测量。*ACM Computing Surveys (CSUR)*，2021年。[91] Akshay Paruchuri, Xin Liu, Yulu Pan, Shwetak Patel, Daniel McDuff 和 Soumyadip Sengupta。运动至关重要：通过神经运动迁移提升摄像头生理感测效果。*arXiv*预印本 *arXiv:2303.12059*，2023年。[92] Yunhao Ba, Zhen Wang, Kerim Doruk Karinca, Oyku Deniz Bozkurt 和 Achuta Kadambi。通过生物逼真外观操控进行风格迁移，实现适用于各种肤色的远程光电容积描记。载于*2022*年*IEEE*国际计算摄影会议（*ICCP*），第1–12页。IEEE，2022年。[93] Curtis P Langlotz。人工智能会取代放射科医生吗？，2019年。[94] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag 和 Adrian V Dalca。使用学习到的变换进行数据增强，实现单样本医学影像分割。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8543–8553页，2019年。

[95] Benjamin Billot、Douglas N Greve、Oula Puonti、Axel Thielscher、Koen Van Leemput、Bruce Fischl、AdrianV Dalca 和 JuanEugenioIglesias。Synthseg：针对任意对比度和分辨率脑磁共振成像扫描的分割的域随机化。*arXiv* 预印本*arXiv:2107.09559*，2021年。[96] Zizhao Zhang、Lin Yang和YefengZheng。使用循环一致性和形状一致性生成对抗网络对多模态医学体积进行转换和分割。在*IEEE* 计算机视觉与模式识别会议论文集中，第9242–9251页，2018年。[97] Ewa Nowara、Daniel Mcduff、AshutoshSabharwal和Ashok Veeraraghavan。利用计算摄影看见皮肤之下。*Communications of the ACM*，65(12)：90–100，2022年。[98] Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和ShmargaretShmitchell。论随机鹦鹉的危险：语言模型是否可能过大？在*2021*年 *ACM* 公平性、问责制与透明度会议论文集中，第610–623页，2021年。[99] Chao Yan、Yao Yan、Zhiyu Wan、Ziqi Zhang、Larsson Omberg、Justin Guinney、Sean D Mooney和Bradley A Malin。对合成电子健康记录生成模型的多层面基准评测。*Nature Communications*，13(1)：1–18，2022年。[100] Michael Schmidt和 Hod Lipson。从实验数据中提炼自由形式的自然定律。*science*，324(5923)：81–85，2009年。[101] Pradyumna Chari、Chinmay Talegaonkar、Yunhao Ba和Achuta Kadambi。视觉物理学：从视频中发现物理定律。*arXiv* 预印本*arXiv:1911.11893*，2019年。[102] Himabindu Lakkaraju、Ece Kamar、Rich Caruana和 EricHorvitz。在开放世界中识别未知的未知：用于引导探索的表征和策略。在*AAAI*人工智能会议论文集中，第31卷，2017年。[103] Junqiao Chen、David Chun、Milesh Patel、EpsonChiang 和Jesse James。合成临床数据的有效性：使用临床质量指标对领先的合成数据生成器（synthea）进行验证研究。*BMC*医学信息学与决策，19(1)：1–9，2019年。[104] WPatrickWalters和 Mark Murcko。评估生成式人工智能对药物化学的影响。*NatureBiotechnology*，38(2)：143–145，2020年。[105] Aditya Grover、JiamingSong、Ashish Kapoor、Kenneth Tran、Alekh Agarwal、Eric JHorvitz 和 StefanoErmon。使用无似然重要性加权对学习到的生成模型进行偏见校正。神经信息处理系统进展，32，2019年。[106] Moritz Hardt、EricPrice 和 Nati Srebro。监督学习中的机会平等。神经信息处理系统进展，29，2016年。[107] Boris van Breugel、Trent Kyono、Jeroen Berrevoets 和Mihaela vander Schaar。Decaf：使用因果感知生成网络生成公平的合成数据。神经信息处理系统进展，34：22221–22233，2021年。[108] Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。用于模型报告的模型卡。在公平性、问责制与透明度会议论文集中，第220–229页，2019年。[109] Dingfan Chen、Ning Yu、Yang Zhang 和MarioFritz。Gan-leaks：针对生成模型的成员推断攻击分类法。在*2020*年*ACM SIGSAC*计算机与通信安全会议论文集中，第343–362页，2020年。