卫生信息化国际发展动态

北京市卫生健康大数据与政策研究中心 2024.2.29

（二） 元数据

1.标题：医学领域使用元数据驱动的方法进行数据协调：范围界定审查

来源： JMIR Med Inform.

时间： 2024 年 2 月.

链接： https://medinform.jmir.org/2024/1/e52967/citations.

概要： 多地点临床研究越来越多地使用真实数据来作真实证据。但是由于数据异质性，很难进行跨医院的统一分析。因此，为保证研究数据质量，有必要实施提取-转换-加载(ETL)或提取-加载-转换(ELT)流程来协调本地健康数据。但是这种进程的开展既耗时又不可持续。解决这一问题的一个好方法是 ETL/ELT 过程泛化。本研究调查了通用 ETL/ELT 过程开展的现有可能性，并通过使用描述性元数据和结构化元数据来特别关注开展复杂度较低的方法。研究是按照 PRISMA 指南进行文献研究，用到 PubMed、IEEE Explore、Web of Science 和 Biomed Center 四个文献数据库，检索 2012-2022 年相关论文。然后使用基于 R 的工具进行可视化 PRISMA 流程，并将论文所有相关内容提取到电子表格中，以便进一步分析和可视化。最终研究共纳入 33 篇论文，共分为 7 个不同专题(即医学、数据仓库、大数据、工业、地理信息学、考古学和军事)，从提取的数据来看，基于本体和基于规则的方法是不同主题类别中使用最多的两种方法，在用例中选择不同方法和工具来实现不同目的。文献研究表明，使用元数据驱动(MDD)方法开展 ETL/ELT 过程可以在不同主题中实现不同目的。应用 MDD 方法实现从快速医疗互操作性资源到观察性医疗结果伙伴关系公共数据模型的数据自动转换，有望实现 ETL/ELT 流程。但是确定一个合适 MDD 方法和工具来实现这样一个 ETL/ELT 过程仍面临挑战，这主要因为对本研究中提出的 MDD 方法的特征还缺乏全面了解，研究的下一步工作就是评估本研究中提出的 MDD 方法，并确定最合适的 MDD 方法以及将其集成到 ETL/ELT 过程中的方法。

2. 标题：医疗保健中合成时间序列生成方法的比较评估：利用患者元数据进行准确数据合成

来源： BMC Med Inform Decis Mak.

时间： 2024 年 2 月.

链接： https://doi.org/10.1186/s12911-024-02427-0

概要： 合成数据是解决医学研究中涉及个人和临床数据的法律和监管问题的一种新兴方法，无论是作为单一工具还是通过与其他增强隐私的技术相结合。生成没有损失的合成数据非常有利于外部研究人员进行二次分析、无限制信息访问提供和同时履行相关规定。但是，如果合成的原始数据(如在 Living Labs 中获得的数据)是由受试者的元数据(静态)和纵向组件(一组与时间相关的测量值)组成，刚可能使生成连贯合成数据面临挑战。本研究定义并比较了三种综合时间序列生成法：仅生成元数据并与原始数据的实时时间序列耦合(A1)、分别生成元数据和时间序列再加入(A2)、同时生成元数据和时间序列(A3)。研究利用两种不同的合成数据（带梯度惩罚的 Wasserstein GAN (WGAN-GP)和 DöppelGANger (DGAN)）生成了模型，并对三种方法进行了比较评估。研究实验是通过三种不同医疗保健相关的纵向数据集来完成：来自 Malaga 大学的跑步机全功测试值(TMET)(1)，来自 MIMIC-III v1.4 数据库的低血压子集(2)，以及名为 PMData 的生活记录数据集(3)。研究在生成的合成数据上评估了三个关键维度：与原始数据的相似性(1)、效用(2)和隐私水平(3)。最佳方法是根据评估的维度和度量而波动。最后研究得出结论：要合成的数据集的初始特征在确定最佳方法上起到至关重要的作用。将合成元数据与实时时间序列(A1)耦合，以及联合生成合成时间序列和元数据(A3)，都各有优势，而单独生成时间序列和元数据(A2)的总体性能则似乎相对差一些。

（徐健编辑）

译文一：

医学领域使用元数据驱动的方法进行数据协调：范围界定审查

Yuan Peng; Franziska Bathelt; Richard Gebler; Robert Gött; Andreas Heidenreich; Elisa Henke; Dennis Kadioglu; Stephan Lorenz; Abishaa Vengadeswaran; Martin Sedlmayr; 徐健(译)

介绍

多中心临床研究越来越多地使用真实世界数据来获得真实世界证据，尤其是在COVID-19大流行期间。但是并非所有诊所都用相同的医院信息系统，导致不同医院信息系统产生的数据存在异质性。这些异构数据在语义和句法上不可互操作。因此，很难跨站点以统一的方式分析此类数据。为此，需要对异构数据进行协调和标准化，比如通过使用通用数据模型（CDM），例如，欧洲医疗机构成立DARWIN EU（欧盟数据分析和真实世界审讯网络）提供有关整个欧盟药物使用和不良事件的真实世界证据。DARWIN EU 使用观察性医疗结果伙伴关系 （OMOP） CDM作为基础模型，由观察健康数据科学和信息学社区。要参与此类网络，需要对本地数据进行转换，常见的方法是研制提取-转换-加载（ETL）或提取-加载-转换（ELT）过程。两者都可以将异构数据协调到目标系统中。它们之间的唯一区别是处理数据的顺序。ETL是在将数据加载到目标系统之前转换数据，而ELT首先将数据加载到目标系统，然后转换数据。由于数据格式和源系统不同，必须实现多个ETL/ELT流程。这项工作既费时又难以维护。使用标准数据交换格式可以降低将异构数据转换为CDM的复杂性，如快速医疗保健互操作性资源（FHIR）格式。FHIR是一种通信标准，由七层卫生信息交换标准（HL7）构成。在德国，医学信息学倡议（MII）采用FHIR格式提供核心数据集（CDS），用于实现所有大学医院之间数据的互操作性。另一种是德国协会“全国法定健康保险医师协会”（KBV，德语：Kassenärztliche Bundesvereinigung）为医疗信息对象（如免疫记录和产妇记录）提供的FHIR格式

的 KBV CDS。尽管 MII CDS 和 KBV CDS 都基于德国 HL7 基础配置文件，但是由于 MII 和 KBV 的要求不同，2 个 CDS 中定义的 FHIR 配置文件也不完全相同。如诊所内科室的代码（比如 0100 代表内科）会使用不同的编码系统。这也使 ETL/ELT 流程的实施和维护变得复杂。

此外，多数国家/地区都试图将其电子健康记录（EHR）数据标准化以进行研究，并提高数据的互操作性。因此，开发了特定国家/地区的 FHIR 配置文件，比如德国 HL7 基础配置文件和美国 CDS。由于不同的语言（即德语与英语）、不同的结构定义（如扩展名和基数）和不同的编码系统（如国际疾病分类的系统 URL 10，修订版：德语版与国际疾病分类的系统 URL 10 临床版）用于 FHIR 配置文件中，需要实现不同的 ETL 过程。虽然这只是几个例子，但可以想象，随着支持用例的扩展，实现 ETL/ELT 过程所需的时间会大幅提升，而可维护性则会下降。因此，实施通用的 ETL/ELT 流程来协调当地卫生数据，可以保证研究数据在各个地点和国家之间的语义和句法互操作性。

正如 David Loshin 所说，使用元数据实现 ETL/ELT 流程是一种很有前途的方法："以分析为目的为组织提供的数据，需要从原始源（源元数据）中提取，并将其转换为与数据库（目标元数据）一致的表示形式，其方式不会因格式和精度的差异而丢失信息（结构元数据），并以有意义的方式表示（语义元数据）。元数据的一个非常广泛的定义是"关于其他数据的数据"。根据具体的使用环境，元数据可分为 3 类：

- **描述性元数据**：发现和识别目标的元数据，如源数据和目标数据的元数据。
- **结构元数据**：管理信息系统中数据的元数据，如数据库中的列名和表名。
- **管理元数据**：数据库中用于提供其他管理信息的元数据，如修改数据库中数据的人员名单。

元数据可以用元数据语言（如资源描述框架和符号）来表示。这种语言也称为本体语言。为实现来自不同源和目标系统的数据的互操作性，可以使用规则语言（如规则标记语言和语义 Web 规则语言）来定义它们之间的转换规则。因此，元数据的使用更有利于将 FHIR 资源转换为 OMOP CDM 的开发和维护。

需要特别指出的是，我们将用于开发 ETL/ELT 流程的任何（描述性和结构性）基于元数据的方法理解为元数据驱动（MDD）方法。本工作重点是概述 MDD 方法的

类型及其在不同主题类别中的应用。该概述旨在确定一种合适的 MDD 方法，以增强从 FHIR 到 OMOP CDM 的数据转换。这将通过回答以下问题来实现：

- Q1：MDD 方法的应用主题是什么？
- Q2：文献中有哪些类型的 MDD 方法？
- Q3：使用 MDD 方法的原因是什么？
- 问题 4：使用什么工具来实施 MDD 方法？

方法

为了回答上述 4 个研究问题，我们进行了文献综述。为确保综述过程的透明度，我们遵循了 PRISMA（系统综述和荟萃分析的首选报告项目）指南。我们用到 4 个论文数据库（即 PubMed、IEEE Explore、Web of Science 和 Biomed Center），检索了 2012 年-2022 年用德语或英语撰写的相关论文（文本框 1)。第一次检索在 2022 年 8 月 11 日进行，第二次检索在 2023 年 3 月 15 日进行，检索 2022 年 12 月 31 日前完成的文献。收集到的论文被加载到 Zotero Citation Management 计划（数字学术公司），并手动删除重复项。为了更顺利地分类排除，我们定义了 8 个排除标准（文本框 2)。

本综述是一个由标题-摘要-筛选（TAS）和全文筛选（FTS）组成的双重过程。两种筛选过程都使用了如文本框 2 的相同排除标准，并根据发表日期将独特的论文分为两组，并上传到研究合作平台 Rayyan（卡塔尔计算研究所和 Cochrane Bahrain），作为 2 个独立项目。每个审阅组都分配 4 名审稿人。通讯作者审阅所有论文。TAS 是背靠背进行，每个审稿人独立标记论文。在所有论文都被标记并讨论并解决冲突后，结束背靠背审阅。之后，所有收录的论文被随机分为 2 组，并作为 FTS 的新项目重新加载到 Rayyan 中。与 TAS 类似，每个审阅组分配 4 名审稿人，通讯作者审阅所有论文。FTS 也是在背靠背下进行，并遵循与 TAS 相同的审查程序。

通讯作者根据文本框 3 提取论文内容，并由 4 位合著者验证。提取的内容存储在电子表格中，以便进一步分析和可视化。

文献综述结果使用基于 PRISMA 2020 开发的基于 R 的工具进行可视化。

文本框 1. 搜索字符串和发布数据库

搜索字符串

- **PubMed**
- ((meta data) OR (meta-data) OR (metadata) OR (ontology) OR (rules)) AND ((extract transform load) OR (ETL) OR (extract load transform) OR (ELT))
- **IEEE Explore**
- ((“All Metadata”:metadata) OR (“All Metadata”:meta-data) OR (“All Metadata”:meta data) OR (“All Metadata”:ontology) OR (“All Metadata”:rules)) AND ((“All Metadata”:ETL) OR (“All Metadata”:extract transform load) OR (“All Metadata”:ELT) OR (“All Metadata”:extract load transform))
- **Web of Science**
- (ALL=(metadata) OR ALL=(meta-data) OR ALL=(“meta data”) OR ALL=(ontology) OR ALL=(rules)) AND (ALL=(ETL) OR ALL=(“extract transform load”) OR ALL=(ELT) OR ALL=(“extract load transform”))
- **Biomed Center (BMC)**
- (“meta data” OR meta-data OR metadata OR ontology OR rules) AND (“extract transform load” OR ETL OR “extract load transform” OR ELT)

文本框 2. 排除标准的标签和描述

不合格缩写

- 发布不包含“提取-转换-加载 （ETL）”作为“提取-转换-加载”。
- 发布不包含“提取-加载-转换 （ELT）”作为“提取-加载-转换”。

不合格定义

- 发布不会在“源或目标中的数据元数据”的上下文中使用元数据。
- 发布不会在“数据转换规则”的上下文中使用规则。

唯一 ETL/ELT

- 论文仅摘要 ETL/ELT。

唯一元数据

- 论文仅描述元数据。

不合适主题

- 论文提到了元数据和 ETL/ELT，但重点不是数据协调

不合适类型

- 论文不是会议论文或期刊论文

外语

- 论文不是英语和德语书写。

不合适的内容

- 论文没有提及 ETL/ELT 或 metadata

文本框 3 数据字段提取

主题

- 工作主题

元数据驱动方法

- 工作中使用的元数据驱动方法

元数据驱动的方法工具

- 用于执行元数据驱动方法的工具

目的

- 使用元数据驱动方法的目的。

结果

文献检索

文献检索共收录了 538 篇论文。在删除 85 份重项后，在 TAS 阶段筛选了 453 份论文。通过使用文本框 2 除去没有全文的论文，FTS 收录了 64 篇论文。最后，我们将 33 篇论文纳入到本项工作中。筛选过程和结果使用 PRISMA 流程图（图 1）。收录论文的完整列表可在多媒体附录 1 中查阅。

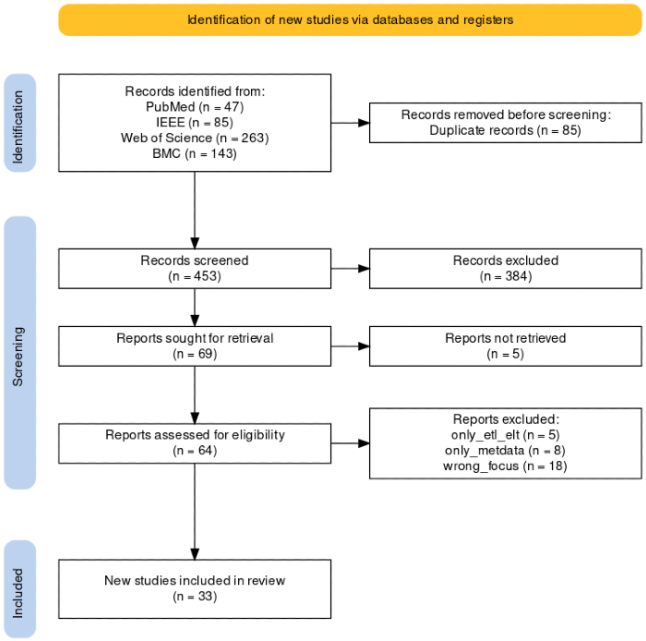


图 1. PRISMA（系统综述和荟萃分析的首选报告项目）流程图

使用基于 R 的工具生成（转载自 Haddaway 等人，并得到 Neal R Haddaway 许可）。

论文发表

为概要了解 MDD 的潜在应用重点（Q1），从而说明这些方法在哪些方面有用，首先评估了应用的重点主题。根据提取的数据，所有纳入论文的重点分为 7 个不同的类别，即医学（n=9， 数据库（n=13），大数据（n=4），工业（n=4），地理信息学（n=1），考古学（n=1）和军事（n=1）。这表明数据库和药物是使用 MDD 方法最多的 2 个类别。

用于各专题类别的 MDD 方法

在各专题类别中使用不同类型的 MDD 方法。为了解这些类型的 MDD 方法在每个类别中的使用情况（Q2），研究 MDD 方法分布。

图 2 不同类型的 MDD 方法在不同主题类别中的应用。最常用的 MDD 方法类型是基于本体的，其中源或目标的本体（如使用资源描述框架）在 ETL/ELT 过程中应用。这种方法用于 6 类，特别是在数据库类中和医学。另一种常用的 MDD 方法是基于规则的，它将基于源和目标生成的转换规则应用于 ETL/ELT 过程。基于规则的方法也被广泛用于数据库类中和医学。除了基于本体论和基于规则的方法外，其他 MDD 方法都被归为“其他”（如表 1）。

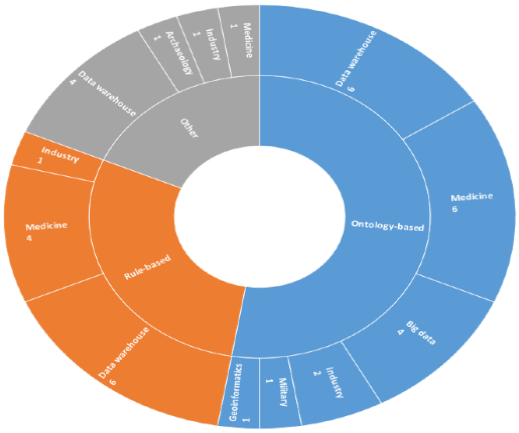


图 2. 每个主题类别都采用元数据驱动的方法

表 1. MDDa 被归类为“其他”的方法

MDD 方法类型和发布	例
基于 UML[b]	
Dhaouadi 等人 [46]	UML 类图用于对转换过程进行建模
基于图形	
Dhaouadi 等人 [46]	BPMN[c] 标准用于对 ETL[d] 过程进行建模
基于临时形式主义	
Dhaouadi 等人 [46]	实体映射图用于表示 ETL 任务
基于 MDA[电子]	
Dhaouadi 等人 [46]	MDA 是一个多层框架，具有多个子模块，用于将功能规范与其实现分离
基于消息	

MDD 方法类型和发布	例
诺瓦克等人 [51]	“普通消息”包含映射和转换信息;“命令消息”配置（执行）系统
基于模板	
McCarthy 等人 [58]	管理复杂转换过程的每个数据源的转换模板
Binding 等人 [62]	模板包含映射模式，然后用于在数据库中进行查询
基于元数据的 [f]	
奥祖尔特和格雷特 [36]	实现通用数据转换语言，将来自多个源的异构数据转换为通用格式
Tomingas 等人 [44]	存储在知识和元数据存储库中的源和目标的元数据
苏莱金和潘菲洛夫 [60]	存储在元数据管理框架中的映射路径的元数据

aMDD：元数据驱动。bUML：统一建模语言。cBPMN：业务流程模型表示法。dETL：提取-转换-加载。eMDA：模型驱动架构。f基于元数据的方法：方法使用元数据，没有任何规范。

使用 MDD 方法进行数据协调的目的

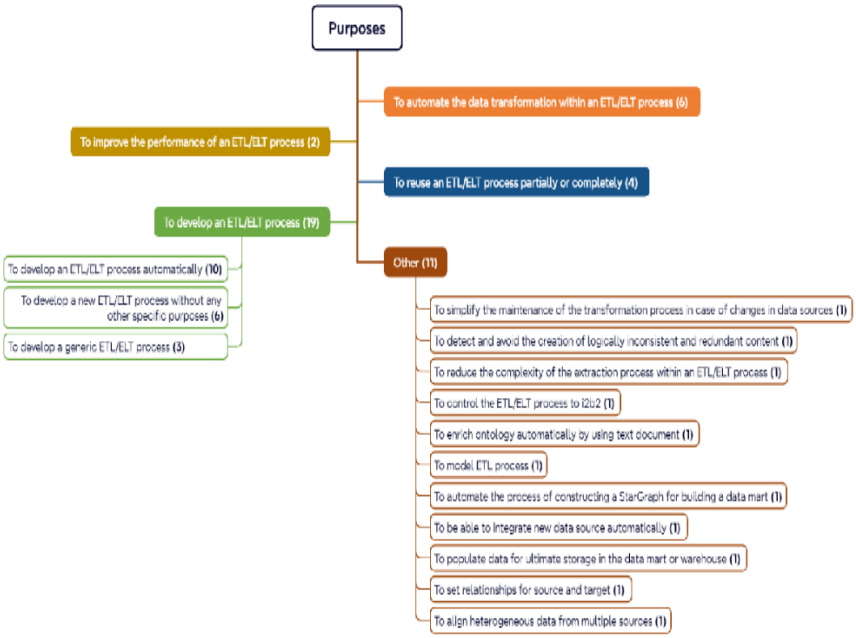


图 3. 在 ETL/ELT 过程中使用 MDD 方法的目的

ELT：提取-加载-转换;ETL：提取-转换-加载;i2B2：整合生物学和床边的信息学;MDD：元数据驱动。

然后调查在每个用例中使用 MDD 方法的目的，以阐明使用 MDD 方法的原因（Q3）。图 3 显示了使用 MDD 方法根据提取的数据开发 ETL/ELT 流程的不同目的。

大多数论文描述了使用 MDD 方法开发 ETL/ELT 过程。此目的可分为三个详细类别：（1） 自动化 ETL/ELT 流程的开发，（2） 开发通用 ETL/ELT 工艺,（3）开发新的 ETL/ELT 工艺，无需任何进一步的技术规范。此外，ETL/ELT 流程的转换部分也可以通过应用 MDD 方法实现自动化。如，Chen 和 Zhao 描述了一种用于自动生成用于数据转换的 SQL 脚本的 MDD 方法。此外，使用 MDD 方法还有助于提高 ETL/ELT 流程的性能或部分或全部重用 ETL/ELT 过程。其他目标（归类为“其他”）。如图 3 所示，简化转换过程的维护并降低提取过程的复杂性，也可以通过在 ETL/ELT 过程中使用 MDD 方法来实现。

用例与使用的 MDD 方法之间的关系

如上一节所示，不同的 MDD 方法被应用于不同目的。为了进一步阐明选择 MDD 方法的原因（Q3），研究了它们之间的关系。表 2 列出了使用一种 MDD 方法实现特定目的的论文数量。基于本体的方法用于实现目的（1）和（2）以及（4）-（7）。例如，Huang 等人为数据转换过程同时创建了本地的本体（基于源数据的本体）和全局本体（查询处理的本体），通过应用本体学习、本体映射和本体规则，可以自动实现从本地本体到全局的数据转换。此外，基于本体的方法还用于实现其他目标，例如控制 ETL 过程以整合生物学和临床并降低提取过程的复杂性。与基于本体的方法类似，基于规则的方法用于实现（1）-（3）和（5）-（7）的目的。由于转换规则的可重用性，还可以通过在流程中应用规则来简化 ETL/ELT 流程的维护。其他 MDD 方法，如基于模板、基于消息和基于元数据的用于实现（1）-（3）和（5）-（7）的目标。基于元数据的方法（如元数据管理框架）可用于自动开发 ETL 任务。详见多媒体附录 1 中表 2。

表 2. 目的与 MDD 之间的关系[a]使用的方法

目的		MDD 方法		
数	描述	基于本体 n/N （%）	基于规则 n/N （%）	其他 n/N （%）
(1)	在 ETL[b]/ELT[c]流程中自动执行数据转换	2/6 (33)	3/6 (50)	1/6 (17)
(2)	重用 ETL/ELT 过程（部分或全部）	1/4 (25)	2/4 (50)	1/4 (25)

目的		MDD 方法		
数	描述	基于本体 n/N （%）	基于规则 n/N （%）	其他 n/N （%）
(3)	提高 ETL/ELT 过程的性能	0/2 (0)	1/2 (50)	1/2 (50)
(4)	开发通用 ETL/ELT 流程	3/3 (100)	0/3 (0)	0/3 (0)
(5)	自动开发 ETL/ELT 流程	5/9 (56)	2/9 (22)	2/9 (22)
(6)	开发新的 ETL/ELT 流程（无任何其他特定目的）	4/6 (67)	1/6 (17)	1/6 (17)
(7)	其他	5/11 (45)	2/11 (18)	4/11 (36)

[a]MDD：元数据驱动。 [b]ETL：提取-转换-加载。 [c]ELT：提取-加载-转换。

用于实施 MDD 方法的工具

图 4. 用于开发元数据驱动方法的工具

MMF：元数据管理框架;OWL：Web 本体语言;YAML：YAML 不是标记语言。

最后，我们重点介绍用于实施 MDD 方法的工具（Q4）。为了实现上一节所示的各种目的，使用了不同的工具。如图 4 所示，每类 MDD 方法都可以通过使用现有工具或特定用例工具来实现。根据纳入的论文，基于本体的方法主要使用 Protégé（斯坦福大学生物医学信息学研究中心）。Protégé 是一种本体编辑器，也是 OntoEdit 器（卡尔斯鲁厄大学 AIFB 研究所和 Ontoprise GmbH）。使用本体编辑

器的主要原因是它的易用性和维护性，以及各种插件。使用特定案例的工具，例如 Kamil 等人，基于关系数据库的数据定义语言生成本体。这两种类型的工具都用于创建和维护本体，然后用于在 ETL/ELT 过程中建立通用映射逻辑。另一种常用的 MDD 方法是基于规则的，用于表述和存储转换规则。转换规则可以存储在映射表、CSV 文件、YAML（YAML Ain't Markup Language）文件或数据库中的表，手动实现。之后，可以在 ETL/ELT 过程中使用转换规则，例如，启用自动转换。其他类型的 MDD 方法也可以通过使用现有工具（如知识和元数据存储库）或特定用例的工具（如元数据存储库和元数据管理框架)，如 Ozyurt 和 Grethe 使用 bioCADDIE 数据标签套件（bioCADDIE 项目）实现了通用转换语言（元数据架构）来统一来自多个来源的异构数据，这为进一步分析查询提供了基础。

讨论

主要发现

我们关于“元数据驱动的 ETL/ELT”主题的文献综述包括 2012 年-2022 年在 PubMed、IEEE Explore、Web of Science 和 Biomed Center 上列出的所有关于 MDDETL/ELT 过程的论文。在某些情况下，元数据的使用专门使用“本体”或“规则”来表示。因此，我们在搜索字符串中添加了“ontology”和“rules”以扩大搜索范围。

通过摘要审查过程，我们能够概述应用 MDD ETL/ELT 流程的主题类别（问题 1），ETL/ELT 流程中使用的 MDD 方法类型（Q2），使用 MDD 方法的目的（Q3），以及用于实施 MDD 方法的工具（Q4）。

在所有主题类别中，基于本体和基于规则的方法是数据库和医学主题类别中最常用的方法。在某些情况下，在 ETL/ELT 过程中使用了不止一种 MDD 方法。例如，Del Carmen Legaz-García 等人同时使用了基于本体和基于规则的方法。因此，此类论文被归类为两种 MDD 方法类型。

可以使用各种工具来实现 MDD 方法。不幸的是，我们无法从所有收录的论文中提取这些信息。其原因是一些论文使用了专有或不可转让的方法（例如，特定于数

据的本体和 Data Vault [DataVaultAlliance]).其他一些论文没有明确提及或描述他们使用的工具。因此这些论文不包括在 MDD 工具分析中。

结果表明，通过应用 MDD 方法，实现通用 ETL/ELT 过程以自动将不同的 FHIR 配置文件转换为 OMOP CDM 是有希望的。但是在结果中并没为此提供一个很小的解决方案。例如，Huang 等人使用基于本体的方法能够在 ETL/ELT 过程中自动执行数据转换，而 Ong 等人使用基于规则的方法来实现相同的目的。在某些情况下，为了完成数据转换，使用了不止一种 MDD 方法作为补充。例如，Pacaci 等人选择了基于本体的方法来自动执行数据转换，并选择了基于规则的方法来简化数据源更改时转换过程的维护。通过结合应用这两种方法，作者能够将异构 EHR 系统中的 EHR 数据转换为 OMOP CDM。因此，确定适当的 MDD 方法和工具来实施通用 ETL/ELT 流程以自动将 FHIR 转换为 OMOP CDM 仍然是一个挑战。

这项工作旨在概述不同类型的 MDD 方法及其工具。因此，本综述缺乏对每种 MDD 方法的具体特征的详细分析。这一差距凸显了全面了解本研究中提出的 MDD 方法的特征的重要性。该分析将在未来进行，为选择最合适的 MDD 方法和工具提供坚实的证据，或考虑结合使用多种 MDD 方法来实施将 FHIR 转换为 OMOP CDM 的通用 ETL/ELT 流程。

结论

本文献综述表明，使用 MDD 方法开发 ETL/ELT 流程可以在不同聚焦群体（即医学、数据库、大数据、工业、地理信息学、考古学和军事）中服务于不同的目的。结果表明，通过应用 MDD 方法实现从 FHIR 到 OMOP CDM 的数据转换的自动化，实现 ETL/ELT 过程是有希望的，但是确定适当的 MDD 方法和工具来实施这种 ETL/ELT 过程仍面临挑战。由于对本研究中提出的 MDD 方法的特征缺乏全面的了解，所以，我们的下一步是评估本研究中提出的 MDD 方法，并确定最合适的 MDD 方法以及将它们集成到 MII CDSFHIR 到 OMOP CDM ETL 流程中的方法。这可以验证使用 MDD 方法推广 ETL 过程以协调医疗数据的能力。

***注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。**

译文二：

医疗保健中合成时间序列生成方法的比较评估：利用患者元数据进行准确的数据合成

Imanol Isasa, Mikel Hernandez, Gorka Epelde, Francisco Londoño, Andoni Beristain, Xabat Larrea, Ane Alberdi, Panagiotis Bamidis & Evdokimos Konstantinidis, 徐健(译)

背景

在医疗保健应用程序数字化日益精简的信息时代，隐私正演变为越来越宝贵的资产。数据保护法，如欧洲的 GDPR(通用数据保护条例)或美国的 HIPAA(健康保险流通与责任法案)，通过定义数据主体的权利、数据处理代理的责任或各种行为准则来规范个人身份信息(PII)的处理。然而，针对理论上安全的私有数据集的攻击正变得越来越复杂，并设法打破安全壁垒，这迫使数据收集者、控制器和处理器实施更严格的隐私保护。

合成数据生成(SDG)是一种相对较新的主流技术，可以帮助应对数据泄漏风险。该技术包括使用生成模型、统计模型或基于专家知识的模型创建真实数据(RD)的人工版本，这些模型可以学习潜在的变量分布和多元相关性。一旦训练 SDG 模型，就可以按照真实数据集的原始模式，合成尽可能多的数据记录。然而，即使可持续发展目标技术提供了一定隐私保障，当前关于该主题的文献仍朝着生成的合成数据(SD)是否考虑了个人数据的方向发展。

在可用于可持续发展目标的不同部门中，医疗保健在敏感信息方面至关重要，因此是需要控制和监测的最受损害的信息源之一。在这种情况下，数据通常有两个组件，第一个组件被分类为主题的元数据(静态组件、不随时间变化)，第二个组件是时间序列或纵向部分(动态组件、跨时间轴测量)。到目前为止，临床可持续发展目标主要集中在生成私有和有用的表格静态数据，而纵向成分以及两者的联合生成则相对滞后。考虑到这种情况，值得注意的是，需要确保上述两个组件的隐私。元

数据与时间序列的结合可能有助于重新识别和受损的数据泄露，但合成时间序列带来了固有的挑战，因为除了通常的变量间相关性之外，数据生成器必须学习时间相关性。

2022 年，我们针对合成时间序列生成（STSG）任务提出了两种不同的方法，其中将真实和合成元数据和时间序列的不同组合合并在一起，以找到实现有用时间序列的最佳方法。在本工作中，对这两种方法进行了更广泛的评估，评估了（1）SD 与 RD 的相似性 SD、（2）对构建下游应用程序的有用性，以及（3）生成的信息的私密性。这些成分的统一和平衡在目前的文献中仍然是一个悬而未决的问题，因为三者之间存在着复杂的关系。强有力的隐私保证与 SD 相对于 RD 的较高偏差有关，这使得前者的效用和相似性降低。相比之下，保持较高的数据效用会导致 SD 受损，因为由于 SD 和 RD 之间的广泛相似性，个人信息可能会从中提取。此外，每个组件的相关性取决于目标应用，因为有一些关键的例子需要解决隐私问题并牺牲 SD 的相似性和实用性，但情况并非总是如此。此外，我们在 Isasa 等人中评估的模型被纳入第三种 STSG 方法，其中元数据和时间序列使用单个模型联合生成。根据我们之前的研究，据我们了解，这是第一个评估三种不同变体以生成结合时间序列和相关元数据的 SD 的工作。采用通用的评估程序来评估结果的相似性、实用性和隐私性，从而为该领域做出重大贡献。

本文档的其余部分组织如下："方法"部分详细介绍了如何使用多个模型和数据集组合生成 SD，以及在生成步骤后如何耦合和评估数据。接下来，介绍评估结果，然后进行讨论，以及在这项研究中发现的局限性列表。最后，给出了比较的总体结论。

方法

本节首先介绍用于这项工作的数据集以及准备数据所需的预处理步骤。然后，介绍了三种不同的 STSG 方法，然后是使用的 SDG 模型。最后，对生成的 SD 进行了评估。

数据源

为了确保结果的一致性，没有数据依赖性，实验使用来自 *Physionet* 的两个不同数据集和第三个名为 PMData 的生命记录数据集进行。

首先，考虑了一个数据集，该数据集由业余和专业运动员进行的 992 次跑步机最大运动测试（Treadmill Maximal Exercise Tests， TMET）中获得的心肺测量值组成。这些实验包括分级运动测试（GET）的纵向测量以及与受试者相关的元数据，它们在 *Powerjog J 系列*跑步机上进行，该跑步机连接了 *CPX MedGraphics* 气体分析仪和 *Mortara* 12 导联心电图设备。有了这些，记录了受试者的心率（HR）以及耗氧量（VO_2）、二氧化碳生成（VCO_2）和肺通气措施（呼吸频率，RR;呼出量，VE）。至于元数据部分，获得了受试者的年龄（M：27.10，Q1-Q3：21.10-36.32）、性别（男女比例为 5.66：1）、身高（M：175.00，Q1-Q3：170.00-180.00）和体重（M：73.00，Q1-Q3：66.00-80.23），以及空气温度（M：22.90，Q1-Q3：20.80-24.40）和湿度（M：47.00，Q1-Q3：42.00-54.00）等环境测量值。GET 包括 5 km/h 速度的预热期，并施加 0.5 至 1 km/h 的速度增量，直到消耗的氧气量饱和。受试者以 5 公里/小时的初始速度恢复了努力。该研究由马拉加大学运动生理学和人类表现实验室发表，符合所有相关国家法规，并遵循《赫尔辛基宣言》的原则进行。研究方案由马拉加大学研究伦理委员会批准，并收集了所有参与者或其法定代表的书面知情同意书。

其次，低血压亚群来自 MIMIC-III v1.4（重症监护医学信息市场）数据库。该数据库由 2001 年至 2012 年间在波士顿贝斯以色列女执事医疗中心重症监护室住院的患者的去识别化健康相关数据组成。使用两种不同的重症监护信息系统来记录患者信息：*飞利浦 CareVue 临床信息系统*（飞利浦医疗保健，马萨诸塞州安多弗）和 *iMDsoft Metavision ICU*（iMDsoft，马萨诸塞州尼德姆）。使用这些，舒张压和收缩压，吸入氧的分数（FiO_2）、尿量和给药血管加压药均有监测。此外，还记录了格拉斯哥昏迷量表 （GCS） 分数，并将纵向数据与随后的性别和种族元数据合并。通过选择平均动脉压（Mean Arterial Pressure， MAP）低于 18mmHg 的成年患者（>65 岁）获得低血压亚群。所有数据均符合 HIPAA 的数据保护要求，该项目已获得 Beth Israel Deaconess 医疗中心（马萨诸塞州波士顿）和麻省理工学院（马萨诸塞州剑桥）的机构审查委员会的批准。

第三，将一个名为 PMData 的生命周期记录数据集纳入分析。2 人（16 名男性

和 13 名女性）携带 *FitBit Versa 3* 智能手表腕带五个月进行数据收集。这些信息丰富了 *PMSys* 运动记录应用程序数据，以及一组对旨在收集静态信息（年龄、性别、身高以及受试者是否具有 A 型或 B 型人格）的问卷的回答。另一方面，在研究期间获得的时间序列范围从每周记录到每五秒测量一次。本研究的所有参与者都签署了知情同意书，该研究是在挪威研究数据中心（NSD）的评估下根据挪威和欧盟数据保护法进行的。

数据预处理

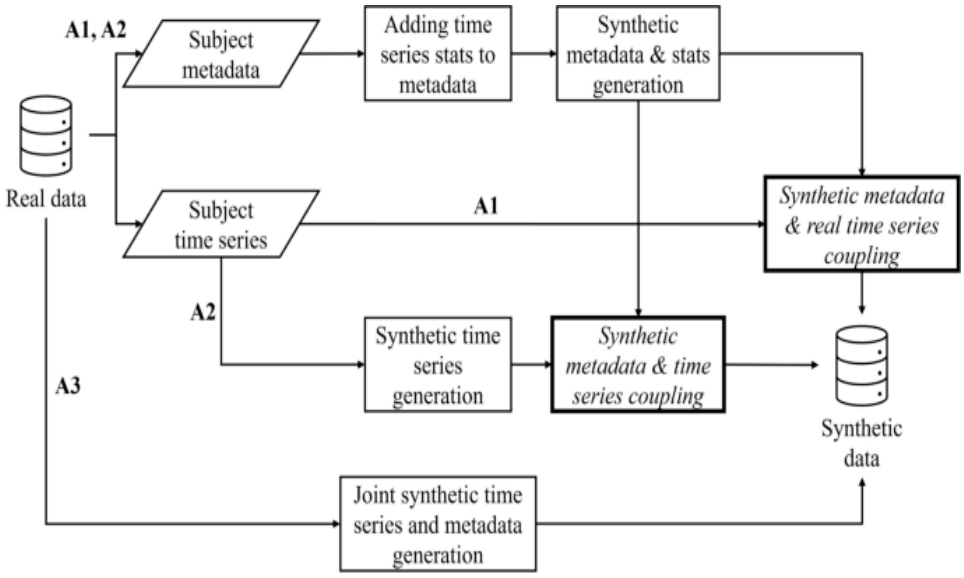
首先，按照 Larrea 等中定义的步骤对 TMET 数据集进行预处理。从 992 个主题数据集开始，其中 30 个由于主题元数据上的缺失值而被直接排除。关于纵向分量，排除了超过 30 个缺乏数据点的实验，而其他实验则进行线性插值以避免缺失值。由于发现 HR 变量缺少 942 个样本，并且两个 VO 都缺少 4,871 个样本和 VCO_2 变量方面，又有 12 个实验被排除在研究之外。最终的 TMET 数据集由 950 个实验组成，总共有 216,600 个时间戳，序列长度为 228 个样本。

其次，从 MIMIC-III 派生的低血压子集是通过直接查询 Google Cloud Platform（GCP）生成的。收缩压或舒张压测量值缺失的受试者被排除在最终数据集之外。假设其余测量值的缺失值不会改变受试者的先前状态，则使用一种称为前向填充的技术来估算时间序列。该技术使用每个变量的最后一个有效测量值来确定缺失单元格的值，直到找到新的有效测量值。此数据集上的分类变量针对 DGAN 进行了标签编码，因为这是特定于模型的要求。对于 WGAN-GP 来说，单热编码被认为更好，不会对模型造成误解。性别变量被二元化，而种族则被聚类为五个类别。最后，数据集中包括 674 名受试者，每个受试者由 48 小时测量组成，总共产生 32,352 个时间戳。

最后，通过从纵向信息中选择心率变量（每五秒采集一次测量值）并将其连接到相关的元数据（年龄、身高、性别和个性）来构建基于 PMData 的实验数据集，开始预处理步骤，数据集大小为 20991392 个样本和 10 个变量。考虑到受试者包含可变长度的心率序列，选择最短序列的长度来截断其余的序列。其余数据集由 155936634 个样本组成，每个受试者的时间序列长度为 74650 个样本。最后，由于计算资源的限制，合成时间序列的生成过程必须限制在 50000 个样本的长度。

合成时间序列生成方法

在这项研究中，提出并比较了三种不同的合成时间序列生成方法，即合成元数据生成和实时序列耦合（A1）、单独的合成元数据和时间序列生成和后期耦合（A2）以及联合合成元数据和时间序列生成（A3）。Larrea 等的前两种方法在手稿中较早地介绍过，Isasa 等也考虑过第三种方法。在本节中，可以找到有关不同方法的更详细解释，如图所示。1，以图形方式表示每个方法生成 SD 所遵循的工作流。



本研究中比较的三种方法的工作流程。加重的方框是指 A1 和 A2 的耦合步骤

第一种方法（A1）假设在实时序列数据中没有私人信息被披露。这种方法的灵感来自 Schiff 等人，该方法基于用实时序列丰富合成主题元数据。为此，从实际测量序列计算有意义的统计测量值（每个时间序列的最大值、最小值和平均值），以将其视为元数据。接下来，生成元数据的合成对应物，由合成主题元数据和合成时间序列统计测量值组成。最后，通过将生成的值与真实值配对，将合成元数据与实时序列相干耦合，使它们之间的距离尽可能小。

对于第二种方法（A2），从实时序列集计算出相同的有意义的统计测量值。但是，在这种情况下，将单独生成一组合成的时间序列。然后，在合成元数据和单独生成的合成时间序列之间进行相同的相干耦合。

最后，第三种方法（A3）假设主题元数据对生成的时间序列的质量产生积极影响。在这种情况下，主题元数据和时间序列将在一个步骤中生成。

合成数据生成模型

使用两个不同的 SDG 模型对前一节中介绍的方法进行了测试，这些模型可以将元数据合并到时间序列生成任务中。两个模型的超参数值均在实验中选择，并且在近似和数据集组合中保持不变。

带有梯度惩罚的 Wasserstein GAN (WGAN-GP)在 Health Gym 项目中首次与附加的对齐损失特征一起使用[22]。对齐损失通过计算真实数据集和合成数据集中变量间相关系数差的 L1 范数来形成生成器损失函数(\({\mathcal{L}}_{G}\))，定义如下:

$${\mathcal{L}}_{G}=-\mathbb{E}\left[D\left(G\left(z\right)\right)\right]+\lambda \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{i-1}{\parallel{\widehat{r}}_{i,j}-{r}_{i,j}\parallel}_{L1}$$

其中\(\lambda\)为对齐系数，\({\widehat{r}}_{i,j}\)为两个综合生成变量之间的相关系数，\({r}_{i,j}\)为实际对应变量。有了这个损失函数，如果生成的数据没有适当地相关，生成器网络就会受到更大的惩罚。

对于 TMET 数据集，WGAN-GP 使用 64 个样本的批量大小进行了 5000 次训练，而对于低血压子集，选择了 32 个样本的批量大小。对于 PMData 数据集，由于受试者数量较少，因此使用批量大小为 2。对于所有数据集，评论家网络的更新速度是生成器网络的 5 倍，其恒定学习率为 0.001。此外，统一系数保持恒定值为 10.0。

本研究中使用的第二个生成模型是 DöppelGANger 或 DGAN。DGAN 首先生成元数据，然后在其上有条件地生成时间序列。多层感知器(multilayer Perceptron, MLP)生成元数据组件，并在每个时间步将输出并入递归神经网络(Recurrent Neural Network, RNN)，RNN 的任务是生成纵向数据。在 DGAN 中，批处理生成用于每次 RNN 迭代生成 S 条记录，该过程旨在从时间序列中捕获长期影响。此外，通过添加一个辅助鉴别器来单独关注元数据，将复杂的鉴别器体系结构合并到模型中。组合鉴别器负责最终将样本分类为真假。用于训练 DGAN 模型的批量大小与用于

WGAN-GP 模型的批量大小相等。TMET 数据集的 S 值为 12，低血压子集的 S 值为 6,PMData 数据集的 S 值为 127。对于这些模型，在 5000 次的训练过程中，也使用了恒定的学习率 0.001，而对于 DGAN 模型，梯度惩罚系数设置为 10.0。

合成数据评估

使用不同方法生成的 SD 根据不同的指标进行了评估，量化了其与原始数据的相似性（1），其用于构建下游应用程序的效用（2）以及合成生成的数据的隐私级别（3）。

由于生成模型每次使用时都会产生不同的数据资产，因此在评估管道中加入了伪交叉验证方法，以消除因这方面而产生的潜在偏差。每个模型和使用的数据集生成八个不同的等维合成数据集，并用作该方法的伪交叉验证折叠。通过计算不同折叠的平均值，将各评价指标和方法的结果结合起来。该分析中使用的评估方法主要基于 Hernandez 等，他们提出了一系列客观和通用的指标来评估和比较 SDG 模型和方法。

这项工作中计算的相似性指标旨在评估合成数据集相对于其真实对应数据集的相似程度，同时考虑它们之间的统计、分布和相关特征。为此，首先进行多变量分析以检查变量间的相关性。使用 *Pearson 相关系*数评估数值变量对，而使用从先前的\（{\chi }^{2}\）检验计算的 *Cramér's V 值*检查分类变量对。使用数值特征相对于分类特征的线性回归的\（{R}^{2}\）系数评估混合类型相关性。使用伪交叉验证过程中每个折叠的平均值从 SD 生成最终的相关矩阵。使用余弦相似度指标将最终矩阵与对应于真实数据集的相关性进行比较，余弦相似度量定义为\（1-{d}_{cos}\），其中值越高表示相关性保留越好。

$${d}_{cos}（x，x'）= \frac{\sum _{i=1}^{n} {x}_{i}\cdot{x}_{i}^{'}} {\sum _{i=1}^{n}{x}_{i}\cdot\sum _{i=1}^{n}{x}_{i}^{'}}$$

其中\（{x}_{i}\）和\（{x}_{i}^{{\prime }}\）都是指要比较的样本。

评估 SD 与真实数据相似的另一种方法是数据标记分析（DLA）。这个过程包括训练几个分类模型（随机森林、k 最近邻、决策树、支持向量机和 MLP，在本例中）以查看它们是否可以区分合成样本和真实样本，模拟 GAN 判别器。有关如何配置每个分类模型的详细信息，请参见表 1。使用受试者工作特征曲线下面积

（AUROC）评估预测性能，随机预测意味着值约为 0.5，并且模型无法区分来自不同来源的样本。

表 1 分类模型参数说明

Model	Configuration
RF	Number of trees: 100, split criteria: Gini index
KNN	Number of neighbors: 10
DT	Split criteria: Gini index
SVM	Regularization parameter: 100, maximum epochs: 300, kernel type: linear
MLP	Maximum epochs: 100, hidden layers: (a) 64 neurons and (b) 32 neurons, activation: relu, solver: adam, batch size: 200, learning rate: 0.001

根据 Sajjadi 等人的提议，还生成了精确召回分布（PRD）图，以便使用单个数值对轻松量化模型的质量。在他们的出版物中，提出了精确度\（\alpha\）和召回率\（\beta\）的概念来比较两个分布，前者近似于生成样本的质量，后者测量它们之间的重叠比例。\（\alpha\）和\（\beta\）都可以通过计算最大 F 分数合并到一个指标上，如下式所示：

$${F}_{\gamma} = \left（1+{\gamma }^{2}\right）\frac{\alpha \cdot \beta }{\left（{\gamma}^{2}{\alpha} \right） + \beta}$$

其中\（\gamma\）是一个正实数因子，它根据精度来权衡召回率。对于该分析，根据经验选择了\（\gamma\）值 8 进行评估，对召回的权重更高，尽管它并不重要，因为相对分析优先于绝对评估。此外，\（\gamma\）值的调整将导致值的绝对重新定位，而不会改变方法或模型之间的相对差异。

数据效用是指可持续发展局为分析或建模目的而有效模仿原始数据的程度。关于该定义，在这项工作中，对真实综合测试车（TSTR）设置与真实车真实测试（TRTR）设置进行了测试。在这些模型中，各种机器学习（ML）模型孪生体首先在 RD 上训练，然后分别在 SD 上训练。每个双胞胎都根据相等的真实测试集进行测试。在这两种情况下，都为 ML 模型定义了一个回归任务，将 VE 和 GCS 变量作为 TMET 数据集和低血压子集的目标。在这项工作中，使用平均绝对误差（MAE）评估每个模型的性能，并使用先前提出的余弦相似度指标将 SD 上训练的模型与 RD 对应模型进行比较。此相似性指标与 MAE 相似度成正比，MAE 相似度提供有关合成数据效用的信息。

使用 *t* 检验评估观察到的相似性和方法之间效用度量差异的显着性。对于设置为 0.05 的显著性水平（SL），低于该值的 *p* 值表示原假设被拒绝，这意味着所测试

的差异具有统计显著性。

最后，在可持续发展目标的背景下，隐私可以定义为生成的数据通过防止真实个人的成员资格在原始数据集中被披露来保护敏感信息的程度。在此分析中，使用成员资格推理攻击（MIA）设置评估了隐私。在这次攻击中，攻击者被认为拥有用于训练生成模型的原始数据的一部分，范围从 10%到 50%不等。此外，SD 被认为是公开可用的，因此对手可以访问，因为它是 SD 的优势之一。对于每个生成的合成样本，计算相对于原始数据集中每个真实样本的余弦距离。距离值 0.2 用作阈值，以考虑足够相似的合成样本，以将其标记为来自用于训练生成模型的 RD 的样本。作为纵向数据设置，两个数据集之间的单一匹配导致受试者被标记为已公开。

结果

在本节中，将介绍相似性、效用和隐私分析的结果。为了清晰和简单起见，我们在本节中引用的模型下标为字母 T、M 或 P，指的是 TMET、MIMIC 或 PMData，它们表示每种情况下用于训练模型的数据集。此外，在表 2 和表 3 中可以观察到从分析期间执行的所有 *t* 检验计算得出的 *p* 值。

表 2 相似度量在方法对之间表现的 *T 检验 p* 值

Model	Approach comparison	DLA AUROC[a]	Correlation Similarity[a]	Autocorrelation MAE[a]	\(\boldsymbol{F}_{\boldsymbol{\gamma}}\)[a]
$DGAN_T$	A1-A2	~ 1.00	3.39×10^{-8}	9.33×10^{-4}	3.47×10^{-8}
	A2-A3	7.46×10^{-1}	7.25×10^{-12}	1.10×10^{-4}	1.80×10^{-6}
	A3-A1	7.90×10^{-1}	5.83×10^{-6}	2.04×10^{-1}	1.26×10^{-9}
$WGAN\text{-}GP_T$	A1-A2	8.36×10^{-1}	5.56×10^{-11}	~ 0.00	2.30×10^{-11}
	A2-A3	4.44×10^{-1}	2.22×10^{-16}	~ 0.00	9.11×10^{-12}
	A3-A1	2.64×10^{-1}	~ 0.00	~ 0.00	5.37×10^{-1}
$DGAN_M$	A1-A2	4.20×10^{-13}	9.17×10^{-6}	5.67×10^{-2}	1.82×10^{-10}
	A2-A3	4.19×10^{-4}	5.50×10^{-2}	~ 0.00	3.15×10^{-1}
	A3-A1	8.90×10^{-14}	3.34×10^{-7}	~ 0.00	8.88×10^{-16}
$WGAN\text{-}GP_M$	A1-A2	1.36×10^{-6}	1.18×10^{-7}	1.46×10^{-9}	3.42×10^{-3}
	A2-A3	5.60×10^{-5}	2.20×10^{-11}	~ 0.00	7.31×10^{-1}
	A3-A1	1.01×10^{-9}	~ 0.00	~ 0.00	9.92×10^{-12}
$DGAN_P$	A1-A2	1.00	1.379×10^{-1}	4.001×10^{-10}	1.811×10^{-6}
	A2-A3	1.00	1.806×10^{-6}	4.563×10^{-10}	3.659×10^{-3}
	A3-A1	1.00	2.114×10^{-5}	4.911×10^{-12}	1.029×10^{-9}
$WGAN\text{-}GP_P$	A1-A2	1.00	3.718×10^{-3}	3.044×10^{-12}	8.882×10^{-16}
	A2-A3	1.00	8.677×10^{-1}	~ 0.00	1.222×10^{-7}
	A3-A1	1.00	1.510×10^{-4}	4.441×10^{-16}	4.996×10^{-14}

关于相似性，DGANT 模型在 A1 PRD 图上产生了近乎完美的精度(\(\alpha\): 0.975-0.987)和召回率(\(\beta\): 0.974-0.994)值，以及 DGANM 模型(\(\alpha\): 0.983-0.990, \(\beta\): 0.984-0.994)。DGANP 模型在很大程度上匹配了之前模型的召回率(\(\beta :\) 0.953-0.985)和精度指标，即使后者在不同迭代之间的可

变性很高(\(\alpha :\) 0.643-0.941)。A2 对应模型在每个伪交叉验证 fold 中显示出更高的偏差,导致 TMET 数据集的精度值在 0.918 至 0.970 之间,召回值在 0.785 至 0.893 之间，低血压子集的精度值在 0.906 至 0.939 之间，召回值在 0.705 至 0.813 之间,PMData 数据集的精度值在 0.40 至 0.877 之间,召回值在 0.472 至 0.794 之间。DGAN 的 A3 方法注意到数据集之间的进一步差异,导致 TMET 数据集(\(\alpha\): 0.984-0.987, \(\beta\): 0.933-0.950)相对于低血压子集(\(\alpha\): 0.966-0.976, \(\beta\): 0.720-0.763)的性能更好，PMData 数据集(\(\alpha\): 0.160-0.297, \(\beta\): 0.425-0.704)的性能更好。

表 3 效用度量的方法对之间表现的 T 检验 *p* 值

Model	Approach comparison	TSTR metric[a]
$DGAN_T$	A1-A2	6.46×10^{-1}
	A2-A3	9.31×10^{-1}
	A3-A1	7.23×10^{-2}
$WGAN\text{-}GP_T$	A1-A2	2.83×10^{-1}
	A2-A3	3.65×10^{-1}
	A3-A1	7.22×10^{-1}
$DGAN_M$	A1-A2	1.04×10^{-1}
	A2-A3	5.51×10^{-1}
	A3-A1	3.08×10^{-1}
$WGAN\text{-}GP_M$	A1-A2	6.10×10^{-1}
	A2-A3	$\mathbf{3.56 \times 10^{-3}}$
	A3-A1	$\mathbf{8.98 \times 10^{-3}}$
$DGAN_P$	A1-A2	$\mathbf{2.02 \times 10^{-2}}$
	A2-A3	1.10×10^{-1}
	A3-A1	$\mathbf{6.20 \times 10^{-4}}$
$WGAN\text{-}GP_P$	A1-A2	**~ 0.00**
	A2-A3	$\mathbf{5.02 \times 10^{-3}}$
	A3-A1	$\mathbf{2.33 \times 10^{-4}}$

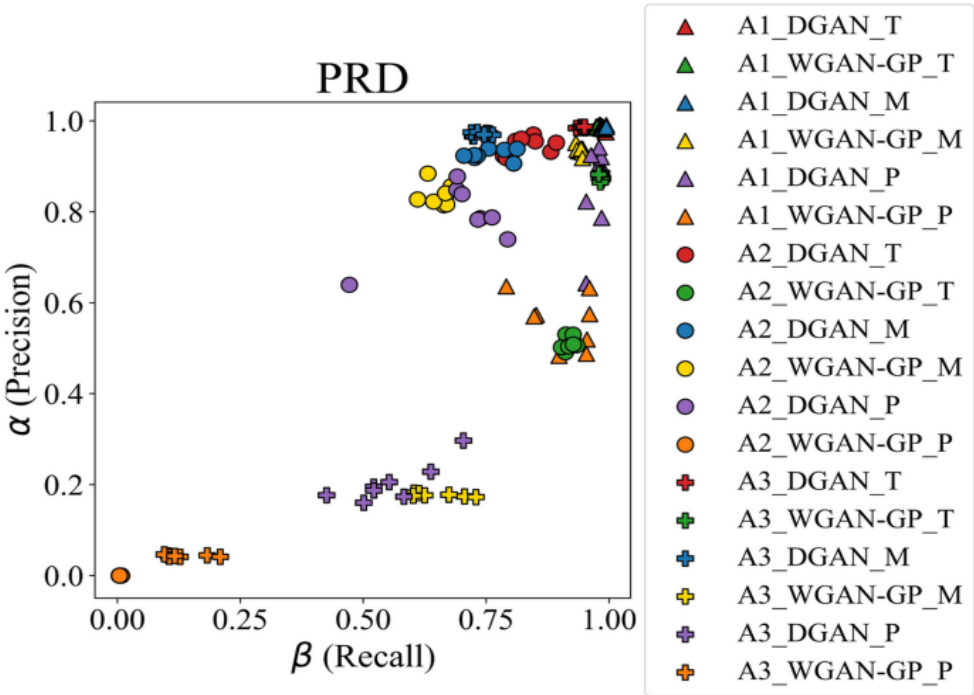
对于 WGAN-GPT 模型，经 A1 方法的伪交叉验证评价，召回值在 0.974 ～ 0.987 之间，精度值在 0.987 左右。A2 的\(\beta\)值在 0.903 ～ 0.934 之间，\(\alpha\)在 0.491 ～ 0.530 之间波动，A3 的值在 0.979 ～ 0.986 之间，A3 的值在 0.865 ～ 0.886 之间波动。对于 WGAN-GPM 模型，\(\beta\)的 A1 耦合结果在 0.933 ～ 0.947 之间，\(\alpha\)的 A1 耦合结果在 0.918 ～ 0.951 之间。A2 方法的查准率和查全率都较差，前者在 0.814 ～ 0.884 之间，后者在 0.610 ～ 0.679 之间。第三，A3 的联合生成导致\(\beta\)的值在 0.932 ～ 0 之间。0.947 和\(\alpha\)值在 0.918 和 0.951 之间。

表 4 相似度指标结果

Generation Method	Model[a]	DLA AUROC	Correlation Similarity	Autocorrelation MAE	\({\boldsymbol{F}_{\boldsymbol{\gamma}}}\)
A1	$DGAN_T$	0.902 ± 0.010	0.880 ± 0.010	2.601 ± 0.176	**0.987 ± 0.006**
	$WGAN\text{-}GP_T$	0.900 ± 0.004	**0.979 ± 0.004**	2.575 ± 0.058	0.980 ± 0.004
	$DGAN_M$	**0.744 ± 0.012**	**0.929 ± 0.014**	1.526 ± 0.043	**0.989 ± 0.005**
	$WGAN\text{-}GP_M$	**0.814 ± 0.016**	**0.942 ± 0.015**	1.527 ± 0.046	**0.942 ± 0.005**
	$DGAN_P$	1.000 ± 0.000	0.624 ± 0.091	1.004 ± 0.174	**0.970 ± 0.015**
	$WGAN\text{-}GP_P$	1.000 ± 0.000	**0.743 ± 0.049**	0.998 ± 0.074	**0.893 ± 0.063**
A2	$DGAN_T$	0.902 ± 0.008	0.823 ± 0.011	3.190 ± 0.358	0.836 ± 0.039
	$WGAN\text{-}GP_T$	0.903 ± 0.040	0.951 ± 0.002	**4.160 ± 0.023**	0.908 ± 0.010
	$DGAN_M$	0.890 ± 0.011	0.645 ± 0.118	**1.582 ± 0.063**	0.758 ± 0.040
	$WGAN\text{-}GP_M$	0.898 ± 0.025	0.797 ± 0.039	**1.842 ± 0.045**	0.656 ± 0.230
	$DGAN_P$	1.000 ± 0.000	0.550 ± 0.097	**1.944 ± 0.005**	0.699 ± 0.097
	$WGAN\text{-}GP_P$	1.000 ± 0.000	0.657 ± 0.050	**1.573 ± 0.004**	0.000 ± 0.000
A3	$DGAN_T$	**0.901 ± 0.003**	**0.906 ± 0.003**	**2.518 ± 0.006**	0.945 ± 0.006
	$WGAN\text{-}GP_T$	**0.890 ± 0.024**	0.894 ± 0.003	5.789 ± 0.074	**0.981 ± 0.002**
	$DGAN_M$	0.871 ± 0.004	0.742 ± 0.057	3.985 ± 0.036	0.742 ± 0.017

最后，在 WGAN-GPP 模型中，A1 方法获得了最好的结果，在\(\alpha\)的 0.483-0.636 和\(\beta\)的 0.791-0.960 之间变化。A3 是 PMData 数据集的第二好方法(\(\alpha\): 0.041-0.047, \(\beta\): 0.095-0.209)，其次是 A2，未能达到覆盖真实分布和原始数据质量(\(\alpha\): 0, \(\beta\): 0.003-0.008)。在图 2 中，可以观察到每个伪交叉验证\(\alpha\)和\(\beta\)值对的 PRD 图，结合这些值的\({F}_{8}\) -得分如表 4 所示，可以观察到 A1 优于其他两种方法。

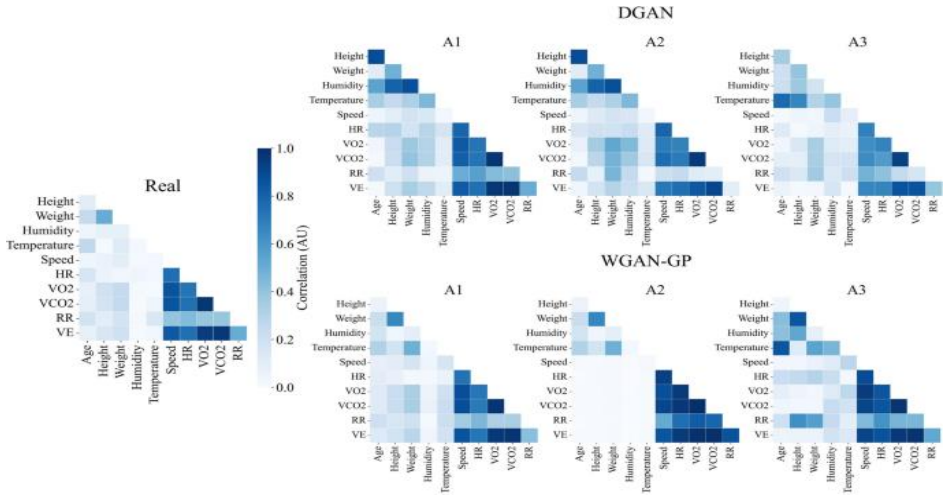
图 2



每个生成的数据集的精度和召回率值对。每个模型和数据集组合有 8 个标记。召回率指标相当于实分布和合成分布之间的重叠面积，而精度是指样品质量

DGAN 的多变量分析 $_T$ 模型表明，最真实的相关矩阵是使用 A3 方法实现的，如图 3 所示和在表 4 中进行了数字确认。关于 WGAN-GP$_T$ 模型中，A1 方法的相关性最好，A2 方法的相关性最差。图 3 显示了 A2 中受试者的元数据如何与时间序列集不相关。

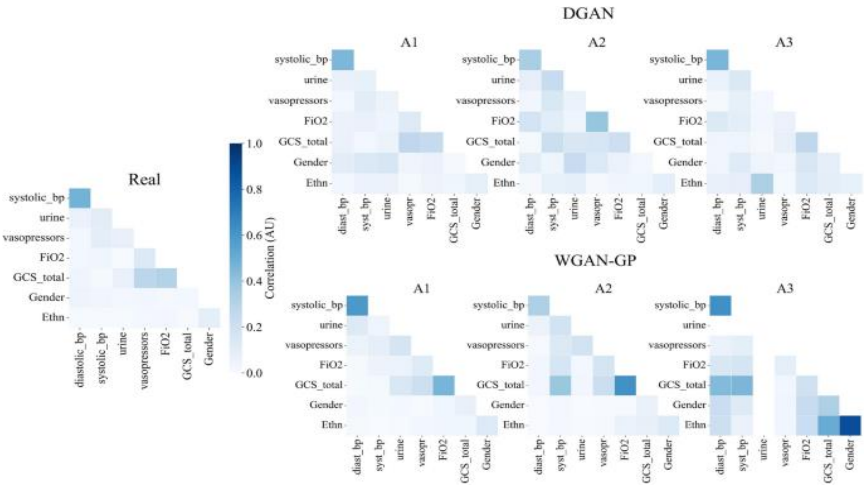
图 3



真实（左）和合成 TMET 数据集的多元相关性。上行显示使用 DGAN 生成的数据，而下行显示使用 WGAN-GP 生成的数据。右边的三列指的是合成过程中采用的不同方法。HR：心率;配音 $_2$：氧气量;VCO 系列 $_2$：二氧化碳量;RR： 呼吸频率;VE： 呼出量

至于 DGAN$_M$ 模型中，A1 是性能最好的模型，如图 4 所示并在表 4 中得到证实。TMET 数据集上的 WGAN-GP 对应物显示出与数据合成性成正比的变相关性恶化，在 A1 上表现最好，在 A3 上表现最差，就像 WGAN-GP 模型在低血压子集上训练一样。

图 4

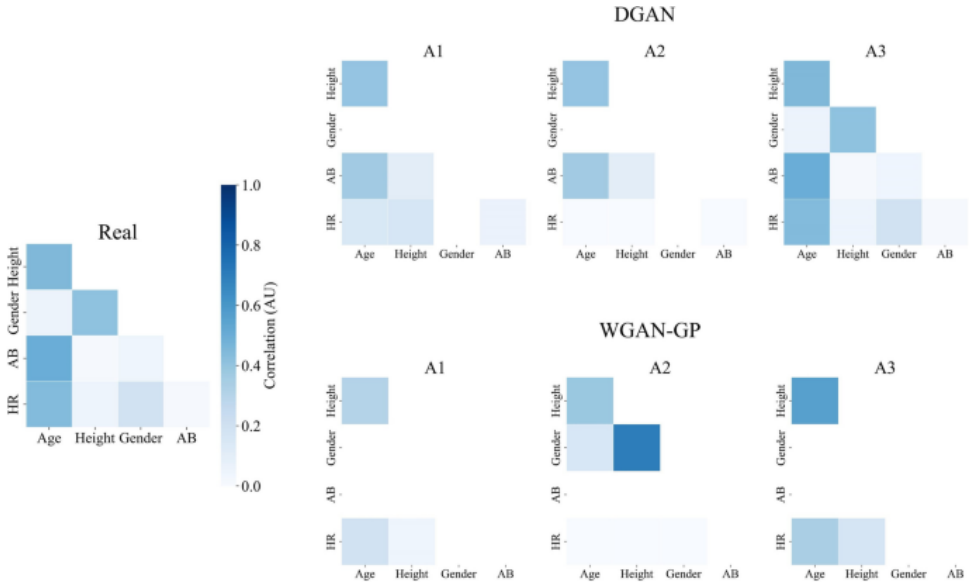


真实（左）和合成低血压数据集的多变量相关性。上行显示使用 DGAN 生成的数据，而下

行显示使用 WGAN-GP 生成的数据。右边的三列指的是合成过程中采用的不同方法。GCS：格拉斯哥逗号量表。请注意，对于标准差等于零的要素，将显示空白值，这会导致相关计算中的空值

至于第三个多变量分析，如图 5 所示，为 PMData 数据集计算的相关性。DGAN 的 A3 方法 P 是唯一一个设法产生原始数据集相关性的模型，因为其余模型错过了捕获涉及二元变量的模型。A1 方法是 WGAN-GP 的最佳方法 P 模型，但表 4 中的相似度指标显示结果较差。

图 5



真实（左）和合成 PMData 数据集的多变量相关性。上行显示使用 DGAN 生成的数据，而下行显示使用 WGAN-GP 生成的数据。右边的三列指的是合成过程中采用的不同方法。HR：心率。请注意，标准差等于零的要素将显示空白值，这会导致相关计算中的空值

事实证明，对 TMET 数据集的 DLA 分析对于每种方法都是相似的，因此每次实验的值都接近 0.9。关于低血压子集，第一种方法（A1）取得了最好的结果，因为预测值接近 0.5。性能最佳的 WGAN-GP$_T$DLA 评估的模型是用 A3 方法获得的，即使紧随其后的是其他方法。此外，对 WGAN-GP 的评估 M 模型表明，当遵循 A1 方法时，样本的可分类性最低。PMData 数据集的 DLA 分析没有差异，因为在每种情况下，分类器都正确区分了样本。

为了完成相似性评估，考虑了每个数据集中的时间相关变量进行了自相关分析。由于评估 A1 方法没有意义（生成的数据和原始数据都包含相同的实时序列），因此将这些值设置为将它们用作其他两种方法的基线。A2 方法是最好地模

拟 A1 中低血压子集、PMData 数据集和 WGAN-GP 的自相关指标的方法 $_T$ 模型，而 A3 是 DGAN 中表现最好的模型 $_T$。

从 DGAN 派生的效用指标 $_T$ 模型显示，最有用的数据是使用 A3 方法生成的，而 A1 是表现最差的数据，即使统计显着性没有得到证实，因为它可以在表 3 中检查。另一方面，$DGAN_M$ 模型显示低血压亚群的情况正好相反，第一种方法是表现最好的方法（P>0.05）。关于 WGAN-GP 模型，使用 TMET 数据集训练的模型显示，即使与其他两种方法相差不远，最佳方法也是 A3。The $WGAN\text{-}GP_M$ 组合使用 A2 方法表现最佳。与这些模型和数据集组合相关的统计检验均未产生统计学意义。关于 PMData 数据集，A1 方法在 DGAN 情况下表现最好（p<0.05），以及 $WGAN\text{-}GP_P$(p<0.05）。在表 5 中，可以比较为 TSTR 评估获得的值。

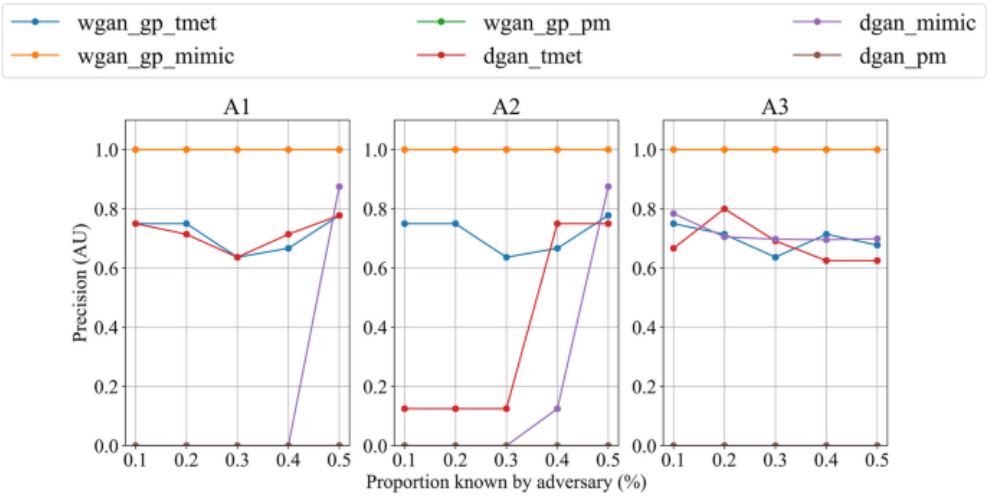
表 5 效用指标结果

Generation Method	Model[a]	TSTR metric
A1	$DGAN_T$	0.802 ± 0.033
	$WGAN\text{-}GP_T$	0.880 ± 0.002
	$DGAN_M$	**0.931 ± 0.064**
	$WGAN\text{-}GP_M$	0.909 ± 0.065
	$DGAN_P$	**0.983 ± 0.005**
	$WGAN\text{-}GP_P$	**0.986 ± 0.002**
A2	$DGAN_T$	0.900 ± 0.589
	$WGAN\text{-}GP_T$	0.837 ± 0.109
	$DGAN_M$	0.869 ± 0.078
	$WGAN\text{-}GP_M$	**0.924 ± 0.049**
	$DGAN_P$	0.906 ± 0.083
	$WGAN\text{-}GP_P$	0.919 ± 0.000
A3	$DGAN_T$	**0.919 ± 0.167**
	$WGAN\text{-}GP_T$	**0.900 ± 0.156**
	$DGAN_M$	0.893 ± 0.079
	$WGAN\text{-}GP_M$	0.758 ± 0.125
	$DGAN_P$	0.957 ± 0.016
	$WGAN\text{-}GP_P$	0.778 ± 0.120

最后要展示的结果是隐私评估结果。关于 WGAN-GP 模型，两个数据集在方法之间产生了相似的值：$WGAN\text{-}GP_T$ 导致精度值约为 0.7，在攻击者已知的最高比例下，A3 和 WGAN-GP 的结果略好 $_M$ 在任何情况下都无法对抗 MIA。对于 $WGAN\text{-}GP_P$ 模型中，所有方法都产生了私有数据，因为在 MIA 仿真期间实现了零精度。关于 $DGAN_T$，则在 A2 和对手已知的数据集的低比例（0.1-0.3）中取得了最好的结果。对于较高比例（0.4-0.5），每种测试方法都获得了可比的结果。DGAN 公司

MA1 和 A2 中的模型在最低比例 （3.0–1.0） 下优于 A3 方法，MIA 设置的精度为零，即使最高比例的精度显着下降。最后，使用 PMData 数据集训练的 DGAN 模型再次产生私有数据，因为所有方法的精度都为零。MIA 仿真结果如图 6 所示。

图 6



成员推理攻击（MIA） 精度结果。每个子图表示合成数据所遵循的方法，每条曲线表示 MIA 的伪交叉验证平均精度。AU：任意单位

讨论

相似性评估

在评估部分呈现的相似性结果被发现遵循依赖于要合成的数据集的初始特征的模式。具体来说，时间序列与元数据的比率（指纵向变量相对于数据集中静态变量的数量）被发现预先确定了相似性指标以及时间序列的长度，因为这些是 A1 方法中保持真实的数据量的决定因素。事实上，TMET 数据集中的时间序列与元数据的比率为 6：5，从中可以推断出 54.5% 的数据集变量是时间序列，而对于低血压子集，6：2 的比率构成了 75%的数据集变量是纵向的，而对于 PMData 数据集，该比率为 1：4。考虑到这一事实，当使用更高的比率和/或更长的序列时，A1 似乎更受青睐，因为生成的 SD 的百分比更高，从而实现了改进的相似性指标。原则上，方法 A2 和 A3 不应依赖于此比率，因为元数据和时间序列都是在这些方法中合成的。

在相似度量结果的统计分析方面，大多数相关性相似性检查、自相关测量和 F

分数度量都存在统计学上的显著差异。DLA 仅对 MIMIC 数据集具有统计学意义，而 TMET 和 PMData 数据集没有发现差异。

效用评估

关于生成数据的效用方面，方法 A3 对于 TMET 数据表现最好，而方法 A1 对于 PMData 和 $DGAN_M$ 表现更好。这可能会导致类似于相似性部分讨论的基本原理，其中大量真实数据可能会增强结果数据集的结果。

至于比较的意义，既没有使用 TMET 数据集训练的模型，也没有 DGAN 训练 M 发现模型具有统计学意义。相比之下，DGAN 和 WGAN-GP 都使用 PMData 数据集和 WGAN-GP 进行训练 $_M$ 发现模型在很大程度上不同。

隐私评估

本研究中用于评估的 MIA 设置很严格，但在文献中也没有发现标准化的特定时间序列 MIA 设置。隐私评估领域的未来工作将包括研究和纳入不同的攻击和阈值，以创造包含更广泛可能性的潜在场景。

WGAN-GP 模型被发现容易受到模式坍缩的影响，这或许可以解释对手在攻击 $WGAN\text{-}GP_M$ 时所达到的精度生成的数据。在 $WGAN\text{-}GP_T$ 中没有发现显著差异表明在隐私保护方面，没有一种方法比另一种方法更好。

对使用 $DGAN_T$ 生成的数据的攻击建议最保护隐私的方法是 A2，这与使用它获得的效用指标一致（使用 A3 实现最佳效用）。与我们最初的假设相反，$DGAN_M$ 的隐私评估导致方法 A1 和 A2 在使用 DGAN 模型时优于 A3，这可能与 MIA 参数的配置方式有关。

关于使用 PMData 数据集执行的 MIA 模拟，原始数据集中的受试者数量较少可能影响了结果。由于执行的计算较少，受试者数量越少，在记录之间找到潜在匹配的可能性就越低。

贡献和限制

这项工作的主要贡献是详细定义和系统地比较了三种生成时间序列的方法以及不同评估维度的元数据，将它们呈现在一个共同的框架中，并大大扩展了作者先前在发表的分析。

该研究的个局限性与以下事实有关：在文献中找不到标准化的指标来评估纵向合成数据集。即使效用和相似性指标可以很容易地适应这个用例，但考虑到评估过程中引入的不同阈值和参数，以及所用数据集间的差异，也很难客观地评估数据的隐私性。此外，即使纵向变量与元数据联合评估，自相关也是唯一计算的时间序列特定指标。未来，计划通过分析时间序列特征，如集中趋势（如均值、众数和中位数）、变异性指标（如范围和方差）、趋势指标（如斜率）或季节性指标（如季节性指数和周期）等，对这方面进行更深入的研究。

此外，生成的 SD 没有通过任何临床验证，因此，仅通过使用数据模拟可能的下游应用来评估数据在真实场景或用例中的效用。临床专家应检查和讨论从 SD 得出的分析得出的结论，以便为在真实场景或用例中使用 SD 提供足够的证据。

结论

在这项研究中，对三种不同的 STSG 方法进行了比较评估，评估了 SD 的三个关键维度：相似性 （1）、实用性（2）和隐私性（3）。尽管它对于设定黄金标准可能并不完全具有决定性意义，但该研究有助于说明 A2 方法总体表现最差，因此，对于任何用例，都应该优先考虑 A1 和 A3 方法。

根据数据集特征、用例和最终用户要承担的风险，该分析的结果为用户指明了在相似性和效用指标方面实施 A1 或 A3 的方向。在推广之前，应更谨慎地分析此建议，因为进一步的指标计算可能会促使方向转变。隐私评估是未来的主要工作重点，包括评估具有不同特征（如主题数量少）的数据集，因为它对可持续发展共享或发布至关重要。

***注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。**