## 通用数据模型

1. 简介

所谓通用数据模型（common data model, CDM）是一种从多种来源电子数据库中提取特定信息的结构和框架，通过建立标准化的变量表单，从海量数据中准确、快速、有效地提取科学研究或管理所需要的关键信息。目前应用CDM较多的是药物安全性和有效性监测领域，用于从临床各种来源、各类格式的数据（多源异构数据）中用抽提标准化、药物安全性研究相关的数据。

**SHARPn项目开发的标准化流程（使用了临床元素模型Clinical Element Modell）**

1990年Inmon数据仓库、

2001年–疫苗安全数据链路（VSD）共享数据网络（SDN）、

2003年– CRN发布了虚拟数据仓库（HCSRN-VDW）通用数据模型（CDM）、

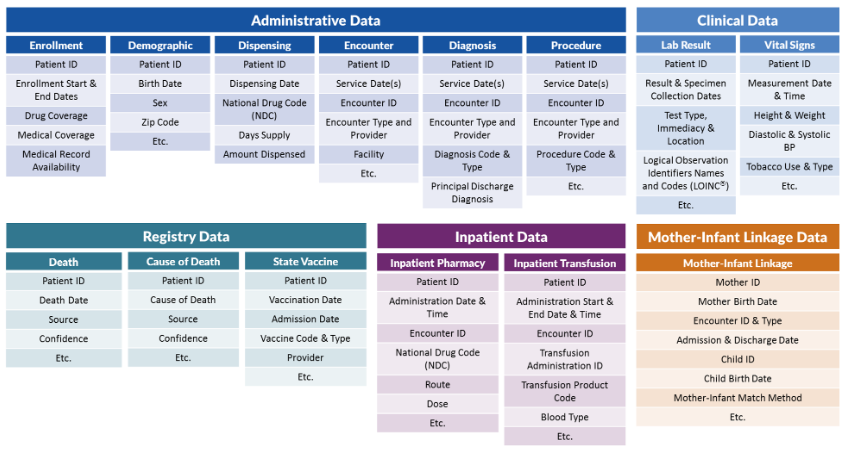
2004年研究患者数据注册中心（RPDR）CDM，该仓库由一个数据模型和一个基于Web的用户界面组成，用于查询和检查数据。i2b2提供了比VSD提供更多访问共享数据的工具

2010年– CESR虚拟数据仓库（CESR-VDW）

1. **前哨通用数据模型（SCDM）--** <https://www.sentinelinitiative.org/>

SCDM是一种标准数据结构，允许数据合作伙伴针对本地数据快速执行分布式程序。数据合作伙伴在其现有的环境中维护对电子数据的物理和操作控制。SOC数据核心使用SCDM管理Sentinel分布式数据库（SDD）的创建，并维护有关SDD实现和特征的完整文档。SDD是指数据合作伙伴以SCDM格式保存和维护的数据。

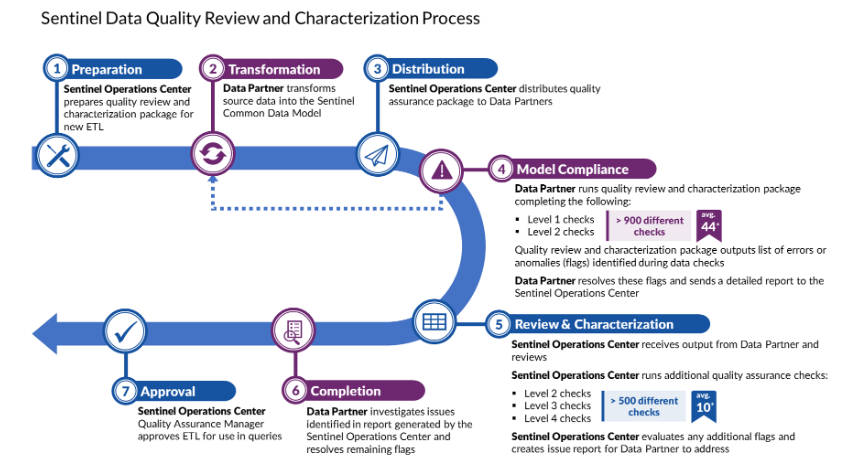
模型包括管理性数据（登记表、人口统计表、处方表、诊断表、进程表）、临床数据（实验室检查表、生命体征表）、注册登记数据（死亡表、死亡原因表、疫苗状态表）、住院病人的数据（住院药物表、住院输液表）、母-婴相关数据。



模型的分布式方法是通过使用成为Sentinel通用数据模型的标注化数据结构来实现的。所有数据合作伙伴之间的数据集合成为Sentinel分布式数据库（SDD）。分布式方法的主要好处是，它最大程度地减少了共享可识别患者信息的需求。此外，每个医疗保健数据系统都有其独特的特性，并且使用分布式系统可以更好地使数据合作伙伴参与运行分析，以确保以明智的方式来解释结果。

**数据质量审查**

数据质量是指与度量或改善数据质量相关的特征以及用于度量或改善数据质量的过程。如果数据符合预先指定的标准并且适合其预期用途，则被认为是高质量的。下图表示Sentinel Data质量审查的流程图。



1. SOC首先准备一个更新的质量审查和特征描述程序包，以评估数据合作伙伴的新ETL数据。
2. 数据合作伙伴将其源数据转换为Sentinel通用数据模型，以便可以将其包含在Sentinel分布式数据库中。
3. 在本地数据运行质量检查和特征描述包（官方提供）
4. 程序成功完成运行后，质量检查和特征描述软件包将创建标志输出。然后，数据合作伙伴将输出和已识别标志的报告以及所有相关注释发送给SOC。
5. SOC接收数据合作伙伴的输出和报告，并进行审核。
6. 数据合作伙伴从SOC收到报告，调查问题，并解决所有剩余标志。
7. SOC确认所有标志均已解决，并批准ETL以集成到Sentinel分布式数据库中。
8. **FDA发布的小规模安全警戒通用数据模型（mini-sentinel CDM, MSCDM）**

**观察性医疗成果合作伙伴关系和小型前哨通用数据模型和分析的比较评估（**<https://www.ncbi.nlm.nih.gov/pubmed/26055920>**）**

针对2007年美国食品药物管理局（FDA）提出的主动检测系统的要求，FDA于2009年与哈佛朝圣者保健研究所建立了FDA“小规模安全警戒计划”，其主要有两个特点:

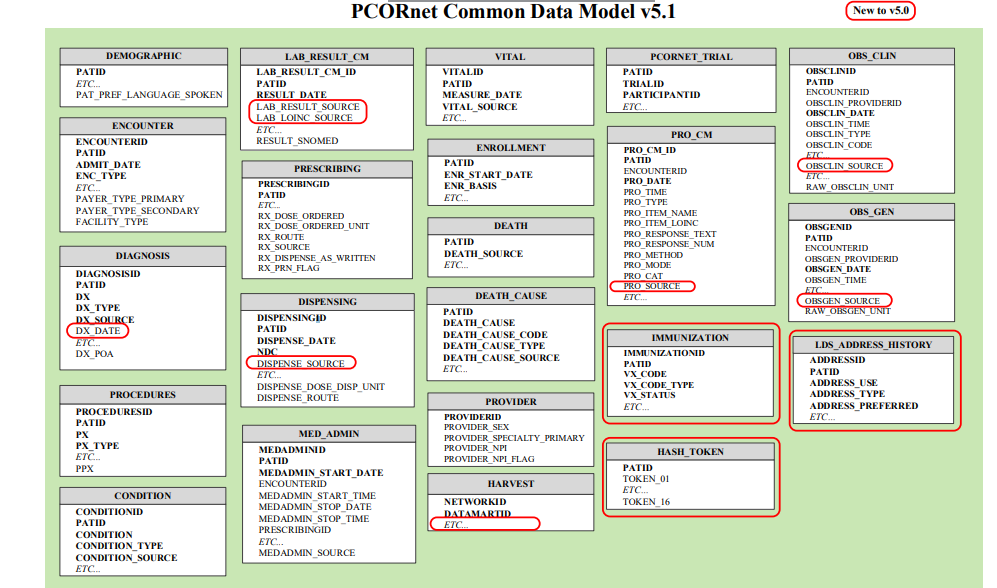
一个是分布式网络，不同的数据提供者将数据保存的本地计算机系统，而不是输送到统一的数据中心。这样能更有效的保护数据隐私，并且发挥不同数据提供者对数据了解特长，更有效的处理和更新这些数据。

一个是通用数据模型。包括：登记表、基本数据表、药物处方表、就医史、诊断史、手术史、死亡情况、死亡原因、实验室检查结果、重要体征和疫苗状态表

**3. PCORNet CDM (**[**http://www.pcornet.org/resource-center/pcornet-common-data-model/**](http://www.pcornet.org/resource-center/pcornet-common-data-model/)**)**

2014年，以患者为中心的结果研究所（PCORI）启动了一个通用数据模型/ SHDN，称为PCORnet：以患者为中心的国家临床研究网络。就像之前的Sentinel一样，PCORI的CDM扩展了VDW数据模型，其中包含影响患者结果的数据实体，尤其是患者报告的数据实体。PCORnet CDM借鉴了其他几种数据模型，并努力保持与标准代码和其他CDM的互操作性。开源PCORnet CDM基于Sentinel CDM。PCORnet是由临床数据研究网络（CDRN）和患者驱动的研究网络（PPRN）组成的研究伙伴关系。目前有33个网络参与（13个CDRN和20个PPRN）。 为了参与研究，网络合作伙伴必须符合PCORnet协调中心制定的数据质量标准。其中包括评估数据模型的一致性，所需表和变量中缺少数据以及日期和重要指标字段中的数据合理性。

PCORNet CDM通用数据模型 以患者为中心的结果研究所的所有网络均提供支持，. PCORnet CDM是传统的关系型数据库，包含15个表格对应与临床应用领域。表格有很多列，包括表关键字（患者标识符，遭遇标识符等）以及其他详细信息（例如用药频率）等，还添加了实验室值和吸烟状态等新领域信息。其所包含的信息如图所示：



**4.CDISC SDTM模型（****<https://www.cdisc.org/standards/foundational/sdtm>）**

SDTM提供了一种用于组织和格式化数据以简化收集，管理，分析和报告过程的标准。实施SDTM支持数据聚合和仓库；促进采矿和再利用；促进分享；帮助进行尽职调查和其他重要数据审查活动；并改善监管审批流程。SDTM还用于非临床数据（SEND），医疗设备和药物基因组学/遗传学研究中。

**5.OMOP通用数据模型**

健康观测数据科学和信息学组织（OHDSI）开发的观察医疗结果合作项目通用数据模型（OMOP CDM ）是一个为医学数据标准化而设计的数据模型，其基本思想是通过统一的数据模型与医学概念词汇表示，使得不同来源的医学数据以统一的标准进行整合。由于 OMOP CDM 中所有的医学数据均会匹配到标准词汇表中的概念上，因此只要根据标准词汇制定标准化的查询语句，就可以实现不同数据库、甚至不同机构之间的标注化数据查询与提取，既实现了异构数据的标准化查询，又节省了编写查询语句的人力与时间

## 一：OMOP CDM通用数据模型

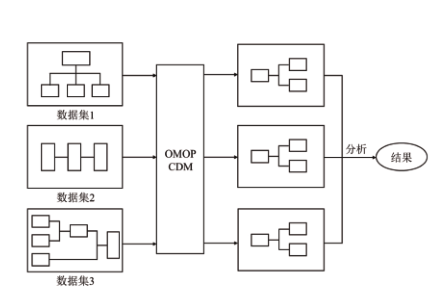
HL7：标准化的卫生信息传输协议，是医疗领域不同应用之间电子传输的协议

**问题**：但是由于各数据资源的产生单位不同、建设时期不同，采用的数据标准也不尽相同，导致在数据资源的整合方面存在很大难度。为了更有效的利用医学数据，需要针对资源的异构现象，建立统一的数据交汇标准，实现多个异构数据源的整合。

**方法**：OMOP组织建立的通用数据模型（CDM），多元数据向CDM转换流程及每一步的实现方法。

**OMOP CDM 的整体架构**：

OMOP CDM定义了统一的数据标准，以规范多源异构的观察性数据的格式和内容。



多源异构数据向CDM转化

**1.OMOP CDM 的主要模块**：

1. 标准化词汇表：包含了不同的标准术语以及源数据编码与标准术语的映射信息。
2. 标准化元数据表：储存了从源数据中派生的元数据的相关信息。

源数据：来自于多个异构数据库的原始数据

元数据：关于数据的数据，即用于说明某一数据在特定环境下的意义的一组数据。

1. 标准化临床数据表包含了每个受试者在有效观察期内的纵向临床数据以及相应的人口统计学信息。
2. 标准化健康系统数据表：描述了负责管理患者医疗保健事项的医疗保健提供者和医疗场所的相关信息。
3. 标准化健康经济数据表：包含了医疗保健的成本信息。
4. 标准化派生元素表：包含了从 CDM 的其他域表中获得的患者临床数据的相关信息(非源数据获得)。
5. 结果架构表：是CDM 6．0 版本的新模块，目前包含“队列”和“队列定义”两个域表。

**2.OMOP CDM中数据表：**

包含7张核心数据表（**Person、Condition Occurrence、Drug Exposure、Procedure Occurrence、Visit、Observation、Observation Period**）及两张分别从condition occurrence 和 drug exposure data汇总的数据表（**Condition ERA、Drug ERA**）

**3，整体而言，OMOP CDM主要有一下特点**：

1. 通用数据模型：OMOP 的核心部分是CDM的开发，意在对不同来源的医疗数据建立起统一标准的形式。通过 CDM 数据模型对信息的梳理，达到不同信息的编码和信息之间的关系进行详细的分类和定义的目的
2. 分析方法与分析工具开发

ETL工具的开发：White Rabbit、Usagi、Rabbit in a hat等。

分析工具的开发：

ACHILLES：ACHILLES 是一个可视化数据浏览工具，在提取数据之前，可以利用 ACHILLES 对数据进行初步的统计分析。ACHILLES 有 2个主要的组成部分，第一个组成部分是应用**Ｒ包**，对数据进行初步的分析，第二部分是将初步统计结果形成报告。多数据库的分析都可以在 ACHILLES 上实现。

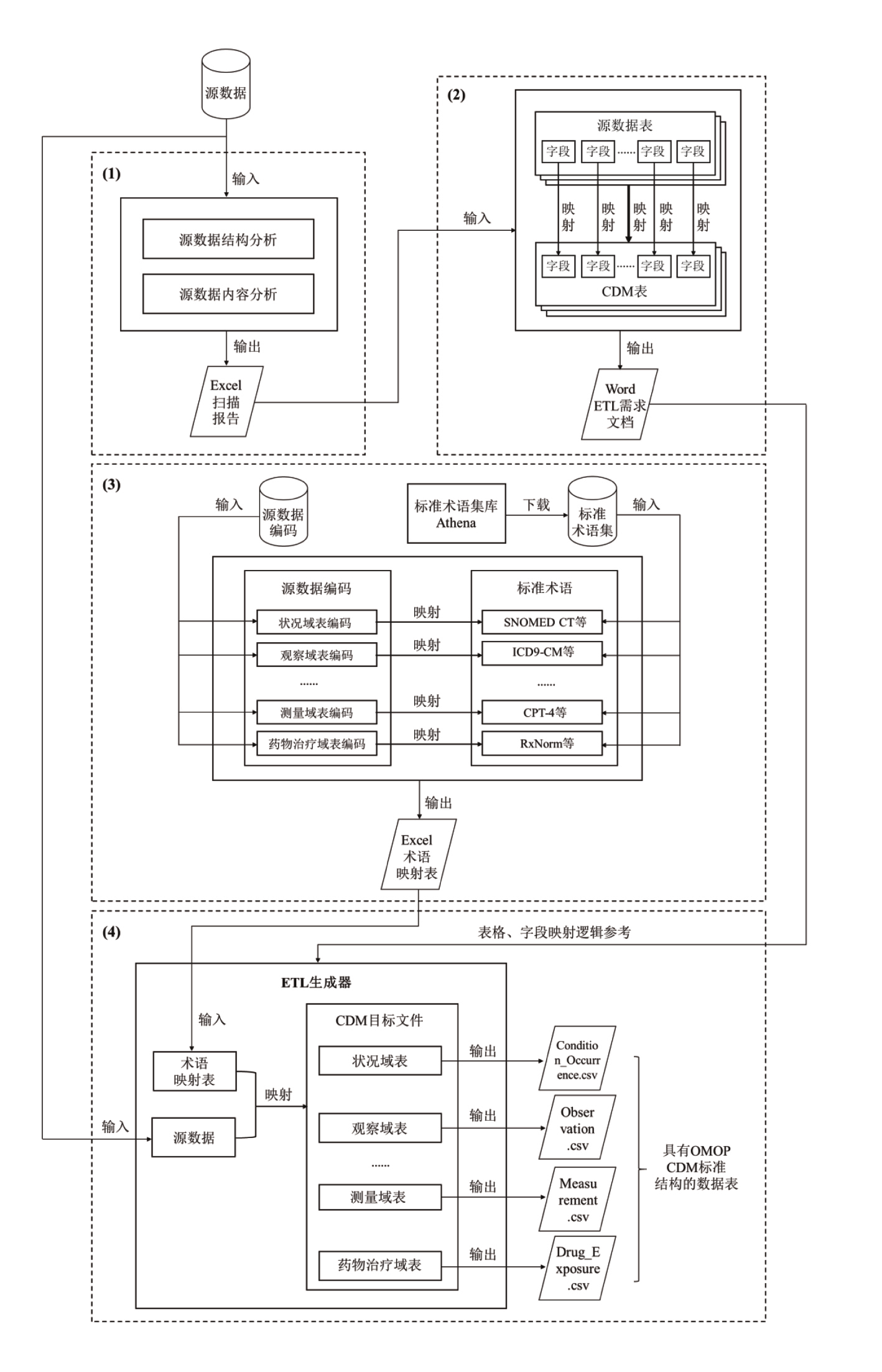
HERMES：用于对特定关键词进行检索，并寻找关键词及其相关概念的联系。

HOMER: 可以对风险进行识别，同时给出比较效果研究的结果. 例如可以对一种药物和不良反应进行关系分析

PLATO :可以用于对患者的某种结局指标进行预测和估计

1. 信息共享与交流模式

在 CDM 的基础下，数据的结果可以分享到多中心的合作组织，利于数据结果的整合。打破了大量合作者合作的阻碍，并且通过研究者之间的讨论和审阅，提高研究质量.



图二：源数据到 OMOP CDM 的转换流程

**源数据到 OMOP CDM 的转换方法如图二所示,整个 ETL 过程可分为源数据分析、数据表与字段映射、标准术语映射、ETL 实现四个部分：**

1. **源数据分析**： **White Rabbit（白兔子）软件生成扫描报告**

通过源数据结构和内容分析了解数据集的结构特征。将源数据的数据表导入 **White Rabbit**工具，进行数据扫描，生成扫描报告，该Excel扫描报告包含各个数据表的信息，每个数据表的字段信息，各个字段不同值的列表以及各个值的出现频率。数据表和字段信息可以帮助了解数据结构，数据值和值的出现频率可以帮助识别具体信息的记录方式。

1. **数据表与字段映射 Rabbit in a Hat软件得到ETL需求文档**

将得到的扫描报告导入到Rabbit－In－a－Hat 工具，进行数据映射的需求设计。

首先，建立数据表的链接。源数据表映射到CDM域表。一个源数据表可以映射到多个 CDM 域表，一个 CDM 域表也可以接收多个源数据表的映射。然后建立各字段到CDM字段的链接。

完成全部表格和字段的连接建立和映射逻辑标注后，Rabbit－In－a－Hat 可生成并导出构建 ETL 的需求设计文档。

1. **标准术语映射 Usagi工具**

**Usagi** 支持导入源数据编码文件和 OMOP 标准术语文件，通过术语相似度的方法自动的建立各个源编码到 OMOP 标准术语的建议映射，并获得每一个映射的匹配得分(通常是 0～1 的数值，1 为自信匹配)。生成术语映射表。

1. **ETL实现**

在数据表与字段映射时生成的ETL需求文档作为实现ETL的整体数据映射规范。跟据需求文档中记录的表和字段的映射逻辑编写程序，构建ELT生成器。

将数据源和通过Usagi得到的术语映射表输入ETL生成器中。不同数据源有不同的内容结构和编码规则，因此ETL生成器也具有独特性。

通过上述完整的ETL流程，将多源异构的科学数据统一转化为OMOP CDM的标准格式。

1. **ETL工具**

开发ETL工具脚本，例如将FDA不良事件报告系统（FAERS）数据集加载到OHDSI（OMOP）CDM中而开发的adepedia-on-ohdsi,以及OHDIS开发的ETL-CDMBuilder工具。

Apatar、kettle、Talend、Scriptella等纯java开源工具

Kettle 是一款国外开源的 ETL 工具，纯 java 编写，绿色无需安装，数据抽取高效稳定(数据迁移工具)。Kettle 中有两种脚本文件，transformation 和 job，transformation 完成针对数据的基础转换，job 则完成整个工作流的控制。

## 二：具体应用

科研数据模型的建立流程有数据清洗（数据缺失、数据重复、数据混乱、非结构化数据）、数据抽取。应用到具体数据库中，需跟据数据库中记录的具体情况，进行数据预处理操作。

如对药物不良反应数据库（FAERS https://github.com/adepedia/adepedia-on-ohdsi）进行去重和药物标准化等（运用AEOLUS工具）。数据提取、转化与加载内容具体包括：数据类型的转换、医学概念数据的标准化、数据计算、遗失数据插补、数据加载等流程。对数据标准化结果的评价：对数据的转化率、医学概念匹配的正确率、数据计算插补的准确率等进行了评估，用来反应具体数据库标准化的结果。

通常临床研究分为：临床RCT实验和观察性研究。然而实践过程中成本高、设计

实施复杂，有严格的“纳入标准”和“排除标准”，不能很好地反映真实的临床情况等，限制了RCT的应用范围。观察性研究是在自然状态下进行的非干预的评估和观察 [3] ，从某种程度上看，观察性研究提供的资料，不仅可以进一步在临床实践中验证RCT取得的结果，而且更能反映临床实践中的真实情况。

图三：技术流程图

## 三.细化ETL流程

1.OHDIS给出的OMOP CDM模型最佳的ETL流程包括以下4部分：

数据专家和ETL专家共同设计ETL；

有医学背景知识的人创建代码映射；

技术人员实施ELT；

·所有人都参与质量控制。

2.设计ETL：

White Rabbit扫描源数据，生成以下信息：

源数据库中的表列表

每个表的字段列表

在字段中找到的特殊值的列表（不包括唯一或不经常出现的值）

值出现的频率。

表格和字段列表可以深入了解数据的结构，而值及其频率可以帮助识别信息的记录方式（例如，性别编码为“ m”和“ f”还是“ 1”和“ 2” ，或者ICD-9代码是否存储有句点），包括任何可能的特殊值（例如“ 999”，可用于标识丢失的数字）。

Rabbit in a hat建立源数据和CDM之间的映射：

最好先建立表与域表之间的关系，然后再建立字段之间的关系。同时，不光要记录最佳的匹配结果，还要记录最佳匹配结果的推到分析，否则容易遗忘，以便于将来进行讨论。

3.创建映射代码。Usagi **java -jar Usagi\_v1.2.8.jar**

1）.专注最常用的代码字段。使用过一次的代码或者未使用的代码不必映射，研究中用不到它们。

2）.尽可能利用已有的信息。比如对于药物编码，OMOP标准词汇表是RxNorm，很多国家药品编码系统已被映射到ATC，可以利用ATC和RxNorm 之间的关系确定正确的RxNorm编码。

使用工具来Usagi来完成标准术语映射。如果源代码仅以外语提供，可以讲源代码翻译成英文进行映射。

4.实施ETL。

5.质量控制

* 审查ETL设计文档，计算机代码和代码映射。任何人都可能犯错，因此总是至少要有另一个人来复习第一人的所作所为。
* 手动比较源数据和目标数据中有关人员样本的所有信息。
* 比较源数据和目标数据中的总计数。
* 复制已经对CDM版本上的源数据执行的研究
* **住院和护理**

四：OMOP CDM应用于临床医疗过程难点及思考

1.数据的结构化程度-----通过NLP讲HER中的非结构化数据转化为结构化数据。

2.数据质量的控制：

临床科研数据的抽取、转换、加载过程是伴随着数据标准化的一个过程，将源数据库中的数据转换成目标数据库中的标准数据格式和内容，首先应保证源数据库中数据的质量，包括数据完整性、规范性以及数据的结构化程度等；其次，数据的ETL过程既需要信息技术人员的参与，也需要具有医疗知识背景的人员的参与，保证数据及编码系统映射的一致性和准确性。

3.临床标准术语的映射：

国外的医学术语和编码标准相对于国内发展快、应用较为广泛，而国内医学术语和编码标准研发和应用还不完善。OMOPCDM 使用的标准术语均为外文标准，国内医学数据中虽然有些直接使用英文术语和编码，如ICD、SNOMEDCT、LOINC等，但仍有部分需要进行中文标准转换的数据和很多缺乏标准描述的数据，这些数据无法很好地实现 OMOP CDM 的映射。

4.患者敏感信息处理-----拿到的EHR脱敏后的

五．相关文献调研。

1.通过内部和外部验证，评估英国EHR数据库（THIN）的OMOP-CDM结构的适当性，并通过实施三种统计方法，评估其在安全监督方面的潜在用途。（<https://link.springer.com/article/10.1007%2Fs40264-012-0009-3>）。论文遵循OMOP CDM规范，将THIN数据库标准化转为CDM中。

1）转换原始THIN数据以符合OMOP CDM表的数据结构

2）使用OMOP术语词典中的相应标准概念代码增强源数据

3）药物和条件在预定的持久性窗口内聚集，OMOP对此进行了详细描述：

在这项研究中，如果下一个处方的开始时间是在上一个处方结束后的30天内（持续时间窗口），那么药物的使用将聚合为一个药物周期。

4）文档：在映射过程中用于指导数据提取、转换和加载（ETL）的过程、假设和标准使用OMOP提供的标准ETL规范进行了文档化。

**2. OMOP CDM 在药品不良反应监测中的应用**

目的： OMOP CDM用于药品不良反应（ADR）的主动检测可行性分析。

方法：通过文献分析等方法，对于目前应用大数据进行 ADR 主动监测的方法进行分析，总结经验，并发现当前还需解决的问题.

ADR主动监测是药物警戒的重要组成部分，目的是监测药物投放市场后的临床不良反应，进行关联性研究。CDM 格式可以有效地将不同格式数据统一化，甚至外推于建立分布安全监测网络中，可以快速地在大规模观察性数据中发现药品的不良反应信号。

**3. 在OHDSI网络中大规模预测急性缺血性中风有症状出血转变的预后模型的开发和验证**

论文的目的是开发和验证模型，以预测患者在局部缺血性卒中后30天内出现HT的风险，开发了具有大型美国电子病历数据集的Lasso Logistic回归预测模型，数据集的构建复合OMOP CDM模型标准。并且在10个其他实际医疗保健数据集上进行了外部验证，其中包括来自美国，欧洲和亚洲的患者的EHR记录，来检查该模型在临床可应用性。结果表明OMOP CDM是用于临床多中心研究中的EMR二次使用的合适数据标准，用于预测模型的开发和验证

**4. CDM模型用于药物安全监控**（https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000752//）

使用PHARMetrics和GE两个数据库的电子病历数据作为数据集（去隐私化），对罗非昔布和心肌梗塞患者两种疾病药品研究

1）文章的CDM创新模型与OMOP模型相似，包含进行药物安全性分析所需要的基本数据元素，并讲CDM模式描述为实体关系图。

2）文章提出以病人为中心，将病人的医疗经历组织到“Person Timeline”表中，以促成纵向分析。记录病人的基本信息Person（Patient） >>> Person Timeline

3）术语标准化：病症—ICD编码 药物术语—SNOMED-CT、MedDRA

药物数据—字符串归一化工具，来适应临床词汇的需求

condition data—UMLS工具

4）将转换后的数据加载到关系型数据库中（数据的大容量以及对数据访问和查询效率的需求）

5）数据转换过程中每个步骤都进行数据统计，来评估数据转换对数据内容的影响。

在转换过程中，人数保持不变，并且本地数据和转换后的数据相同。例如，患某疾病的人数在源数据库与转换后数据库中比例应该相同。量化，事先统计

6）规则制定需要跟据具体的疾病来具体分析。例如急性疾病可以选取30天连续时间窗来观察，而例如恶性肿瘤等慢性疾病则需要调整。

7）如果源数据库中包含CMD模型不支持的数据，则CMD中不会出现此类数据，会造成数据缺失，质量控制的时候需要注意。例如，文章值未提取实验室观察值，lab项目可能包含与主题相关信息，但是未统计在内。所示设计CDM的时候要考虑可扩展性。

**5.. 使用OHDSI OMOP CDM进行生存分析：以肝内胆管癌为例，收集分析了ICC患者的HER数据。**（<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284277/>）

1） 采集ICC患者的：人口统计学数据、病史、吸烟状况、实验室结果和肿瘤特征相关数据映射到OHDSI概念。

2） 使用概念图将患者数据加载到CDM中去（考虑OMOP工具）

3） 开发基于OPOM接口，用于生存分析。

4） 将基于CDM的分析结果与传统统计分析结果进行对比---对比实验

5） 问题：OMOP词汇表对于成像数据、病理数据和模块化数据词汇表示不完整（扩充词汇表）。映射和值规范化不准确造成语义缺失（提前处理，将非结构化数据转化为结构化数据）

6） 缺失值处理：使用R Mice（https://cran.r-project.org/web/packages/mice/index.html）来估算缺失值

7） 标准化术语：Athena（http://athena.ohdsi.org/），两人审查映射概念并达成共识。

1. R'survival'软件完成生存分析接口

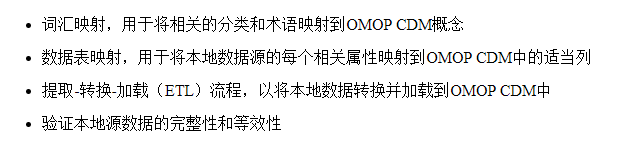
**6.PatientExploreR：一种可扩展的应用程序，用于从OMOP通用数据模型中的电子健康记录动态显示患者的临床病史**

论文展示了一种基于R / Shiny框架构建的可扩展应用程序PatientExploreR（可以从git获取https://github.com/BenGlicksberg/PatientExploreR），可与观察性医学成果合作伙伴关系CDM格式的EHR数据的关系数据库进行交互，可以生成患者级别的交互式和动态报告，并且无需任何编程即可促进临床数据的可视化。

1. **将奥地利健康声明数据映射到OMOP通用数据模型的可行性**

论文分析了观察性医疗结果合作伙伴关系（OMOP）通用数据模型（CDM）的适用性，并将数据转换为OMOP CDM。1,023（99.7％）个药物代码和3,812（99.2％）个诊断代码与OMOP词汇表一致。最大的障碍缺少针对本地词汇。

流程符合OMOP标准流程：



针对本地词汇与OMOP词汇表(包含81个标准词汇)的对应关系，论文将本地词汇分为三类：

S1 : 本地词汇与OMOP词汇一一对应。

S2：本地词汇与OMOP词汇不一一对应，但是语义上等效。

S3：本地词汇不包含在OMOP词汇表中，并且语义上也不等效。

针对S1, 工具Athena可以在 OMOP-Vocabularies查询本地词汇，可以直接进行映射；

针对S2, 工具USAGI用于查找潜在的适当概念 ，进行语义匹配；

针对S3，必须与OHDSI社区协调根据相关域的定义结构（关系和层次结构）对新的标准概念进行建模。

开发了Python脚本和SQL查询以根据数据表映射来处理ETL过程。将患者，药物，住院时间和诊断的源记录转换为临时数据结构，然后进行转换并将其写入OMOP CDM中的指定数据字段。使用SQL查询从OMOP CDM分别收集有关患者，药物，住院时间和诊断的信息到临时表中。然后将这些表记录记录地与相应的源表进行比较，来验证ETL过程的正确性。

1. 将20年的电子病历数据转换为用于临床研究的OMOP CDM

论文使用巴西圣保罗心脏研究所（Incor）的电子病历（ePR）数据.按照OMOP的操作流程：

* 1. 使用WhiteRabbit对表和字段进行采样。可以检查和忽略那些不相关或者缺失值严重的列。
  2. 使用Rabbit in a hat来对生成的excel文档绘制关系数据流图。文章将源数据转换为json格式，包含文档、注释、约束、外键。用来分析匹配表格之间的隐式关系。
  3. 病人信息去隐私化。
  4. 编码对照
  5. 质量评估：使用软件Achilles进行分析，报告数据质量问题，入错误、警告或通知。使用软件Atlas进行质量评估。验证了他汀类药物对心血管疾病患者生存率的影响，用来评估Incor-CDM模型。评估标准为计算在不同环境下执行12次的ROC曲线下的面积。

结果表明，所采用的方法能够产生一个既符合原始数据又符合OMOP-CDM标准的新数据库。