Capítulo 1

OMOP/OHDSI Descomplicado: da Teoria à Construção de Evidências em Saúde

Maria Tereza Fernandes Abrahão, Pablo Jorge Madril, Mateus de Lima Freitas, Marcos Silva de Mendonça

Abstract

This chapter presents the growing role of observational research based on real-world data in the generation of reliable health evidence. In this context, the OMOP/OHDSI ecosystem is described as a well-established international infrastructure for the standardization, harmonization, and analysis of observational data from multiple sources. The text addresses the concepts of Real-World Data (RWD) and Real-World Evidence (RWE), as well as the scientific principles underpinning OHDSI. It presents the structure of the OMOP Common Data Model, including standardized tables and vocabularies, and outlines the implementation workflow with a focus on the structural and semantic data mapping processes. Also, it explores the use of Artificial Intelligence to support the definition and validation of clinical phenotypes. Finally, it discusses the impact of OMOP/OHDSI on multicenter research, the generation of real-world evidence, and the advancement of digital health.

Resumo

Este capítulo apresenta o papel crescente da pesquisa observacional baseada em dados do mundo real na geração de evidências confiáveis em saúde. Nesse contexto, descreve-se o ecossistema OMOP/OHDSI como uma infraestrutura internacional consolidada para a padronização, harmonização e análise de dados observacionais de múltiplas fontes. O texto aborda os conceitos de Real-World Data (RWD) e Real-World Evidence (RWE), bem como os princípios científicos da OHDSI. Apresenta-se a estrutura

do OMOP Common Data Model, incluindo tabelas e vocabulários padronizados, além do fluxo de implementação com foco no mapeamento estrutural e semântico dos dados. Também explora o uso da Inteligência Artificial no apoio à definição e validação de fenótipos clínicos. Por fim, discute o impacto do OMOP/OHDSI na pesquisa multicêntrica, na produção de evidências do mundo real e no avanço da saúde digital.

1. Introdução

A pesquisa observacional em saúde vem ganhando destaque como uma abordagem essencial para a geração de evidências em larga escala, especialmente em contextos em que ensaios clínicos randomizados não são viáveis, éticos ou suficientes para responder às perguntas científicas necessárias. Esse tipo de pesquisa permite explorar dados do mundo real coletados rotineiramente em sistemas de saúde, favorecendo análises reprodutíveis, comparáveis e capazes de refletir cenários clínicos complexos e heterogêneos.

Nesse contexto, o ecossistema OMOP/OHDSI (*Observational Medical Outcomes Partnership / Observational Health Data Sciences and Informatics*) consolidou-se como o padrão global para a harmonização de dados clínicos e execução de estudos multicêntricos em saúde. O modelo de dados comum OMOP (CDM OMOP) oferece uma estrutura unificada que padroniza informações provenientes de diversas fontes, como prontuários eletrônicos, sistemas administrativos, registros farmacêuticos e bancos de dados de pesquisa, permitindo que instituições ao redor do mundo alinhem seus dados de forma consistente.

A comunidade OHDSI complementa essa infraestrutura ao disponibilizar metodologias, ferramentas analíticas e uma rede colaborativa internacional que possibilita a realização de estudos em larga escala, de maneira distribuída e tecnicamente robusta. Isso promove a comparabilidade entre instituições e países, fortalece a reprodutibilidade científica e acelera o processo de geração de insights clínicos, atuando como um sistema em que em cada etapa possui mecanismos de validação *a priori*, que garante a qualidade da pesquisa.

Nestes mecanismos de validação temos abertura para a incorporação de tecnologias de Inteligência Artificial e Deep Learning que permitem escalar o processo, sem interferir com os resultados, criando uma forma de adoção destas tecnologias com baixo risco.

Como resultado, o ecossistema OMOP/OHDSI tem desempenhado papel central na evolução da pesquisa observacional moderna, ampliando a capacidade de transformar dados em conhecimento e contribuindo diretamente para avanços em vigilância epidemiológica, farmacovigilância, predição de riscos e avaliação de efetividade de intervenções no mundo real.

O capítulo está estruturado como se segue. Primeiro, a seção 1.1 oferece uma introdução à pesquisa observacional, aos conceitos básicos de RWD e RWE. A seção 1.2 apresenta a iniciativa OHDSI, sua origem e os princípios científicos e a seção 1.3 o

ecossistema OMOP/OHDSI. A seção 1.4 apresenta o modelo comum de dados, sua estrutura, tabelas e vocabulários e a seção 1.5 o fluxo de implementação para a preparação da base OMOP apresentando os mapeamentos dos dados, dos vocabulários e a ferramenta de Inteligência Artificial para detecção de fenótipos. A seção 1.6 apresenta as considerações finais e conclusões e a seção 1.7 as referências consultadas.

Na elaboração deste capítulo, consultou-se o livro da OHDSI, *The book of OHDSI*, de domínio público, sob a licença *Creative Commons Zero v1.0 Universal,* (16/04/2020). O livro é um documento vivo, mantido pela comunidade por meio de ferramentas de desenvolvimento de código aberto e evolui continuamente. A versão online, disponível gratuitamente[1], sempre representa a versão mais recente. Os exemplos foram elaborados a partir dos conhecimentos adquiridos pelos autores.

1.1 Introdução a pesquisa observacional

A pesquisa observacional é uma abordagem central para a compreensão de fenômenos em saúde a partir do uso de dados do mundo real, analisando informações tal como são geradas rotineiramente nos serviços de saúde, sem intervenção direta dos pesquisadores sobre a exposição ou o tratamento dos indivíduos. Diferentemente dos ensaios clínicos randomizados, que seguem protocolos rígidos e operam em ambientes controlados, a pesquisa observacional permite capturar a complexidade, a diversidade populacional e as condições reais da prática clínica, oferecendo uma visão mais próxima do cotidiano dos sistemas de saúde. Por essa razão, ela desempenha um papel estratégico na produção de evidências científicas, especialmente em cenários nos quais estudos experimentais seriam inviáveis, excessivamente onerosos ou eticamente inadequados (Hernán & Robins, 2025; Ricotta, EE, et al, 2025).

No contexto atual, marcado pela crescente digitalização dos sistemas de saúde, a disponibilidade de grandes volumes de dados estruturados abriu novas possibilidades para a pesquisa observacional. Esses dados, conhecidos como Real World Data (RWD), incluem informações provenientes de prontuários eletrônicos, registros administrativos, sistemas de faturamento, bancos de dados farmacêuticos, registros de doenças, entre outras fontes, que refletem a utilização real dos serviços de saúde e o comportamento clínico de grandes populações ao longo do tempo.

O uso sistemático desses dados permite avaliar desfechos clínicos em populações amplas e diversas, acompanhar trajetórias de cuidado ao longo do tempo e identificar padrões que dificilmente seriam observados em estudos altamente controlados.

A análise sistemática e metodologicamente rigorosa do RWD possibilita a geração de Real World Evidence (RWE), definida como a evidência clínica obtida a partir da análise rigorosa desses dados, contribuindo para avaliações de segurança, efetividade e uso de intervenções no mundo real (FDA, 2023; FDA, 2025).

Os estudos observacionais assumem diferentes delineamentos, como estudos de coorte, caso-controle, transversais e séries temporais. Cada um desses formatos responde

[1] The book OHDSI http://book.ohdsi.org

a tipos específicos de perguntas cientificas, permitindo investigar associações entre exposições e desfechos, acompanhar trajetórias de cuidado e avaliar variações clínicas e assistenciais. Embora não tenham como principal objetivo estabelecer causalidade nos mesmos moldes dos ensaios clínicos randomizados, estudos observacionais bem conduzidos são essenciais para complementar evidências clínicas e orientar decisões em saúde pública, avaliação de tecnologias em saúde e farmacovigilância, especialmente quando aliados a métodos analíticos avançados e boas práticas metodológicas (Hernán & Robins, 2025; Sterrantino, 2024).

Entretanto, a pesquisa observacional está sujeita a uma série de vieses metodológicos que podem comprometer a validade interna e externa dos resultados caso não sejam devidamente identificados, controlados e discutidos. Entre os principais vieses, destaca-se o viés de seleção, que ocorre quando a população analisada não é representativa da população alvo, podendo distorcer associações observadas. O viés de informação refere-se a erros sistemáticos ou inconsistências no registro, mensuração ou classificação das exposições, desfechos ou covariáveis. Já o viés de confusão ocorre quando fatores externos estão associados simultaneamente à exposição e ao desfecho, levando a interpretações causais equivocadas se não forem adequadamente ajustados.

Adicionalmente, problemas recorrentes como dados ausentes, heterogeneidade na qualidade das fontes, mudanças nos sistemas de informação ao longo do tempo e variações nas práticas clínicas representam desafios frequentes em estudos baseados em dados do mundo real. O reconhecimento explícito dessas limitações é essencial tanto para a interpretação crítica dos achados quanto para o desenho de estudos observacionais mais robustos e transparentes, em consonância com as boas práticas metodológicas em epidemiologia e ciência de dados em saúde (Porta et al., 2014; Dekkers et al., 2012).

Para mitigar tais limitações, a adoção de boas práticas em pesquisa observacional é indispensável. Essas práticas incluem a definição clara da pergunta científica, a especificação prévia do protocolo analítico, o uso de métodos estatísticos apropriados para controle de confundimento, a transparência metodológica e a reprodutibilidade das análises. Iniciativas internacionais, como a declaração STROBE, fornecem diretrizes consolidadas para melhorar a qualidade e a transparência no relato de estudos observacionais, fortalecendo a credibilidade das evidências geradas (STROBE Initiative). A documentação completa das decisões metodológicas e a reprodutibilidade dos resultados são princípios centrais para garantir a robustez científica dos estudos baseados em RWD.

Em síntese, a pesquisa observacional desempenha um papel estratégico no uso de dados do mundo real para a geração de evidências relevantes, complementando os ensaios clínicos e ampliando a compreensão da efetividade das intervenções na prática cotidiana.

O uso adequado de RWD para gerar RWE, aliado à atenção aos vieses, à adoção de boas práticas metodológicas e à padronização dos dados, permite ampliar a capacidade de compreender fenômenos complexos, avaliar intervenções no mundo real e apoiar decisões mais informadas. Dessa forma, a pesquisa observacional consolida-se como um instrumento indispensável para o avanço do conhecimento, a melhoria da qualidade do

cuidado e o fortalecimento dos sistemas de saúde, constituindo-se como a base para uma ciência mais reprodutível, colaborativa.

Nesse cenário, a padronização de dados em saúde emerge como um fator crítico para a qualidade, a comparabilidade e a escalabilidade da pesquisa observacional. Sistemas de saúde frequentemente utilizam diferentes estruturas de dados, formatos e terminologias clínicas, o que dificulta a integração de múltiplas fontes, compromete a reprodutibilidade dos estudos e limita a realização de análises multicêntricas. A ausência de padrões comuns aumenta o risco de erros analíticos, dificulta a interoperabilidade entre instituições e reduz a confiabilidade dos achados científicos. Em contrapartida, a adoção de modelos padronizados de dados, associada ao uso de vocabulários clínicos comuns e definições semânticas consistentes, contribui para maior qualidade analítica, facilita o compartilhamento de informações, estimula a colaboração entre pesquisadores e amplia o potencial de reutilização dos dados para diferentes finalidades de pesquisa e geração de evidências do mundo real (McDonald C.J. et al, 2022; Kahn et al., 2016; Hripcsak et al., 2015).

Além disso, a padronização é um pilar fundamental para a ciência aberta e colaborativa, permitindo que métodos e análises sejam replicados em diferentes contextos, sem a necessidade de compartilhamento direto de dados sensíveis.

Ao alinhar dados de diferentes origens sob uma mesma estrutura conceitual, torna-se possível executar análises distribuídas, comparar resultados entre instituições e replicar estudos em diferentes populações, preservando a privacidade dos dados. Isso promove estudos distribuídos, comparabilidade internacional e maior confiança nas evidências produzidas. Dessa forma, a padronização não apenas fortalece a pesquisa observacional do ponto de vista técnico, mas também acelera a produção de conhecimento científico com impacto direto na prática clínica e na formulação de políticas públicas (FDA, 2025).

1.2 A Iniciativa OHDSI: Origem e Princípios Científicos

A iniciativa Observational Health Data Sciences and Informatics (OHDSI) tem sua origem diretamente relacionada aos esforços conduzidos no âmbito do Observational Medical Outcomes Partnership (OMOP), um projeto iniciado formalmente em 2008 como uma parceria público-privada coordenada pela Foundation for the National Institutes of Health (FNIH), com forte participação da Food and Drug Administration (FDA), da indústria farmacêutica, da academia e de organizações de saúde (Overhage et al., 2012; Hripcsak et al., 2015). O surgimento do OMOP ocorreu em um contexto regulatório específico, marcado pela promulgação do FDA Amendments Act de 2007, que ampliou as exigências por sistemas mais eficazes de monitoramento pós comercialização de medicamentos. O projeto OMOP foi concebido com duração aproximada de cinco anos (2008–2013) e teve como objetivo central avaliar empiricamente a confiabilidade de métodos analíticos aplicados a dados observacionais em saúde, especialmente no contexto da farmacovigilância (Stang et al., 2010; Overhage et al., 2012). Diferentemente de iniciativas orientadas à obtenção direta de resultados clínicos, o OMOP concentrou seus esforços na comparação sistemática de abordagens

metodológicas, examinando sua capacidade de identificar associações causais verdadeiras e minimizar falsas descobertas em bases de dados heterogêneas.

Uma das principais contribuições científicas decorrentes do OMOP foi o desenvolvimento do OMOP Common Data Model (CDM), criado para padronizar a estrutura e o conteúdo de dados observacionais provenientes de diferentes fontes, como prontuários eletrônicos e bases administrativas. A adoção de um modelo de dados comum e dos vocabulários padronizados permitiu a execução de análises comparáveis em múltiplas bases de dados, reduzindo a variabilidade estrutural e semântica e aumentando a reprodutibilidade dos estudos (Hripcsak et al., 2015). https://www.ohdsi.org/data-standardization/

Com o encerramento formal do projeto OMOP em 2013, seus principais ativos científicos, incluindo o Common Data Model, os vocabulários padronizados e os métodos analíticos desenvolvidos, passaram a ser mantidos e ampliados por uma nova iniciativa colaborativa: o OHDSI (Hripcsak et al., 2015). Diferentemente do OMOP, estruturado como um projeto com prazo definido, a OHDSI foi concebida como uma iniciativa contínua, aberta e global, orientada à evolução permanente de métodos e ferramentas para análise de dados observacionais em saúde.

Desde sua constituição, a OHDSI adotou como princípio estruturante a separação entre dados e métodos, promovendo um modelo no qual as instituições participantes mantêm seus dados sob custódia local, enquanto compartilham definições de coorte, protocolos analíticos e código-fonte (Hripcsak et al., 2015). Tal abordagem viabiliza a replicação sistemática de estudos em diferentes contextos assistenciais, fortalecendo a reprodutibilidade científica e favorecendo a validação externa dos resultados obtidos. Entre os princípios científicos fundamentais da iniciativa, destaca-se a transparência metodológica, que se manifesta na exigência de documentação explícita de hipóteses, critérios de inclusão e exclusão, estratégias analíticas e suposições estatísticas.

A OHDSI também reconhece explicitamente as limitações inerentes aos dados observacionais, promovendo o controle sistemático de vieses e fatores de confusão por meio de desenhos de estudo rigorosos e metodologias estatísticas avançadas (Schuemie et al., 2014).

Outro pilar conceitual da OHDSI é o paradigma de análise distribuída, no qual apenas o código analítico é compartilhado entre instituições, enquanto os dados permanecem localmente armazenados (Hripcsak et al., 2015). Esse modelo responde diretamente às exigências éticas e legais associadas à proteção de dados em saúde, permitindo a realização de estudos multicêntricos internacionais em conformidade com legislações como o General Data Protection Regulation (GDPR) e a Lei Geral de Proteção de Dados (LGPD), sem necessidade de transferência de dados sensíveis.

Sob uma perspectiva científico-metodológica, a iniciativa OHDSI desempenha um papel central na reconfiguração da pesquisa observacional, ao afirmá-la como uma fonte legítima, rigorosa e complementar de evidência clínica e populacional. Ao articular padronização estrutural e semântica dos dados, métodos analíticos explicitamente documentados e práticas de colaboração científica em larga escala, a OHDSI contribui

para mitigar limitações historicamente atribuídas aos estudos observacionais, como baixa reprodutibilidade e elevado risco de vieses. Dessa forma, a iniciativa reduz a assimetria tradicional entre estudos observacionais e ensaios clínicos randomizados, não por equipará-los indiscriminadamente, mas por posicionar cada abordagem de maneira contextualizada, especialmente em cenários nos quais os ensaios randomizados são inviáveis, eticamente limitados ou insuficientes para representar a complexidade e a heterogeneidade da prática clínica real (Hripcsak et al., 2015).

Em síntese, a transição do OMOP (2008–2013) para o OHDSI representa a passagem de um projeto metodológico delimitado para um ecossistema científico sustentado, fundamentado em princípios de ciência aberta, rigor metodológico e colaboração global. Essa evolução consolidou as bases técnicas e conceituais para a produção contínua de evidências em saúde a partir de dados observacionais em escala populacional.

Ao estabelecer uma comunidade global estruturada de pesquisadores, instituições acadêmicas e bases de dados observacionais em saúde, com um centro de coordenação científica sediado na Universidade de Columbia, a iniciativa OHDSI viabilizou a colaboração científica em escala verdadeiramente internacional (Hripcsak et al., 2015). Atualmente, a comunidade OHDSI reúne mais de 4.700 colaboradores distribuídos por aproximadamente 88 países, integrando uma rede de mais de 540 fontes de dados padronizadas, que representam cerca de 970 milhões de registros de pacientes únicos mapeados ao OMOP Common Data Model (cerca de 12% da população mundial) (Columbia DBMI, 2025; OHDSI, 2026). Esse arranjo colaborativo e distribuído permite a execução de estudos observacionais multicêntricos, reprodutíveis e metodologicamente consistentes, transcendendo fronteiras geográficas e institucionais. Ao promover ciência aberta, padronização analítica e validação cruzada em muitos contextos assistenciais, a OHDSI vem transformando a forma como a pesquisa médica é conduzida, com impacto direto na geração de evidências para apoiar decisões clínicas, regulatórias e de políticas públicas, e com o objetivo explícito de contribuir para a melhoria da saúde em escala global.

1.3 Ecossistema OMOP/OHDSI

A interoperabilidade entre sistemas de informação em saúde é um requisito fundamental para o uso integrado e eficiente de dados provenientes de múltiplas fontes. Segundo Mucheroni (Mucheroni 2011), essa interoperabilidade depende essencialmente de dois fatores complementares:

- Interoperabilidade sintática (forma)
- Interoperabilidade semântica (conteúdo)

A interoperabilidade sintática refere-se à padronização da estrutura, formatos e modelos de dados, garantindo que diferentes sistemas consigam trocar informações de maneira técnica e operacional. Já a interoperabilidade semântica assegura que o significado dos dados seja preservado e compreendido de forma consistente entre os sistemas, permitindo interpretações corretas e comparáveis das informações trocadas.

No contexto da pesquisa em saúde baseada em dados observacionais, esses dois níveis de interoperabilidade são indispensáveis para possibilitar comparações, integrações e análises estatísticas robustas entre conjuntos de dados heterogêneos, oriundos de diferentes instituições, regiões ou países.

Para viabilizar análises reprodutíveis e em larga escala de dados observacionais em saúde, a OHDSI se fundamenta em três componentes principais:

Modelo Comum de Dados (Common Data Model - CDM OMOP) O CDM OMOP define uma estrutura padronizada para o armazenamento de dados clínicos e administrativos, promovendo interoperabilidade sintática ao uniformizar tabelas, campos e relacionamentos. A especificação completa do modelo, incluindo a descrição detalhada das tabelas, campos, regras de preenchimento e versões suportadas, está disponível na documentação oficial da comunidade OHDSI: https://ohdsi.github.io/CommonDataModel/index.html

Vocabulários Padronizados - O uso de vocabulários controlados e padronizados (como SNOMED, LOINC, RxNorm, entre outros) assegura a interoperabilidade semântica, permitindo que os mesmos conceitos clínicos sejam representados de forma consistente, independentemente da origem dos dados. No ecossistema OMOP/OHDSI, esses vocabulários são integrados ao modelo por meio de um mecanismo unificado de conceitos e relacionamentos, permitindo a padronização semântica dos dados observacionais. Os vocabulários oficiais utilizados pelo CDM OMOP podem ser obtidos e atualizados por meio da plataforma Athena, disponível em: https://athena.ohdsi.org

Ferramentas para preparação do CDM e análises - A OHDSI disponibiliza um ecossistema de ferramentas que apoia desde o processo de transformação dos dados fonte para o CDM OMOP até a execução de análises estatísticas avançadas e a definição padronizada de estudos observacionais.

As ferramentas disponibilizadas pela OHDSI operam de maneira coesa e padronizada, permitindo que análises no nível do paciente sejam conduzidas de forma distribuída, transparente e reprodutível. Esse conjunto integrado de soluções viabiliza a realização de estudos comparáveis entre diferentes bases de dados, sem a necessidade de compartilhamento direto dos dados sensíveis.

Os principais pontos em destaque são:

- Padronização estrutural e semântica, reduzindo ambiguidades e aumentando a qualidade analítica dos dados.
- Reprodutibilidade científica, uma vez que métodos e códigos analíticos podem ser reutilizados em diferentes bases convertidas para o CDM OMOP.
- Escalabilidade e colaboração internacional, permitindo estudos multicêntricos com grande volume de dados.
- Transparência metodológica, com definição clara e compartilhável de coortes, critérios de inclusão e métodos estatísticos.
- Aceleração da geração de evidências do mundo real (Real World Evidence), apoiando decisões clínicas, regulatórias e estratégicas em saúde.

No geral, as ferramentas da OHDSI desempenham um papel crucial na promoção da colaboração, reprodutibilidade e transparência na análise de dados observacionais de saúde, permitindo avanços significativos na pesquisa e na prática clínica. O conjunto de ferramentas disponibilizadas pela OHDSI auxiliam a preparação da base no modelo CDM OMOP, na validação do processo de mapeamento, na verificação da qualidade dos dados que compõem o CDM, na elaboração e análise dos diferentes tipos de estudos, facilitando a exploração dos dados e a geração de evidências (Hripcsak et al., 2015; Kahn et al., 2016). Podemos citar:

- Ferramentas para geração do banco CDM OMOP
  - ETL: White Rabbit, Rabbit-In-A-Hat[2]
  - Vocabulários: Athena[3]/Usagi[4]
  - Qualidade: Achilles[5] e *Data Quality Dashboard* (DQD)[6]
- Ferramentas de análise
  - Geração de Coortes/estudos - Atlas[7]
  - Análises estatísticas: HADES[8]

1.3.1 Onde achar: principais referências

As informações a respeito dos componentes da OHDSI podem ser classificadas em:

- ➢ Informações gerais: http://www.ohdsi.org - Site principal;
- ➢ Código e instalações: https://github.com/OHDSI/ - Aqui está disponibilizado o código fonte de todas as ferramentas. Em particular destacamos: *Common Data Model* (https://github.com/OHDSI/CommonDataModel), com a definição completa do modelo e as implementações para os diversos bancos suportados;
- ➢ Tutoriais e vídeos: Procure no Google por YouTube OHDSI (https://www.google.com/search?q=youtube+ohdsi), existe muita documentação e tutoriais em vídeo dos eventos anuais do grupo;
- ➢ Fórum: Para resolver dúvidas mais frequentes, consulte (http://forums.ohdsi.org/).
- ➢ Livro OHDSI: The book of OHDSI

1.4 Common Data Model: estrutura, tabelas essenciais e vocabulários

O OMOP Common Data Model (CDM) constitui a infraestrutura técnica fundamental que sustenta análises reprodutíveis e metodologicamente consistentes sobre dados do mundo real. Seu desenvolvimento responde diretamente à necessidade de lidar com a heterogeneidade estrutural e semântica dos dados de saúde, criando uma base comum que

[2] White Rabbit, Rabbit-In-A-Hat http://ohdsi.github.io/WhiteRabbit/index.html
[3] ATHENA http://athena.ohdsi.org
[4] Usagi http://ohdsi.github.io/Usagi/
[5] ACHILLES http://www.ohdsi.org/web/achilles
[6] DQD https://ohdsi.github.io/DataQualityDashboard/articles/DataQualityDashboard.html
[7] Atlas https://github.com/OHDSI/Atlas/wiki/Atlas-Setup-Guide
[8] HADES https://github.com/OHDSI/Hades

permita comparabilidade analítica entre diferentes sistemas, instituições e contextos geográficos.

O CDM OMOP foi concebido com um objetivo central: permitir que dados observacionais de saúde, provenientes de fontes diversas como prontuários eletrônicos, bases administrativas e de faturamento, registros farmacêuticos, sistemas laboratoriais e registros de mortalidade, sejam analisados de forma padronizada por meio da aplicação de métodos analíticos comuns. Diferentemente de modelos voltados à interoperabilidade clínica em tempo real ou à troca transacional de informações, o CDM OMOP é explicitamente orientado à análise retrospectiva e longitudinal de dados populacionais. Ao padronizar simultaneamente a estrutura e o significado dos dados, o modelo possibilita que uma mesma pergunta científica seja investigada em diferentes contextos institucionais e sistemas de saúde, utilizando protocolos analíticos idênticos.

O desenho do CDM OMOP reflete princípios que atendem tanto a requisitos técnicos quanto a exigências científicas. O modelo adota padronização estrutural rigorosa, por meio da definição de esquemas de tabelas, campos e relacionamentos uniformes, o que permite a reutilização de códigos analíticos sem necessidade de adaptações específicas para cada base de dados. Paralelamente, incorpora padronização semântica mediante o uso de vocabulários clínicos internacionais, assegurando que conceitos médicos mantenham significado consistente entre diferentes instituições e países. Um princípio adicional fundamental é a separação entre a estrutura do modelo e o conteúdo semântico, característica que confere flexibilidade e permite a evolução controlada do modelo ao longo do tempo, sem comprometer análises previamente implementadas. O CDM OMOP é fortemente orientado à preservação da dimensão temporal dos eventos clínicos, viabilizando análises longitudinais e a reconstrução das trajetórias assistenciais completas dos indivíduos.

A arquitetura do CDM OMOP baseia-se em um modelo relacional organizado em domínios clínicos bem definidos, nos quais cada domínio representa um tipo específico de evento ou informação em saúde. Essa organização contempla informações demográficas dos indivíduos, períodos de observação, atendimentos ambulatoriais e internações, diagnósticos e condições clínicas, exposição a medicamentos, procedimentos clínicos e cirúrgicos, exames laboratoriais, medições clínicas e registros de óbito. Tal estrutura suporta tanto análises populacionais de grande escala quanto estudos em nível individual, mantendo coerência clínica e temporal dos dados.

A estrutura prática do CDM OMOP foi concebida para padronizar dados de saúde de diferentes origens, permitindo análises consistentes, reprodutíveis e comparáveis entre instituições, países e sistemas distintos. Uma vez que os dados são transformados e carregados no modelo OMOP, passam a obedecer a uma lógica estrutural muito bem definida, voltada principalmente para a análise clínica e epidemiológica em larga escala.

Estrutura centrada no paciente

O princípio fundamental do CDM OMOP é ser patient-centric. Isso significa que o paciente é o eixo central de toda a modelagem de dados. Todas as tabelas clínicas

possuem relacionamento direto ou indireto com a tabela PERSON, por meio do campo PERSON_ID, que identifica de forma única cada indivíduo no banco de dados.
Essa abordagem permite reconstruir a linha do tempo clínica do paciente, organizando eventos como visitas, diagnósticos, tratamentos e medições de forma cronológica. Datas de início e fim (por exemplo, *start_date* e *end_date*) são elementos essenciais do modelo, garantindo análises longitudinais, estudos de trajetória de cuidado, identificação de recorrência de eventos e avaliação de desfechos ao longo do tempo.

Organização temporal dos eventos clínicos
No CDM OMOP, eventos não são registros isolados: eles fazem parte de uma sequência temporal. Cada condição diagnosticada, medicamento administrado ou exame realizado está associado a:

- Um paciente (*PERSON_ID*),
- Um período no tempo,
- Um contexto assistencial, geralmente representado por uma **visita** (*VISIT_OCCURRENCE*).

Essa estrutura possibilita análises como:

- Progressão de doenças,
- Comparação entre períodos pré e pós-intervenção,
- Identificação de padrões de uso de medicamentos,
- Avaliação de respostas clínicas a tratamentos.

Principais tabelas clínicas
PERSON: A tabela PERSON armazena as informações demográficas básicas do indivíduo, como sexo, data de nascimento, data de óbito (quando aplicável), raça e etnia, sempre codificadas por meio de vocabulários padronizados. Essa padronização é essencial para estudos populacionais e comparações entre bases distintas.
VISIT_OCCURRENCE: representa os contatos do paciente com o sistema de saúde, como internações, atendimentos ambulatoriais, consultas de emergência ou teleatendimentos. Ela funciona como um elemento organizador dos demais eventos clínicos, permitindo que diagnósticos, procedimentos, medicamentos e exames sejam associados a um mesmo episódio de cuidado.
CONDITION_OCCURRENCE: registra os diagnósticos e condições clínicas identificadas no paciente. Cada registro indica quando a condição foi observada, o tipo de diagnóstico (principal, secundário, histórico, suspeito, confirmado) e, sempre que possível, o contexto da visita. As condições são codificadas com terminologias padronizadas, como SNOMED CT, o que aumenta a interoperabilidade e a precisão analítica.
DRUG_EXPOSURE: descreve a exposição do paciente a medicamentos, incluindo prescrições, dispensações ou administrações. Essa tabela permite representar informações como princípio ativo, dose, duração do tratamento e via de administração. Ela é

fundamental para análises de segurança medicamentosa, adesão terapêutica, estudos de efetividade e farmacovigilância.

MEASUREMENT: concentra dados quantitativos e qualitativos provenientes de exames laboratoriais, sinais vitais e outros testes clínicos. Além do valor medido, são armazenados unidades, métodos, intervalos de referência e datas, possibilitando avaliações clínicas detalhadas, análises de tendência e estratificação de risco.

OBSERVATION: utilizada para registrar achados clínicos diversos que não se enquadram diretamente em diagnósticos, medicamentos ou medições. Exemplos incluem hábitos de vida, fatores sociais, respostas a questionários, escala de dor ou informações clínicas observacionais relevantes. Essa tabela amplia a capacidade analítica do modelo, incorporando dados contextuais e complementares.

A figura a 1.1 ilustra as principais tabelas clínicas do modelo OMOP.

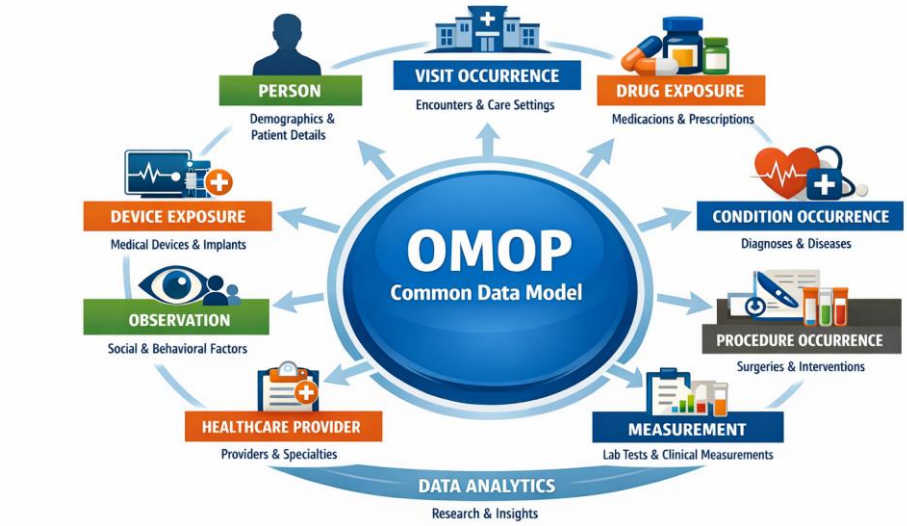

Figura 1.1 Visão geral das principais tabelas clínicas do CDM OMOP

Padronização e escalabilidade

Em conjunto, a estrutura do CDM OMOP oferece uma base sólida, coerente e orientada ao paciente, preparada para suportar análises descritivas, preditivas e causais, sendo amplamente utilizada em projetos de saúde digital, real world evidence e pesquisa clínica avançada.

Um dos aspectos mais distintivos do CDM OMOP é o suporte explícito à interoperabilidade semântica por meio do uso de vocabulários clínicos padronizados, integrados no chamado OMOP Standardized Vocabulary. Durante o processo de conversão dos dados, códigos locais utilizados pelos sistemas de origem são mapeados para conceitos padronizados, ao mesmo tempo em que os códigos originais são preservados. Essa estratégia garante rastreabilidade, flexibilidade analítica e comparabilidade internacional, promovendo harmonização semântica mesmo em ambientes altamente heterogêneos.

A adoção do CDM OMOP envolve um processo estruturado de extração, transformação e carga (ETL) dos dados de origem, que inclui a análise do modelo original, o mapeamento estrutural para o CDM, o mapeamento semântico para vocabulários padronizados, a validação técnica e clínica dos dados transformados e a avaliação sistemática da qualidade dos dados. Embora esse processo seja tecnicamente complexo e demande conhecimento especializado, ele representa um investimento estratégico, pois viabiliza a reutilização contínua dos dados para múltiplos estudos, sem necessidade de reestruturações adicionais.

Um dos impactos mais relevantes do CDM OMOP na pesquisa observacional reside em sua contribuição para a reprodutibilidade científica. Ao garantir que bases de dados distintas compartilhem a mesma estrutura e semântica, o modelo permite que protocolos analíticos padronizados sejam executados em diferentes contextos, produzindo resultados diretamente comparáveis. Essa característica é particularmente importante para estudos multicêntricos internacionais, validação externa de achados científicos, análises distribuídas em ambientes regulados e auditorias metodológicas, posicionando o CDM OMOP como uma ponte entre dados fragmentados e práticas científicas sistematizadas.

Apesar de seus benefícios amplamente reconhecidos, o CDM OMOP apresenta desafios relevantes. A complexidade do processo de ETL, a necessidade de governança robusta de dados e a dependência da qualidade dos dados de origem são limitações frequentemente discutidas na literatura. Além disso, a adoção institucional do modelo exige maturidade organizacional, capacitação técnica e compromisso estratégico de longo prazo. Esses desafios, entretanto, não diminuem o valor científico do modelo, mas evidenciam que sua implementação bem-sucedida depende da integração entre fatores técnicos, humanos e institucionais.

Ao fornecer dados longitudinalmente estruturados e semanticamente consistentes, o CDM OMOP também se consolida como uma base sólida para inovação analítica em saúde digital. Suas aplicações incluem modelos preditivos clínicos, aprendizado de máquina e inteligência artificial, avaliação de tecnologias em saúde, estudos de efetividade no mundo real, saúde populacional e iniciativas de medicina personalizada. Dessa forma, o CDM OMOP transcende sua função original como modelo de dados e se afirma como uma infraestrutura estratégica para a pesquisa observacional baseada em dados e evidências. Na prática, CDM OMOP **f**unciona como uma *camada de padronização* que transforma dados de saúde heterogêneos em uma estrutura comum, permitindo análises comparáveis, reprodutíveis e em larga escala.

Em síntese, o CDM OMOP representa um marco na evolução da pesquisa observacional em saúde. Ao integrar padronização estrutural, interoperabilidade semântica e suporte metodológico, o modelo viabiliza análises reprodutíveis, escaláveis e cientificamente rigorosas sobre dados do mundo real. Inserido no ecossistema OHDSI, o CDM OMOP não apenas enfrenta os desafios técnicos da heterogeneidade de dados, mas contribui para a redefinição de práticas científicas, promovendo colaboração global,

transparência e geração consistente de evidências, reforçando sua relevância estratégica para o presente e o futuro da ciência em saúde.

O ponto de partida para iniciativas de análise de dados em saúde é, em geral, um cenário caracterizado por dados de fonte heterogêneas. As organizações de saúde armazenam informações clínicas e administrativas em diversos sistemas distintos, como prontuários eletrônicos do paciente (EHR/PEP), sistemas administrativos e de faturamento, registros laboratoriais, bases de dados de farmácia e bancos de sinistros.

Esses diferentes repositórios refletem a diversidade de objetivos operacionais e contextos institucionais, resultando em dados que variam significativamente quanto à estrutura, incluindo tabelas, campos e relacionamentos, bem como quanto à codificação, que pode envolver classificações padronizadas como CID-10, códigos locais ou mesmo texto livre. Além disso, há variações importantes em termos de granularidade, completude e qualidade da informação.

Nesse contexto, o CDM OMOP não tem o objetivo de substituir os sistemas de origem. Seu papel é criar uma representação padronizada dos dados, preservando os registros originais, mas organizando-os de forma consistente para viabilizar análises comparáveis, reprodutíveis e escaláveis sobre dados observacionais em saúde.

O OMOP adota um princípio fundamental que sustenta todo o seu modelo analítico: toda análise é realizada exclusivamente sobre conceitos padronizados (*standard concepts*). Esse princípio é essencial para garantir a interoperabilidade semântica e a comparabilidade dos dados entre diferentes instituições e sistemas de origem.

Na prática, isso significa que cada dado clínico armazenado no CDM OMOP mantém duas representações complementares. A primeira é o conceito fonte (*source concept*), que corresponde ao código ou à descrição original utilizada no sistema de origem, preservando fielmente a informação conforme foi registrada. A segunda é o conceito padrão (*standard concept*), selecionado a partir dos vocabulários padronizados da OHDSI, e que representa o significado clínico unificado daquele dado.

Essa dupla representação permite que o CDM OMOP concilie dois objetivos fundamentais: preservar a rastreabilidade e a integridade dos dados originais e, ao mesmo tempo, viabilizar análises consistentes, reprodutíveis e independentes da codificação de origem. Como resultado, diferentes bases de dados podem ser analisadas de forma uniforme, mesmo quando utilizam vocabulários distintos nos sistemas fonte.

O mapeamento de vocabulários é o processo que viabiliza a interoperabilidade semântica no CDM OMOP. Na prática, ele garante que conceitos clínicos oriundos de diferentes sistemas, códigos e idiomas passem a ter o mesmo significado computacional, permitindo comparações e análises consistentes entre bases distintas. Sistemas de saúde utilizam múltiplos vocabulários e codificações. Esses códigos representam o mesmo conceito clínico com rótulos diferentes ou representam conceitos parecidos, mas não idênticos. Sem mapeamento, não é possível garantir que por exemplo, o diagnóstico de "Diabetes tipo 2" em um sistema seja tratado como o mesmo evento em outro.

A OHDSI mantém um repositório central de vocabulários integrados, que constitui a base para a interoperabilidade semântica no CDM OMOP. Esse repositório

consolida e organiza diferentes sistemas de codificação amplamente utilizados na área da saúde, definindo papéis claros entre vocabulários padrão e vocabulários fonte.

Nesse contexto, alguns vocabulários são adotados como padrões para análise, como o SNOMED CT, utilizado para representar condições clínicas, o RxNorm, empregado na padronização de medicamentos, e o LOINC, utilizado para exames laboratoriais e medições clínicas. Outros sistemas de codificação, como ICD-10, ICD-9, CPT e diversos códigos locais, são tratados como vocabulários fonte, preservando a forma original com que os dados foram registrados nos sistemas de origem.

Cada conceito presente nesse repositório possui um CONCEPT_ID único, que funciona como identificador universal dentro do modelo OMOP. Além disso, todo conceito pertence a um domínio específico, como *Condition*, *Drug*, *Measurement* ou *Procedure*, definindo o tipo de evento clínico ao qual ele se refere. Os conceitos também são conectados por relacionamentos semânticos explícitos, que descrevem, por exemplo, equivalências, hierarquias ou mapeamentos entre códigos fonte e conceitos padrão.

Essa organização estruturada dos vocabulários permite que dados provenientes de sistemas distintos sejam interpretados de forma uniforme, garantindo consistência semântica, reprodutibilidade das análises e comparabilidade entre diferentes bases de dados observacionais.

1.5 Mapeamentos dos dados e dos vocabulários

O mapeamento dos dados para o CDM OMOP é o processo de transformação dos dados brutos de sistemas de origem (prontuário eletrônico, sistemas administrativos, laboratórios, farmácias etc.) para a estrutura padronizada e semântica do OMOP Common Data Model. Esse processo é fundamental para garantir interoperabilidade, comparabilidade e reuso analítico dos dados. De forma prática, o mapeamento ocorre em etapas bem definidas, que combinam conhecimento técnico, clínico e semântico. As etapas estão descritas a seguir:

1. Entendimento da fonte de dados (source data)

Esta é a etapa inicial e uma das mais críticas no processo de mapeamento para o CDM OMOP. Antes de qualquer transformação técnica, é indispensável conhecer profundamente o sistema de origem, pois é nele que estão as regras implícitas, as práticas clínicas locais e as limitações que influenciarão todo o processo de ETL. Um mapeamento bem-sucedido depende diretamente da qualidade desse diagnóstico inicial.

Esse entendimento começa pela análise detalhada do modelo de dados da fonte, incluindo a identificação das tabelas existentes, seus campos, os relacionamentos entre elas e a forma como os dados são organizados. É fundamental compreender onde cada informação clínica ou administrativa está armazenada e como diferentes tabelas se conectam, evitando interpretações equivocadas ou associações incorretas durante a transformação.

Paralelamente, é necessário avaliar os tipos de dados utilizados, como datas, campos textuais, códigos estruturados ou textos livres. Datas podem ter granularidade

variada ou formatos inconsistentes, textos livres podem conter informações clínicas relevantes não estruturadas, e códigos podem seguir padrões distintos ou misturar conceitos diferentes em um mesmo campo. Identificar essas características desde o início evita perdas de informação e facilita a definição de regras adequadas de transformação.

Outro ponto essencial é o levantamento dos padrões locais de codificação. Sistemas de origem podem utilizar classificações como CID-10 para diagnósticos, catálogos internos para medicamentos, tabelas proprietárias para procedimentos ou combinações desses padrões. Compreender quais terminologias são usadas, como elas são mantidas e se existem variações locais é indispensável para o posterior mapeamento semântico para os vocabulários padronizados do OMOP.

A etapa também envolve uma análise cuidadosa da qualidade dos dados, identificando campos nulos, registros inconsistentes, duplicidades e lacunas frequentes. Essas características impactam diretamente as decisões de transformação e ajudam a definir regras de limpeza, consolidação ou inferência que serão aplicadas nas fases seguintes do ETL.

Durante esse processo de entendimento, busca-se responder a perguntas fundamentais, como onde exatamente estão registrados os diagnósticos, de que forma as visitas e episódios de cuidado são representados, se os dados de medicamentos correspondem a prescrições, dispensações ou administrações, e se existem datas confiáveis de início e fim para os eventos clínicos. As respostas a essas questões orientam tanto o mapeamento estrutural quanto a aplicação de regras clínicas e de negócio.

Identificar padrões locais de codificação é um passo essencial para garantir um mapeamento correto e semanticamente consistente para o CDM OMOP. Esse processo exige uma combinação de análise técnica, interpretação clínica e investigação operacional do sistema de origem. Na prática, ele ocorre por meio de várias abordagens complementares.

O primeiro passo é a análise exploratória do banco de dados de origem. Nessa etapa, examinam-se tabelas e campos que armazenam códigos clínicos, como diagnósticos, procedimentos, medicamentos e exames. A inspeção dos nomes das colunas, do tipo de dado e da distribuição dos valores frequentemente revela indícios claros de padrões utilizados, como códigos no formato CID-10, sequências numéricas internas ou combinações alfanuméricas proprietárias. Consultas simples de contagem e frequência ajudam a identificar quais códigos aparecem com maior recorrência e se seguem algum padrão reconhecível.

Em seguida, é fundamental analisar tabelas de domínio ou cadastros mestres. Muitos sistemas mantêm tabelas auxiliares que descrevem os códigos utilizados, com campos como descrição, categoria, status ou data de vigência. Essas tabelas costumam indicar explicitamente se um código pertence a um padrão internacional (como CID-10) ou se trata de um identificador interno criado pela própria instituição. Quando existem múltiplos catálogos para um mesmo tipo de informação, isso já sinaliza a presença de padrões locais distintos coexistindo.

Outro aspecto importante é a avaliação do contexto de uso dos códigos. O mesmo código pode representar coisas diferentes dependendo de onde é utilizado. Por exemplo, um código numérico pode significar um diagnóstico em uma tabela médica e um motivo administrativo em outra. Analisar como os códigos se relacionam com visitas, especialidades, tipos de atendimento ou faturamento ajuda a entender seu real significado clínico e evita interpretações equivocadas.

A consulta a especialistas do domínio, como equipes clínicas, de faturamento, farmácia ou TI do sistema de origem, é uma etapa indispensável. Muitas regras de codificação são implícitas e nunca foram formalmente documentadas. Profissionais que utilizam o sistema no dia a dia conseguem explicar, por exemplo, se determinados códigos representam diagnósticos confirmados ou suspeitos, se medicamentos correspondem a prescrições ou a dispensações, ou se certos códigos internos são equivalentes a padrões internacionais específicos.

Também é importante investigar documentações internas e manuais do sistema, quando disponíveis. Especificações técnicas, dicionários de dados, guias de integração ou materiais de treinamento costumam descrever os padrões de codificação adotados, inclusive exceções e extensões locais. Mesmo documentação desatualizada pode fornecer pistas valiosas sobre a origem e a intenção dos códigos.

Por fim, a identificação de padrões locais se consolida por meio de testes de mapeamento piloto. Ao tentar mapear um conjunto de códigos para os vocabulários OMOP (como SNOMED, RxNorm ou LOINC), torna-se evidente quais códigos possuem correspondência direta que exigem interpretação clínica e quais não encontram equivalência. Esse exercício ajuda a classificar os códigos em padrões reconhecidos, extensões locais ou lacunas semânticas.

Em conjunto, essas práticas permitem identificar com clareza quais padrões de codificação estão presentes na fonte de dados, como eles são utilizados e quais cuidados precisam ser adotados no processo de ETL. Esse entendimento é crucial para evitar perdas de significado, reduzir erros de mapeamento e garantir que os dados transformados para o CDM OMOP reflitam fielmente a realidade clínica capturada nos sistemas de origem.

Em síntese, este é o alicerce de todo o processo de transformação para o CDM OMOP e garante que os dados sejam interpretados corretamente, preserva o significado clínico original e evita erros que poderiam comprometer análises futuras. Sem essa etapa bem executada, há alto risco de perdas de informação, distorções semânticas e conclusões analíticas incorretas.

2. Mapeamento estrutural (ETL – Extract, Transform, Load)

O mapeamento estrutural, realizado por meio do processo de ETL (Extract, Transform, Load), é a etapa responsável por alinhar fisicamente os dados de origem ao desenho lógico do OMOP Common Data Model, que possui tabelas e campos rigidamente pré-definidos. Diferentemente do mapeamento semântico, que trata do significado dos dados, o mapeamento estrutural responde à organização e ao posicionamento correto de cada informação dentro do modelo, garantindo conformidade com a arquitetura do CDM.

A figura 1.2 ilustra o modelo CDM OMOP completo.

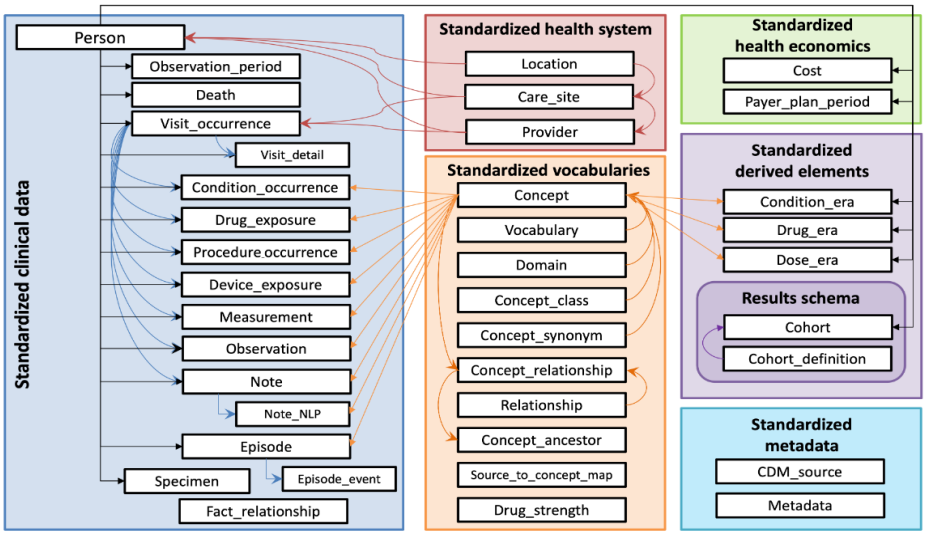

Figura 1.2 Visão geral da estrutura do *CDM OMOP V5* [9]

Nesse processo, define-se de forma explícita qual tabela do sistema de origem alimenta qual tabela do OMOP e como cada atributo da origem deve ser transformado e carregado nos campos correspondentes do CDM. Em outras palavras, trata-se de decidir "onde cada pedaço de dado deve morar dentro do OMOP". Por exemplo, informações cadastrais de pacientes são direcionadas para a tabela **PERSON**, enquanto registros de consultas, atendimentos ambulatoriais e internações são convertidos em registros da **VISIT_OCCURRENCE**. Da mesma forma, diagnósticos clínicos passam a compor a tabela **CONDITION_OCCURRENCE**, dados relacionados ao uso de medicamentos são estruturados na **DRUG_EXPOSURE**, e exames laboratoriais ou sinais vitais são representados na **MEASUREMENT**.

Durante essa etapa, não ocorre apenas a realocação direta de dados entre tabelas. São aplicadas diversas transformações estruturais essenciais para garantir consistência e integridade do modelo. Entre elas está a padronização de datas, assegurando formatos homogêneos e coerentes com o CDM, bem como a resolução de inconsistências comuns nos sistemas de origem, como datas parciais ou campos com múltiplos significados. Também ocorre a normalização de identificadores, na qual chaves oriundas de sistemas distintos são traduzidas para identificadores únicos e consistentes dentro do OMOP.

Um ponto central do mapeamento estrutural é a criação e gerenciamento dos identificadores do CDM, como PERSON_ID, VISIT_OCCURRENCE_ID e demais chaves primárias e estrangeiras. Esses identificadores garantem que todos os eventos clínicos estejam corretamente relacionados ao paciente e, quando aplicável, ao contexto da visita, permitindo análises longitudinais e preservando a integridade referencial do

[9] Fonte: https://ohdsi.github.io/CommonDataModel/

banco. A geração desses IDs segue regras claras e documentadas, pois qualquer inconsistência nessa etapa compromete todo o ecossistema analítico.

Em conjunto, o mapeamento estrutural assegura que os dados transformados respeitem o desenho técnico do CDM OMOP e estejam preparados para receber a camada semântica e as validações de qualidade. Ele é a base que permite que as ferramentas da comunidade OHDSI funcionem corretamente e que análises padronizadas sejam executadas de forma consistente, escalável e reprodutível.

3. Mapeamento semântico (vocabulários padronizados)

O mapeamento semântico por meio de vocabulários padronizados é um dos componentes mais críticos e conceitualmente relevantes do CDM OMOP. Diferentemente de uma simples padronização estrutural, essa etapa garante que os dados carregados compartilhem o mesmo significado clínico, independentemente da origem, do sistema ou do país onde foram gerados. Sem essa padronização terminológica, análises comparáveis e reprodutíveis seriam inviáveis.

Na prática, os códigos locais existentes nos sistemas de origem, muitas vezes criados para fins assistenciais, administrativos ou financeiros, precisam ser convertidos em conceitos padrão (standard concepts) definidos pelo OMOP. Esses conceitos padrão pertencem a vocabulários clínicos internacionais amplamente aceitos, que funcionam como uma "língua franca" para os dados de saúde. Exemplos comuns incluem o mapeamento de códigos CID-10 para SNOMED CT no caso de condições clínicas, a conversão de códigos locais de medicamentos para RxNorm, a padronização de exames laboratoriais em LOINC e a representação de procedimentos clínicos em SNOMED CT ou outros vocabulários suportados pelo modelo.

Esse processo de mapeamento é sustentado pelas tabelas centrais de vocabulário do OMOP. A tabela **CONCEPT** armazena todos os conceitos disponíveis, incluindo seus identificadores, nomes, domínios e a indicação de quais são conceitos padrão. Já a tabela **CONCEPT_RELATIONSHIP** descreve as relações entre conceitos, como mapeamentos de equivalência, hierarquia ou associação entre códigos não padronizados e seus correspondentes padrão. Complementarmente, durante o ETL, costumam ser criadas tabelas auxiliares de mapeamento, que documentam de forma explícita como cada código da fonte foi traduzido para o universo OMOP, servindo como referência técnica e de governança.

No entanto, a tabela **CONCEPT_RELATIONSHIP** é um dos elementos centrais da camada semântica do CDM OMOP, pois é responsável por descrever como os conceitos clínicos e administrativos se relacionam entre si dentro do ecossistema de vocabulários padronizados. Enquanto a tabela **CONCEPT** define cada conceito individualmente, indicando seu identificador, nome, domínio e se ele é padrão ou não, a **CONCEPT_RELATIONSHIP** explicita o significado das conexões entre esses conceitos, permitindo interoperabilidade semântica real entre diferentes sistemas e terminologias.

Na prática, essa tabela funciona como uma grande rede de relacionamentos direcionais, na qual cada registro conecta um conceito de origem a um conceito de destino por meio de um tipo de relacionamento bem definido. Além dos identificadores dos conceitos envolvidos, cada relacionamento possui um período de validade, garantindo que mudanças nos vocabulários ao longo do tempo sejam corretamente representadas e respeitadas durante análises e processos de ETL.

Um dos usos mais importantes da **CONCEPT_RELATIONSHIP** é o mapeamento de códigos não padronizados para conceitos padrão. Códigos locais, CID-10 ou terminologias proprietárias não são utilizados diretamente nas análises OMOP. Em vez disso, eles são conectados a conceitos padrão por meio de relacionamentos como "Maps to", que indica a equivalência semântica entre um conceito não padrão e seu correspondente padrão. Dessa forma, mesmo que dados distintos tenham origens diferentes, eles passam a ser interpretados de maneira uniforme pelas ferramentas analíticas do OMOP.

Além do mapeamento, a **CONCEPT_RELATIONSHIP** também sustenta hierarquias clínicas e relações semânticas complexas. Relacionamentos como "Is a" permitem identificar que um conceito específico pertence a uma categoria mais ampla, viabilizando análises que incluem automaticamente todos os subtipos de uma condição, procedimento ou medicamento. No domínio de medicamentos, por exemplo, relações como "Has ingredient" conectam produtos farmacêuticos aos seus princípios ativos, permitindo análises tanto no nível de formulação quanto no nível de substância.

Ferramentas da comunidade OHDSI, como ATLAS e ACHILLES, dependem fortemente da CONCEPT_RELATIONSHIP para funcionar corretamente. Ao definir uma coorte ou uma análise, essas ferramentas utilizam os relacionamentos para incluir conceitos equivalentes ou descendentes, garantindo consistência e reprodutibilidade entre diferentes bases OMOP. É essa lógica que permite que uma mesma definição analítica seja executada em múltiplas instituições e produza resultados comparáveis, mesmo quando os dados de origem são heterogêneos.

Por fim, a CONCEPT_RELATIONSHIP é parte de um conjunto de vocabulários que evolui continuamente. Por isso, seus relacionamentos são versionados e possuem validade temporal, o que reforça a importância de alinhar o processo de ETL à versão específica dos vocabulários utilizados. Em síntese, essa tabela é o que transforma o OMOP de um modelo estrutural em um modelo verdadeiramente semântico, permitindo tradução entre terminologias, navegação hierárquica e análises padronizadas em escala global.

Nem sempre, porém, existe uma correspondência direta e perfeita entre um código local e um conceito padrão. Nesses casos, o OMOP oferece diferentes estratégias para preservar o dado sem comprometer a integridade analítica. Um código pode ser mapeado para um conceito mais genérico, quando não há equivalência exata, mantendo o significado clínico essencial. Alternativamente, o dado pode ser mantido como um non-standard concept, permitindo sua preservação e rastreabilidade, mesmo que ele não seja utilizado diretamente nas análises padronizadas. Por fim, quando não há mapeamento

adequado, a situação deve ser explicitamente documentada como uma lacuna semântica, garantindo transparência sobre os limites da base de dados.

Em conjunto, o mapeamento semântico transforma dados locais e heterogêneos em informações semanticamente interoperáveis, capazes de sustentar estudos multicêntricos, análises preditivas e geração de Real World Evidence. É essa camada semântica que diferencia o OMOP de um simples modelo relacional e o torna uma base sólida para análises científicas confiáveis e comparáveis em escala global.

4. Aplicação de regras clínicas e de negócio

A aplicação de regras clínicas e de negócio é uma etapa essencial no processo de transformação dos dados para o CDM OMOP, pois nem toda informação proveniente do sistema de origem pode ser transferida de forma direta ou literal. Os dados operacionais refletem realidades locais, diferentes práticas assistenciais e limitações dos sistemas transacionais, o que exige interpretações cuidadosas para que o significado clínico seja corretamente representado no modelo padronizado.

Nesse contexto, são definidas e aplicadas regras de transformação que traduzem a realidade da fonte para a lógica analítica do OMOP. Um exemplo frequente é a necessidade de definir qual diagnóstico é primário ou secundário, especialmente em cenários onde o sistema de origem não diferencia explicitamente essa informação. Essa definição é crucial para diversas análises clínicas e epidemiológicas, como estudos de causa principal de internação ou estratificação de pacientes.

Outro ponto comum é a unificação de registros duplicados. Sistemas de origem podem gerar múltiplos registros para um mesmo evento clínico devido a correções, reprocessamentos ou registros paralelos. As regras de negócio determinam quando esses registros devem ser consolidados, evitando supercontagem de eventos e distorções analíticas.

A inferência de datas de término também é uma prática recorrente, sobretudo para medicamentos, internações ou condições clínicas onde apenas a data de início está disponível. Nesses casos, regras clínicas bem definidas permitem estimar períodos de exposição ou duração de eventos de forma consistente, sempre com critérios claros e documentados.

Além disso, muitas fontes de dados não classificam explicitamente o tipo de visita. Cabe às regras de transformação identificarem se um atendimento deve ser representado como internação, atendimento ambulatorial ou emergência, a partir de atributos como especialidade, local de atendimento, duração ou tipo de faturamento. Essa classificação é fundamental para análises que dependem do contexto assistencial.

Todas essas regras têm impacto direto nas análises futuras e, por isso, devem ser claramente documentadas e rigorosamente versionadas. A documentação garante transparência, reprodutibilidade e auditabilidade, enquanto o versionamento permite acompanhar evoluções, comparar resultados entre cargas e manter consistência ao longo do tempo. Dessa forma, a aplicação consciente de regras clínicas e de negócio transforma

dados brutos em informações analiticamente confiáveis, preservando o significado clínico e a robustez do CDM OMOP.

5. A carga para o CDM OMOP (Load)

Corresponde à etapa em que os dados, já devidamente transformados do ponto de vista estrutural e semântico, são inseridos nas tabelas finais do modelo. Nesse momento, todo o trabalho prévio de entendimento da fonte, aplicação de regras de negócio e mapeamento para vocabulários padronizados se materializa em um banco de dados que segue rigorosamente o desenho do OMOP Common Data Model.

Durante a carga, é essencial respeitar as chaves primárias, que garantem a unicidade dos registros, e as chaves estrangeiras, que asseguram a integridade referencial entre as tabelas. Relacionamentos como aqueles entre PERSON, VISIT_OCCURRENCE e as tabelas clínicas (por exemplo, CONDITION_OCCURRENCE, DRUG_EXPOSURE e MEASUREMENT) devem estar corretamente estabelecidos, permitindo que os eventos clínicos sejam corretamente associados ao paciente e ao contexto assistencial correspondente.

Outro aspecto crítico dessa etapa é o respeito aos tipos de dados definidos pelo modelo, como formatos de datas, campos numéricos, identificadores e flags conceituais.

A conformidade com esses padrões evita inconsistências técnicas, falhas em ferramentas analíticas e problemas de interoperabilidade. Além disso, a carga deve observar regras específicas do CDM, como o uso correto de identificadores de conceitos, a distinção entre conceitos padrão e não padrão e o preenchimento adequado de campos obrigatórios.

Garantir a integridade referencial é um requisito fundamental na implementação do CDM OMOP, pois assegura que os relacionamentos entre pacientes, visitas e eventos clínicos sejam consistentes, rastreáveis e confiáveis ao longo de todo o banco de dados. Na prática, isso envolve uma combinação de boas práticas de modelagem, regras técnicas, processos de ETL bem definidos e validações sistemáticas.

O primeiro passo para garantir integridade referencial é o controle rigoroso dos identificadores. Cada tabela do OMOP possui uma chave primária única (por exemplo, person_id, visit_occurrence_id, condition_occurrence_id) que deve ser gerada de forma consistente durante o ETL. Esses identificadores não podem ser reutilizados nem alterados após a carga, pois servem como base para todos os relacionamentos entre tabelas.

Em paralelo, é essencial assegurar que todas as chaves estrangeiras apontem para registros válidos. Por exemplo, todo person_id presente em tabelas como condition_occurrence, drug_exposure ou measurement deve existir previamente na tabela PERSON. Da mesma forma, quando um evento clínico referência uma visita (visit_occurrence_id), essa visita precisa existir na tabela VISIT_OCCURRENCE. Para isso, o fluxo de carga deve respeitar uma ordem lógica, carregando primeiro as tabelas centrais (como PERSON e VISIT_OCCURRENCE) e só depois as tabelas de eventos.

Outra prática fundamental é a padronização e persistência dos mapeamentos de origem. Durante o ETL, identificadores do sistema fonte deve ser consistentemente traduzidos para os identificadores OMOP por meio de tabelas de correspondência (lookup tables). Essas tabelas garantem que o mesmo paciente, visita ou evento seja sempre representado pelo mesmo identificador no CDM, evitando duplicidades ou quebras de relacionamento.

A validação automática da integridade também é indispensável. Ferramentas da comunidade OHDSI, como o Data Quality Dashboard (DQD), executam centenas de verificações que identificam registros órfãos, referências inválidas ou violações das regras estruturais do CDM. O ACHILLES, por sua vez, ajuda a detectar indiretamente problemas de integridade ao expor inconsistências de volume, lacunas inesperadas e padrões atípicos nos dados.

Além disso, é altamente recomendável implementar restrições de integridade no próprio banco de dados, sempre que tecnicamente possível. Constraints de chave primária e estrangeira, validações de tipo e regras de não-nulidade reforçam a integridade em nível físico, reduzindo o risco de cargas incorretas ou manuais não controlados.

Por fim, a integridade referencial depende fortemente de governança e documentação. Todas as regras de geração de identificadores, relacionamentos permitidos, exceções conhecidas e decisões de modelagem devem estar documentadas e versionadas. Isso garante que futuras evoluções do ETL ou do CDM mantenham coerência com as cargas anteriores.

A ordem correta de carga dos dados (load order) é um ponto crítico para o funcionamento adequado das *constraints* no CDM OMOP. Para que as chaves estrangeiras sejam respeitadas e a integridade referencial seja mantida, a carga deve seguir rigorosamente a hierarquia lógica do modelo, iniciando pelas tabelas centrais e, somente depois, avançando para as tabelas de eventos clínicos.

O processo deve começar pela tabela PERSON, que representa o paciente e é a base para todos os demais relacionamentos. Em seguida, deve-se carregar a tabela VISIT_OCCURRENCE, responsável por registrar os episódios de cuidado e servir de contexto para muitos eventos clínicos. Somente após essas duas etapas é que as tabelas de eventos clínicos devem ser populadas, como CONDITION_OCCURRENCE, DRUG_EXPOSURE, MEASUREMENT, PROCEDURE_OCCURRENCE e OBSERVATION, todas dependentes direta ou indiretamente das tabelas anteriores.

Seguir essa ordem é essencial porque os registros das tabelas de eventos fazem referência a pacientes e, frequentemente, a visitas já existentes. Caso a carga seja executada fora dessa sequência, o banco rejeitará os dados por violação de chave estrangeira, tornando o processo de ETL instável e comprometendo a integridade do CDM. Portanto, respeitar o load order não é apenas uma boa prática, mas um pré-requisito para um ambiente OMOP consistente, robusto e governado.

Ao final do processo de carga, o banco de dados já passa a refletir fielmente a arquitetura lógica e semântica do OMOP, tornando-se plenamente compatível com as ferramentas da comunidade OHDSI e apto para validações, análises descritivas e estudos

observacionais. Essa etapa marca a transição definitiva dos dados operacionais para um ativo analítico padronizado, pronto para uso científico e analítico em larga escala.

6. Validação e controle de qualidade

A validação e o controle de qualidade são etapas indispensáveis após a carga dos dados no CDM OMOP. Nessa fase, é fundamental assegurar que os dados estejam completos, que sejam clinicamente plausíveis, que mantenham coerência temporal entre os eventos e que respeitem rigorosamente as regras estruturais e semânticas do CDM. Esse processo garante que a transformação dos dados preservou tanto a integridade técnica quanto o significado clínico das informações. Para isso, normalmente são utilizadas ferramentas consolidadas da comunidade OHDSI.

O Data Quality Dashboard (DQD) aplica um amplo conjunto de validações padronizadas, avaliando conformidade, plausibilidade e completude de cada tabela e atributo do modelo. Já o ACHILLES é amplamente utilizado para análises descritivas e exploratórias, permitindo identificar padrões, outliers e possíveis inconsistências nos dados carregados. Em conjunto, essas ferramentas fornecem evidências objetivas da qualidade do banco, garantindo que ele esteja pronto para análises reprodutíveis, confiáveis e comparáveis em diferentes contextos e instituições.

7. Documentação e governança

A documentação e a governança são etapas fundamentais no processo de mapeamento para o CDM OMOP. Todo o trabalho realizado durante a transformação dos dados deve ser cuidadosamente documentado, incluindo as decisões clínicas adotadas, as regras de transformação aplicadas, as limitações conhecidas dos dados e o nível de cobertura semântica alcançada durante o mapeamento. Esse registro garante que o significado original dos dados, bem como as escolhas feitas ao longo do processo, seja compreendido e rastreáveis.

Essa documentação é essencial para assegurar a auditabilidade do processo, permitindo verificar a consistência e a qualidade dos dados ao longo do tempo. Além disso, ela viabiliza a evolução do modelo, facilitando ajustes, refinamentos e reprocessamentos futuros. Do ponto de vista científico, contribui para a transparência e reprodutibilidade dos estudos, e, do ponto de vista organizacional, permite o reuso do trabalho por outras equipes, reduzindo esforços, aumentando a confiabilidade e promovendo escalabilidade nas iniciativas baseadas em dados.

A governança de dados no contexto do CDM OMOP é sustentada por um conjunto integrado de ferramentas que atuam de forma complementar para garantir padronização, qualidade, rastreabilidade e transparência ao longo de todo o ciclo de vida dos dados. As ferramentas da comunidade OHDSI desempenham um papel central nesse processo, especialmente o ATLAS, que permite a definição, versionamento e compartilhamento de coortes, conceitos e estudos observacionais, assegurando que as análises sejam reproduzíveis e auditáveis. O ACHILLES complementa esse ecossistema ao gerar perfis descritivos dos dados, contribuindo para a compreensão da estrutura e da distribuição das

informações carregadas no CDM, enquanto o Data Quality Dashboard aplica verificações sistemáticas de conformidade, completude e plausibilidade, fortalecendo a governança da qualidade dos dados.

Além das ferramentas específicas do OMOP, a governança é ampliada por soluções de gestão de metadados e documentação, como catálogos de dados e repositórios colaborativos. Essas ferramentas são essenciais para registrar regras de ETL, decisões clínicas, limitações conhecidas, versões do CDM e dos vocabulários utilizados, criando uma memória institucional sólida e facilitando o reuso das informações por diferentes equipes. O controle de versão, geralmente apoiado por plataformas como Git e seus derivados, garante rastreabilidade técnica, permitindo acompanhar mudanças em scripts, mapeamentos semânticos e regras de negócio, além de viabilizar auditorias e a evolução controlada do modelo.

A governança semântica também é um pilar fundamental, apoiada pelos vocabulários oficiais do OMOP e por mecanismos de gestão dos mapeamentos entre códigos locais e conceitos padrão, assegurando consistência clínica e alinhamento com terminologias internacionais. Complementarmente, ferramentas de monitoramento operacional e observabilidade dos pipelines de dados reforçam a governança ao permitir a detecção de falhas, desvios de qualidade e impactos entre versões de carga. Em conjunto, essas ferramentas garantem que o CDM OMOP não seja apenas um repositório técnico, mas um ativo confiável, sustentável e preparado para suportar análises científicas e estratégias analíticas avançadas em escala.

1.5.1 Como o mapeamento dos vocabulários acontece na prática

O mapeamento de vocabulários no contexto do CDM OMOP ocorre principalmente durante o processo de ETL (Extract, Transform, Load) e envolve etapas bem definidas que garantem a interoperabilidade semântica dos dados.

Na primeira etapa, realiza-se a identificação do código fonte presente no sistema de origem. Isso inclui o reconhecimento do código original, como, por exemplo, o CID-10 E11.9 (Diabetes tipo 2 sem complicações), a identificação do vocabulário de origem ao qual esse código pertence e a contextualização do dado dentro do cenário clínico, como diagnóstico, medicamento ou exame laboratorial.

Essas informações são preservadas no CDM OMOP por meio de campos específicos que asseguram a rastreabilidade do dado original. O valor do código é registrado no campo *_source_value, enquanto o identificador do conceito correspondente no vocabulário fonte é armazenado no campo *_source_concept_id. Dessa forma, o modelo mantém um vínculo explícito entre o dado padronizado e sua origem.

Após a identificação do código fonte, o próximo passo é a busca do conceito padrão correspondente, que será utilizado nas análises. Esse mapeamento é realizado, em geral, por meio de relacionamentos semânticos do tipo *"Maps to"*, definidos e mantidos no repositório de vocabulários da OHDSI, especificamente na tabela CONCEPT_RELATIONSHIP.

Por exemplo, o código CID-10 E11.9 é mapeado para o conceito SNOMED CT 44054006, que representa *Type 2 diabetes mellitus*. Esse conceito SNOMED CT passa a ser o conceito padrão armazenado no campo analítico do CDM OMOP e é sobre ele que todas as consultas, definições de coorte e análises estatísticas serão realizadas. Esse mecanismo assegura que diferentes códigos de origem, ainda que provenientes de sistemas distintos, sejam interpretados de maneira uniforme, possibilitando análises consistentes, comparáveis e reprodutíveis em ambientes distribuídos.

No CDM OMOP, cada evento clínico armazenado passa a conter, de forma explícita, tanto a representação padronizada quanto a informação original do sistema de origem. O campo condition_concept_id registra o conceito padrão, geralmente proveniente de vocabulários como o SNOMED CT, e é esse identificador que serve de base para todas as análises e definições de coortes. Paralelamente, o campo condition_source_concept_id armazena o conceito original, como um código CID-10, enquanto o campo condition_source_value preserva o código ou texto exatamente como registrado no sistema fonte.

Dessa forma, o OMOP garante que nenhuma informação original seja perdida, ao mesmo tempo em que assegura que as análises sejam realizadas de maneira uniforme, utilizando conceitos padronizados e semanticamente consistentes.

1.5.2 Mapeamento automático versus mapeamento manual

O processo de mapeamento de vocabulários pode ocorrer de maneira automática ou manual, dependendo da disponibilidade e da qualidade dos relacionamentos existentes nos vocabulários padronizados.

O mapeamento automático utiliza relacionamentos previamente definidos nos vocabulários do ecossistema OHDSI, especialmente aqueles do tipo *"Maps to"*. Esse tipo de mapeamento apresenta boa cobertura para codificações amplamente utilizadas, como CID, LOINC e RxNorm, sendo, em geral, um processo rápido, escalável e consistente entre diferentes bases de dados.

Já o mapeamento manual torna-se necessário em situações mais complexas, como quando não há uma correspondência direta nos vocabulários padrão, quando os dados utilizam códigos locais ou quando os conceitos clínicos são extremamente específicos. Nesses casos, especialistas clínicos e analistas de dados avaliam cuidadosamente o significado clínico do código de origem, selecionam o conceito padrão mais adequado e documentam as decisões tomadas. Essa documentação é essencial para garantir rastreabilidade, transparência e governança dos dados, especialmente em estudos multicêntricos ou regulatórios.

1.5.3 Análises padronizadas sobre o CDM

Uma vez que os dados são transformados e organizados no CDM OMOP, abre-se a possibilidade de realizar análises padronizadas e reprodutíveis, eliminando a necessidade de reescrever códigos específicos para cada fonte de dados. Essa padronização permite que diferentes bases, originalmente heterogêneas em estrutura, terminologia e

granularidade, passem a falar a mesma "linguagem analítica", reduzindo significativamente o esforço técnico e aumentando a confiabilidade dos resultados.

Nesse contexto, a definição de coortes torna-se mais clara e consistente, pois é baseada em conceitos clínicos padronizados e amplamente validados pela comunidade internacional. Assim, critérios como diagnósticos, procedimentos, medicamentos e eventos clínicos seguem vocabulários comuns, possibilitando a criação de coortes complexas, como, por exemplo, "pacientes adultos com diabetes tipo 2 em uso de metformina", de forma uniforme e comparável entre instituições e países. Isso não apenas melhora a qualidade das análises, mas também favorece a transparência e a replicabilidade dos estudos.

As ferramentas desenvolvidas pela comunidade OHDSI operam diretamente sobre o CDM OMOP, o que viabiliza a execução de uma ampla gama de estudos analíticos avançados. Entre eles, destacam-se os estudos de segurança de medicamentos, voltados à identificação de eventos adversos e riscos associados a terapias; as análises de efetividade comparativa, que permitem avaliar diferentes intervenções em condições do mundo real; a caracterização detalhada de coortes, essencial para compreender perfis populacionais; e o desenvolvimento de modelos preditivos, que apoiam a tomada de decisão clínica e estratégica por meio de técnicas de ciência de dados e aprendizado de máquina.

Um dos maiores diferenciais desse ecossistema é a possibilidade de reutilização integral dos scripts analíticos. O mesmo código pode ser executado em um hospital no Brasil, em uma seguradora de saúde na Europa ou em uma base nacional dos Estados Unidos, mantendo a lógica analítica e os critérios de estudo inalterados. Dessa forma, não é necessário mover ou centralizar os dados, o que fortalece a governança, a privacidade e a conformidade regulatória. Apenas o código analítico é compartilhado, promovendo colaboração global, escalabilidade e geração de evidências robustas em larga escala.

1.5.4 Funcionamento em rede distribuída

O funcionamento do OMOP no contexto de redes distribuídas é um dos seus pilares mais importantes e o que viabiliza, na prática, a colaboração entre múltiplas instituições de forma segura, escalável e regulatoriamente adequada. Esse modelo é baseado em uma arquitetura federada, na qual cada organização mantém total controle sobre seus próprios dados, preservando sua autonomia institucional e suas responsabilidades legais. Os dados clínicos, administrativos ou assistenciais não são transferidos ou centralizados em um repositório único; eles permanecem armazenados localmente, dentro dos ambientes seguros de cada instituição participante.

Nesse contexto, o que efetivamente circula entre os participantes da rede é o código analítico padronizado, desenvolvido de acordo com o CDM OMOP e com as ferramentas da comunidade OHDSI. Esse código é compartilhado e executado localmente em cada base de dados, respeitando exatamente a mesma lógica analítica, definições de coorte, critérios de inclusão e exclusão, métricas e modelos estatísticos. Após a execução, apenas os resultados agregados, previamente definidos e anonimizados, são consolidados

para análise conjunta, sem exposição de dados identificáveis ou sensíveis em nível individual.

Esse modelo federado traz ganhos diretos e estratégicos. Do ponto de vista regulatório, ele permite a conformidade com legislações rigorosas de proteção de dados, como a LGPD no Brasil, o GDPR na União Europeia e o HIPAA nos Estados Unidos, uma vez que não há compartilhamento de dados pessoais entre instituições ou países. A governança da informação permanece local, enquanto a colaboração científica ocorre de forma coordenada e transparente, baseada em métodos comuns e auditáveis.

Além disso, o funcionamento em rede distribuída torna viáveis estudos multicêntricos globais, envolvendo hospitais, sistemas de saúde, seguradoras, agências regulatórias e centros de pesquisa, mesmo quando esses participantes estão sujeitos a diferentes contextos legais e operacionais. Todos contribuem para a geração de evidências utilizando seus próprios dados, mas respondendo às mesmas perguntas de pesquisa, o que fortalece a comparabilidade e a robustez dos achados.

Por fim, essa abordagem possibilita a escala massiva de análises com dados do mundo real, algo fundamental para avaliações de segurança de medicamentos, efetividade comparativa, vigilância pós-comercialização, estudos epidemiológicos e modelos preditivos. Ao combinar grandes volumes de dados distribuídos geograficamente, sem comprometer privacidade ou soberania institucional, o OMOP viabiliza uma ciência colaborativa em larga escala, alinhada às necessidades da saúde baseada em evidências e à transformação digital dos sistemas de saúde.

1.5.5 A conexão com a IA: OHDSI Keeper

Numa época com destaque aos avanços contínuos nas tecnologias de Deep Learning e Inteligência Artificial (IA), é fácil esquecer que nada disso seria possível sem dados e evidências confiáveis que alimentem e treinem estes algoritmos.

Para gerar evidências confiáveis precisamos superar os obstáculos que se apresentam no caminho dela, representados como preocupações tanto metodológicas quanto técnicas. A figura 1.3 apresenta o caminho e as perguntas para as quais devemos dar resposta.

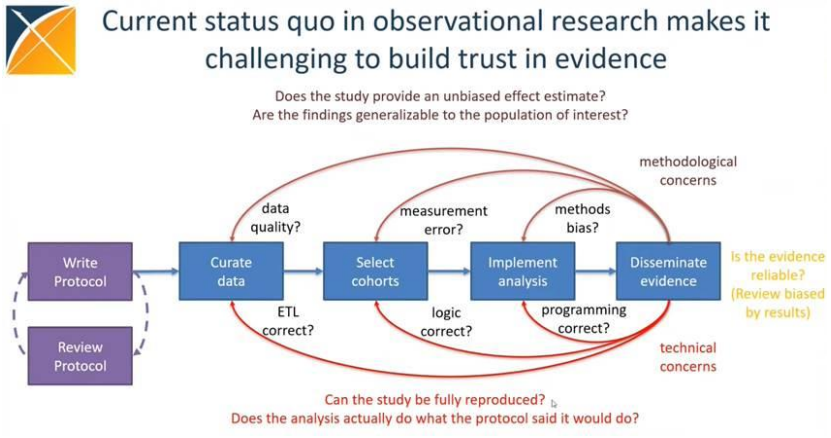

Figura 1.3 O caminho da evidência confiável e seus obstáculos

Evidências confiáveis devem ser **repetíveis**, o que significa que os pesquisadores devem esperar produzir resultados idênticos ao aplicar a mesma análise aos mesmos dados para qualquer questão específica. Implícita nesse requisito mínimo está a noção de que a evidência é o resultado da execução de um processo definido com uma entrada específica e deve estar livre de intervenção manual ou tomada de decisão posterior ao longo do processo.

Idealmente, evidências confiáveis devem ser **reproduzíveis**, de modo que um pesquisador diferente possa realizar a mesma tarefa de executar uma determinada análise em um determinado banco de dados e esperar produzir um resultado idêntico ao do primeiro pesquisador. A reprodutibilidade exige que o processo seja totalmente especificado, geralmente em formato legível por humanos e executável por computador, de forma que nenhuma decisão do estudo fique a critério do investigador.

A solução mais eficiente para alcançar repetibilidade e reprodutibilidade é usar rotinas analíticas padronizadas que tenham entradas e saídas definidas e aplicar esses procedimentos a bancos de dados com controle de versão.

É mais provável que tenhamos confiança nas nossas evidências se pudermos demonstrar que elas são **replicáveis**, ou seja, se a mesma questão abordada usando a mesma análise com dados semelhantes produzir resultados semelhantes. A Figura 1.4 apresenta estes conceitos.

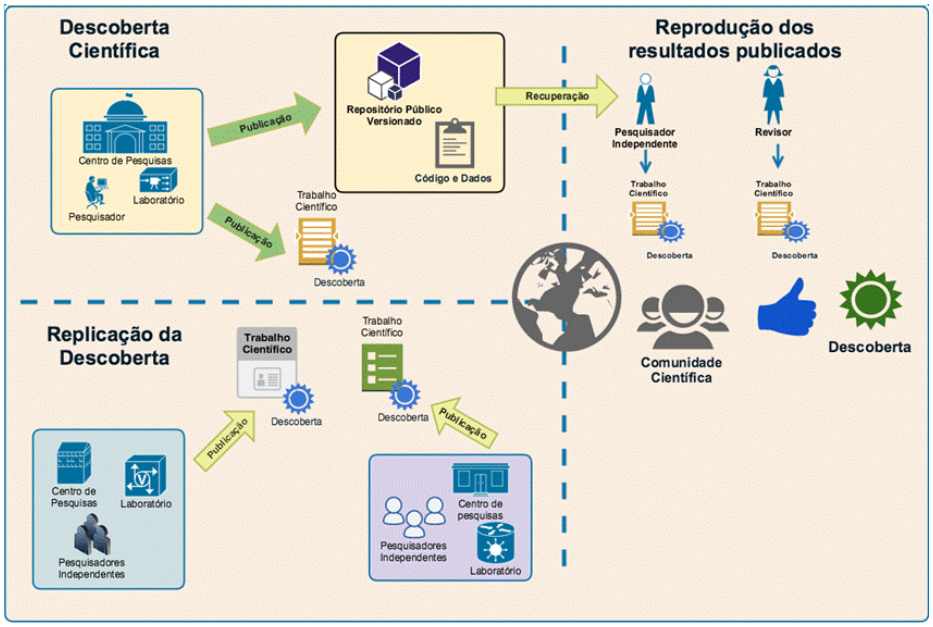

Figura 1.4 Pesquisa reprodutível - Replicação e reprodução

No contexto da predição em nível de paciente, a replicabilidade destaca o valor da validação externa e a capacidade de avaliar o desempenho de um modelo treinado em um banco de dados, observando sua acurácia discriminativa e calibração quando aplicado a um banco de dados diferente.

Em circunstâncias em que análises idênticas são realizadas em diferentes bancos de dados e ainda apresentam resultados consistentemente semelhantes, aumentamos nossa confiança na **generalização** das evidências. Um valor fundamental da rede de pesquisa OHDSI é a diversidade representada por diferentes populações, regiões geográficas e processos de coleta de dados.

Evidências confiáveis devem ser **robustas**, o que significa que as conclusões não devem ser excessivamente sensíveis às escolhas subjetivas que podem ser feitas em uma análise. Se houver métodos estatísticos alternativos que possam ser considerados potencialmente razoáveis para um determinado estudo, pode ser reconfortante verificar se os diferentes métodos produzem resultados semelhantes ou, inversamente, alertar caso sejam encontrados resultados discordantes. (Madigan, Ryan e Schuemie 2013) Para a estimativa do efeito em nível populacional, as análises de sensibilidade podem incluir escolhas de alto nível no desenho do estudo, como a aplicação de um desenho de coorte comparativa ou de séries de casos autocontrolados, ou podem se concentrar em considerações analíticas inerentes ao desenho, como a realização de pareamento por escore de propensão, estratificação ou ponderação como estratégia de ajuste de fatores de confusão na estrutura de coorte comparativa.

Por último, mas talvez o mais importante, as evidências devem ser **calibradas**. Não basta ter um sistema gerador de evidências que produza respostas para perguntas desconhecidas se o desempenho desse sistema não puder ser verificado. Espera-se que um sistema fechado tenha características operacionais conhecidas, que devem poder ser medidas e comunicadas como contexto para a interpretação de quaisquer resultados produzidos pelo sistema. Artefatos estatísticos devem poder ser demonstrados empiricamente como tendo propriedades bem definidas, como um intervalo de confiança de 95% com probabilidade de cobertura de 95% ou uma coorte com probabilidade prevista de 10% com uma proporção observada de eventos em 10% da população. Um estudo observacional deve sempre ser acompanhado por diagnósticos que testem as premissas relativas ao delineamento, aos métodos e aos dados. Esses diagnósticos devem se concentrar na avaliação das principais ameaças à validade do estudo: viés de seleção, fatores de confusão e erros de medição. Os controles negativos têm se mostrado uma ferramenta poderosa para identificar e mitigar erros sistemáticos em estudos observacionais (Schuemie et al. 2016; Schuemie, Hripcsak, et al. 2018; Schuemie, Ryan, et al. 2018). A figura 1.5 resume os atributos desejáveis de uma evidência robusta.

Desired attribute	Question	Researcher	Data	Analysis		Result
Repeatable	Identical	Identical	Identical	Identical	=	Identical
Reproducible	Identical	Different	Identical	Identical	=	Identical
Replicable	Identical	Same or different	Similar	Identical	=	Similar
Generalizable	Identical	Same or different	Different	Identical	=	Similar
Robust	Identical	Same or different	Same or different	Different	=	Similar
Calibrated	Similar (controls)	Identical	Identical	Identical	=	Statistically consistent

Figura 1.5 Atributos da evidência robusta

O sistema da OHDSI

O processo de converter RWD (Real World Data) em RWE (Real World Evidence) levou a OHDSI a construir um sistema de ciência aberta no lugar de gerar apenas componentes que precisam ser combinados posteriormente.

Este sistema, mostrado na figura 1.6, é o que possibilita a colaboração entre os parceiros da rede OHDSI e é a solução para criar confiança na geração de evidências do mundo real (RWE).

Ele consiste das seguintes etapas:

- Avaliação da qualidade dos dados
- Desenvolvimento e avaliação de Fenótipos
- Avaliação da confiabilidade das análises

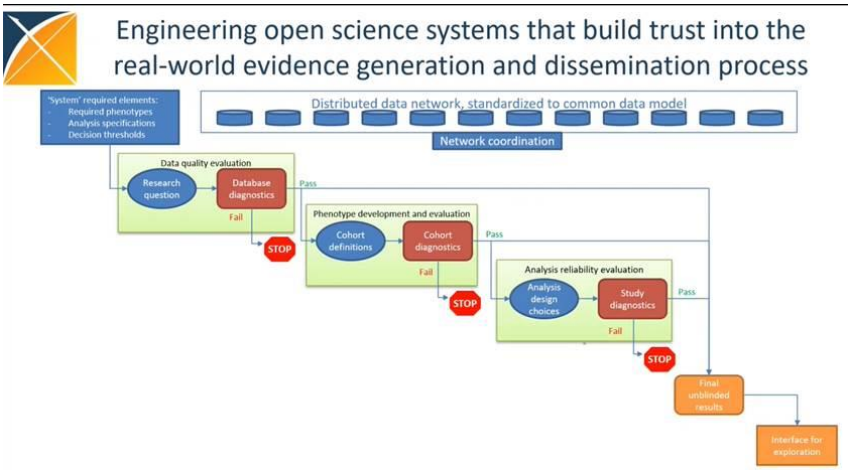

Figura 1.6 Sistema OHDSI de geração de evidências

Para assegurar confiança no processo de conversão de RWD (Real Worl Data) em RWE (Real World Evidence) é necessário estabelecer critérios e controles em cada etapa que validem os resultados e não deixem o processo continuar caso estes critérios não sejam cumpridos.

A solução para assegurar a confiabilidade na produção de evidências, reside na criação de sistemas que tenham entradas e saídas definidas de forma consistente, que possam ser avaliadas objetivamente a priori, mitigando a incidência de vieses e a subjetividade inerente ao processo de interpretação.

Tal processo basicamente exige que, a priori, se defina o conjunto de questões de pesquisa, seja uma única questão sobre uma exposição e um desfecho, ou um conjunto de questões, como por exemplo, para estudar a segurança e a eficácia comparativa dos tratamentos para diabetes.

Seja qual for o conjunto de questões de pesquisa, precisamos desenvolver critérios para avaliar a qualidade dos dados, decidindo sobre critérios objetivos de

aprovação/reprovação, quando um banco de dados é ou não adequado para uma determinada questão.

E para aqueles que são adequados, precisamos passar por um processo para desenvolver e avaliar fenótipos, para dispor de critérios objetivos de aprovação/reprovação sobre quais desfechos e exposições podem ou não ser estudados de forma confiável em nossos bancos de dados.

E para aqueles que são aprovados, precisamos ter critérios para avaliar a confiabilidade de nossa análise. Precisamos provar que nossos métodos são imparciais, que realmente produzem características operacionais aceitáveis. E se produzirem, somente nesse momento então, estaremos no ponto de estarmos confiantes em gerar resultados.

O desenvolvimento e avaliação de um sistema que gera evidências de qualidade é o objetivo e desafio principal da OHDSI como comunidade.

Como parte desse sistema, um conjunto de componentes em R foi desenvolvido para auxiliar as tarefas de validação de cada uma das etapas. O projeto HADES (Health-Analytics Data to Evidence Suite) agrupa todos estes componentes.

Um dos componentes é o ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems) (Huser et al. 2018). O ACHILLES é uma ferramenta de software que fornece caracterização e visualização de um banco de dados em conformidade com o CDM e gera métricas que permitem validar a qualidade dele.

Validação de coortes: OHDSI Keeper

A validação de fenótipos é importante porque nem sempre uma seleção de uma coorte pode se basear apenas na presença ou ausência de um determinado diagnóstico. Temos que lembrar que a pesquisa observacional utiliza dados coletados com outras finalidades que não a pesquisa.

O uso secundário do dado assistencial se constitui como uma fonte de informação importante para pesquisa de desfechos, porém apresenta uma série de desafios metodológicos peculiares a este tipo de fonte de dados (Abrahão M.T, et al, 2019)

Muitas vezes dados de exames laboratoriais ou outras evidências como prescrições de medicamentos específicos indicam um possível candidato a uma coorte mesmo que o diagnóstico correspondente não tenha sido registrado. Desfechos classificados incorretamente podem enviesar os resultados das análises.

É recomendado revisar cada caso potencial (FDA - regulatory information) e no caso de isso ser inviável, selecionar uma amostra para mensurar o desempenho operacional (PPV, sensibilidade) do fenótipo e realizar análises quantitativas de viés para ajustar os resultados.

Precisamos garantir métodos confiáveis para medir o poder de seleção da nossa definição de coorte (fenótipo), comparar com um Gold Standard e calcular o seu Valor

Preditivo Positivo (PPV), quer dizer, quantos casos a nossa corte selecionou do Gold Standard (PPV = True Positives / (True Positives + False Positives)). A Figura 1.7 ilustra a matriz de confusão para cálculo do PPV.

		Gold Standard: True	Gold Standard: False
Cohort Definition	True	True Positive	False Positive
	False	False Negative	True Negative

Figura 1.7 Matriz de confusão

Em particular, para desenvolver e principalmente validar fenótipos e cuidar destas situações, foi construído o Knowledge-Enhanced Electronic Profile Review (KEEPER).

O **OHDSI Keeper** é uma ferramenta (pacote R) desenvolvida pela comunidade OHDSI, projetada para otimizar e escalar o processo de validação de casos em estudos observacionais de saúde (OHDSI 2023 Plenary). Ele auxilia na validação de fenótipos (definição de coortes), permitindo a revisão de perfis de pacientes. Ele suporta a avaliação da concordância entre a revisão humana e métodos automatizados (como LLMs ou algoritmos) utilizando métricas como o Kappa de Cohen.

Tradicionalmente, validar se um paciente realmente teve uma condição específica (como diabetes mellitus tipo II), exige que revisores humanos leiam prontuários médicos complexos, o que é lento e caro. O Keeper atua como uma ponte eficiente entre os dados estruturados e a revisão detalhada.

Os princípios do KEEPER são: (Schuemie, 2025)

- Adesão ao raciocínio clínico: O KEEPER aplica princípios e etapas gerais do raciocínio clínico diagnóstico.
- Padronização: Tanto a entrada quanto a saída são padronizadas entre fontes de dados e diagnósticos.
- Redução da dimensionalidade: Extrair apenas informações relevantes.

Funcionalidades do Keeper:

- Extrai dados em nível de paciente para:
  a) uma amostra aleatória de pacientes em uma coorte ou
  b) pacientes em uma lista especificada pelo usuário
  e os formata de acordo com os princípios do KEEPER.
- Permite a revisão de perfis de pacientes por humanos por meio de um aplicativo Shiny interativo.
- Permite a revisão de perfis de pacientes por ferramentas de IA (LLMs).

Roteiro do funcionamento do OHDSI Keeper:

1. **Definição de Conceitos (Concept Sets)**: O profissional define quais diagnósticos, sintomas e medicamentos são relevantes para o fenótipo que deseja validar.

2. **Extração de dados do Modelo OMOP**: O sistema utiliza dados já padronizados no CDM OMOP. Extrai informações relevantes de um paciente, como diagnósticos, medicamentos, procedimentos e medições laboratoriais.
3. **Geração de Relatórios Padronizados**: Em vez de obrigar o revisor a navegar por sistemas de prontuário eletrônico (PEP) nativos e confusos, o Keeper gera um "resumo" estruturado e padronizado do histórico do paciente.
4. **Apoio à Decisão com LLMs**: Versões modernas do Keeper estão integradas a Modelos de Linguagem de Grande Escala (LLMs). Isso permite que a ferramenta analise os dados e forneça uma "sugestão" de diagnóstico para o revisor, apontando evidências nos dados que confirmam ou descartam o caso.
5. **Aumento de Escala**: Ao fornecer um formato de saída padronizado e o auxílio de IA, o Keeper permite que os pesquisadores validem centenas de casos em uma fração do tempo que levaria uma revisão manual tradicional, mantendo um nível de concordância similar ao de especialistas humanos (Schuemie, 2025).
6. **Fluxo de Consenso:**
   Para estudos de alta qualidade (FDA - regulatory information), recomenda-se:
   - **Revisão em Dupla**: Dois revisores analisam os mesmos IDs de pacientes.
   - **Adjudicação**: Um terceiro revisor (mais experiente) resolve as divergências quando os dois primeiros discordam.
7. Para centralizar as decisões de múltiplos revisores, o **OHDSI Keeper** geralmente utiliza uma estrutura de banco de dados ou arquivos de log compartilhados. Isso permite o cálculo da taxa de concordância (inter-rater reliability).

Principais Benefícios do Keeper

- **Eficiência**: Reduz a necessidade de exploração extensiva de prontuários brutos.
- **Escalabilidade**: Permite a realização de estudos com coortes muito maiores.
- **Reprodutibilidade**: Como os dados vêm do modelo OMOP, a metodologia de validação pode ser aplicada em diferentes instituições que usem o mesmo padrão. (Abrahão, M. T., 2019)

A primeira etapa, a definição e seleção de conceitos (Concept Sets), é quem guia a qualidade final do fenótipo. Para estruturar os conjuntos de conceitos, precisa focar em capturar os elementos que um médico usaria para confirmar um diagnóstico. O Keeper não precisa apenas do código da doença, mas de todo o contexto clínico ao redor dela para facilitar a revisão. (Schuemie, 2025; Ostropolets, 2023).

Estrutura Recomendada para o Keeper

É recomendado dividir os Concept Sets em categorias que reflitam o raciocínio clínico: (Schuemie, 2025)

1. **Apresentação Clínica (Sintomas/Sinais)**: Inclua conceitos de queixas e achados físicos relacionados (ex: "dor abdominal", "febre").
2. **Histórico e Comorbidades**: Fatores de risco ou condições prévias que tornam o diagnóstico plausível.

3. **Procedimentos Diagnósticos**: Exames de imagem, biópsias ou testes laboratoriais específicos (ex: "colonoscopia", "níveis de troponina").
4. **Tratamentos e Medicamentos**: Medicamentos usados para tratar a condição, que servem como evidência de suporte.
5. **Diagnósticos Diferenciais**: Conceitos que podem "confundir" o revisor e precisam ser monitorados para exclusão.

Como Criar os Concept Sets

Existem três formas principais de gerar esses conjuntos:

- **Uso do Atlas OHDSI**: É a forma mais visual. A partir de uma busca dos termos no Atlas (Atlas Tutorial) seleciona os conceitos padrão (Standard Concepts) e exporta no formato JSON para usar no R.
- **Geração Automática (Função generateKeeperConceptSets)** (OHDSI Keeper Doc.): O próprio pacote Keeper possui uma função experimental que usa LLMs para sugerir os conjuntos de conceitos baseados no nome da doença.
- **Via R (Capr)**: O pacote Capr (Concept Sets in Capr) permite criar conjuntos de conceitos programaticamente usando IDs do OMOP.

A estrutura dos Concept Sets deve abranger os seguintes critérios:

1. Diagnóstico Principal (Critério de Entrada)

Incluir o código principal e seus descendentes para capturar variações (ex: DM2 com complicações renais).

2. Laboratórios (Evidência Objetiva)

Estes são cruciais para o Keeper mostrar ao revisor se existe algum exame que indique presença do diagnóstico. Por exemplo, Hemoglobina Glicada para diagnóstico de diabetes.

3. Medicamentos (Tratamento)

A presença desses fármacos ajuda a confirmar que o paciente está sendo tratado especificamente para uma doença em particular.

4. Diagnósticos Diferenciais (Exclusão)

Importante para o Keeper sinalizar se o paciente pode ter sido classificado incorretamente.

O Keeper então processará esses dados e gerará uma interface (ou planilhas) onde o revisor médico poderá marcar "Sim", "Não" ou "Inconclusivo" para cada paciente.

Na pasta de saída podemos encontrar:

- **Arquivos HTML Individuais**: Cada paciente terá uma página exclusiva mostrando a jornada com o diagnóstico selecionado.
- **Timeline Visual**: Uma linha do tempo com os diagnósticos e as prescrições correspondentes.

- **Tabela de Laboratórios**: Os exames relevantes aparecerão destacados no topo da lista.

Tudo isto facilita e padroniza o processo de revisão de casos, a validação por revisores e a apresentação de resultados.

Para gerar o **relatório de concordância** (Kappa de Cohen ou concordância percentual), o **OHDSI Keeper**, através da função:
computeCohortOperatingCharacteristics, analisa as divergências entre os votos dos revisores armazenados na sua tabela de resultados, juntamente com sensibilidade, especificidade, valor preditivo positivo e AUC.

O uso de LLM no sistema de adjudicação do Keeper obteve uma performance razoável, com um grau de concordância com o Gold Standard semelhante aos revisores humanos (Schuemie, 2025).

O uso combinado do KEEPER com LLMs facilita a rápida avaliação de grandes volumes de casos, permitindo o cálculo do PPV e da sensibilidade para inúmeros fenótipos e fontes de dados. Devido a preocupações com a privacidade dos dados dos pacientes, para não transmitir perfis para fora da instituição, são utilizados LLMs hospedados localmente.

Embora o uso de Modelos de Linguagem de Grande Escala (LLMs) na prática clínica ainda seja controverso, este caso de uso que visa aumentar a confiabilidade das evidências a partir de dados observacionais, parece promissor e de baixo risco.

1.6 Considerações finais e conclusões

Em síntese, a pesquisa observacional baseada em dados do mundo real assume um papel cada vez mais estratégico na produção de evidências científicas aplicáveis à prática clínica e à gestão em saúde. Nesse cenário, o ecossistema OMOP/OHDSI destaca-se como uma iniciativa fundamental para enfrentar os desafios de heterogeneidade, escalabilidade e reprodutibilidade inerentes aos dados observacionais, ao oferecer um modelo comum de dados, vocabulários padronizados e ferramentas analíticas consolidadas. Ao converter dados heterogêneos, provenientes de diferentes sistemas e formatos, em uma estrutura comum e padronizada, o OMOP cria uma base sólida para análises consistentes e comparáveis.

Essa padronização viabiliza a realização de análises reprodutíveis, estudos multicêntricos e modelos preditivos em larga escala, permitindo que diferentes instituições utilizem a mesma lógica analítica sobre bases distintas. Dessa forma, o CDM OMOP se estabelece como um pilar essencial para iniciativas de Real World Evidence, projetos de saúde digital e estratégias de análises avançadas.

Numa época com destaque aos avanços contínuos nas tecnologias de aprendizado de máquina e Inteligência Artificial (IA), fica fácil esquecer que nada disso seria possível sem dados e evidencias confiáveis que alimentem e treinem estes algoritmos.

A iniciativa da OHDSI se integra em perfeita sincronia com o avanço destas tecnologias fornecendo o que elas mais precisam: dados padronizados e evidências confiáveis, para poder realizar o potencial destes avanços. Se hoje podemos integrar a IA no fluxo do atendimento com confiabilidade, é porque o esforço da OHDSI para melhorar a qualidade de informação médica começou há mais de 10 anos atrás.

Veremos num futuro próximo, sistemas médicos que incluem estas tecnologias de IA no dia a dia do atendimento clínico, sem perceber o imenso valor e importância da infraestrutura da OHDSI gerando evidências confiáveis e seguras para manter o treinamento da IA atualizado.

A Inteligência Artificial no ecossistema OHDSI é utilizada como apoio à detecção, definição e validação de fenótipos clínicos, permitindo identificar padrões complexos em grandes volumes de dados observacionais padronizados no CDM OMOP. Por meio de técnicas de aprendizado de máquina, a IA auxilia na descoberta de fenótipos, na construção de modelos preditivos mais sensíveis e específicos e na avaliação do desempenho das definições fenotípicas. Integrada ao ATLAS, a IA pode sugerir conjuntos de conceitos e regras temporais, sempre com validação por especialistas clínicos. Além disso, possibilita análises distribuídas e reprodutíveis, respeitando a governança e a privacidade dos dados. Dessa forma, a IA atua como um acelerador da fenotipagem computacional, complementando o rigor metodológico da OHDSI e fortalecendo a geração de evidências do mundo real.

Por fim, conclui-se que o ecossistema OMOP/OHDSI representa uma infraestrutura essencial para o presente e o futuro da pesquisa observacional em saúde. Sua relevância tende a se intensificar à medida que sistemas de saúde incorporam, de forma cada vez mais integrada, análises avançadas e soluções baseadas em inteligência artificial no cuidado assistencial. Muitas dessas aplicações ocorrerão de maneira quase invisível ao usuário final, mas apoiadas em uma base sólida, confiável e sustentável de padronização de dados e geração de evidências, justamente o legado e a contribuição contínua da comunidade OHDSI para a ciência e a saúde global.

1.7 Referências bibliográficas

Abrahão, M. T., Nobre, M. R., Madril, P. J., O estado da arte em pesquisa observacional de dados de saúde: A iniciativa OHDSI 2019.
https://books-sol.sbc.org.br/index.php/sbc/catalog/book/29
https://books-sol.sbc.org.br/index.php/sbc/catalog/view/29/98/248

ATLAS Tutorial: Explore Concept Sets
Video: https://www.youtube.com/watch?v=mfjxNwn3KkM

Ricotta EE, Bustos Carrillo FA, Angelli-Nichols S, Barugahare J, Benton A, Carlson CJ, et al. Observational research in epidemic settings: a roadmap to reform. BMJ Global Health. 2025;10:e017981. https://doi.org/10.1136/bmjgh-2024-017981

Concept Sets in Capr: https://ohdsi.github.io/Capr/articles/Capr-conceptSets.html

Columbia DBMI: OHDSI Sets Path for Enhancing Trust in Science, Engaging Global Community at 2025 Symposium. COLUMBIA UNIVERSITY, Department of Biomedical Informatics (DBMI). 2025.
Disponível em: https://www.dbmi.columbia.edu/ohdsi-2025/

Dekkers OM, Egger M, Altman DG, Vandenbroucke JP. Distinguishing case series from cohort studies. Ann Intern Med. 2012 Jan 3;156(1 Pt 1):37-40. doi: 10.7326/0003-4819-156-1-201201030-00006. PMID: 22213493.

FDA: Food and Drug Administration. Real-World Evidence. 2023–2025.
https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

FDA – regulatory information: Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry
Final version: https://www.fda.gov/media/152503/download
More info: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

Hernán, M. A., & Robins, J. M. Causal Inference: What If. Chapman & Hall/CRC, 2025. https://miguelhernan.org/whatifbook chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://static1.squarespace.com/static/675db8b0dd37046447128f5f/t/691fb7706ce66332f0b44467/1763686256720/hernan robins_WhatIf_21nov25.pdf

Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8. PMID: 26262116; PMCID: PMC4815923.

Huser V, Kahn MG, Brown JS, Gouripeddi R. Methods for examining data quality in healthcare integrated data repositories. Pac Symp Biocomput. 2018;23:628-633. PMID: 29218922.

Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. doi: 10.13063/2327-9214.1244. PMID: 27713905; PMCID: PMC5051581.

Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. Ther Adv Drug Saf. 2013 Apr;4(2):53-62. doi: 10.1177/2042098613477445. PMID: 25083251; PMCID: PMC4110833.

McDonald CJ, Humphreys BL. The U.S. National Library of Medicine and standards for electronic health records: One thing led to another. Inf Serv Use. 2022 May 10;42(1):81-94. doi: 10.3233/ISU-210142. PMID: 35600128; PMCID: PMC9108563.

OHDSI (OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS). Who We Are. 2026. Disponível em: https://www.ohdsi.org/who-we-are/

OHDSI Keeper. Repositório Github: https://github.com/OHDSI/Keeper/ Documentação: https://ohdsi.github.io/Keeper/ Manual: https://raw.githubusercontent.com/OHDSI/Keeper/main/extras/Keeper.pdf

OHDSI 2023 Plenary: Improving the reliability and scale of case validation Anna Ostropolets, Martijn Schuemie, Patrick Ryan.
Apresentação:
https://www.ohdsi.org/wp-content/uploads/2023/10/OHDSI2023-Plenary-1.pdf
video: https://www.youtube.com/watch?v=scUssf863TI

Ostropolets A, Hripcsak G, Husain SA, Richter LR, Spotnitz M, Elhussein A, Ryan PB. Scalable and interpretable alternative to chart review for phenotype evaluation using standardized structured data from electronic health records. J Am Med Inform Assoc. 2023 Dec 22;31(1):119-129. doi: 10.1093/jamia/ocad202. PMID: 37847668; PMCID: PMC10746303.

Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):54-60. doi: 10.1136/amiajnl-2011-000376. Epub 2011 Oct 28. PMID: 22037893; PMCID: PMC3240764.

Porta, M. (Ed.). A Dictionary of Epidemiology. Oxford University Press, 2014. ISBN: 9780197663639

Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2014 Jan 30;33(2):209-18. doi: 10.1002/sim.5925. Epub 2013 Jul 30. PMID: 23900808; PMCID: PMC4285234.

Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of p-values using observational data. Stat Med. 2016 Sep 30;35(22):3883-8. doi: 10.1002/sim.6977. PMID: 27592566; PMCID: PMC5108459.

Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci U S A. 2018 Mar 13;115(11):2571-2577. doi: 10.1073/pnas.1708282114. PMID: 29531023; PMCID: PMC5856503.

Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philos Trans A Math Phys Eng Sci. 2018 Sep 13;376(2128):20170356. doi: 10.1098/rsta.2017.0356. PMID: 30082302; PMCID: PMC6107542.

Schuemie, M.J., Ostropolets, A., Zhuk, A. et al. Standardized patient profile review using large language models for case adjudication in observational research. npj Digit. Med. 8, 18 (2025). https://doi.org/10.1038/s41746-025-01433-4
Apresentação:http://www.airis.or.kr/file/S4_4_Martijn%20Schuemie.pdf
Video: https://www.youtube.com/watch?v=_fl6v46BjIA&t=3734s

Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Annals of Internal Medicine. 2010 Nov;153(9):600-606. DOI: 10.7326/0003-4819-153-9-201011020-00010. PMID: 21041580.

Sterrantino, A. Observational studies: practical tips for avoiding common statistical pitfalls. The Lancet Regional Health - Southeast Asia, 2024; 25 https://www.thelancet.com/journals/lansea/article/PIIS2772-3682(24)00065-9/fulltext

STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology. https://www.strobe-statement.org/

