Capítulo

4

O estado da arte em pesquisa observacional de dados de saúde: A iniciativa OHDSI

Maria Tereza Fernandes Abrahão (USP), Moacyr Roberto Cuce Nobre (USP), Pablo Jorge Madril

Abstract

The Observational Medical Outcomes Partnership (OMOP) initiative, together with its follow-up, Observational Health Data Sciences and Informatics (OHDSI), redefined the area of observational research in health data, bringing the possibility of systematic analysis in large masses of data (big data), from a variety of sources, through the definition of a common data model, mechanisms to treat different vocabularies, and the availability of a set of free software tools to analyze them. These tools are expected to leverage evidence-gathering in the medical field and evaluation of therapies and procedures in the real world to support clinical research in general. The chapter gives an overview of the OHDSI platform and tools.

Resumo

A iniciativa da Observational Medical Outcomes Partnership (OMOP) em conjunto com a sua sequência, a Observational Health Data Sciences and Informatics (OHDSI), redefiniram a área de pesquisa observacional em dados de saúde trazendo a possibilidade de realizar análises sistemáticas em grandes massas de dados (big data), provindas de diversas fontes, através da definição de um modelo comum de dados, de mecanismos para tratamento de diferentes vocabulários, e da disponibilidade de um conjunto de ferramentas de software livre para análise dos mesmos. Com estas ferramentas espera-se alavancar o levantamento de evidências na área médica e na avaliação de terapias e procedimentos no mundo real, para apoiar a pesquisa clínica em geral. O capítulo apresenta uma visão geral da plataforma e ferramentas OHDSI.

4.1. Introdução

Com a disponibilidade das bases de dados de saúde em formato eletrônico, surgiu a possibilidade de gerar evidências e discernimentos sistemáticos em larga escala sobre a aplicação dos cuidados de saúde aos pacientes.

Os dados de assistência médica podem variar muito de uma organização para outra e são coletados para diferentes propósitos, como reembolso de provedor, pesquisa clínica e atendimento direto ao paciente. Esses dados podem ser armazenados em diferentes sistemas de banco de dados, formatos e modelos de informação e, apesar do uso crescente de terminologias padrão na área da saúde, o mesmo conceito pode ser representado de várias maneiras, de um ambiente para outro. A padronização de dados é o processo de trazer dados para um formato comum que permita a pesquisa colaborativa, análises em grande escala e compartilhamento de ferramentas e metodologias.

No entanto, a pesquisa na área de saúde requer informações de qualidade que estejam disponíveis no formato adequado e com a devida estruturação, para que efetivamente auxiliem os pesquisadores com informações relevantes acerca do universo de sua pesquisa. É necessário uma definição correta e clara sobre o acesso aos dados, as etapas de busca, limpeza, análise e uso das informações, bem como, a recuperação sistemática dos dados, informações e conhecimentos relevantes que se encontram distribuídos de modo disperso. 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.

Os estudos observacionais e análises secundárias de conjuntos de dados existentes na área da saúde, podem ser de grande valor na geração de evidências e eficácia em contextos do mundo real. Embora os estudos observacionais não possam fornecer evidências definitivas de segurança ou eficácia, eles podem: fornecer informações sobre o uso e a prática do "mundo real"; detectar sinais sobre os benefícios e riscos do uso de terapias na população em geral; ajudar a formular hipóteses a serem testadas em experimentos subsequentes; fornecer parte dos dados a nível populacional necessários para os ensaios clínicos; e informar a prática clínica.

A iniciativa da *Observational Medical Outcomes Partnership* (OMOP) em conjunto com a sua sequência, a *Observational Health Data Sciences and Informatics* (OHDSI)[1], redefiniram a área de pesquisa observacional em dados de saúde trazendo a possibilidade de realizar análises sistemáticas em grandes massas de dados (*big data*). Através da definição de um modelo comum de dados (*Common Data Model* - CDM-OMOP)[2], de mecanismos para tratamento de diferentes vocabulários, e da disponibilidade de um conjunto de ferramentas de software livre para análise dos dados,

[1] OMOP/OHDSI https://ohdsi.org/
[2] CDM-OMOP https://www.ohdsi.org/data-standardization/the-common-data-model/

espera-se alavancar o levantamento de evidências na área médica e na avaliação de terapias e procedimentos no mundo real, para apoiar a pesquisa clínica em geral.

O objetivo deste capítulo é apresentar os fundamentos da pesquisa observacional e os tipos de estudos disponíveis em bases de dados de saúde; o modelo comum de dados (CDM); e as ferramentas de geração e análise do OHDSI de forma prática com o uso de exemplos de estudos reais. Espera-se que, ao final do capítulo, o leitor tenha uma visão geral da plataforma e ferramentas OHDSI e seja capaz de: i) compreender o modelo comum de dados (CDM-OMOP), ii) ter uma visão dos vocabulários e da definição e utilização dos conceitos, e iii) entender o processo da definição de uma coorte e a visualização dos dados assistenciais, fomentando assim uma pesquisa observacional de alto nível.

O capítulo está estruturado como se segue. Primeiro, a seção 4.1 oferece uma breve fundamentação sobre o uso dos sistemas de registros eletrônicos de saúde como fontes de dados, os desafios que envolvem a reprodutibilidade da pesquisa científica e a iniciativa OMOP/OHDSI. Em seguida, a seção 4.2 apresenta a arquitetura do sistema OHDSI, a seção 4.3 o modelo comum de dados (CDM - OMOP) e a seção 4.4 os vocabulários. A seção 4.5 apresenta a definição de uma coorte e a seção 4.6 os tipos de análises. As considerações finais e conclusões são apresentadas na seção 4.7 e a seção 4.8 apresenta um glossário de termos e na sequência, as referências consultadas.

4.1.1. Sistemas de registros eletrônicos de saúde (RES) como fontes de dados

A disponibilidade de bases de dados de saúde em formato eletrônico abriu a possibilidade de gerar evidências e discernimentos sistemáticos em larga escala sobre a aplicação dos cuidados de saúde aos pacientes. Essa pesquisa é denominada Estudos Observacionais de Desfechos, e utiliza dados clínicos longitudinais no nível do paciente para descrever e compreender a patogênese da doença e o efeito de eventos clínicos, bem como intervenções de tratamento, sobre a progressão da doença. Constitui uso secundário dos dados o aproveitamento de informações que estão sendo coletadas normalmente para outros fins que não a pesquisa: dados administrativos, solicitações de saúde complementar, faturamento, contabilidade geral, fontes de saúde pública, biobancos, fontes farmacêuticas e o Registro Eletrônico de Saúde (RES).

O uso desses dados têm como vantagens o baixo custo de coleta e a possibilidade de utilizar amostras maiores, proporcionando maior força para detectar pequenas diferenças ou eventos raros, dispensando o contato direto com o paciente. Possibilita também, maior diversidade metodológica, permitindo que modificações no estudo possam ser implementadas diferentemente dos projetos de pesquisa de maior rigor de protocolos, como os ensaios randomizados. Além disso, um relatório do *New England Journal of Medicine*[3] aponta que "foram encontradas poucas evidências de que

[3] Benson K. A Comparison of Observational Studies and Randomized, Controlled Trials. The New England Journal of Medicine. 2000;9.

as estimativas dos efeitos do tratamento em estudos observacionais relatados após 1984 sejam consistentemente maiores ou qualitativamente diferentes daquelas obtidas em ensaios controlados e randomizados".

As desvantagens estão relacionadas à falta da padronização na coleta dos dados, que afeta a qualidade dos dados registrados; a perda potencial de seguimento após um determinado ponto no tempo; a ausência de informações clínicas relativas a um paciente que podem ser importantes para as análises de interesse.

Os conjuntos de dados de origem diferem uns dos outros em formato e representação de conteúdo e por vezes introduzem viés nos dados. Tudo isso torna a pesquisa robusta, reproduzível e automatizada um desafio significativo. Uma solução é a padronização dos dados e da sua representação. Isso permite que métodos e ferramentas operem em dados de origem díspar, e que métodos analíticos desenvolvidos para um determinado conjunto de dados possam ser aplicados a qualquer outro conjunto de dados no mesmo formato.

4.1.2. Estudos de coorte observacionais retrospectivos

O uso secundário de registros eletrônicos de saúde são indicados para estudos observacionais, retrospectivos e comparativos, a partir de uma coorte extraída da base assistencial. Definimos "coorte" para significar um conjunto de pacientes que satisfazem um ou mais critérios de inclusão por um período de tempo; "observacional" para significar que não há intervenção ou atribuição de tratamento imposta pelo estudo; "retrospectivo" para significar que o estudo será conduzido usando dados já coletados antes do início do estudo; "design de coorte comparativo" para significar a comparação formal entre duas coortes, uma coorte alvo e uma coorte de referência, para o risco de um desfecho durante um período de tempo definido após a entrada da coorte.

4.1.3. Desafios da pesquisa científica na atualidade: reprodutibilidade

A reprodutibilidade é a possibilidade de um experimento ou estudo poder ser repetido, tanto pelo mesmo pesquisador quanto por um pesquisador atuando independentemente. Reproduzir um experimento é chamado de replicar o mesmo. A reprodutibilidade é um dos princípios fundamentais do método científico.

A pesquisa reprodutível exige que os conjuntos de dados, os códigos, software ou outros tipos de instruções de computador que foram usados para computar os resultados publicados devem ser fornecidos para a verificação ou a realização de análises alternativas.

4.1.4. A iniciativa OMOP e OHDSI

A necessidade de se gerar evidências clínicas acessíveis e confiáveis, acessando a experiência do atendimento de saúde de milhões de pacientes em todo o mundo, é uma realidade. A *Observational Health Data Sciences and Informatics* (OHDSI, pronunciado "Odyssey") baseou-se em aprendizados da iniciativa da *Observational Medical Outcomes Partnership* (OMOP), para transformar métodos de pesquisa em um conjunto de aplicativos e ferramentas de exploração que aproximam a pesquisa de campo do objetivo final de gerar evidências sobre todos os aspectos de saúde para atender às necessidades de pacientes, clínicos e todos os outros tomadores de decisão.

4.1.4.1. Os cegos e o elefante[4]

Como na parábola hindu onde um grupo de cegos tentam descrever um elefante apenas apalpando partes dele, cada prestador de serviço de saúde possui na sua base uma visão parcial da população.

Bases clínicas de hospitais possuem informação dos pacientes e de suas doenças, mas, não tem dados de pessoas sadias, e, mesmo dentro de um grupo de hospitais, a especialização influencia o perfil das bases. Por exemplo, em uma maternidade vou achar fundamentalmente informações de mulheres grávidas, num centro de especialidades cardiológicas terei predominância deste grupo de doenças.

Bases de administradoras de seguros de saúde tem dados de sinistros e também de pessoas sadias mas não possuem dados clínicos. Bases laboratoriais armazenam resultados de exames e o seu histórico para cada paciente, porém sem relação com um diagnóstico (ou a ausência de um). No fim, ninguém possui uma visão completa da população.

Estudos observacionais restritos a bases independentes estão sujeitos a vieses e deformações por não termos condições de capturar amostras randômicas da população completa.

A iniciativa OMOP/OHDSI se propõe a montar uma imagem que incorpore as diversas fontes de informações médicas através do uso de um modelo comum de dados para conseguir ter uma visão de toda a população, tanto atual quanto histórica.

4.1.4.2. Resumo histórico

Em 2007, reconhecendo o aumento da utilização de registros eletrônicos de saúde, o Congresso Americano requisitou ao *Food and Drug Administration* (FDA)[5] criar um novo programa de vigilância farmacológica para identificar de forma mais agressiva, potenciais problemas de segurança. O FDA lançou várias iniciativas para alcançar esse objetivo, incluindo o programa “Sentinela”, para criar uma rede de dados a nível nacional para o monitoramento de drogas, utilizando dados eletrônicos de detentores de informações de saúde.

---

[4] https://en.wikipedia.org/wiki/Blind_men_and_an_elephant
[5] FDA https://www.fda.gov/

Em particular, a *Pharmaceutical Research and Manufacturers of America* (PhRMA)[6], o FDA e a *Foundation for the National Institutes of Health* (FNIH)[7] criaram a *Observational Medical Outcomes Partnership* (OMOP), uma parceria público-privada estabelecida nos EUA. Este grupo de pesquisa interdisciplinar abordou uma tarefa que é fundamental para os objetivos mais amplos da comunidade de investigação: identificar os métodos mais confiáveis para a análise de grandes volumes de dados extraídos de fontes heterogêneas.

Empregando uma variedade de abordagens das áreas de epidemiologia, estatística e ciências da computação, OMOP procura responder a um desafio crítico: o que podem pesquisadores médicos aprender com a avaliação dessas novas bases de dados de saúde? Poderia uma abordagem única ser aplicada a várias doenças e poderiam as suas conclusões serem provadas? A comunidade de pesquisa médica poderia fazer mais estudos em menos tempo, utilizando menos recursos e obtendo resultados mais consistentes. No final, isso significaria um melhor sistema de monitoramento de drogas, dispositivos e procedimentos para a comunidade de saúde poder identificar com segurança os riscos e oportunidades para melhorar o atendimento ao paciente.

A peça central do projeto OMOP foi o desenvolvimento do modelo comum de dados (CDM), que representa dados de saúde de diversas fontes heterogêneas de uma forma consistente e padronizada. Este CDM é um modelo de informação no qual a codificação e as relações entre os conceitos são explícita e formalmente especificadas. Em conjunto com o CDM, foram construídos os vocabulários padronizados para a condução dos experimentos OMOP.

A OHDSI surgiu na continuação do projeto de cinco anos desenvolvido pela OMOP (2009-2013). A OHDSI realizou sua primeira reunião anual no Columbia University Medical Center em Nova York em outubro de 2014. Cinquenta e oito participantes revisaram a visão e os objetivos que deram origem à criação do OHDSI e formaram grupos de trabalho para abordar o modelo de dados comum, vocabulário, bases de conhecimento, métodos de estimativa, geração de fenótipo, caracterização clínica e definição de coorte. Os investigadores da pesquisa OMOP iniciaram o esforço OHDSI, e o laboratório de pesquisa mudou-se para a Fundação *Reagan-Udall* no âmbito do *Innovation in Medical Evidence Development and Surveillance* (IMEDS)[8].

A equipe do OHDSI adotou e continuou a manutenção deste modelo e de seus serviços de vocabulário associados. A abordagem geral do OHDSI é criar uma rede aberta de detentores de dados observacionais a partir da tradução dos dados para o CDM - OMOP. Cada elemento no banco de dados do participante deve ser mapeado para o vocabulário de CDM aprovado e colocado no esquema de dados. Em troca, essa abordagem possibilita implementar várias ferramentas existentes de exploração e geração de evidências e participar de estudos em todo o mundo, visto que, qualquer consulta pode ser executada em qualquer site sem necessidade de modificações.

[6] PhRMA https://www.phrma.org/
[7] FNIH https://fnih.org/
[8] IMEDS http://reaganudall.org/innovation-medical-evidence-development-and-surveillance

Análises globais e multicêntricas podem ser executadas de forma rápida e eficiente usando aplicativos ou programas desenvolvidos em um único site.

A OHDSI conta com mais de 140 colaboradores em 16 países e é composta por pesquisadores clínicos, cientistas da computação, bioestatísticos e profissionais do setor de saúde.

4.1.4.3. Onde achar: principais referências

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

- Informações gerais: http://www.ohdsi.org - É o 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; Broadsea (https://github.com/OHDSI/Broadsea), que disponibiliza uma versão em containers Docker do conjunto de ferramentas, ver também: repositório Docker com os componentes dockerizados (https://hub.docker.com/u/ohdsi/); OHDSI-In-a-Box (https://github.com/OHDSI/OHDSI-in-a-Box) que disponibiliza uma máquina virtual pronta para testes e demonstrações;
- 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 o fórum (http://forums.ohdsi.org/).

4.2. Arquitetura do sistema OHDSI

A arquitetura do sistema disponibilizado pela iniciativa OHDSI consiste em um modelo comum, o CDM-OMOP v5, e um conjunto de ferramentas construídas ao redor deste modelo para oferecer acesso e suporte a consultas e atualizações das informações nele contidas. O CDM-OMOP consiste em:

- O esquema CDM v5;
- Dados dos pacientes migrados de outras fontes de informação;
- Um conteúdo de Vocabulários;

Esse conjunto é armazenado em uma base relacional. Atualmente são suportadas as seguintes tecnologias SQL: BigQuery, Impala, Netezza, Oracle, Parallel Data Warehouse, Postgres, Redshift, e SQL Server.

Para acesso ao modelo são construídas as seguintes ferramentas:

- OHDSI Web API: fornece acesso padronizado ao modelo via serviços REST;
- ATLAS: Aplicação Web, que utiliza a API para permitir o acesso aos usuários comuns as informações contidas no modelo;

- WhiteRabbit e Rabbit in a Hat, ferramentas de ETL (Extração, Transformação e carga de dados) que auxiliam no processo de migração de outras fontes para o modelo;
- ATHENA: Aplicação web que permite o acesso e manutenção dos vocabulários utilizados no modelo;
- ARACHNE: Permite gerar e executar o código de um estudo (R e SQL), coletar estatísticas e inferência de evidências;
- ACHILLES: Fornece estatísticas descritivas da base OMOP.

Atualmente, estas ferramentas estão sendo incorporadas no ATLAS. A Figura 4.1. apresenta os diversos componentes da arquitetura:

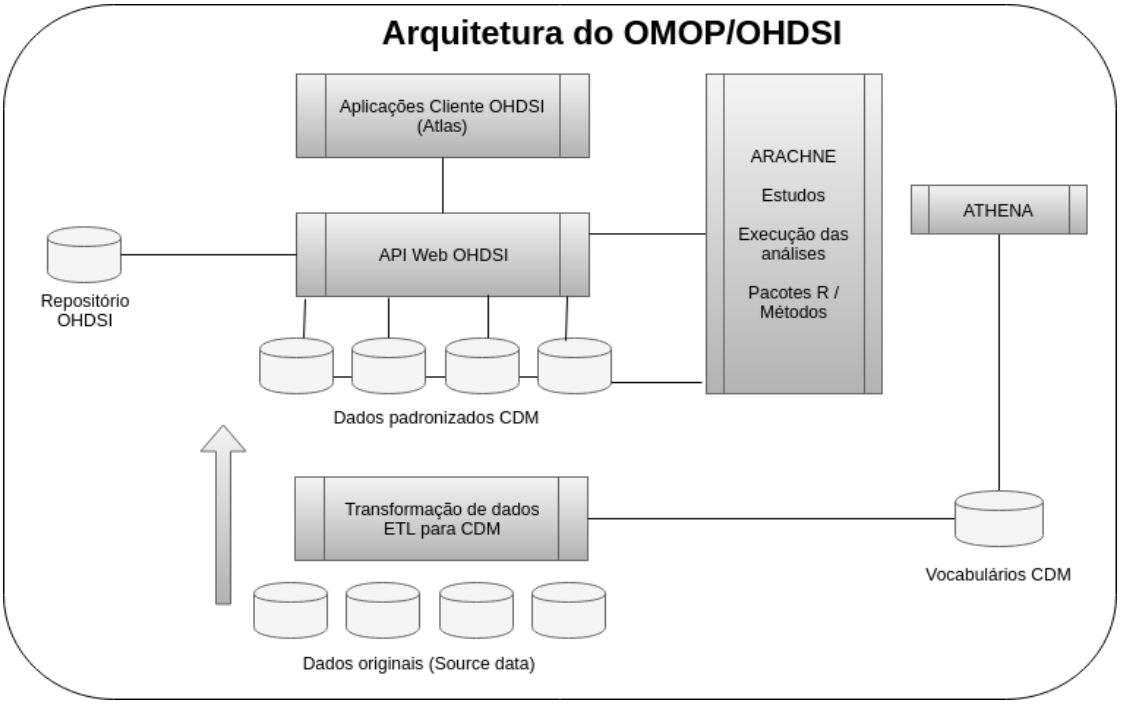

Figura 4.1. Componentes da arquitetura OHDSI

4.3. Modelo comum de dados (*Common Data Model* CDM-OMOP)

Nenhuma fonte única de dados observacionais fornece uma visão abrangente dos dados clínicos que um paciente acumula ao receber cuidados de saúde e, portanto, nenhuma delas é suficiente para atender a todas as necessidades da pesquisa observacional. Uma peça central do projeto OMOP foi o desenvolvimento do modelo comum de dados (CDM). A finalidade do modelo comum de dados é padronizar o formato e o conteúdo dos dados observacionais, oriundos de sistemas heterogêneos, para que aplicativos, ferramentas e métodos padronizados possam ser aplicados. O CDM representa dados de saúde de diversas fontes de uma forma consistente e padronizada.

O CDM é projetado para apoiar a realização de pesquisas, identificar e avaliar associações entre intervenções (exposição a medicamentos, procedimentos, mudanças na política de saúde, etc.) e os resultados causados por essas intervenções (ocorrências de condições, procedimentos, exposição a drogas etc.). Os resultados podem ser eficazes (benefício) ou adversos (risco de segurança). Muitas vezes, coortes específicas de pacientes (por exemplo, aqueles que tomam determinado medicamento ou sofrem de uma determinada doença) podem ser definidos para tratamentos ou resultados, usando eventos clínicos (diagnósticos, observações, procedimentos) que ocorrem em relacionamentos temporais pré-definidos. Com o seu conteúdo padronizado (através dos vocabulários padronizados), assegurará que os métodos de pesquisa possam ser sistematicamente aplicados para produzir de forma significativa resultados comparáveis e reprodutíveis.

O CDM é flexível o suficiente para armazenar dados do RES, dados de sinistros, bem como o vocabulário padronizado. Cada tabela contém um conjunto mínimo de campos que devem ser preenchidos. A rede de pacientes disponível para pesquisa no OHDSI inclui aproximadamente 84 bancos de dados, tanto clínicos quanto informativos, totalizando mais de 650 milhões de pacientes.

4.3.1. Modelo comum de dados para pesquisa

O CDM foi elaborado para incluir todos os elementos observacionais de dados de saúde que sejam relevantes para análise de casos de uso para apoiar a geração de evidências científicas confiáveis sobre a história natural da doença, cuidados médicos, identificação de informação demográfica, intervenções de saúde e resultados.

Portanto, o CDM é projetado para armazenar dados observacionais para permitir a pesquisa, sob os seguintes princípios:

- Adequação à finalidade: O CDM visa fornecer dados organizados de maneira ideal para a análise;
- Proteção de dados: Todos os dados que possam comprometer a identidade e a proteção dos pacientes, como nomes, datas, etc., são limitados. Exceções são possíveis quando a pesquisa exige expressamente informações mais detalhadas, como datas de nascimento precisas para o estudo de bebês;
- Domínios: os domínios são modelados em um modelo de dados relacionais centrados na pessoa e são identificados e definidos separadamente em um modelo de relacionamento de entidade;
- Vocabulários Padronizados: Para padronizar o conteúdo desses registros, o CDM se baseia nos vocabulários padronizados contendo todos os conceitos de saúde padrão correspondentes necessários e apropriados, explicitamente representando todos os fatos e eventos clínicos. Com poucas exceções, não há informações textuais nas tabelas do CDM;
- Reutilização de vocabulários existentes: Se possível, esses conceitos são aproveitados de organizações ou iniciativas de padronização nacional ou de

indústria ou definição de vocabulário, como a *National Library* of *Medicine*[9], o *Department of Veterans' Affairs*[10], o Centro de Controle e Prevenção de Doenças, etc;

- Manutenção de códigos-fonte: Embora todos os códigos sejam mapeados para os Vocabulários Padronizados, o modelo também armazena o código-fonte original para garantir que nenhuma informação seja perdida;
- Neutralidade da tecnologia: O CDM não requer uma tecnologia específica. Ele pode ser realizado em qualquer banco de dados relacional, como Oracle, SQL Server etc., ou como conjuntos de dados analíticos do SAS;
- Escalabilidade: O CDM é otimizado para processamento de dados e análise computacional para acomodar fontes de dados que variam em tamanho, incluindo bancos de dados com milhões de pessoas e bilhões de observações clínicas;
- Compatibilidade retroativa: Todas as alterações dos CDMs anteriores são claramente delineadas no repositório do github[11]. Versões mais antigas do CDM podem ser facilmente criadas a partir do CDM versão 5 sem perda de nenhuma informação.

O CDM foi elaborado para incluir todos os elementos observacionais de dados de saúde que são relevantes para análise de casos de uso para apoiar a geração de evidências científicas confiáveis sobre a história natural da doença, assistência médica, efeitos de intervenções médicas, identificação de informações demográficas, intervenções de saúde e resultados. A versão 5 do CDM foi apresentada em 14 de outubro de 2014, disponível em[12]. A Figura 4.2. apresenta uma cópia traduzida do CDM-OMOP na versão 5.

Além dos dados da pessoa, da condição, droga, procedimento e informações de visitas, o modelo provê informações de custo e do provedor do atendimento. Esta proposta tende a apoiar a economia da saúde e estudos de casos de uso de tratamento médico, incluindo a segurança de dispositivos médicos, eficácia comparativa e qualidade de saúde.

Em 11 de outubro de 2018 foi publicada uma especificação do modelo comum de dados CDM-OMOP, versão 6.0. O manual técnico e as diferenças entre as versões estão disponível em[13].

4.3.2. Outros modelos

Existem outros modelos de dados utilizados para organizar as informações médicas. Dentre eles podemos destacar os seguintes:

[9] NLM https://www.nlm.nih.gov/
[10] VA https://www.va.gov/
[11] CDM repositório github https://github.com/OHDSI/CommonDataModel
[12] CDM v5 https://github.com/OHDSI/CommonDataModel/blob/v5-historical/OMOP%20CDM%20v5.pdf
[13] CDM v6 https://github.com/OHDSI/CommonDataModel/blob/master/OMOP_CDM_v6_0.pdf

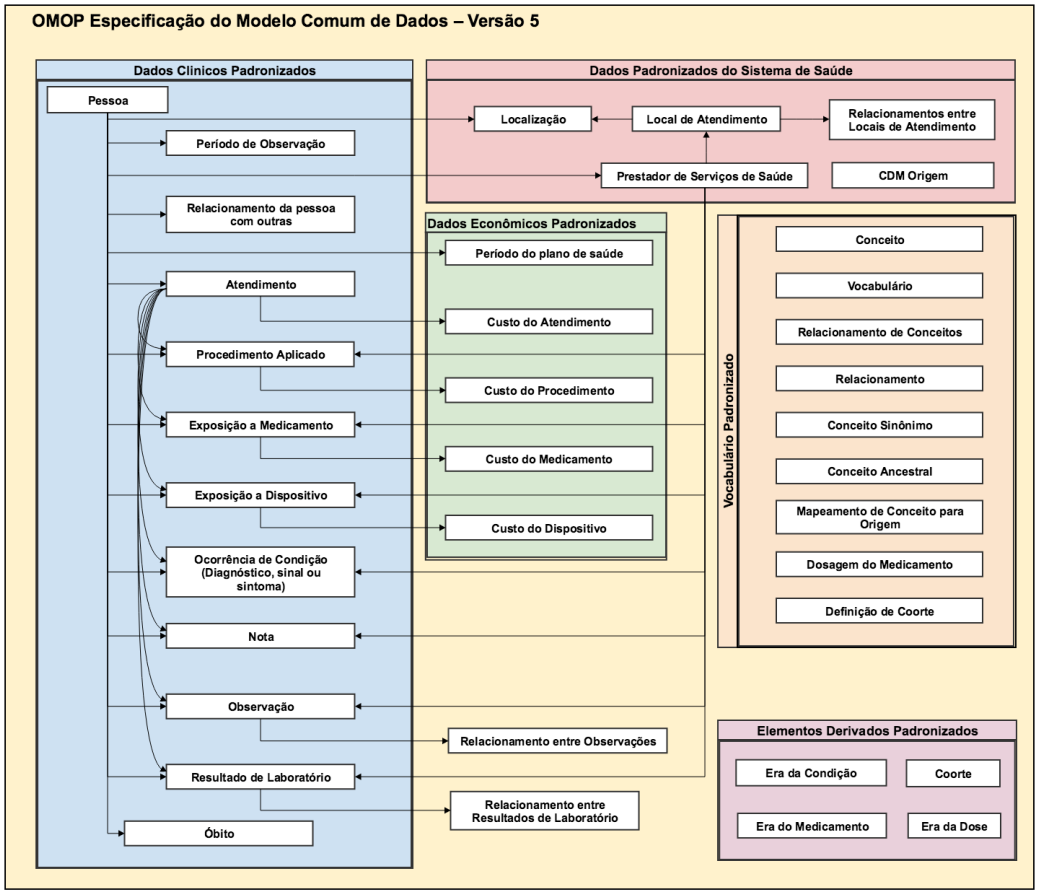

Figura 4.2. Modelo CDM OMOP versão 5

- Sentinel Common Data Model: A Sentinel Initiative da Food and Drug Administration (FDA) dos EUA é um esforço de longo prazo para melhorar a capacidade do FDA de identificar e avaliar questões de segurança de produtos médicos que utiliza dados de saúde eletrônicos pré-existentes de várias fontes para monitorar a segurança de produtos médicos regulamentados; [https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model]
- PCORnet Common Data Model: A rede do National Patient-Centered Clinical Research Network financiada pelo Patient-Centered Outcomes Research Institute (PCORI), integra dados de 11 redes heterogêneas para permitir pesquisa de efetividade comparativa em larga escala. [https://pcornet.org/pcornet-common-data-model/]
- i2b2: O projeto Informatics for Integrating Biology and the Bedside (i2b2) [https://www.i2b2.org/index.html] suporta a interoperabilidade entre as fontes de informação por meio de uma abordagem orientada pela ontologia para o armazenamento de dados. [https://academic.oup.com/jamia/article/23/5/909/2379861]

- Base Pauá: Trabalho de doutorado que resultou na elaboração um modelo comum de dados para o sistema de informações hospitalares do InCor-HCFMUSP, hospital de referência em doenças cardiovasculares, na cidade de São Paulo, Brasil. Este modelo, centrado no paciente, foi pensado para poder ser utilizado em outros sistemas e para isolar as particularidades do sistema subjacente sendo possível elaborar consultas genéricas que são aplicadas ao modelo. A falta de vocabulários padronizados fez com que as consultas genéricas se limitassem a diagnósticos codificados em CID-10, na atualidade, o único vocabulário padronizado de uso obrigatório no Brasil. [http://www.teses.usp.br/teses/disponiveis/5/5131/tde-04082016-160129/pt-br.php]

4.3.3. Estrutura das tabelas CDM

O CDM define as estruturas de tabela centrada na pessoa. No lugar de uma tabela central com fatos, temos uma coleção dos mesmos diferenciada por domínio: procedimentos, condição, droga, medição, observação, etc.
Isso permite uma visão longitudinal de todos os eventos de saúde de um indivíduo. Estes eventos estão ligados aos prestadores de cuidados de saúde, como médicos, consultórios médicos, postos de saúde, ambulatórios, hospitais, departamentos hospitalares, etc.).

Para representar os domínios relevantes, o CDM contém as 39 tabelas descritas a seguir:

Vocabulários Padronizados:

CONCEPT - Essa tabela contém registros que identificam exclusivamente cada unidade usada para expressar informações clínicas. Os conceitos são derivados de vocabulários de origem, que representam informações clínicas em diferentes domínios (por exemplo, condições, medicamentos, procedimentos) por meio do uso de códigos e descrições associadas. Alguns conceitos são designados como conceitos padrão, o que significa que esses conceitos podem ser usados no CDM e em análises padronizadas. Cada conceito padrão possui um domínio principal, que define a localização em que o conceito deveria ser observado dentro do CDM.

VOCABULARY - Essa tabela inclui uma lista dos vocabulários coletados de várias fontes ou criados pela comunidade OMOP. Essa tabela de referência é preenchida com um único registro para cada fonte de vocabulário e inclui um nome descritivo e outros atributos associados ao vocabulário.

DOMAIN - Essa tabela inclui uma lista dos domínios dos elementos de dados que estão contidos no CDM. Um domínio define o conjunto de conceitos permitidos para cada

campo padronizado. Essa tabela de referência é preenchida com um único registro para cada domínio e inclui um nome descritivo para o domínio.

CONCEPT_CLASS - Essa tabela inclui uma lista das classificações usadas para diferenciar conceitos dentro de um determinado vocabulário. Essa tabela de referência é preenchida com um único registro para cada classe conceitual e inclui um nome descritivo para a classe conceitual.

CONCEPT_RELATIONSHIP - Essa tabela contém registros que definem relacionamentos entre dois conceitos e a natureza do relacionamento. O tipo de relacionamento é definido na tabela RELATIONSHIP e é geralmente classificado como hierárquico (pai-filho) ou não hierárquico (lateral). Todos os relacionamentos são direcionais e cada relação de conceito é representada duas vezes simetricamente na tabela de relacionamento de conceito. Por exemplo, os dois conceitos SNOMED de 'infarto agudo do miocárdio da parede anterior' e 'infarto agudo do miocárdio' têm duas relações conceituais: 1- 'infarto agudo do miocárdio da parede anterior' 'é um' 'infarto agudo do miocárdio', e 2- 'Infarto agudo do miocárdio' 'agrupa o' 'Infarto agudo do miocárdio da parede anterior'.

RELATIONSHIP - Essa tabela fornece uma lista de referência de todos os tipos de relacionamentos permitidos que podem ser usados para associar quaisquer dois conceitos na tabela CONCEPT_RELATIONSHIP. Os relacionamentos são classificados como hierárquicos (pai-filho) ou não-hierárquicos e são usados para determinar quais registros de relacionamento de conceito devem ser incluídos na tabela CONCEPT_ANCESTOR.

CONCEPT_SYNONYM - Essa tabela é usada para armazenar nomes alternativos para um conceito. Cada sinônimo é atribuído a seu próprio identificador exclusivo e contém o texto de uma descrição e o identificador do conceito que ele representa.

CONCEPT_ANCESTOR - Essa tabela contém registros que definem os relacionamentos hierárquicos entre todos os conceitos padrão. A tabela CONCEPT_ANCESTOR permite a identificação de relacionamentos hierárquicos em várias etapas, como medicamentos de marca que se enquadram em uma classe terapêutica ou diagnóstico específico que são classificados dentro de um sistema específico de classes.

SOURCE_TO_CONCEPT_MAP - Essa tabela é uma estrutura de dados legada no CDM, recomendada para uso em processos de extração, transformação e carregamento (ETL), para manter códigos fonte locais que não estão disponíveis como conceitos nos vocabulários padronizados e para estabelecer mapeamentos para cada código-fonte em um conceito padrão que pode ser usado para preencher as tabelas do CDM.

DRUG_STRENGTH - Essa tabela contém conteúdo estruturado sobre a quantidade ou concentração e unidades associadas de um ingrediente específico dentro de um determinado medicamento. A tabela de composição do medicamento é um arquivo suplementar para apoiar a análise padronizada da utilização de medicamentos usando conceitos do vocabulário RxNorm[14]. Um conceito clínico de medicamento que contenha múltiplos ingredientes ativos resultará em um registro do medicamento para cada ingrediente ativo.

COHORT_DEFINITION - Essa tabela contém registros para definir cada coorte derivada por meio de uma descrição e sintaxe associadas. Coortes são elementos derivados de um conjunto de assuntos que satisfazem um determinado conjunto de critérios de inclusão por um período de tempo. A tabela COHORT_DEFINITION fornece uma estrutura padronizada para manter as regras de inclusão de um assunto em uma coorte, e pode armazenar código para instanciar a coorte dentro do CDM.

ATTRIBUTE_DEFINITION - Essa tabela contém registros para definir cada atributo por meio de uma descrição e sintaxe associadas. Atributos são elementos derivados que podem ser selecionados ou calculados para um assunto dentro de uma coorte. A tabela ATTRIBUTE_DEFINITION fornece uma estrutura padronizada para manter as regras do cálculo de covariáveis para um sujeito em uma coorte, e pode armazenar código para instanciar os atributos para uma dada coorte dentro do CDM.

Meta-dados normalizado:

CDM_SOURCE - Essa tabela contém detalhes sobre a fonte de dados e o processo utilizado para transformar os dados para o CDM. Se um banco de dados de origem for derivado de vários fontes de dados, espera-se que a integração dessas diferentes fontes seja documentada nas especificações de ETL.

Dados clínicos padronizados:

Essas tabelas contêm as informações básicas sobre os eventos clínicos que ocorreram longitudinalmente durante os períodos de observação válidos para cada pessoa, bem como as informações demográficas.

PERSON - Essa tabela contém registros que identificam exclusivamente cada paciente nos dados de origem que tem tempo em risco para ter eventos clínicos registrados nos sistemas de origem. Uma pessoa deve ter pelo menos um período de observação para definir o tempo em risco, mas pode ou não ter quaisquer eventos clínicos registrados nos outros domínios de dados. Cada registro pessoal tem atributos demográficos

[14] RxNorm https://www.nlm.nih.gov/research/umls/rxnorm/

associados que são considerados constantes para o paciente ao longo de seus períodos de observação. Todos os outros domínios de dados no nível do paciente têm uma referência de chave estrangeira ao domínio da pessoa.

OBSERVATION_PERIOD - Essa tabela contém registros que definem com exclusividade os períodos de tempo em que uma pessoa está em risco de ter eventos clínicos registrados nos sistemas de origem. Uma pessoa pode ter um ou mais períodos de observação disjunta, durante os quais as análises podem assumir que os eventos clínicos seriam capturados se observados, e fora do qual nenhum evento clínico poderia ser registrado.

SPECIMEN - Essa tabela contém os registros que identificam cada amostra biológica de uma pessoa.

DEATH - Essa tabela contém o evento clínico de como e quando uma pessoa morre. Uma pessoa pode ter até um registro se os sistemas de origem contiverem evidências de que ele é falecido. Todas as pessoas que estavam vivas durante todos os períodos de observação não devem conter nenhuma informação na tabela DEATH.

VISIT_OCCURRENCE - Essa tabela contém os períodos de tempo em que uma pessoa recebe continuamente serviços médicos de um ou mais prestadores de serviços em uma instalação em um determinado ambiente dentro do sistema de assistência médica. As visitas são classificadas em quatro configurações: atendimento ambulatorial, internação, sala de emergência e cuidados de longa duração. As pessoas podem fazer a transição entre essas configurações ao longo de um episódio de atendimento. As visitas de internação são definidas pelo período de tempo entre a admissão e a alta de uma instalação hospitalar específica. As consultas ambulatoriais são definidas como período de tempo dentro do consultório de um provedor específico, que é esperado para menos de 1 dia. Visitas de cuidados de longo prazo são definidas como o período de tempo em que uma pessoa é tratada dentro de uma instalação específica de cuidados de longo prazo.

PROCEDURE_OCCURRENCE - Essa tabela contém registros de atividades ou processos solicitados e / ou executados por um profissional de saúde para que o paciente tenha uma finalidade diagnóstica e / ou terapêutica.

DRUG_EXPOSURE - Essa tabela captura registros sobre a utilização de uma substância bioquímica com um efeito terapêutico fisiológico quando ingerida ou de outra forma introduzida no corpo. As drogas incluem medicamentos prescritos e de venda livre, vacinas e terapias biológicas. A exposição a medicamentos é inferida a partir de eventos clínicos associados a pedidos, prescrições escritas, dispensas de

farmácia, administrações de procedimentos e outras informações relatadas pelo paciente.

DEVICE_EXPOSURE - Essa tabela captura registros sobre a exposição de uma pessoa a um objeto físico ou instrumento que é utilizado para fins de diagnóstico ou terapêuticos. Os dispositivos incluem objetos implantáveis (marcapassos, stents, articulações artificiais), equipamentos e suprimentos médicos duráveis (bandagens, muletas, seringas) e outros instrumentos usados em procedimentos médicos (suturas, desfibriladores, etc.).

CONDITION_OCCURRENCE - Essa tabela captura os registros de uma doença ou de uma condição médica com base na avaliação de um provedor ou relatada por um paciente.

MEASUREMENT - Uma medida é a captura de um valor estruturado (numérico ou categórico) obtido através do exame sistemático de uma pessoa ou amostra. A tabela MEASUREMENT captura ordens de medição e resultados de medição. O domínio de medição pode conter resultados laboratoriais, sinais vitais ou descobertas quantitativas de relatórios de patologia.

NOTE - Essa tabela captura informações não estruturadas que foram gravadas por um provedor ou paciente em notas de texto livre em uma determinada data.

OBSERVATION - Essa tabela capta qualquer fato clínico sobre um paciente obtido no contexto de um exame, questionamento ou procedimento. O domínio de observação suporta a captura de dados não representados por outros domínios, incluindo medidas não estruturadas, histórico médico e histórico familiar.

FACT_RELATIONSHIP - Essa tabela contém registros para detalhar as relações entre fatos em um domínio ou entre dois domínios e a natureza do relacionamento. Exemplos de tipos de relacionamentos de fatos incluem: relacionamentos pessoais (ligação mãe-filho), relacionamentos no local de cuidados (representando a estrutura organizacional hierárquica de instalações dentro dos sistemas de saúde), exposições de medicamentos fornecidas devido a condição indicada associada, dispositivos usados durante o curso de um procedimento e as medidas derivadas de uma amostra. Todos os relacionamentos são direcionais e cada relacionamento é representado duas vezes simetricamente na tabela de relacionamento de fatos. Por exemplo, duas pessoas (PERSON_ID = 1 é a mãe de PERSON_ID = 2) têm dois relacionamentos de fatos: 1- 'PERSON_ID 1' 'mãe de' 'PERSON_ID 2' e 2 'PERSON_ID 2' 'filho de' 'PERSON_ID 1'.

Dados padronizados do sistema de saúde:

LOCATION - Essa tabela representa uma maneira genérica de capturar localização física ou informações de endereço. Os locais são usados para definir os endereços de pessoas e locais de atendimento.

CARE_SITE - Essa tabela contém uma lista de unidades organizacionais onde a prestação de cuidados de saúde é praticada (consultórios, alas, hospitais, clínicas, etc.).

PROVIDER - Essa tabela contém uma lista provedores de assistência à saúde identificados exclusivamente. Estes são tipicamente médicos e enfermeiros.

Dados padronizados de economia em saúde: Essas tabelas contêm informações de custo sobre assistência médica. Dependem do sistema de prestação de cuidados de saúde em que a população de doentes está envolvida, que pode variar significativamente em diferentes países. No entanto, o modelo atual do CDM está focado no sistema de saúde dos EUA.

PAYER_PLAN_PERIOD - Essa tabela captura registros que detalham o período de tempo em que uma pessoa está continuamente inscrita em uma estrutura de benefício de plano de saúde específico de um determinado pagador. Cada pessoa que recebe cuidados de saúde e está coberta por benefícios de saúde está sujeita a um plano definido pelo pagador para a pessoa ou sua família. Para uma dada política de benefícios, pode haver um ou mais planos ativos por determinados períodos de tempo, definindo o custo dos serviços de saúde fornecidos.

VISIT_COST - Essa tabela captura os custos da visita de saúde de um paciente que não estão relacionados a procedimentos, medicamentos ou dispositivos específicos usados no encontro.

PROCEDURE_COST - Essa tabela captura o custo de um procedimento executado em uma pessoa. As informações sobre o custo são derivadas apenas dos valores pagos pelo procedimento.

DRUG_COST - Essa tabela captura registros que indicam o custo de uma droga de exposição. A informação sobre o custo é definida pela quantidade de dinheiro pago pela pessoa e pagador pelo medicamento, bem como pelo custo cobrado do medicamento.

DEVICE_COST - Essa tabela captura o custo de um dispositivo médico ou fornecimento usado em uma pessoa. As informações sobre o custo são derivadas apenas dos valores pagos pelo dispositivo.

Elementos derivados padronizados: São tabelas montadas a partir dos dados já descritos através de algoritmos ou de seleções feitas utilizando os dados. Por exemplo, a partir das informações de DRUG_EXPOSURE, são geradas as eras, intervalos contínuos de exposição ao medicamento ou intervenção.

COHORT - Essa tabela contém registros derivados como um conjunto de assuntos que satisfazem determinados critérios de inclusão por um período de tempo. A definição da coorte está contida na tabela COHORT_DEFINITION. Exemplos de coortes podem incluir pacientes diagnosticados com uma condição específica, pacientes expostos a um determinado medicamento ou provedores que realizaram um procedimento específico.

COHORT_ATTRIBUTE - Essa tabela contém atributos associados a cada assunto dentro de uma coorte, conforme definido por um determinado conjunto de critérios de inclusão por um período de tempo. A definição do atributo de coorte está contida na tabela ATTRIBUTE_DEFINITION. Exemplos de atributos de coorte podem ser idade, índice de massa corpórea ou pontuação de comorbidades.

Eras: Uma era é definida como o intervalo de tempo durante o qual se presume que a pessoa tem uma determinada condição ou ficou exposta a um determinado princípio ativo. As eras são calculadas utilizando algoritmos padronizados a partir das informações de datas. Cada era corresponde a uma ou mais exposições que formam um intervalo contínuo. As eras são calculadas no momento da transformação da base. Temos DRUG_ERAS para medicamentos, DOSE_ERAS para doses constantes de medicamentos e CONDITION_ERAS para condições.

Por exemplo, no caso de medicamentos, as DRUG_ERAS são calculadas a partir das informações das datas de dispensação, ou DRUG_EXPOSURE. Uma pessoa tem 4 prescrições para a droga A (A1, A2, A3, A4), válida para 60 dias de dispensa. A pessoa também tem duas prescrições para a droga B (B1, B2). O diagrama da Figura 4.3. mostra a situação.

Para definir a era da droga para o medicamento A, o momento, a duração, a sobreposição e a persistência das prescrições do medicamento A devem ser consideradas. A3 foi preenchida antes do final esperado de A2. A4 foi preenchida após a conclusão do A3, mas dentro da janela de persistência para o medicamento A. Portanto, as quatro prescrições do medicamento A serão consolidadas em uma única era de medicamentos (Drug Era1), com o início da receita A1 registrado como o início a data do registro consolidado e a data final da prescrição A4 registrada como a data final.

Como a janela de persistência foi excedida entre o preenchimento das duas prescrições para a droga B, elas são definidas como duas eras de drogas distintas. As datas de início e término de Drug Era2 e Drug Era3 são as datas de início e término das prescrições B1 e B2, respectivamente. Observe que nenhuma janela de persistência adicional está sendo adicionada no final da última exposição à droga.

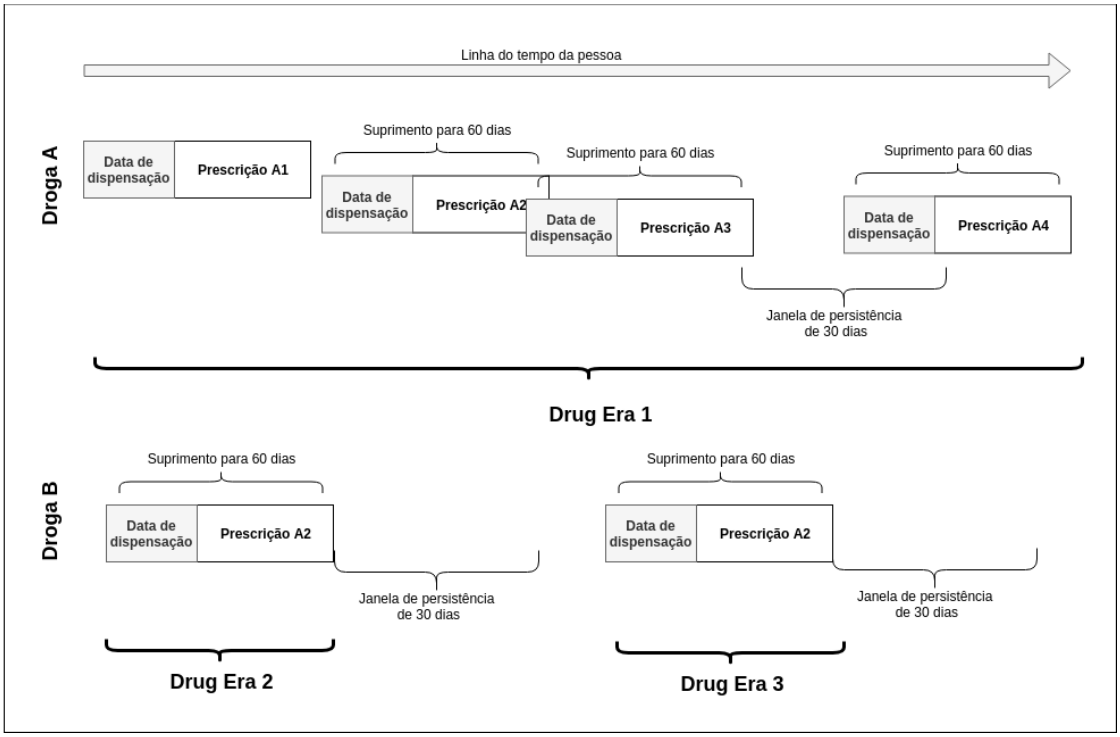

Figura 4.3. Definição de Eras

DRUG_ERA - É definida como um intervalo de tempo em que a pessoa é considerada exposta a um ingrediente ativo em particular.

DOSE_ERA - É definida como um intervalo de tempo em que a pessoa é considerada exposta a uma dose constante de um ingrediente ativo específico. É derivada das informações das tabelas de DRUG_EXPOSURE e DRUG_STRENGTH usando um algoritmo padronizado.

CONDITION_ERA - É definida como um período de tempo quando a pessoa é considerada como tendo uma determinada condição. Similar ao DRUG_ERA, CONDITION_ERA são períodos cronológicos da ocorrência da condição. Têm dois propósitos:

- Permite a agregação de condições crônicas que exigem frequentes cuidados contínuos, em vez de tratar cada ocorrência como um evento independente;
- Permite a agregação de múltiplas visitas médicas programadas para atender a mesma condição.

Por exemplo, considere uma pessoa que visita seu médico de cuidados primários e é encaminhado para um especialista. A pessoa visita o especialista, que define o diagnóstico e fornece o tratamento adequado para resolver a condição. Estas duas consultas médicas independentes devem ser agregadas em uma CONDITION_ERA.

4.3.4. Convenções do Modelo de Dados

Há um número de convenções que foram adotadas no CDM, sendo:

4.3.4.1. Convenções gerais das tabelas de dados

O CDM é independente de plataforma. Os tipos de dados são definidos genericamente usando ANSI SQL (VARCHAR, INTEGER, FLOAT, DATE, DATETIME, CLOB). O CDM não impõe o formato de data e data/hora.

Na maioria dos casos, o primeiro campo em cada tabela termina em '_ID', contendo um identificador de registro que pode ser usado como uma chave estrangeira em outra tabela.

4.3.4.2. Convenções gerais dos campos

Os nomes de variáveis em todas as tabelas seguem uma convenção:

- <entity>_SOURCE_VALUE:
  - informações dos dados de origem, geralmente usadas em ETL para mapear para CONCEPT_ID e não para serem usadas por nenhuma análise padrão
  - Ex: condition_source_value = '787.02' foi o código ICD-9 capturado como um diagnóstico da declaração administrativa.
- <entity> _ID:
  - Identificadores exclusivos para entidades chave, que podem servir como chaves estrangeiras para estabelecer relações entre entidades
  - Ex: person_id identifica exclusivamente cada indivíduo. visit_occurrence_id identifica exclusivamente um encontro PERSON em um ponto de atendimento.
- <entity> _CONCEPT_ID:
  - Chave estrangeira nos Vocabulários Padronizados (ou seja, o atributo standard_concept para o termo correspondente é true), que serve como base principal para todas as análises padronizadas
  - Ex: condition_concept_id =31967 contém valor de referência para o conceito SNOMED de ' Náusea '
- <entidade> _SOURCE_CONCEPT_ID:
  - Chave estrangeira nos vocabulários padronizados representando o conceito e terminologia utilizados nos dados de origem, quando aplicável
  - Ex: condition_source_concept_id = 35708202 denota o conceito de 'Náusea' na terminologia da terminologia MedDRA; o análogo condition_concept_id pode ser 31967, uma vez que SNOMED-CT é o vocabulário padronizado para a maioria dos diagnósticos e descobertas clínicas
- <entidade> _TYPE_CONCEPT_ID:
  - delineia a origem da informação da fonte, dentro dos vocabulários padronizados

- Ex: drug_type_concept_id pode permitir discriminar entre 'distribuição pharmacy' e 'receita escrita'

4.3.4.3. Representação do conteúdo através de conceitos

Os vocabulários padronizados contêm registros, ou conceitos, que identificam com exclusividade cada unidade fundamental de significado usada para expressar informações clínicas. Conceitos são derivados de vocabulários, que representam informações clínicas em diferentes domínios (por exemplo, condições, drogas, procedimentos) através do uso de códigos e descrições associadas. Alguns conceitos são designados como conceitos padrão, o que significa que esses conceitos podem ser usados como expressões normativas de uma entidade clínica dentro do CDM e dentro de análises padronizadas. Cada conceito padrão possui um domínio principal, que define a localização em que o conceito deveria ocorrer dentro do CDM.

Os conceitos podem representar categorias amplas (como "doença cardiovascular"), elementos clínicos detalhados ("infarto do miocárdio da parede anterolateral") ou características modificadoras e atributos que definem conceitos em vários níveis de detalhe (gravidade de uma doença, morfologia associada, etc.).

Os registros nas tabelas de vocabulários padronizados são derivados de vocabulários nacionais ou internacionais, como SNOMED-CT[15], RxNorm e LOINC[16], ou conceitos personalizados definidos para cobrir vários aspectos da análise de dados observacionais.

Nas tabelas de dados CDM o significado do conteúdo de cada registro é representado usando conceitos. Os conceitos são armazenados com seu concept_id como chaves estrangeiras para a tabela CONCEPT nos vocabulários padronizados, que contém conceitos necessários para descrever a experiência de assistência médica de um paciente. Se um conceito padrão não existir ou não puder ser identificado, é usado um conceito com o concept_id = 0, representando um conceito que não existente ou não é mapeável.

Os registros na tabela CONCEPT contêm todas as informações detalhadas (nome, relacionamentos, tipos etc.). A Tabela 4.1. apresenta a tabela CONCEPT, campos, se requerido ou não, tipo de dado e descrição.

[15] SNOMED-CT http://www.snomed.org/
[16] LOINC https://loinc.org/

Tabela 4.1. Tabela CONCEPT

Campo	Requerido	Tipo	Descrição
concept_id	sim	inteiro	Um identificador exclusivo para cada conceito em todos os domínios
concept_name	sim	varchar (255)	Um nome inequívoco, significativo e descritivo para o conceito
domain_id	sim	varchar (20)	Uma chave estrangeira para a tabela DOMAIN a qual o conceito pertence
vocabulary_id	sim	varchar (20)	Uma chave estrangeira para a tabela VOCABULARY indicando de qual fonte o conceito foi adaptado
concept_class_id	sim	varchar (20)	O atributo ou classe de conceito do Conceito. Exemplos são “Droga Clínica”, “Ingrediente”, “Localização Clínica” etc
standard_concept	não	varchar (1)	Esse sinalizador determina se o Conceito é um Conceito Padrão, um Conceito de Classificação ou um Conceito de Origem não padrão. Os valores permitidos são 'S' (Conceito Padrão) e 'C' (Conceito de Classificação), caso contrário, o conteúdo é NULL
concept_code	sim	varchar (50)	O código conceitual representa o identificador do Conceito no vocabulário de origem, como os IDs de conceitos SNOMED-CT, RxNorm, etc. Observe que os códigos conceituais não são exclusivos entre os vocabulários
valid_start_date	sim	date	A data em que o conceito foi gravado pela primeira vez. O valor padrão é 1-jan-1970, significando que o Conceito não possui data (conhecida) de início
valid_end_date	sim	date	A data em que o Conceito se tornou inválido porque foi excluído ou substituído (atualizado) por um novo conceito. O valor padrão é 31-Dec-2099, ou seja, o Conceito é válido até que se torne obsoleto
invalid_reason	não	varchar (1)	Motivo porque o conceito foi invalidado. Os valores possíveis são D (excluídos), U (substituídos por uma atualização) ou NULL quando valid_end_date tem o valor padrão

4.3.4.4. Conceito padrão, de classificação e de origem

Dentro de um Domínio, os códigos vêm de vários vocabulários, e frequentemente, têm significados idênticos ou sobrepostos. Para organização, para cada um deles é atribuído uma das três designações:

- Conceito padrão (standard_concept = 'S'): O conceito padrão é o conceito “oficial” que deve ser usado para representar uma entidade clínica única nas tabelas de dados clínicos padronizados. Seu código é gravado nos respectivos

campos concept_id. Normalmente, o conceito padrão é originado de vocabulários estabelecidos que têm uma cobertura abrangente e são bem definidos. Por exemplo, conceito obtido através do vocabulário SNOMED.

- Conceito de classificação (standard_concept = 'C'): conceito que têm um relacionamento hierárquico com o conceito padrão e, portanto, podem ser usados para consulta usando os registros da tabela CONCEPT_ANCESTOR. No entanto, eles próprios não podem aparecer nas tabelas de dados. Por exemplo, o conceito MedDRA para "COPD (chronic obstructive pulmonary disease)" têm relações hierárquicas com os conceitos padrão SNOMED-CT, que são todas as formas desta doença. Da mesma forma, o conceito 4283987 "ANTICOAGULANTES" do vocabulário VA Class[17] não pode aparecer nas tabelas DRUG_EXPOSURE ou DRUG_ERA, mas seus conceitos descendentes que possuem a classe de conceito "Ingrediente", "Droga Clínica" ou "Droga de Marca" podem.

  Os conceitos de classificação podem ser originados de diferentes vocabulários e não são exclusivos. Por exemplo, para a classe de medicamentos 'Anticoagulantes' há conceitos provenientes dos vocabulários NDF-RT[18], VA Class e ATC[19]. Observe também que a associação depende do vocabulário. Na maioria dos casos, a classificação é semelhante ou idêntica, mas não fornece uma definição padrão.
- Conceitos de origem (standard_concept = NULL): São todos os conceitos restantes que não são conceitos padrão ou de classificação. Observe que os conceitos podem alterar sua designação ao longo do tempo: se eles forem invalidados (valid_end_date for menor que 31-12-2099 e invalid_reason = 'D' (excluídos), ou 'U' (substituídos por uma atualização)), os antigos conceitos padrão ou de classificação se transformarão em conceitos de origem. Conceitos de origem só podem aparecer nos campos source_concept_id das tabelas de dados. Eles representam o código nos dados de origem. Cada conceito de origem é mapeado para um ou mais conceitos padrão durante o processo ETL. Se nenhum mapeamento estiver disponível, o conceito padrão com o concept_id = 0 será gravado no campo concept_id.

4.4. Vocabulários

Um dos principais problemas no agrupamento de fontes de dados diversas é a procura por uma definição comum do significado das informações nelas armazenadas.

Os vocabulários padronizados contêm registros, ou conceitos, que identificam de forma exclusiva cada unidade fundamental de significado usada para expressar

[17] VA Class (Veterans Affairs Drug Class) https://www.pbm.va.gov/clinicalguidance/drugclassreviews.asp
[18] NDF-RT (National Drug File - Reference Terminology) https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/
[19] ATC (Anatomical Therapeutic Chemical) https://www.whocc.no/atc_ddd_index/

informações clínicas em todas as tabelas de domínio CDM. São construídos com alguns princípios que representam o processo contínuo de melhoria e desenvolvimento:

1. Padronização: vários vocabulários usados em dados observacionais são consolidados em um formato comum. Isso alivia os pesquisadores de ter que entender e lidar com vários formatos e convenções de ciclo de vida diferentes dos vocabulários de origem;
2. Conceito padrão único: para cada entidade clínica existe apenas um conceito representando-o, denominado 'conceito padrão'. Outros conceitos equivalentes ou similares são designados como não-padrão e mapeados para os padrão;
3. Domínios: cada conceito recebe um domínio. Conceitos não-padrão também podem pertencer a mais de um domínio. Isso também define em qual tabela CDM uma entidade clínica deve ser colocada no momento da consulta;
4. Cobertura abrangente: todos os eventos relevantes de assistência médica do paciente (por exemplo, condições, procedimentos, exposições a medicamentos, etc.) e alguns dos artefatos administrativos do sistema de saúde (por exemplo, visitas, locais de atendimento, etc.) são cobertos pelo conceito de um domínio;
5. Hierarquia: dentro de um domínio, todos os conceitos são organizados em uma estrutura hierárquica. Isso permite consultar todos os conceitos (por exemplo, medicamentos) que são subordinados hierarquicamente sob um conceito de nível superior (por exemplo, uma classe de medicamentos). Isso implica abordar dois problemas distintos:
    - Cada conceito deve ter uma ou mais classificações (de baixo para cima);
    - Cada classificação deve conter todos os conceitos relevantes (de cima para baixo).
6. Relacionamentos entre conceitos dentro e entre vocabulários e mapeamentos de conceitos não padronizados para conceitos padrão;
7. Ciclo de vida mantendo a representação de dados atualizada, mas suportando o processamento de conceitos descontinuados e atualizados.

É importante notar que esses critérios têm o objetivo de servir à pesquisa observacional. Nesse sentido, os vocabulários padronizados diferem de grandes coleções com mapeamentos de equivalência de conceitos como o UMLS[20] que suporta indexação e pesquisa de toda a literatura biomédica. Os recursos da UMLS têm sido usados como base para a construção de muitos dos componentes do Vocabulário Padronizado, mas esforços adicionais significativos foram feitos para ajustar a estrutura:

- São estabelecidos vocabulários adicionais, principalmente para fins de metadados;
- Estão sendo adicionados mapeamentos e relacionamentos para obter uma cobertura abrangente. Se não for possível uma equivalência, serão criados relacionamentos de conceitos-padrão mais granulares, não padronizados, para níveis mais elevados;

[20] UMLS https://www.nlm.nih.gov/research/umls/

- É estabelecida uma estrutura de domínio abrangente e cada conceitos recebe um domínio ou uma combinação de domínios;
- É construída uma árvore hierárquica dentro dos domínios representando as classificações usadas na ciência médica e na prática clínica.

Descrevemos aqui a solução adotada pela iniciativa OHDSI, a forma como os diversos vocabulários são incorporados dentro da plataforma e como estes afetam as pesquisas.

4.4.1. Estrutura dos Vocabulários Padronizados

Os vocabulários padronizados contêm todos os conjuntos de códigos, terminologias, vocabulários, nomenclaturas, léxicos, tesauros, ontologias, taxonomias, classificações, abstrações e outros dados que são necessários para:

- Geração dos dados transformados (padronizados) do conjunto de dados brutos para o CDM;
- Pesquisar, consultar e extrair dados transformados e navegar pelas hierarquias de classes e abstrações inerentes aos dados transformados;
- Interpretar os significados dos dados.

Os dados em nível de paciente disponíveis no CDM exigem explicitamente a representação de todos os fatos e eventos clínicos usando conceitos dos vocabulários padronizados. Com poucas exceções, não há informações textuais nas tabelas do CDM. Portanto, os vocabulários padronizados são parte integrante do CDM. Geralmente, todos os componentes são *Open Source* - a menos que especificado de outra forma para alguns vocabulários comerciais.

A Tabela 4.2. apresenta os termos e descrições usados na estrutura dos vocabulários.

Os vocabulários padronizados fornecem uma representação padronizada de dados nos seguintes domínios clínicos:

- Dados demográficos: gênero, etnia, raça
- Condição
- Droga
- Procedimento
- Medição
- Observação
- Nota
- Dispositivo
- Espécime
- Unidade
- Visita
- Óbito
- Fornecedor
- Custo

Tabela 4.2. Tabela de termos

Termos	Descrição
Vocabulários padronizados	Contém um sistema de vocabulários, classificações, domínios e conceitos, todos consolidados em um formato comum e armazenados em um conjunto de tabelas CDM
Vocabulário	Um conjunto de códigos ou conceitos, incluindo, se disponíveis, relações entre eles, incluindo, se disponível, uma hierarquia, ontologia ou taxonomia dos conceitos. Muitos vocabulários são adotados de organizações nacionais ou internacionais, como o ICD-9-CM[1], o SNOMED-CT, o RxNorm, o Read[2]
Terminologia	Semelhante ao vocabulário e frequentemente usado como sinônimo
Esquema de codificação	Semelhante ao vocabulário e frequentemente usado como sinônimo
Classificação	Um sistema hierárquico de conceitos e relações conceituais que define classes semanticamente úteis, como estruturas químicas para drogas
Domínio	Uma categoria semântica clínica, como droga, condição, procedimento definido para todos os conceitos nos vocabulários padronizados
Conceito	Unidade básica de informação definida em cada vocabulário
Classe Concept	Um atributo de um conceito que caracteriza sua classificação dentro de um vocabulário. A diferença para a classificação é que uma classe concept é um único atributo sem qualquer estrutura hierárquica

4.4.2. Vocabulários padronizados

Os Vocabulários Padronizados estão organizados em domínios e vocabulários. Os domínios referem-se à natureza ou tipo de uma entidade clínica. Ele também define a tabela de dados do CDM onde um registro de dados deve ser armazenado. Vocabulários são conjuntos de conceitos importados de um padrão externo nacional ou internacional existente, ou criados pela equipe dos vocabulários padronizados, se nenhum padrão adequado estiver disponível. Não existe uma relação de um para um entre domínios e vocabulários. Alguns vocabulários são muito amplos, como SNOMED ou Read, e contêm conceitos de todos os domínios médicos. Da junção do SNOMED RT e do Read Codes surgiu o SNOMED CT. Outros vocabulários são específicos para um determinado domínio, como RxNorm for Drugs ou ICD9CM. Em muitos casos, os vocabulários são geralmente assumidos na comunidade como sendo de um único domínio, quando na verdade eles não são. Por exemplo, CPT4[21] e HCPCS[22] são esperados para conter apenas códigos de procedimento, mas na realidade contêm conceitos de observação, condição, dispositivo e droga. A Figura 4.4. apresenta um

[21] CPT (Current Procedural Terminology) https://www.ama-assn.org/
[22] HCPCS (Healthcare Common Procedure Coding System) https://www.ama-assn.org/

esquema dos domínios e a amplitude das relações de alguns vocabulários com cada domínio[23].

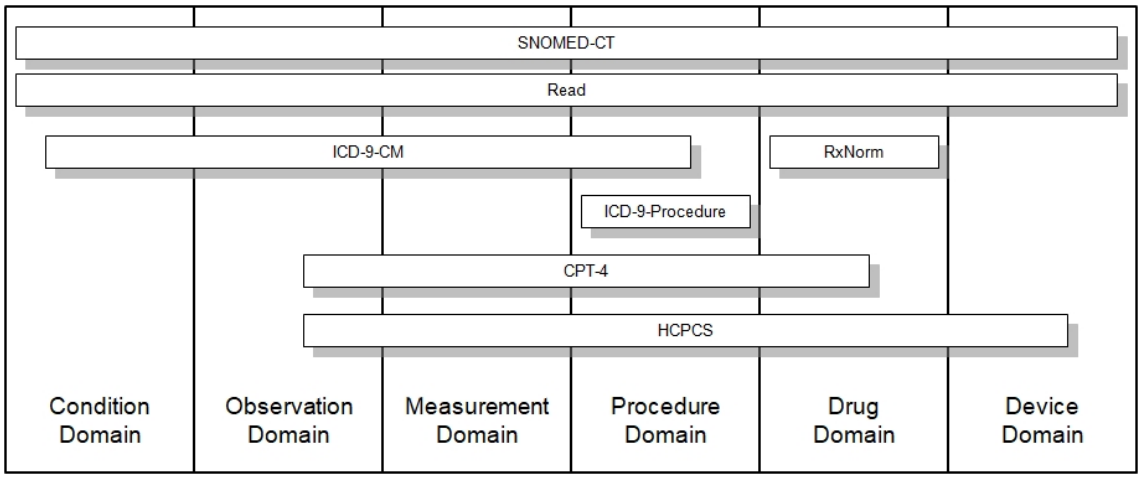

Figura 4.4. Domínios e vocabulários

As tabelas contêm informações detalhadas sobre os conceitos usados em todas as tabelas de fatos do CDM. O conteúdo das tabelas de vocabulários padronizados são mantidas centralmente como um serviço para a comunidade. Várias proposições foram feitas para o projeto das tabelas de vocabulários padronizados, sendo:

- Existe um esquema que acomoda todas as diferentes terminologias e classificações de origem;
- Todas as terminologias são carregadas na tabela CONCEPT;
- Cada termo carregado recebe um código novo como chave (concept_id). O código original da terminologia não é utilizado como identificador porque ele não é exclusivo entre terminologias;
- Alguns conceitos são declarados conceitos padrão, isto é, são usados para representar uma determinada entidade clínica nos dados. Todos os conceitos podem ser conceitos de fonte; eles representam como a entidade foi codificada na fonte. Os conceitos padrão são identificados por meio do campo standard_concept na tabela CONCEPT;
- Registros na tabela CONCEPT_RELATIONSHIP definem relações semânticas entre conceitos. Essas relações podem ser hierárquicas ou laterais;
- Os registros na tabela CONCEPT_RELATIONSHIP são usados para mapear códigos fonte para conceitos padrão, substituindo o mecanismo da tabela SOURCE_TO_CONCEPT_MAP usada em versões anteriores de vocabulários padronizados. A tabela SOURCE_TO_CONCEPT_MAP é mantida como um

[23] Documentação Vocabulários
http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:domains_and_vocabularies

auxílio opcional aos códigos de contabilidade não encontrados nos vocabulários padronizados;

- As cadeias de relacionamentos hierárquicos são registradas na tabela CONCEPT_ANCESTOR. Os relacionamentos de ancestralidade são registrados apenas entre os conceitos padrão que são válidos (não preteridos) e são conectados por meio de relacionamentos válidos e hierárquicos na tabela RELATIONSHIP (sinalizador define_ancestry). A vantagem dessa abordagem reside na preservação de códigos e relacionamentos entre eles, sem a adesão às várias estruturas de dados de origem diferentes, um design simples para acesso padronizado e a otimização do desempenho para a análise OHDSI. Navegação entre conceitos padrão não requer conhecimento do vocabulário de origem. Finalmente, a abordagem é escalável e vocabulários futuros podem ser facilmente integrados. Por outro lado, é necessária uma transformação extensiva de dados de origem para o vocabulário e nem toda estrutura de dados de origem e hierarquia de origem podem ser retidas.

Atualmente, 81 vocabulários fazem parte dos vocabulários padronizados. Muitos deles são adotados de fontes de terceiros, que os desenvolvem e os mantêm para fins específicos, como, ICD10[24] ou SNOMED-CT. A consolidação dos vocabulários em uma forma padronizada requer uma série de decisões e convenções. Um grupo trata da organização dos vocabulários dentro dos domínios clínicos, a implementação específica de cada vocabulário nos vocabulários padronizados, e o mapeamento de conceitos dentro e entre os vocabulários. Também fornece orientação sobre como aplicá-los para transformação de dados de origem no CDM e sobre a consulta de dados, uma vez estabelecidos no formato CDM. Os vocabulários padronizados estão na versão 5.x. e todos os conceitos nas versões anteriores ainda estão disponíveis e identificados usando os mesmos IDs de conceito.

4.4.2.1. Mapeamento de conceitos

O mapeamento é o processo para transformar um conceito em outro. As tabelas de dados clínicos do CDM permitem apenas conceitos padrão. Todos os outros códigos usados nos bancos de dados de origem precisam ser traduzidos para os conceitos padrão. O mapeamento é feito por meio de registros na tabela CONCEPT_RELATIONSHIP. Eles conectam cada conceito a um conceito padrão através de um número especial de relationship_id (maps to e maps to value).

Relacionamentos **'Maps to'**: Os conceitos que participam do mapeamento 'mapear para' são conceitos de origem e conceitos padrão. O mapeamento tenta mapear para o conceito de destino equivalente. Equivalente significa que ele carrega o mesmo significado e, mais importante, os filhos na hierarquia (se houver algum) também são equivalentes ou cobrem o mesmo espaço semântico. Se um conceito equivalente não

[24] ICD10 https://www.who.int/classifications/icd/icdonlineversions/en/

estiver disponível, o mapeamento tentará corresponder a um conceito mais amplo. Isso garante que uma consulta no vocabulário de destino recupere os mesmos registros, como se tivessem sido consultados no vocabulário de origem original.

Geralmente, conceitos de origem e conceitos padrão são mapeados:

- Conceitos de origem são mapeados para um ou vários conceitos padrão. Se eles forem mapeados para mais de um conceito padrão, então, na tabela CDM resultante, mais de um registro será gravado para cada registro na origem;
- Os conceitos padrão também são mapeados para os conceitos padrão, geralmente este é um mapa para si mesmo.

Os conceitos de classificação (standard_concept = 'S') não possuem um mapeamento para um conceito padrão.

Relacionamentos '**Maps to value'**: essas relações são projetadas para distinguir entre observação e medidas e seus resultados. Por exemplo, ICD9CM V12.71 concept_id 44820383 "História pessoal de doença ulcerosa péptica" tem uma relação "Maps to" para SNOMED 4214956 "História de descoberta clínica em questão" (Domínio de Observação) e outro relacionamento "Maps to value" para SNOMED 4027663 "Úlcera péptica" (domínio de condição).

'Perdas' devido ao mapeamento: há uma preocupação significativa sobre perdas do mapeamento de um vocabulário para outro, sobre a qualidade dos dados e a capacidade de identificar com segurança pacientes para uma coorte, dado os critérios de inclusão e exclusão. Essa perda pode ocorrer por vários motivos:

- Está faltando o mapeamento. Informe os mapeamentos perdidos para o fórum OHDSI[25], para que possam ser adicionados;
- O código fonte está mal definido, por exemplo, CID9CM 799 "Outras causas de morbidade e mortalidade mal definidas e desconhecidas";
- Afirmação negativa, por exemplo ICD9CM V64.0 "Vacinação não realizada";
- Código irrelevante para o paciente, por exemplo, ICD9CM V65 "Outras pessoas que procuram consulta";
- Hierarquias de origem e destino incongruentes.

O último efeito não é incomum para conceitos altamente pré-coordenados (conceitos complexos e refinados que são combinações de diferentes dimensões), a coordenação depende da estrutura da topologia dos vocabulários e para os quais o conceito é mapeado. Se não forem equivalentes, nenhum mapeamento direto poderá ser estabelecido. Mas, para manter a capacidade de recuperar esses conceitos ao pesquisar usando conceitos hierárquicos, dois ou mais mapeamentos são fornecidos em seu lugar. Por exemplo, o ICD10 conceito 45755355 "Diabetes mellitus não dependente de insulina com coma" (código E10.0) não pode mapear diretamente para SNOMED. A Figura 4.5. apresenta o fluxo do mapeamento desse conceito para o seu conceito padrão.

[25] Fórum OHDSI http://forums.ohdsi.org/c/cdm-builders

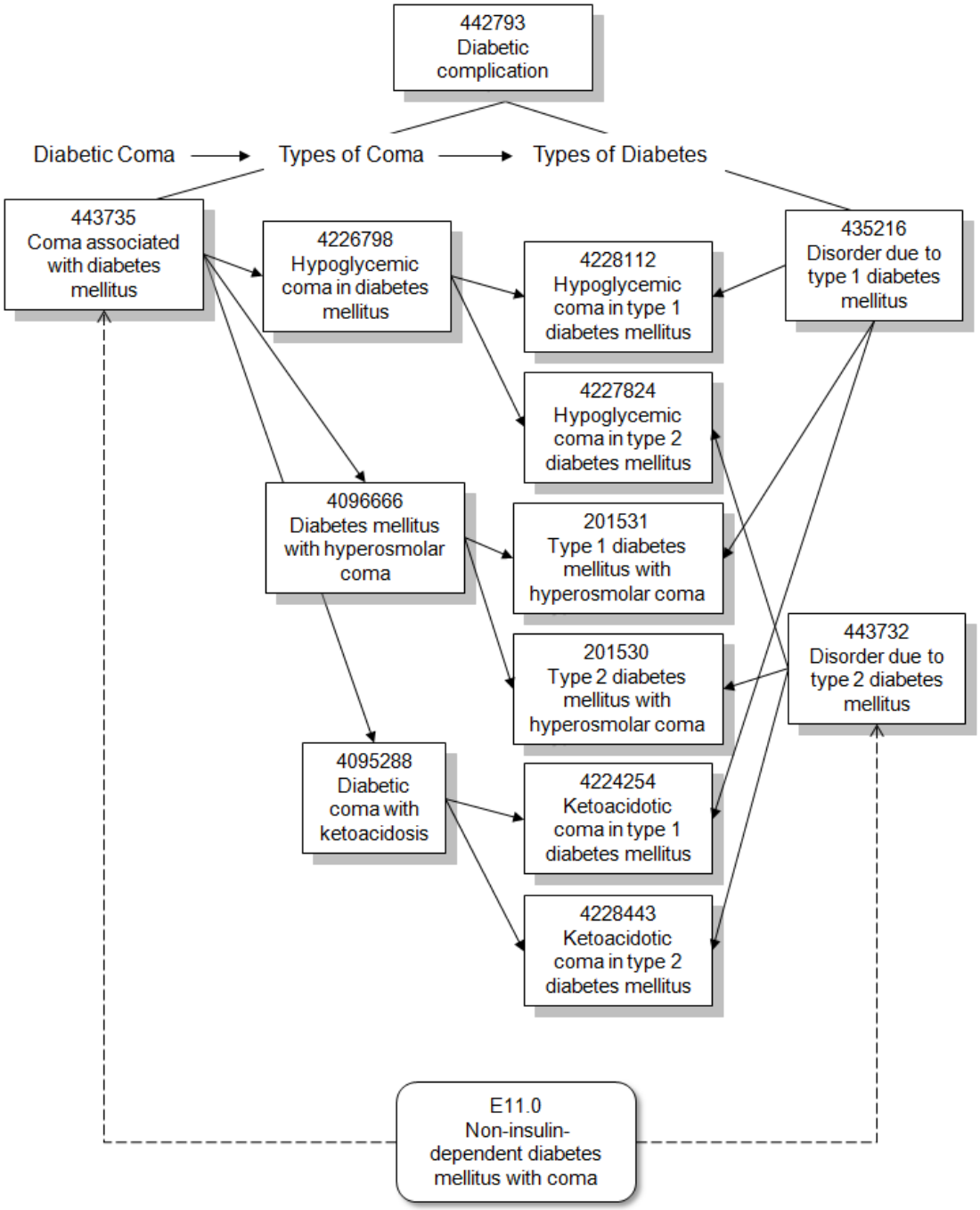

Figura 4.5. Mapeamento de um código ICD10 para um conceito padrão

A hierarquia de destino no SNOMED tem uma topologia diferente para a estrutura organizacional no ICD10. Aqui, diabetes tipo 1 (dependente de insulina, E10) e tipo 2 (independente de insulina, E11) são as principais características distintivas. A combinação de diabetes tipo 2 com a combinação coma E11.0 define diabetes tipo 2 com o coma de complicação. No SNOMED, as complicações do diabetes têm sua própria hierarquia, que se ramifica no conceito de diabetes com coma, ou diabetes tipo 1 ou 2 com complicação. No nível seguinte, distinguem-se as formas distintas de coma

(hipoglicêmico, hiperosmolar e cetoacidótico). No entanto, não há combinação de diabetes tipo 2 com coma. A solução é mapear o E11.0 para o diabetes mellitus tipo 2 44054006 e para o coma associado com diabetes mellitus 420662003. Mesmo que não haja um equivalente distinto no SNOMED, as consultas hierárquicas para o diabetes 2 em combinação com o coma recuperarão os registros corretos.

4.4.2.2. Condições (diagnósticos e achados clínicos)

Condições são registros de uma pessoa sugerindo a presença de uma doença ou condição médica declarada como um diagnóstico, um sinal ou um sintoma, que é observado por um provedor ou relatado pelo paciente. Note que o domínio de condição não faz distinção entre uma mera observação de um sintoma, um diagnóstico totalmente elaborado ou qualquer coisa entre eles. Conceitos de domínio de condição devem ser usados somente nos campos condition_concept_id das tabelas CONDITION_OCCURRENCE e CONDITION_ERA, bem como no campo value_as_concept_id na tabela OBSERVATION (para observações "History of", "Family history of" etc.).

Os conceitos padrão são retirados do vocabulário SNOMED. Todos os conceitos SNOMED válidos (invalid_reason is null) do domínio de condição são conceitos padrão e, portanto, podem ser usados no campo condition_concept_id nas tabelas CONDITION_OCCURRENCE e CONDITION_ERA.

SNOMED é um vocabulário hierárquico. Portanto conceitos SNOMED também podem ser usados como conceitos de classe: descendentes de qualquer conceito SNOMED na tabela CONCEPT_ANCESTOR podem ser usados como uma correspondência semântica correta em uma consulta.

Regras para o mapeamento de condição: Há muitas condições que não são originárias da função biológica do corpo humano, mas ainda requerem atenção médica. Ou são verdadeiras condições, mas não no momento do levantamento da 'história de' ou na 'história familiar de'. As regras de mapeamento para estas condições são as seguintes:

- Cuidados posteriores após os procedimentos: estes são mapeados para um único conceito 413467001 "Aftercare", e com um segundo relacionamento "Maps to value" para o procedimento adequado. Por exemplo, ICD10CM Z47.1 "Cuidados posteriores à cirurgia de substituição articular" mapeia para o 413467001 "Aftercare" e para o procedimento 4189532 "Implantation of joint prótese". Note que "Aftercare" não é uma condição, mas sim do domínio de observação;
- Efeitos tardios ou sequelas de outras condições: se possível, eles são mapeados para conceitos únicos, descrevendo-os como efeitos tardios. Por exemplo, ICD10CM S82.874S concept_id 45589204 "Fratura de pilão não desdentado da tíbia direita, sequela" é mapeada para 197150 "Efeito tardio da fratura das extremidades inferiores". O domínio é condição;

- História de uma condição: estas são condições, mas não no momento da condição_start_date. Portanto, eles são mapeados para o conceito 4214956 “História de descoberta clínica em questão”, que está no domínio de observação; A condição em si é registrada por meio do relacionamento "Mapear para avaliar". Por exemplo, ICD10CM Z87.820 concept_id “História pessoal de traumatismo cranioencefálico” mapeia para 4214956 “Histórico de achados clínicos em questão” e com “Maps to value” 4132546 “Lesão cerebral traumática”;
- História familiar de uma condição: estes, também, têm um mapeamento duplo para 4167217 "História familiar de descoberta clínica" e um para a condição real através de um relacionamento "Maps to value". Por exemplo, ICD10 / ICD10CM Z80.0 concept_id 45542462 “História familiar de neoplasia maligna de órgãos digestivos” mapeia para 4167217 “Histórico familiar de constatação clínica”, bem como 443568 “Neoplasia maligna do trato gastrointestinal”;
- História do tratamento médico: similarmente, estes são mapeados para o conceito de observação 4207283 “História da terapia medicamentosa”, e através de um link “Maps to value” para a terapia real. Por exemplo, ICD10 Z92.0 concept_id 45605174 “História pessoal de contracepção” vai para 4207283 “História da terapia medicamentosa” história de terapia medicamentosa e 4027509 “Contracepção”. Se o ingrediente exato ou o medicamento clínico / de marca for conhecido, o mapa será direcionado a esses conceitos de medicamentos. Mas isso não é típico;
- Efeito adverso da medicação: se possível, eles são mapeados para um equivalente direto. No entanto, isso geralmente não existe e, em seguida, eles são tratados como a história acima de tratamentos médicos. Por exemplo, tanto ICD10CM T36.8X5A concept_id 45551127 “efeito adverso de outros antibióticos sistêmicos, encontro inicial” e T36.8X5D “efeito adverso de outros antibióticos sistêmicos, encontro posterior” mapear para 437191 “reação adversa a medicamentos antibacterianos”, enquanto CID10CM T36.8X5S concept_id 45560654 “Efeito adverso de outros antibióticos sistêmicos sequela” mapeia para 4207283 “História da terapia medicamentosa”;
- Subdosagem de medicação: essas informações são tratadas de maneira semelhante aos efeitos adversos e os conceitos equivalentes do SNOMED geralmente não existem. Por exemplo, ICD10CM T36.0X6A concept_id 45565479 “Subdosagem de penicilinas, encontro inicial” e T36.0X6D concept_id 45565480 “Subdosagem de penicilinas, encontro subsequente” vá para 40488434 “Dose de medicação muito baixa”, enquanto T36.0X6S concept_id 45565481 “Subdosagem de penicilinas, sequela ”aponta em 4207283“ História da terapia medicamentosa ”;
- Status de ausência do órgão ou transplante / presença de prótese: órgãos ausentes são devidos a um procedimento que os removeu (a menos que sejam condições inatas, que são mapeadas como tal). Portanto, eles são mapeados para

4215685 “Histórico passado de procedimento” e o procedimento apropriado que removeu o órgão por meio de um link “Maps to value”. Por exemplo, ICD10CM Z94.0 concept_id 35225404 “Situação do transplante renal” mapeia para 4215685 “Histórico passado de procedimento” e 4322471 “Transplante de rim”. Z95.5 concept_id 35225418 “Presença de implante de angioplastia coronariana e enxerto” é apontada para 4215685 “História pregressa de procedimento” e 4184832 “Angioplastia coronariana”;

- Conceitos pré-coordenados que listam dois ou mais componentes semânticos através de AND ou OR: esses conceitos são tratados com a seguinte ordem de precedência:
    - Para um conceito de combinação equivalente que também é bem conectado hierarquicamente
    - Para ambos os componentes separadamente
    - Para o ancestral mais comum

Exemplos para estas possibilidades são:

- ICD10 A01 concept_id 45576225 “Febre tifoide e paratifoide” tem uma relação única “Maps to” para 4022808 “Febre tifoide humana E / OU paratifoide”;
- IC10CM F12.22 concept_id 45591098 “Dependência de cannabis com intoxicação” tem dois relacionamentos “Maps to” para 4052690 “intoxicação por cannabis” e 440387 “dependência de cannabis”;
- ICD10 L02.0 concept_id 45596354 “Abscesso cutâneo, furúnculo e carbúnculo da face” tem um único relacionamento “Maps to” de 400082007, “Transtorno da pele da cabeça”;

- Cuidado materno: muitas condições requerem atenção não por causa de uma condição de uma mulher grávida, mas do feto. No entanto, todas essas condições estão sendo mapeadas para a mãe de qualquer maneira. Por exemplo, ICD10 O35.6 concept_id 45567927 “Cuidados maternos para (suspeita) danos ao feto por radiação” tem dois relacionamentos “Maps to” para 199553006 “Feto com dano por radiação” e 289908002 “gravidez”. Ambas as condições são registradas com a mãe;
- Necessidade de imunização: esses conceitos são mapeados para uma observação indicando essa lacuna de imunidade. Um segundo mapeamento com o relationship_id “Maps to value” é então direcionado para a condição (representada como um conceito SNOMED) contra a qual a imunização é inoculada. Observe que ele não está mapeado para a vacina em si (que seria representado como um conceito de RxNorm). Por exemplo, ICD10 Z23 concept_id 45556822 “Necessidade de imunização contra doenças bacterianas únicas” é mapeado para 170536002 “Vacinação necessária” e mapeia para o valor 87628006 “Doença infecciosa bacteriana”;
- Condições que indicam níveis anormais de um teste: estes são divididos em conceitos de medição e resultado. Por exemplo, ICD10 R77.1 concept_id

45553745 “Anormalidade da globulina” tem uma relação de “Maps to” com a medição “Globulina” da medição 4353510 e uma relação “Maps to value” para 4135493 “Anormal”.

4.4.2.3. Medicamentos/drogas

Os conceitos de domínio de exposição a medicamentos capturam registros sobre a utilização de um medicamento quando ingeridos ou introduzidos de uma forma no corpo humano. Uma droga é uma substância bioquímica formulada de tal maneira que, quando administrada a uma pessoa, ela exerce certo efeito fisiológico ou bioquímico.

Os seguintes produtos não são considerados drogas, mas dispositivos:

- Radiofármacos
- Material de contraste para geração de imagens
- Produtos nutricionais e suplementos, incluindo fórmulas infantis. Na realidade, isso resulta na situação ligeiramente arbitrária e, em alguns casos, difícil de verificar que soluções de sais para uso parental são Drogas (hidratar pacientes e manter o equilíbrio iônico), enquanto a adição de nutrientes como glicose ou vitaminas os torna dispositivos (alimentação de pacientes)
- Produtos diretamente derivados do sangue (por exemplo, eritrócitos ou plasma)

Os conceitos do domínio medicamentos devem ser usados no drug_concept_id das tabelas DRUG_EXPOSURE, DRUG_ERA e DOSE_ERA (ambos apenas no nível do ingrediente) ou no campo value_as_concept_id da tabela OBSERVATION ou MEASUREMENT (por exemplo, para medições como "Nível Plasma").

4.4.2.4. Medidas (valores quantitativos)

A tabela MEASUREMENT contém registros de medição, ou seja, valores estruturados (numéricos ou categóricos) obtidos por meio de exame sistemático e padronizado ou teste de uma amostra de uma pessoa. Contém tanto pedidos quanto resultados de tais medidas, como testes de laboratório, sinais vitais, resultados quantitativos de relatórios de patologia, etc. A tabela MEASUREMENT requer algumas convenções, descritas a seguir:

- As medições diferem das observações, na medida em que exigem um teste padronizado ou alguma outra atividade para gerar um resultado quantitativo ou qualitativo. Por exemplo, LOINC 1755-8 concept_id 3027035 'Albumina [Massa / tempo] em urina de 24 horas' é o teste de laboratório para medir um determinado produto químico em uma amostra de urina;
- Mesmo que cada medida tenha sempre um resultado, os campos value_as_number e value_as_concept_id não são obrigatórios. Quando o resultado não é conhecido, o registro de medição representa apenas o fato de que a medição correspondente foi realizada, o que em si já é uma informação útil para alguns casos de uso;

- Conceitos válidos de medição (measurement_concept_id) pertencem ao domínio 'Measurement', mas podem se sobrepor ao domínio 'Observation'. Isso se deve ao fato de que existe uma sobreposição entre o exame ou teste sistemático (medição) e uma simples determinação de fato (observação). Quando o valor da fonte de medição do código não pode ser convertido em um ID de conceito de medição padrão, uma entrada de medição é armazenada com apenas o source_concept_id e measurement_source_value correspondente e um measurement_concept_id de 0;
- As medições são armazenadas como pares de valores de atributo, com o atributo como o conceito de medição e o valor representando o resultado. O valor pode ser um conceito (armazenado em value_as_concept) ou um valor numérico (value_as_number) com uma unidade (unit_concept_id);
- Conceitos válidos para o campo value_as_concept pertencem ao domínio 'Meas Value';
- Para alguns conceitos de medição, o resultado é incluído no teste. Por exemplo, ICD10 concept_id 45595451 “Presença de álcool no sangue, nível não especificado” indica uma medição e o resultado (presente). Nessas situações, a tabela CONCEPT_RELATIONSHIP além do registro “Maps to” contém um segundo registro com o relationship_id definido como “Maps to value”. Neste exemplo, o relacionamento "Maps to" direciona para 4041715 "Medição de etanol de sangue", bem como um registro "Maps to value" para 4181412 "Presente";
- O operator_concept_id é fornecido opcionalmente para medições relativas, em que o valor preciso não está disponível, mas sua relação com um determinado valor está. Por exemplo, isso pode ser usado para limites mínimos de detecção de um teste;
- O significado do conceito 4172703 para '=' é idêntico à omissão de um valor operator_concept_id. Como o uso desse campo é raro, é importante, ao elaborar análises, não esquecer o teste do conteúdo desse campo para valores diferentes de =;
- Os conceitos válidos para o campo operator_concept_id pertencem ao domínio 'Meas Value Operator';
- A unidade é opcional, mesmo que seja fornecido um value_as_number;
- Se os intervalos de referência para o limite superior e inferior do normal, conforme previsto (normalmente por um laboratório), estes são armazenados nos campos range_high e range_low. Os intervalos têm a mesma unidade que o value_as_number;
- A visita durante a qual a observação foi feita é registrada através de uma referência à tabela VISIT_OCCURRENCE. Esta informação nem sempre está disponível;
- O provedor que faz a observação é registrado por meio de uma referência à tabela PROVIDER. Esta informação nem sempre está disponível.

4.4.3. Vocabulários Locais

Existem três abordagens para manipular códigos fonte que não estão no vocabulário OMOP (em ordem de complexidade):

1. Utilizando o SOURCE_TO_CONCEPT_MAP: No vocabulário OMOP existe uma tabela vazia chamada SOURCE_TO_CONCEPT_MAP. É uma estrutura de tabela simples que permite estabelecer mapeamento(s) para cada código-fonte com um conceito padrão no vocabulário OMOP (TARGET_CONCEPT_ID). Esse trabalho pode ser facilitado pela ferramenta USAGI da OHDSI, que verifica a semelhança de texto entre as descrições do código-fonte e o vocabulário OMOP e mapeia os resultados em uma estrutura de tabela SOURCE_TO_CONCEPT_MAP. Exemplos de arquivos SOURCE_TO_CONCEPT_MAP podem ser encontrados em [26]. Esses arquivos SOURCE_TO_CONCEPT_MAP gerados são carregados no SOURCE_TO_CONCEPT_MAP vazio do vocabulário do OMOP antes de incorporar os dados nativos ao CDM, para que o processo de ETL do CDM possa utilizá-los.

2. Adicionando CONCEPT.CONCEPT_IDs: Quando um código-fonte não é suportado pelo vocabulário OMOP, pode-se criar novos registros na tabela CONCEPT, porém os CONCEPT_IDs devem iniciar > 2000000000 para que seja fácil distinguir entre os conceitos de vocabulário OMOP e os conceitos específicos do site. Quando esses conceitos existirem CONCEPT_RELATIONSHIPS podem ser gerados para atribuí-los a terminologias padrão, o USAGI também pode facilitar esse processo.

3. Trabalhe com o *ODYSSEUS Data Services*[27] para adicionar ao Vocabulário OMOP. O vocabulário OMOP está em evolução e novos vocabulários podem ser adicionados.

4.5. Fenotipagem: Definição de uma coorte

Fenótipo é o termo criado pelo pesquisador dinamarquês Wilhelm L. Johannsen (1857 – 1912) e representa as características (parâmetros) que definem um indivíduo, sejam elas morfológicas, fisiológicas ou comportamentais.

A fenotipagem é o processo de identificação de pacientes com uma condição ou característica médica por meio de uma consulta de pesquisa a um sistema RES ou repositório de dados clínicos usando um conjunto definido de elementos de dados e expressões lógicas. O objetivo da fenotipagem é construir coortes identificando pacientes com uma condição médica particular, por exemplo, pacientes com Diabetes Mellitus Tipo 2 (DM2) ou aqueles que sofreram um Infarto do Miocárdio (IM).

Fenotipagem é a seleção dos valores de um conjunto de parâmetros que definem a classificação dos indivíduos participantes de uma coorte. Podemos agrupar o conjunto

[26] https://github.com/OHDSI/ETL-CDMBuilder/tree/master/man/VOCABULARY_ADDITIONS
[27] *ODYSSEUS Data Services* https://odysseusinc.com/

de parâmetros em: eventos de entrada na coorte, critérios de inclusão/exclusão e parâmetros de saída da coorte.

Nesta seção abordamos a definição dos parâmetros que compõem o processo de formação de uma coorte, junto com uma visualização gráfica dos principais elementos da coorte no tempo.

4.5.1. Tabela COHORT_DEFINITION

A tabela COHORT_DEFINITION contém registros definindo uma coorte derivada dos dados através da descrição e sintaxe associadas e mediante instanciação (execução do algoritmo) inserida na tabela COHORT. Coortes são um conjunto de assuntos que satisfazem uma determinada combinação de critérios de inclusão por um período de tempo. A tabela COHORT_DEFINITION fornece uma estrutura padronizada para manter as regras que governam a inclusão de um assunto em uma coorte, e pode armazenar código de programação operacional para instanciar a coorte dentro do CDM.
A tabela requer algumas convenções, descritas a seguir:

- O cohort_definition_syntax não prescreve nenhuma sintaxe específica ou linguagem de programação. Normalmente, seria qualquer SQL, uma linguagem de definição de coorte ou uma descrição de texto livre do algoritmo;
- O subject_concept_id determina em que consistem os sujeitos ou entidades individuais da coorte. Na maioria dos casos, isso seria uma pessoa (paciente). Mas coortes também poderiam ser construídas para provedores, visitas ou qualquer outro domínio. Observe que o domínio não é codificado usando o domain_id alfanumérico como na tabela CONCEPT. Em vez disso, o conceito correspondente é usado. Os conceitos para cada domínio podem ser obtidos na tabela DOMAIN no domínio_concept_id.

4.5.2. A tabela COHORT_ATTRIBUTE

A tabela COHORT_ATTRIBUTE contém atributos associados a cada assunto dentro de uma coorte, conforme definido por um determinado conjunto de critérios por um período de tempo. A definição do atributo de coorte está contida na tabela ATTRIBUTE_DEFINITION. Requer algumas convenções, descritas a seguir:

- Cada registro na tabela COHORT_ATTRIBUTE está vinculado a um registro específico na tabela COHORT, identificado pelos campos correspondentes cohort_definition_id, subject_id, cohort_start_date e cohort_end_date;
- Acrescenta aos registros da coorte co-variáveis calculadas (por exemplo idade, IMC) ou escalas compostas (por exemplo, índice de Charleson);
- A definição unificada ou recurso do atributo de coorte é capturado no atributo_de_definição_id referindo-se a um registro na tabela ATTRIBUTE_DEFINITION;

- O resultado ou valor real do atributo *cohort* (co-variável, valor de índice) é capturado nos campos value_as_number (se o valor for numérico) ou value_as_concept_id (se o valor for um conceito).

4.6. Tipos de análises

Dados observacionais têm potencial para responder a uma miríade de questões importantes na área da saúde:

- Como podemos identificar novos alvos terapêuticos de forma rápida e eficaz?
- Podemos medir o impacto relativo de diferentes intervenções de tratamento?
- Como podemos prever pacientes com um perfil de alto risco para certas doenças antes que apresentem sintomas?
- Como podemos prevenir melhor às condições crônicas?
- Quais são os melhores padrões de cuidado para gerenciar pacientes, especialmente com diferentes combinações de comorbidade?
- Como podemos melhorar o desenho dos estudos clínicos focando nos pacientes com o melhor recrutamento para efetuar o perfil de tamanho?
- Como podemos otimizar a adesão às diretrizes de tratamento e quais fatores influenciam o comportamento dos pacientes?

Cada uma destas questões define uma análise diferente:

- Perfis: linha do tempo que mostra o histórico de um paciente.
    - Quais são todos os eventos associados a um paciente específico ao longo do tempo?
- Estimativas:
    - Qual é o efeito do tratamento A no desfecho X?
- Predições:
    - Em uma população em risco, quais pacientes terão um determinado desfecho?
- Taxas de incidência: proporção e taxa das contagens brutas de pacientes, casos e tempo de risco para determinado evento.
    - Quantos novos desfechos são esperados por um intervalo de tempo?
- Caracterização de populações: geração de estatísticas descritivas da coorte a partir de dados de covariáveis de nível de pessoa.
    - Como podemos melhorar o desenho dos estudos clínicos focando nos pacientes com o melhor recrutamento para efetuar o perfil de tamanho?

4.6.1 Ferramentas de análises

Nesse tópico pretendemos oferecer uma abordagem expositiva da análise de dados de saúde através do conjunto de ferramentas OHDSI. Estas ferramentas auxiliam a

elaboração e análise dos diferentes tipos de estudo, facilitam a exploração dos dados e a geração de evidências. Entre elas podemos citar:

ACHILLES[28] (Caracterização Automatizada de Informações de Saúde em Grande Escala do Sistema de Exploração Longitudinal), é uma ferramenta de visualização baseado em estatísticas resumidas pré-extraídas de conjuntos de dados no formato CDM. Permite a caracterização, avaliação da qualidade e visualização de dados observacionais e fornece aos usuários uma estrutura exploratória e interativa para avaliar a demografia dos pacientes e a prevalência de todas as condições, medicamentos, procedimentos e observações armazenados no conjunto de dados. Possibilita avaliar a qualidade do banco de dados, procurando lacunas que podem significar erros de upload, fazer uma avaliação inicial se o banco de dados conterá um número suficiente de casos de interesse que valham a pena investigar mais. Exibe a prevalência da condição, quantidade de pacientes, distribuição etária, gênero e o tempo de observação. Podemos visualizar os dados a partir da seleção do banco de dados e seleção dos relatórios, que podemos enumerar a seguir:

- *Dashboard:* painel com um sumário da base de dados em análise, população/gênero, idade na primeira observação, pessoas com observações contínuas/mês;
- *Achilles Heel*: mensagens da qualidade dos dados do dados em análise;
- Person: estatística descritiva da população por ano de nascimento, gênero, raça, etnia;
- *Observation Periods*: idade na primeira observação, idade/gênero, tempo de observação, observações contínuas/ano e mês, período/pessoa;
- *Data Density*: total de registros (era de condição, ocorrência da condição, era de drogas, exposição a drogas, período de observação, ocorrência de procedimentos e de visitas) por ano, por pessoa/ano, por conceito/tipo;
- *Conditions/conditions era:* prevalência das condições por número de pessoas e registro por pessoas e tempo (mapa e tabela);
- *Observations*: prevalência das observações por número de pessoas e registro por pessoas (mapa e tabela);
- *Drug eras*: prevalência das drogas por número de pessoas e tempo (mapa e tabela);
- *Drug Exposures*: prevalência de exposição às drogas por número de pessoas e registro por pessoas (mapa e tabela);
- Procedures: prevalência de procedimentos por número de pessoas e registro por pessoas (mapa e tabela);
- *Visits*: prevalência por tipo de visita (internação, consulta, emergência) por número de pessoas e registro por pessoas (mapa e tabela);
- *Death*: prevalência de óbito por idade/gênero/ano, por mês, tipo;

[28] ACHILLES http://www.ohdsi.org/web/achilles

ATHENA[29] Camada intermediária do aplicativo da web para distribuição e navegação nos vocabulários padronizados para o CDM. Site de download dos vocabulários padrão. A ferramenta permite a visualização dos vocabulários, a partir da seleção do domínio (dados demográficos, condição, droga, procedimento, medição, observação, visita, óbito, etc.), os conceitos (padrão, classificação ou não-padrão), a classe (lista das classificações usadas para diferenciar conceitos dentro de um determinado vocabulário. Exemplos: achado clínico, ingredientes, procedimentos, etc.) e o vocabulário (ICD9, ICD10, SNOMED, etc.). Realiza o download dos vocabulários selecionados para serem importados em seu ambiente de construção do CDM. A ferramenta também permite explorar o vocabulário antes de baixá-lo, apresenta os mapeamentos ou um código específico e com qual conceito padrão está associado.

Todos os mapeamentos estão disponíveis na tabela Concept_relationship (que pode ser baixada do ATHENA). Cada valor em uma terminologia de fonte suportada recebe um Concept_id (que é considerado não padrão). Cada Source_concept_id terá um mapeamento para um Standard_concept_id. Por exemplo:

Neste caso, o conceito SNOMED padrão 201826 para diabetes mellitus tipo 2 seria armazenado na tabela Condition_occurrence, pois Condition_concept_id e o conceito ICD10CM 1567956, para diabetes mellitus tipo 2, seriam armazenados como Condition_source_concept_id.

ATLAS[30] é uma das ferramenta publicamente disponível para pesquisadores conduzir análises científicas em dados observacionais padronizados convertidos para o OMOP CDM V5[31]. Permite criar coortes definindo grupos de pessoas com base em uma exposição a um medicamento ou diagnóstico de uma condição específica usando dados de registros de assistência médica. Os perfis dos pacientes podem ser visualizados dentro de uma coorte específica e análises de estimativa de nível populacional permitem a comparação de duas coortes diferentes. Apresenta as seguintes funcionalidades:

- Fontes de dados: opção de seleção da base de dados a ser analisada. Disponibiliza os gráficos de análise da ferramenta Achilles;
- Vocabulário: seleção e importação dos vocabulários necessários para análise dos dados;
- Conjuntos de conceitos: permite definir um novo conceito e listar/exportar;
- Definição de coortes: preparação das coortes para estudo. Define o grupo de pessoas que satisfazem um ou mais critérios de inclusão por um período de tempo. Como consequência desta definição:
    - uma pessoa pode pertencer a múltiplas coortes,
    - pode pertencer a mesma coorte em diferentes períodos de tempo,
    - uma pessoa não pode pertencer mais de uma vez à mesma coorte durante o mesmo intervalo de tempo,

[29] ATHENA http://athena.ohdsi.org
[30] ATLAS http://www.ohdsi.org/web/atlas/#/home
[31] OMOP CDM v5 http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:single-page

    - uma coorte pode ter zero ou mais membros;
- Caracterização de coortes: é definida como o processo de geração de estatísticas descritivas da coorte a partir de dados de covariáveis a nível de pessoa. As estatísticas resumidas dessas covariáveis podem ser contagem, média, sd, var, min, max, mediana, intervalo e quantis. Além disso, as covariáveis durante um período podem ser estratificadas em unidades de tempo para análises de séries temporais, como intervalos fixos de tempo relativos a data de início da coorte (por exemplo, a cada 7 dias, a cada 30 dias, etc.) ou em intervalos absolutos do calendário, como semana, mês, trimestre, ano;
- Caminho da coorte: é definido como o processo de gerar uma sequência agregada de transições entre as coortes de eventos e as pessoas nas coortes alvo;
    - Coortes Alvo: cada uma das coortes-alvo serão analisadas em relação às coortes de eventos;
    - Coortes de Eventos: cada coorte de eventos define o passo em um caminho que pode ocorrer para uma pessoa na coorte de tratamento;
- Taxas de incidência: as taxas de incidência podem ser geradas incluindo as coortes de meta e resultados. A taxa de incidência é reportada como uma proporção e uma taxa. São fornecidas as contagens brutas de pessoas, casos e tempo em risco;
    - Tempo em risco: define a janela de tempo relativa à data de início ou término da coorte com um deslocamento para considerar a pessoa 'em risco' do desfecho em análise;
    - Critérios de estratificação: fornecer critérios de estratificação opcionais para a análise que dividirá a população em grupos únicos em relação aos critérios definidos;
- Perfis: o Atlas fornece a capacidade de pesquisar e explorar perfis individuais de pacientes em um banco de dados. Essa funcionalidade pode ser acessada clicando no item de menu de perfis, selecionando o banco de dados de interesse e inserindo um número de identificação do paciente. Dentro do perfil apresentado, o menu à esquerda lista os registros individuais que podem ser condições, medicamentos, procedimentos, etc. A tabela no canto inferior direito lista os domínios individuais, IDs de conceitos, nomes de conceitos e dias de início e fim no registro de um determinado paciente. O gráfico pode ser redimensionado para aumentar o zoom em uma determinada janela de tempo, por exemplo, nos primeiros 100 dias. Alterando a janela de tempo de interesse, as tabelas a esquerda e inferior direita mudam para refletir a janela de tempo de interesse;
- Estimativa: o Atlas tem a capacidade de realizar estudos de estimativa usando o design de coorte comparativo. O procedimento de estimativa no Atlas usa a

metodologia de escore de propensão[32]. Existem 3 principais modelos de resultados: regressão logística; regressão de Poisson; e riscos proporcionais de Cox;

- Predição: o Atlas incorporou a capacidade de gerar modelos de predição usando métodos de aprendizado de máquina para medicina de precisão e interceptação de doenças, incluindo:
    - *Regularized regression*
    - *Random forest*
    - *k-nearest neighbors*

Usa um conjunto de covariáveis, incluindo, por exemplo, todas as drogas, diagnósticos, procedimentos, bem como idade, índices de comorbidade, etc.
Os modelos de resultados suportados são logísticos, Poisson e sobrevivência (tempo até o evento).

4.7. Considerações finais e conclusões

A iniciativa OHDSI surge como resposta a necessidade de aproveitar o enorme volume de informações disponibilizados pelos sistemas de saúde informatizados e da percepção da carência de integrar estas informações de forma confiável e transparente para poder ter uma pesquisa de qualidade e reprodutível.

A força estatística que traz este volume de dados se contrapõe com a dificuldade de integrar informações que utilizam vocabulários diferentes ou não padronizados para poder ter estudos comparativos realmente efetivos. Isso fomenta a discussão a respeito da importância de passar a utilizar vocabulários padronizados para todas as atividades suscetíveis de informatização a risco de perder a possibilidade de fazer estudos com relevância internacional que permitam o acesso a publicações de alto impacto.

No Brasil, as mais importantes iniciativas estão hoje orientadas a estabelecer vocabulários padronizados. Ao ver o impacto que isto têm na elaboração de uma ferramenta de pesquisa, podemos apreciar a enorme importância deste esforço.

Não precisamos reinventar a pesquisa observacional, as ferramentas já estão disponíveis, em código aberto, de forma gratuita, simples de instalar em ambientes fechados ou prontas na nuvem. Precisamos sim, estudá-las e integrar as bases disponíveis ao modelo OMOP, passando a participar dos centros de pesquisa do mundo, somando esforços com a comunidade mundial, melhorando o que já existe, divulgando o conhecimento e aproveitando toda esta infraestrutura para o ensino da epidemiologia aplicada no mundo real.

[32] Um escore de propensão é a probabilidade de uma unidade (por exemplo, pessoa) ser designada para um tratamento particular, dado um conjunto de covariáveis observadas. Os escores de propensão são usados para reduzir o viés de seleção ao equacionar grupos com base nessas covariáveis.

4.8. Glossário de Termos

Termo - Descrição

Ancestor - O conceito de nível superior em um relacionamento hierárquico. Note que os ancestrais e descendentes podem estar muitos níveis separados uns dos outros.

Ambulatory Payment Classification (APC) - O APC é usado como um método de pagamento de serviços ambulatoriais para o programa *Medicare*, que é análogo aos DRGs para serviços de internação.

Average Wholesale Price (AWP) - Os fabricantes de preços estabelecidos para medicamentos controlados devem ser comprados no atacado para as farmácias e prestadores de serviços de saúde.

Anatomical Therapeutic Chemical (ATC) - é a sigla para a classificação Anatômica Terapêutico Química, que, em conjunto com a Dose Diária Definida - DDD (*Defined Daily Dose*), forma o sistema ATC/DDD, que, desde de 1996, passou a ser reconhecido pela Organização Mundial de Saúde como padrão internacional para os estudos de utilização de drogas. No sistema de classificação ATC, as drogas são divididas em diferentes grupos, de acordo com o órgão ou sistema no qual eles atuam e suas propriedades químicas, farmacológicas e terapêuticas. As drogas são divididas em cinco níveis diferentes, sendo o primeiro dividido em quatorze grupos principais, com um subgrupo farmacológico/terapêutico (segundo nível). Os terceiro e quarto níveis correspondem a subgrupos químicos/farmacológicos/terapêuticos, e o quinto nível, à substância química.

Centers for Disease Control and Prevention (CDC) - Os Centros de Controle e Prevenção de Doenças é uma agência federal dos Estados Unidos sob o Departamento de Saúde e Serviços Humanos. Ele trabalha para proteger a saúde pública e a segurança, fornecendo informações para melhorar as decisões de saúde.

Common Data Model (CDM) - O CDM pretende facilitar a análise observacional de diferentes bases de dados de saúde. O CDM define estruturas de tabela para cada uma das entidades de dados (por exemplo, Pessoas, Ocorrência de Visita, Exposição a Medicamentos, Ocorrência de Condição, Observação, Ocorrência de Procedimentos, etc.). Inclui elementos de dados observacionais que são relevantes para identificar a exposição a vários tratamentos e definir a ocorrência da condição. O CDM inclui os vocabulários padronizados de termos e as tabelas de domínio da entidade.

Concept - Um conceito é a unidade básica de informação. Os conceitos podem ser agrupados em um determinado domínio. Um conceito é um termo único que possui um identificador / nome único e estático, pertence a um domínio e pode existir em relação a

outros conceitos. Os relacionamentos verticais consistem em instruções "é um" que formam uma hierarquia lógica. Em geral, os conceitos acima de um determinado conceito são referidos como ancestrais e os abaixo como descendentes.

Conceptual Data Model - Um modelo de dados conceituais é um mapa de conceitos e seus relacionamentos. Isso descreve a semântica de uma organização e representa uma série de afirmações sobre sua natureza. Especificamente, descreve as coisas importantes para uma organização (classes de entidade), sobre as quais ela está inclinada a coletar informações e características de (atributos) e associações entre pares dessas coisas de significância (relacionamentos).

Current Procedural Terminology, 4th edition (CPT-4) - Uma terminologia que é mantida pela *American Medical Association* (AMA). Ele é usado por hospitais para pacientes ambulatoriais do *Medicare* e por médicos para serviços ambulatoriais.

Data mapping - São os mapeamentos de elementos de dados entre dois modelos de dados, terminologias ou conceitos distintos. O mapeamento de dados é o processo de criação de mapeamentos de elementos de dados entre dois modelos de dados distintos. O mapeamento de dados é usado como primeiro passo para uma ampla variedade de tarefas de integração de dados.

Demographics - A demografia refere-se a características selecionadas de pessoas. Os dados demográficos podem incluir dados como raça, idade, sexo, data de nascimento, local etc.

Descendant - O conceito de nível inferior em um relacionamento hierárquico. Note que os ancestrais e descendentes podem estar muitos níveis separados uns dos outros.

Design Principle - Um arranjo organizado de um ou mais elementos ou princípios para um propósito. Ele identifica os principais princípios e as melhores práticas para ajudar os desenvolvedores a produzir software. Entender completamente as metas das partes interessadas e projetar sistemas com essas metas em mente são as melhores abordagens para entregar resultados com êxito.

Diagnosis- Related Group (DRG) - Os DRG são usados como um método de pagamento de serviços de internação para o programa *Medicare*, que é análogo às APCs para serviços ambulatoriais.

Electronic Health Record (EHR) - Registro eletrônico de saúde refere-se ao prontuário de uma pessoa individual em formato digital. Pode ser composto de registros médicos eletrônicos de vários locais e / ou fontes. O EHR é um registro eletrônico longitudinal de informações de saúde de pessoas geradas por um ou mais encontros em qualquer

ambiente de prestação de cuidados. Incluídos nestas informações estão a demografia pessoal, notas de progresso, problemas, medicamentos, sinais vitais, histórico médico, imunizações, dados laboratoriais e relatórios de radiologia.

Electronic Medical Record (EMR) - Um prontuário eletrônico é um registro médico computadorizado criado em uma organização que presta atendimento, como um hospital ou ambulatório. Registros médicos eletrônicos tendem a fazer parte de um sistema de informações de saúde local independente que permite o armazenamento, a recuperação e a manipulação de registros. Este documento fará referência ao EHR, mesmo que uma fonte de dados específica possa usar internamente a definição do EMR.

Extract Transform Load (ETL) - Processo de obtenção de dados de um armazenamento de dados (*Extract*), modificando-o (*Transform*) e inserindo-o em um armazenamento de dados diferente (*Load*).

Generic Product Identifier (GPI) - Um identificador exclusivo patenteado para um medicamento usado pelo banco de dados de formulários comerciais Medi-Span.

Healthcare Common Procedure Coding System (HCPCS) - Os códigos de nível I do HCPCS são gerenciados pela AMA (*American Medical Association*). Os códigos de nível II do HCPCS são gerenciados pelo CMS (*Centers for Medicare & Medicaid Services*). Os códigos de Nível II incluem: procedimento alfanumérico do HCPCS e códigos modificadores, suas descrições e dados administrativos, de cobertura e de preços aplicáveis do *Medicare*. Esses códigos são usados para serviços ambulatoriais do *Medicare*.

Health Insurance Portability and Accountability Act (HIPAA) - Uma lei federal que foi concebida para permitir a portabilidade do seguro de saúde entre empregos. Além disso, exigiu a criação de uma lei federal para proteger informações de saúde pessoalmente identificáveis; se isso não ocorresse em uma data específica (o que não acontecia), a HIPAA orientou o *Department of Health and Human Services* (DHHS) a emitir regulamentações federais com o mesmo objetivo. O DHHS emitiu os regulamentos de privacidade do HIPAA (a Regra de Privacidade do HIPAA), bem como outros regulamentos no âmbito do HIPAA.

Health Level Seven (HL7) - A HL7 é uma organização global sem fins lucrativos dedicada ao fornecimento de uma estrutura abrangente e padrões relacionados para o intercâmbio, integração, compartilhamento e recuperação de informações eletrônicas de saúde que dão suporte à prática clínica e ao gerenciamento, entrega e avaliação. dos serviços de saúde. As especificações do HL7 baseiam-se principalmente em códigos e vocabulários de uma variedade de fontes.

Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (InCor-HCFMUSP) - O InCor é um hospital público universitário de alta complexidade, especializado em cardiologia, pneumologia e cirurgias cardíaca e torácica. Além de ser um polo de atendimento - desde a prevenção até o tratamento -, o Instituto do Coração também se destaca como um grande centro de pesquisa e ensino. O InCor é parte do Hospital das Clínicas e campo de ensino e de pesquisa para a Faculdade de Medicina da USP.

International Classification of Disease, 9th Revision, Clinical Modifications (ICD-9-CM Or ICD-9) - O sistema oficial de atribuição de códigos para diagnósticos e procedimentos associados à utilização hospitalar nos Estados Unidos.

Logical Data Model - Modelos de dados lógicos são representações gráficas dos requisitos de negócios. Eles descrevem as coisas importantes para uma organização e como elas se relacionam umas com as outras, além de definições e exemplos de negócios. O modelo de dados lógicos pode ser validado e aprovado por um representante comercial e pode ser a base do design do banco de dados físico.

Logical Observation Identifiers Names and Codes (LOINC) - Nomes e identificadores universais de códigos para a terminologia médica relacionada ao Registro de Saúde Eletrônico e auxilia na troca eletrônica e coleta de resultados clínicos (como testes de laboratório, observações clínicas, gerenciamento de resultados e pesquisa).

Medical Dictionary for Regulatory Activities (MedDRA) - MedDRA é uma terminologia médica internacional clinicamente validada, usada pelas autoridades reguladoras e pela indústria biofarmacêutica regulamentada. A terminologia é usada durante todo o processo de regulamentação, desde a pré-comercialização até o pós-marketing, e para entrada, recuperação, avaliação e apresentação de dados.

National Drug Codes (NDC) - Identificadores exclusivos atribuídos a medicamentos individuais. Os NDCs são usados principalmente como um código de inventário e para prescrições.

National Drug File - Reference Terminology (NDF-RT) - Uma terminologia de referência de medicamento não-proprietário que inclui o conhecimento de drogas e classifica as drogas, principalmente por mecanismo de ação e efeito fisiológico.

Primary Care Provider (PCP) - Um prestador de cuidados de saúde designado como responsável pela prestação de cuidados médicos gerais a um doente, incluindo avaliação e tratamento, bem como o encaminhamento para especialistas.

Protected Health Information (PHI) - Informações de saúde protegidas sob HIPAA incluem qualquer informação de saúde identificável individualmente. Identificável refere-se não apenas aos dados explicitamente vinculados a um indivíduo em particular (essa é a informação identificada). Ele também inclui informações de saúde com itens de dados que poderiam ser esperados para permitir a identificação individual. A informação não identificada é aquela a partir da qual todas as informações potencialmente identificáveis foram removidas.

Read Codes - Os *Read Codes* foram desenvolvidos pelo Dr. James Read. Eles contém centenas de milhares de termos, sinônimos e abreviações abrangendo todos os aspectos do atendimento ao paciente, incluindo sinais e sintomas, tratamentos e terapias, investigações, ocupações, diagnósticos, medicamentos e dispositivos médicos. O U.K. *National Health Service Centre for Coding and Classification* (NHS CCC) adquiriu e mantém atualmente os *Read Codes*, agora conhecidos como *Clinical Terms*.

RxNorm - Uma nomenclatura padronizada para drogas clínicas e dispositivos de dispensa de medicamentos, produzida pela *National Library of Medicine*. Em RxNorm, o nome de um medicamento clínico descreve seus ingredientes, pontos fortes e / ou forma de apresentação. O RxNorm fornece nomes normalizados para medicamentos clínicos e vincula seus nomes a muitos dos vocabulários de drogas comumente usados no gerenciamento de farmácia e no software de interação medicamentosa, incluindo os do *First Data Bank, Micromedix, MediSpan, Gold Standard Alchemy e Multum*. Ao fornecer links entre esses vocabulários, o RxNorm pode mediar mensagens entre sistemas que não usam o mesmo software e vocabulário.

SNOMED CT - *Systematized Nomenclature of Medicine - Clinical Terms*: Nomenclatura Sistematizada de Medicina - Termos Clínicos. A versão 3 dos *Clinical Terms* (CTV3) (*Read Codes*) foi incorporada na *Systematized Nomenclature of Medicine - Reference Terminology* (SNOMED RT), resultando na criação do SNOMED - *Clinical Terms* (SNOMED CT). O SNOMED-CT é um de um conjunto de padrões designados para uso em sistemas do Governo Federal dos EUA para o intercâmbio eletrônico de informações clínicas de saúde e também é um padrão exigido em especificações de interoperabilidade do *US Healthcare Information Technology Standards Panel*. Esta terminologia está sendo implementada internacionalmente como padrão dentro de outros países Membros do IHTSDO - *International Health Terminology Standards Development Organisation*.

Referências consultadas

Abrahão, Maria Tereza Fernandes. Método de extração de coortes em bases de dados assistenciais para estudos da doença cardiovascular [tese]. São Paulo:, Faculdade de Medicina; 2016 [citado 2019-04-22] doi:10.11606/T.5.2016.tde-04082016-160129.

Abrahão MTF, Nobre MRC, Gutierrez MA. A method for cohort selection of cardiovascular disease records from an electronic health record system. International Journal of Medical Informatics [Internet] Volume 102, June 2017, Pages 138-149 https://doi.org/10.1016/j.ijmedinf.2017.03.015

Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017, 48-57. Published 2017 Jul 26 Overhage, J. Marc & B Ryan, Patrick & G Reich, Christian & Hartzema, Abraham & E Stang, Paul. (2011). Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association : JAMIA. 19. 54-60. 10.1136/amiajnl-2011-000376.

Ferramenta estatística R https://www.r-project.org/ Acesso: 17/01/2019

Hripcsak G, Duke JD, Shah NH., et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in health technology and informatics. vol. 216 (2015): 574-578. MEDINFO'15; August 19–23, 2015; São Paulo, Brazil.

Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the Impact of Database Heterogeneity on Observational Study Results. American Journal of Epidemiology [Internet]. 2013 Aug 15 [cited 2015 Nov 2];178(4):645–51. http://aje.oxfordjournals.org/cgi/doi/10.1093/aje/kwt010

Madigan D, Ryan P. Commentary: What Can We Really Learn From Observational Studies?: The Need for Empirical Assessment of Methodology for Active Drug Safety Surveillance and Comparative Effectiveness Research. Epidemiology [Internet]. 2011 Sep [cited 2015 Nov 2];22(5):629–31. http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001648-201109000-00006

Madigan D, Stang PE, Berlin JA, Schuemie M, Overhage JM, Suchard MA, et al. A Systematic Statistical Approach to Evaluating Evidence from Observational Studies. Annual Review of Statistics and Its Application [Internet]. 2014 Jan 3 [cited 2019 Apr 20]; 1(1):11–39. Available from: http://www.annualreviews.org/doi/10.1146/annurev-statistics-022513-115645

OHDSI-Vocabulary-CDM-Tutorial-2018-V1.pdf https://www.ohdsi-europe.org/images/symposium-2018/tutorials/OHDSI Vocabulary-CDM-Tutorial-2018-V1.pdf Acesso: 17/01/2019

OHDSI https://www.ohdsi.org/ Acesso: 17/01/2019

Vashisht R, Jung K, Schuler A, Banda JM, Park RW, Jin S, Li L, Dudley JT, Johnson KW, Shervey MM, Xu H, Wu Y, Natrajan K, Hripcsak G, Jin P, Van Zandt M, Reckard A, Reich CG, Weaver J, Schuemie MJ, Ryan PB, Callahan A, Shah NH. Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative. JAMA Network Open 2018, 1(4), pp.e181755-e181755.

Yoon D, Ahn EK, Park MY, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research. Healthc Inform Res. 2016;22(1):54-8.