

Capítulo

6

Fenótipos no contexto da pesquisa observacional:

OHDSI Phenotype Phebruary 2022

Maria Tereza Fernandes Abrahão (HIAE, OHDSI Latam), Pablo Jorge Madril (OHDSI Latam)

Abstract

The Observational Health Data Sciences and Informatics (OHDSI) initiative has redefined the field of observational research into real-world healthcare evidence. This initiative standardizes the structure, content, and analytics by adopting a common data model (CDM-OMOP), standard vocabularies, and routines to generate reliable evidence. Phenotypes are the fundamental elements of the analysis and form a bridge with the standardized data. The reliability of the evidence generated derives from the quality of the phenotypes used in the definition of population cohorts. OHDSI held the Phenotype Phebruary 2022 event, enabling the construction of 28 phenotypes in 28 days. The objective of this chapter is to present the results of this collaborative network, the concepts involved, its influence on the generation of cohorts, and the process of result validation.

Resumo

A iniciativa Observational Health Data Sciences and Informatics (OHDSI) redefiniu o campo das pesquisas observacionais em dados de saúde do mundo real. Essa iniciativa prioriza um modelo de dados comum (CDM-OMOP), utilizando vocabulário e rotinas padronizadas para a geração de evidências confiáveis. Os fenótipos são os elementos fundamentais das análises e a ponte com os dados padronizados. A confiabilidade da evidência gerada deriva da qualidade dos fenótipos utilizados na definição das coortes populacionais. A OHDSI realizou o evento Phenotype Phebruary 2022, possibilitando a construção de 28 fenótipos em 28 dias. O objetivo deste capítulo é apresentar os resultados desta rede colaborativa, os conceitos envolvidos, sua influência na geração das coortes e o processo de validação de resultados.

5.1 Introdução

Por mais de uma década, muitos defenderam a visão de um sistema de saúde ser “projetado para gerar e aplicar as melhores evidências para as escolhas colaborativas de saúde de cada paciente e provedor; para conduzir o processo de descoberta como uma consequência natural do atendimento ao paciente; e garantir inovação, qualidade, segurança e valor na área da saúde” [Olsen 2007]. Um componente principal dessa visão repousa na perspectiva de que os dados dos pacientes capturados durante a rotina do atendimento clínico, possam ser analisados para produzir evidências do mundo real que possam ser disseminadas por todo o sistema de saúde para informar a prática médica.

Em 2007, o *Institute of Medicine Roundtable on Evidence-Based Medicine*, emitiu um relatório que estabeleceu a meta que “Até o ano de 2020, 90% das decisões clínicas serão apoiadas por informações clínicas precisas, oportunas e atualizadas, e irá refletir a melhor evidência disponível” [Olsen 2007]. Embora várias frentes tenham obtido significativo progresso, ainda não estamos lá. Por quê? Em parte, porque a jornada dos dados no nível do paciente para evidências confiáveis é árdua. Não há um caminho único definido dos dados à evidência, e nenhum mapa que possa ajudar a navegar ao longo do caminho. Na verdade, não existe uma noção única de "dados", nem uma noção única de "evidência".

A *Observational Health Data Sciences and Informatics* (OHDSI) surgiu da na continuação do OMOP (*Observational Medical Outcomes Partnership*), que foi uma parceria público-privada estabelecida nos EUA para informar o uso apropriado de bancos de dados observacionais de saúde para estudar os efeitos de produtos médicos. O projeto de cinco anos desenvolveu novos métodos de pesquisa observacional e estabeleceu um laboratório de pesquisa observacional. Na conclusão deste projeto de cinco anos, os investigadores de pesquisa OMOP iniciaram o esforço OHDSI, e o laboratório de pesquisa mudou-se para a Fundação *Reagan-Udall* sob o Programa de Inovação em Desenvolvimento e Vigilância de Evidências Médicas¹.

Uma peça central do projeto OMOP foi o desenvolvimento de um modelo de dados padronizados (*Common Data Model*) CDM-OMOP, que representa dados de saúde de diversas fontes de forma consistente e padronizada. CDM é um modelo de informação no qual a codificação e as relações entre os conceitos são especificadas de forma explícita e formal. A equipe OHDSI adotou e deu continuidade à manutenção deste modelo e seus serviços de vocabulário associados. A abordagem geral do OHDSI é criar uma rede aberta de detentores de dados observacionais configurados para o CDM-OMOP. Essa abordagem cria a oportunidade de implementar uma série de ferramentas existentes de exploração de dados e geração de evidências e participar de estudos mundiais e multicêntricos, pois qualquer consulta pode ser executada em qualquer local sem modificação.

¹ INEDS <https://reaganudall.org/programs/research>

Porém, para alcançarmos a geração de evidências em larga escala, temos que começar construindo uma base sólida. Os fenótipos são o elemento fundamental em todas as análises com dados do mundo real. São a ponte entre os dados padronizados do modelo comum (CDM-OMOP) e as análises padronizadas.

E como definir fenótipo? Segundo Hripcsak [Hripcsak 2018], o termo fenótipo pode ser aplicado às características do paciente inferidas a partir dos dados dos registros de prontuários eletrônicos (RES). Levando em conta a observação, estamos falando sobre algo que é observável em nossos dados observacionais, que inclui a noção de tempo na especificação do fenótipo, já que o estado de uma pessoa pode mudar. Os fenótipos são utilizados nas definições de coortes, um conjunto de pessoas que satisfazem um ou mais critérios de inclusão por um período, e em todos os estudos na rede OHDSI.

Não é possível falar em evidências de qualidade sem falar em qualidade de dados e nos principais atributos que definem a mesma, bem como, as formas de avaliá-la. Descrevemos aqui a definição, o processo de geração e avaliação dos fenótipos, apresentando as diversas ferramentas que a OHDSI disponibiliza para sustentar estes processos.

O capítulo está estruturado como se segue. Primeiro, a seção 5.1 oferece uma breve fundamentação sobre a iniciativa OHDSI, seus componentes e processos de geração de evidências. A seção 5.2 apresenta as definições de fenótipos e a seção 5.3 a definição de coortes. A seção 5.4 a criação de coortes na ferramenta Atlas, a seção 5.5 apresenta o PheValuator para a avaliação de fenótipos. A seção 5.6, os diagnósticos de coortes utilizando a ferramenta de Análise de Coorte, a seção 5.7 apresenta um resumo do evento Phenotype Phebruary 2022 e a biblioteca de fenótipos. A seção 5.8 apresenta as considerações finais e conclusões e a seção 5.9 as referências consultadas.

Na elaboração deste capítulo, seguiu-se o livro da OHDSI, *The book of OHDSI*, de domínio público, sob a licença *Creative Commons Zero v1.0 Universal*, (16/04/2020). O livro é um documento vivo, mantido pela comunidade por meio de ferramentas de desenvolvimento de código aberto e evolui continuamente. A versão online, disponível gratuitamente², sempre representa a versão mais recente. O texto do livro foi traduzido e complementado na escrita deste capítulo. As figuras e tabelas que ilustram os estudos, foram adaptadas do livro e de apresentações dos tutoriais da OHDSI. Os exemplos foram elaborados a partir dos conhecimentos adquiridos pelos autores na participação em eventos OHDSI.

5.1.1 Uma breve história da OHDSI

Fundada em 2014, a iniciativa OHDSI, redefiniu 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 (*Common Data Model* - CDM-OMOP), de mecanismos para

² The book OHDSI <http://book.ohdsi.org>

tratamento de diferentes vocabulários, e da disponibilidade de um conjunto de ferramentas de software livre para análise dos mesmos. Esta iniciativa pode ser vista como uma comunidade de ciência aberta que busca ativamente a interoperabilidade e a reprodutibilidade na geração de evidências médicas. A OHDSI estabeleceu uma rede internacional de pesquisadores e bancos de dados observacionais de saúde com um centro de coordenação localizado na Universidade de Columbia. Atualmente, conta com centenas de pesquisadores em mais de 30 países e registros de saúde de cerca de 600 milhões de pacientes únicos de todo o mundo. Busca melhorar a saúde capacitando a comunidade para gerar de forma colaborativa e sistemática, evidências em larga escala que promovam melhores decisões e cuidados de saúde [Abrahão 2019].

Desde o início da comunidade OHDSI, o objetivo era estabelecer uma colaboração internacional com base em valores de ciência aberta, como o uso de software de código livre, disponibilidade pública de todos os materiais e geração e publicações de evidências médicas.

O termo 'ciência aberta' é usado desde os anos 90, mas ganhou força na década de 2010, no mesmo período de início do OHDSI. É importante salientar que a ciência aberta é uma mudança de cultura que ocorre mundialmente, sendo recomendada pela UNESCO [UNESCO 2021], a qual define Ciência Aberta como: "uma construção inclusiva que combina vários movimentos e práticas com o objetivo de tornar o conhecimento científico multilíngue disponível abertamente, acessível e reutilizável para todos, para aumentar as colaborações científicas e o compartilhamento de informações para o benefício da ciência e da sociedade, e para abrir os processos de criação de conhecimento científico, avaliação e comunicação aos atores sociais além da comunidade científica tradicional. Inclui todas as disciplinas científicas e aspectos de práticas acadêmicas, incluindo ciências básicas e aplicadas, ciências naturais e sociais e humanidades, e se baseia nos seguintes pilares principais: conhecimento científico aberto, infraestruturas de ciência aberta, comunicação científica, engajamento aberto de atores sociais e diálogo aberto com outros sistemas de conhecimento."

A adesão dos pesquisadores às práticas de Ciência Aberta torna a pesquisa mais eficiente, confiável e criativa.

A partir de março de 2021 foi criado um grupo de trabalho, OHDSI LATAM, para reunir os países latino-americanos no esforço de divulgação da iniciativa OHDSI, sendo os autores representantes pelo Brasil nas reuniões mensais. Neste grupo, além de divulgar o estado da arte em pesquisa observacional em cada um dos países, se discutem as necessidades dos membros em relação a vocabulários, mecanismos para geração e preenchimento do modelo CDM-OMOP e propostas de estudos de interesse comum.

5.1.2 Componentes da OHDSI

A OHDSI disponibiliza uma ampla gama de ferramentas de código aberto para dar suporte a vários delineamentos de projetos em análise de dados observacionais no nível do paciente. O que essas ferramentas têm em comum é que todas podem interagir com

um ou mais bancos de dados estruturados no modelo CDM-OMOP. Além disso, essas ferramentas padronizam as análises para vários casos de uso facilitando a execução e melhorando a reprodutibilidade e a transparência. Por exemplo, existem muitas maneiras de calcular uma taxa de incidência. As ferramentas OHDSI permitem especificá-las com algumas opções, e qualquer pessoa que fizer essas mesmas escolhas calcula as taxas de incidência da mesma maneira.

Os componentes:

- Modelo comum de dados: CDM-OMOP
- Vocabulários Padronizados
- Ferramentas para preparação do CDM, bibliotecas para análises estatísticas e definição de estudos

O conjunto de ferramentas disponibilizadas pela OHDSI auxiliam a preparação da base no modelo CDM-OMOP, na validação do processo de mapeamento, na verificação da qualidade dos dados que compõem o CDM, na elaboração e análise dos diferentes tipos de estudos, facilitando a exploração dos dados e a geração de evidências. Podemos citar:

- Ferramentas para geração do banco CDM-OMOP
 - o ETL: White Rabbit, Rabbit-In-A-Hat³
 - o Vocabulários: Athena⁴/Usagi⁵
 - o Qualidade: Achilles⁶ e Data Quality Dashboard (DQD)⁷
- Ferramentas de análise
 - o Geração de Coortes/estudos - Atlas⁸
 - o Análises estatísticas: HADES⁹

5.1.3 O caminho da geração de evidências

Se pensarmos em um processo padronizado para disseminação e geração de evidências, poderíamos pensar sobre os seguintes passos:

- Em primeiro lugar, temos uma pergunta?

Precisamos descobrir o quanto a pergunta clínica é importante para os pacientes, e a partir deste ponto, fazer uma revisão do que já se conhece e para o qual existe evidência.

- Esta pergunta já foi formulada anteriormente?
- Quais são as deficiências das respostas conhecidas?

3 White Rabbit, Rabbit-In-A-Hat <http://ohdsi.github.io/WhiteRabbit/index.html>

4 ATHENA <http://athena.ohdsi.org>

5 Usagi <http://ohdsi.github.io/Usagi/>

6 ACHILLES <http://www.ohdsi.org/web/achilles>

7 DQD <https://ohdsi.github.io/DataQualityDashboard/articles/DataQualityDashboard.html>

8 Atlas <https://github.com/OHDSI/Atlas/wiki/Atlas-Setup-Guide>

9 HADES <https://github.com/OHDSI/Hades>

Neste ponto publicamos um protocolo onde se descreve qual é o delineamento do estudo, especifica quais são todas as escolhas feitas a fim de responder à pergunta de pesquisa e o que vai ser feito para gerar evidência.

Podemos, então, seguir com a execução do estudo, o que significa implementar exatamente o que está escrito no protocolo. Se ele está escrito corretamente, a execução não deve ser mais do que acompanhar o conjunto de instruções. O próximo passo será avaliar, uma vez que o projeto do estudo foi implementado e que começamos a ter resultados.

- O quanto posso confiar nos resultados?
- Como OHDSI está ajudando neste processo?

Como comunidade queremos levantar perguntas que sejam realmente importantes. Quando falamos de revisar a evidência, temos uma ferramenta de rotulagem literária de relatórios, LAERTES¹⁰, para achar controles positivos e negativos.

Várias ferramentas sustentam o delineamento do estudo (Atlas) e a execução é feita pelas bibliotecas de métodos (HADES).

A síntese é uma das melhores oportunidades de funcionar como comunidade dado que nos permite realizar a análise não apenas na nossa base, mas em toda a rede. A Figura 5.1 apresenta o processo de criação e disseminação de evidências adotado na OHDSI.

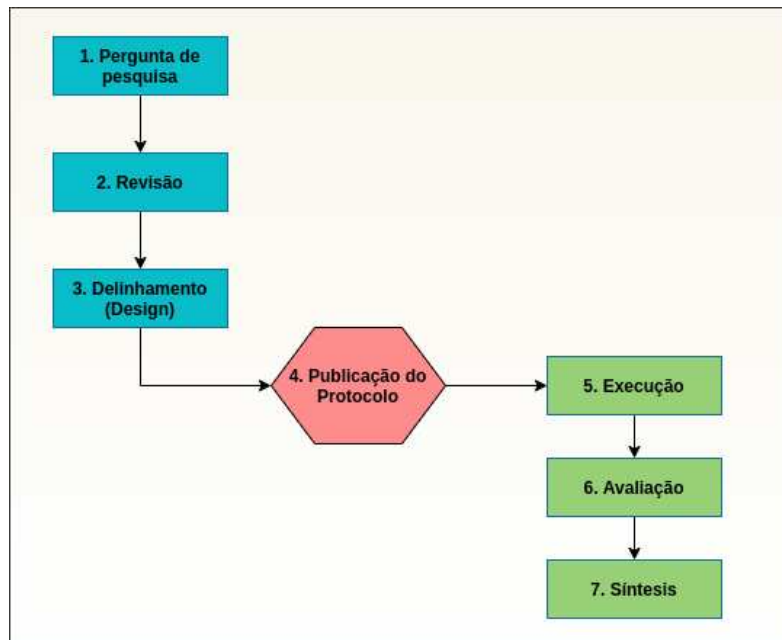


Figura 5.1 Processo padrão de criação e disseminação de evidências¹¹

¹⁰ LAERTES <https://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:laertes>

¹¹ Fonte: <https://www.ohdsi.org/wp-content/uploads/2016/09/PLP-Study-Designer-Track.pdf>

5.1.4 Resultados da OHDSI

O objetivo da Rede de Pesquisa OHDSI é reduzir a barreira para realizar pesquisas colaborativas em larga escala usando dados observacionais e gerar evidências de alta qualidade por meio de revisão por pares em todo o projeto, execução e análise dos dados do estudo. Os proprietários das bases disponibilizam os dados em troca de algoritmos e ferramentas e a capacidade de postar consultas na rede. A OHDSI disponibiliza um processo para o desenvolvimento e execução de estudos de pesquisa que inclui proposta de um colaborador, revisão pela comunidade e promoção a projeto ativo. As propostas iniciais geralmente incluem um protocolo básico, com objetivo, justificativa, população-alvo e código-fonte inicial. Uma vez postado no Fórum de Pesquisa OHDSI¹², o projeto proposto será revisado por outros colaboradores para determinar o nível de interesse, bem como para gerar sugestões de modificação no detalhamento, fenótipos alvo ou métodos analíticos. Projetos que geram interesse de vários sites e produzem um protocolo completo e códigos testados, são promovidos a um projeto ativo. Todos os membros da comunidade são convidados a executar as análises em seu conjunto de dados local e retornar os resultados (anonimizados, agregados) centralmente. As análises em vários locais contam com a padronização do CDM-OMOP, dos domínios de dados para terminologias comuns, como SNOMED para diagnósticos, RxNorm para ingredientes de medicamentos e LOINC para resultados laboratoriais. A escolha de terminologias padrão não limita quem pode participar dos estudos. A OHDSI disponibiliza ferramentas para traduzir as terminologias de uso local para o padrão. Após um período definido, a submissão dos resultados é encerrada e os dados são analisados e apresentados à comunidade. Com base nesses achados, resultam publicações e estudos de acompanhamento.

O CDM-OMOP demonstrou anteriormente que bancos de dados de diferentes fontes podem fornecer respostas muito diferentes. Uma agregação cega dos resultados pode aumentar a variância devido à heterogeneidade em vez de diminuí-la devido ao tamanho da amostra. Portanto, dependendo da questão de pesquisa e dos locais participantes, os resultados coletados podem ser simplesmente relatados sem agregação, resumidos ou, se for demonstrada homogeneidade suficiente, agregados.

A OHDSI está implementando essa colaboração internacionalmente, aproveitando a experiência do CDM-OMOP para pesquisa observacional. A existência de vários bancos de dados que usam o CDM-OMOP, demonstra que uma conversão mundial de dados clínicos é de fato viável. O sucesso no desenvolvimento de algoritmos, distribuição de software e geração de evidências aponta para um potencial futuro em que estudos observacionais informarão a prática clínica, fornecendo evidências baseadas em dados do mundo real, usando a quantidade sem precedentes de informações de pacientes disponíveis e o uso de sistemas computadorizados para processar os dados.

12 <http://forums.ohdsi.org/>

Esta jornada na busca de evidências começa com os dados brutos passando pela conversão para o CDM-OMOP, aplicação das análises e geração das evidências que são disponibilizadas livremente para a comunidade. Esta abordagem que a OHDSI como rede distribuída de centros de dados utiliza obteve sucesso no seu objetivo de gerar evidências. A Figura 5.2, já muito conhecida e disponível no livro da OHDSI, apresenta este modelo:

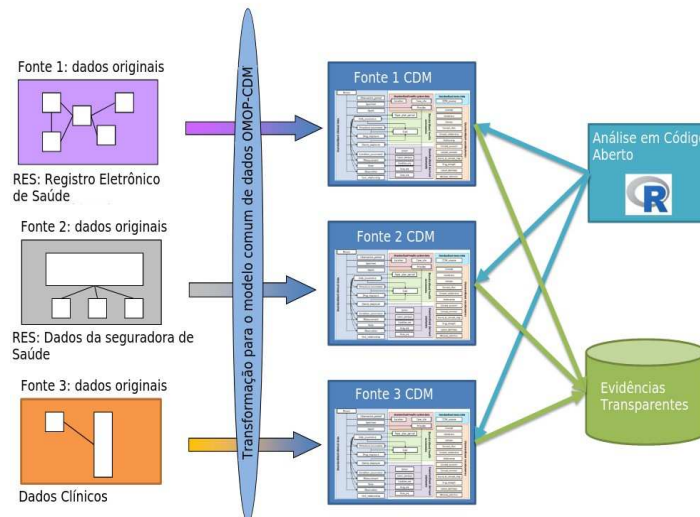
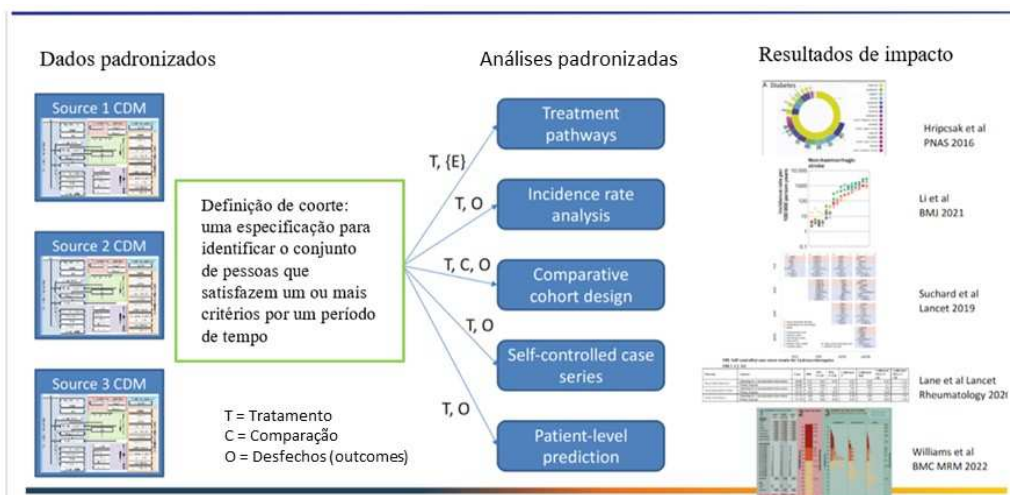


Figura 5.2 Resumo funcional¹³

Inúmeros estudos já foram implementados na rede OHDSI e estão publicados em periódicos de ampla divulgação, gerando evidências importantes na condução da prática clínica. A Figura 5.3 ilustra a jornada para a publicação de resultados de impacto internacional.



¹³https://www.ema.europa.eu/en/documents/presentation/presentation-session-2-overview-omop-common-data-model-peter-rijnbeek_en.pdf

Figura 5.3 A jornada para a evidência¹⁴

O que podemos observar na Figura 5.3, é que o modelo de abordagem da OHDSI está baseado na padronização de dados e isso possibilitou a oportunidade de padronizar as análises tendo como consequência, gerar resultados impactantes.

Desenvolvemos análises padronizadas para históricos de tratamentos, onde o artigo “*Characterizing Treatment Pathways at Scale Using the OHDSI Network*” [Hripcsak 2016] mostra a importância de um dos primeiros estudos feitos na rede OHDSI.

Desenvolvemos análises padronizadas para caracterização de taxas de incidência que resultou no trabalho “*Characterising the Background Incidence Rates of Adverse Events of Special Interest for Covid-19 Vaccines in Eight Countries: Multinational Network Cohort Study*” [Li 2021], publicado na BMJ para apoiar as vacinas para a COVID-19, calculando as taxas de antecedentes de eventos adversos. Estes são casos de sucesso no uso dos algoritmos padronizados de caracterização.

Vemos o uso de delineamentos de estudos comparativos de coorte como uma ferramenta de estimativa de efeito no nível populacional com resultados de impacto como o trabalho de Marc Suchard “*Comprehensive Comparative Effectiveness and Safety of First-Line Antihypertensive Drug Classes: A Systematic, Multinational, Large-Scale Analysis*” [Suchard 2019], publicado no Lancet, sobre hipertensão.

Temos também outros usos de estimativas de efeito a nível populacional como estudos delineados como séries de casos de autocontrole publicados na *Lancet Rheumatology* “*Risk of Hydroxychloroquine Alone and in Combination with Azithromycin in the Treatment of Rheumatoid Arthritis: A Multinational, Retrospective Study*” [Lane 2020], onde destacamos o risco da hidroxicloroquina que foi fundamental em algumas das primeiras tomadas de decisão durante a pandemia. Estes são bons exemplos de ferramentas de estimativa de nível populacional aplicadas no mundo real gerando evidências confiáveis.

No nível de previsão a nível de paciente, temos a recente publicação “*Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network*” [Williams 2022].

Todos estes resultados, tão diversos e importantes, do uso de diferentes tipos de delineamentos de análise tem algo em comum, que normalmente passa despercebido e menosprezado, que é a definição da coorte.

A definição de coorte é uma especificação para identificar um conjunto de pessoas que satisfazem um ou mais critérios por um período. E para cada uma dessas análises padronizadas que geram resultados impactantes, as entradas são as definições de coortes de alta qualidade.

Como comunidade, percebemos que este é um problema importante, mas não disponibilizamos tempo, esforço e ênfase suficiente, neste elemento fundamental. Foi isto que definiu um esforço concentrado, com a conscientização da comunidade, numa abordagem mais sistemática para a geração de evidências e na construção de um sistema

¹⁴ Fonte: <https://www.ohdsi.org/wp-content/uploads/2022/02/PhenotypePhebruary-1feb2022.pdf>

de alta qualidade, sendo os fenótipos a primeira entrada necessária. Portanto, os fenótipos são essenciais para este processo geral e, portanto, precisamos realmente colocar a mesma quantidade de rigor e energia nesta atividade.

5.1.5 Processo de criação de evidências confiáveis

As novas ferramentas que estamos desenvolvendo e a enorme disponibilidade de dados que existe atualmente, está gerando uma pressão para ter um sistema mais eficiente de geração de evidências confiáveis, que no lugar de estar gerando de um passo por vez, passe a ser um processo escalável criando passos de alta qualidade ao longo de toda a cadeia de geração, desde a validação da qualidade dos dados, a validação da qualidade dos fenótipos até a validação da confiabilidade dos resultados da análise. A Figura 5.4 apresenta graficamente este sistema. É justamente nesta conjuntura crítica do desenvolvimento de fenótipos que a comunidade de pesquisa em geral ainda está amadurecendo. Sabemos que muita da pesquisa atual está baseada em fenótipos que possuem um erro de medição substancial, mas não conseguimos ainda incorporar corretamente o erro de medição na análise. Estes são desafios que temos que superar para poder ter o sistema de geração de evidências de qualidade e capacidade que almejamos.

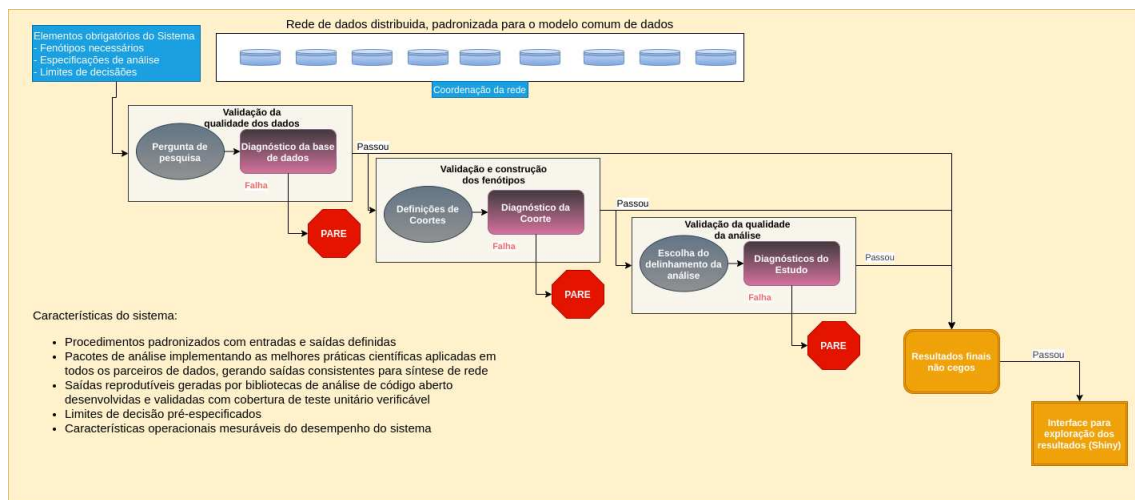


Figura 5.4 Fluxo de criação de evidências¹⁵

5.1.6 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*

¹⁵ Fonte: <https://www.ohdsi.org/wp-content/uploads/2022/02/PhenotypePhebruary-1feb2022.pdf>

Model (<https://github.com/OHDSI/CommonDataModel>), com a definição completa do modelo e as implementações para os diversos bancos suportados;

- Tutoriais e vídeos: Procure no Google por YouTube OHDSI (<https://www.google.com/search?q=youtube+ohdsi>), existe muita documentação e tutoriais em vídeo dos eventos anuais do grupo;
- Fórum: Para resolver dúvidas mais frequentes, consulte (<http://forums.ohdsi.org/>).

5.2 Definições de fenótipos

Fenótipo é o termo criado pelo pesquisador dinamarquês Wilhelm L. Johannsen [Johannsen 1909] e representa as características (parâmetros) que definem um indivíduo, sejam elas morfológicas, fisiológicas ou comportamentais.

Dentro da OHDSI se adota a seguinte definição de fenótipo, ditada por George Hripcsak [Hripcsak 2018]:

“Um fenótipo é uma especificação de um estado observável e potencialmente mutável de um organismo, distinto do genótipo, que é derivado da composição genética de um organismo. O termo fenótipo pode ser aplicado às características do paciente inferidas a partir dos dados do prontuário eletrônico (RES). O objetivo é extrair conclusões a respeito de um determinado conceito de interesse com base nos dados brutos do prontuário, ou dados de seguro saúde, ou quaisquer outros dados clínicos relevantes.”

Por exemplo, se o conceito é “paciente diabético”, quais são os dados brutos, extraídos do prontuário do paciente, que definem e suportam este conceito?

Algoritmos de fenótipos são algoritmos que podem ser criados para identificar ou caracterizar fenótipos. Podem ser gerados por especialistas da área, engenheiros de conhecimento ou através de diversas formas de aprendizado de máquina, como aprendizado ativo, aprendizado profundo e outros, para gerar novas representações de dados.

5.2.1 O que são os fenótipos e o seu papel na pesquisa observacional

A seguir ressaltamos alguns pontos nesta definição de fenótipo, essenciais para a pesquisa observacional.

Falamos de coisas que são observáveis, e vamos observar dados de pacientes baseados em quaisquer dados que tenhamos disponíveis no modelo comum de dados CDM-OMOP. São estas observações que serão usadas para encontrar pacientes.

Ressaltamos o estado potencialmente mutável do organismo. Uma pessoa inicialmente pode não ter diabetes, desenvolver diabetes e com exercícios e cuidados pode chegar num ponto no qual não pode mais ser classificado como diabético. Esta noção de ter um estado potencialmente mutável, significa que uma coorte ou fenótipo é

algo que realmente muda com o tempo, que um indivíduo pode fazer parte de um fenótipo num determinado momento e deixar de fazer parte dele num outro momento.

Enfatizamos a diferença entre fenótipo e algoritmo de fenótipo. O fenótipo é a ideia clínica que existe a respeito do que se quer estudar. Não é algo que está definido em código para ser aplicado diretamente na base de dados, é uma ideia definida abstratamente. Se falo de pessoas com diabetes, temos uma ideia clínica do que isso significa, e o que foi descoberto na OHDSI e debatido ativamente nos fóruns, é que as pessoas têm interpretações diferentes a respeito do que uma doença significa. Este é um dos problemas fundamentais a que somos desafiados. Se falo de diabetes, estou me referindo a pessoas com um diagnóstico clínico de diabetes? Ou a pessoas com resultado de hemoglobina A1c maior que 6.5? Ou pessoas que são suspeitas de ter diabetes por causa de terem uma prescrição de metformina dada por um médico?

Estas são diferentes interpretações a respeito do que é ou não é a diabetes. Nenhuma delas está certa ou errada, são apenas diferentes, mesmo assim, usamos a linguagem como uma forma de descrever ideias abstratas e falamos de “pessoas com diabetes”.

O importante é que o fenótipo é aquele estado observável que estamos tentando representar. Nos interessa ser o mais preciso possível na descrição clínica da ideia que queremos estudar, porque iremos traduzir esta ideia clínica em uma definição que possa ser aplicada numa base de dados para identificar pessoas. Esta definição é o algoritmo do fenótipo, a definição da coorte.

Esta definição de coorte agora não mais é uma definição clínica abstrata de algo, agora é uma definição operacional a respeito de quais observações vou procurar na base de dados para identificar pessoas que espero sejam consistentes com a descrição clínica inicial.

Para chegar nesta definição, existem diversos passos de tradução e em cada passo de tradução posso perder informação. É necessário descrever e articular exatamente o que se tem em mente a respeito da ideia clínica porque ela é o ponto de partida para pensar como vamos achar os pacientes que se encaixam nessa seleção, procurando nos dados e identificando as observações.

O fenótipo, no que se refere à pesquisa observacional, é um conjunto acordado de instruções codificadas que é a melhor aproximação para encontrar membros de uma coorte dentro dos dados de saúde.

O conjunto de instruções para identificação do paciente pode ser: 1) uma expressão de conceitos e regras de consulta (heurística) ou 2) um modelo preditivo (probabilístico). Este conjunto de instruções é independente da fonte de dados, porém, o seu desempenho depende da fonte de dados subjacente e deve ser avaliado em termos da mesma.

Conjunto de Conceitos vs Fenótipo vs Coorte

Muitas vezes confundimos um conjunto de códigos, sejam de doenças, exames ou procedimentos, com a definição de um fenótipo ou um fenótipo com uma coorte. A Tabela 5.1 mostra a diferença entre os termos.

Tabela 5.1 Definição de termos

Termos	Descrição
Conjunto de conceitos (Concept Set)	Lista de códigos que define uma característica de interesse
Fenótipo (Phenotype ou Phenotype Algorithm)	Conjunto(s) de Conceitos + Algoritmo de seleção
Coorte (Cohort)	Fenótipo instanciado sobre um banco de dados específico. O conjunto concreto de pessoas identificado, em um banco de dados em particular, pela aplicação do fenótipo

5.2.2 A abordagem da OHDSI aos fenótipos

A abordagem da OHDSI para a construção de fenótipos está baseada em três princípios básicos:

- Elaboração consciente
- Implementação padronizada
- Avaliação reprodutível

Elaboração consciente:

Uma elaboração consciente leva em conta dois pontos principais. Primeiro precisamos definir quais serão os domínios que vamos utilizar para achar os nossos pacientes. Isto é, em quais domínios se manifesta melhor o evento inicial que determina a entrada na corte de um paciente. Pode ser um diagnóstico, ou a primeira prescrição de um medicamento, ou qualquer outro indicador dos outros domínios.

O importante é que a seleção do evento inicial tem efeito direto no desempenho de nosso fenótipo. Na fenotipagem baseada em regras, podemos ter muitas regras, porém o evento inicial é o determinante, por exemplo, de um diagnóstico de diabetes. Preciso satisfazer o evento inicial para poder aplicar o resto das regras, além disto, é a ocorrência do evento inicial que determina a data de indexação.

O segundo ponto é qual a abordagem que queremos ter em função da seleção de pacientes. Queremos todos, ou a maior quantidade possível, dos pacientes elegíveis? Ou queremos apenas ficar com aqueles nos quais temos maior certeza de pertencerem à corte? A Figura 5.5 ilustra graficamente as várias situações que podemos encontrar.

Precisamos achar um balanceamento entre ter uma definição mais restrita que nos transmita confiança no paciente selecionado e que mesmo assim nos permita recuperar uma quantidade suficiente de pacientes.



Figura 5.5 Critérios para seleção de pacientes

Implementação padronizada:

A OHDSI fornece através de suas ferramentas (Atlas e CapR, Aphrodite) um mecanismo padronizado para a definição de fenótipos e uma implementação que não depende da infraestrutura de banco de dados subjacente. Enquanto o banco de dados implementa o modelo OMOP, a geração da coorte será a mesma para qualquer um dos gerenciadores de banco de dados suportados, quer dizer, dado o mesmo conteúdo em diferentes gerenciadores de banco de dados (SQL Server, Oracle, Postgres, Mysql, etc), a execução do fenótipo recupera os mesmos pacientes.

Avaliação reprodutível:

Um fenótipo precisa poder ser compartilhado entre diferentes sistemas. Quando fazemos isto vemos que o desempenho dos fenótipos depende das informações de cada fonte de dados. A definição que na nossa fonte se revelou abrangente, pode se tornar muito específica em outra fonte de dados diferente, ou o conjunto de conceitos que foi escolhido para o fenótipo não consegue ser aplicado em outra fonte por falta de informações.

Portanto, a qualidade ou desempenho de uma definição de fenótipo como preditor, depende da fonte de dados. Uma definição de fenótipo em si não pode ser caracterizada ou avaliada quanto ao desempenho. Depois que uma definição de fenótipo foi aplicada a uma fonte de dados e os indivíduos são identificados, o fenótipo pode ser avaliado em relação a população do banco de dados, tempo de observação, disponibilidade de recursos etc.

Resumindo, uma entrada completa de fenótipo em uma biblioteca de fenótipos, inclui a definição do fenótipo (regras heurísticas ou probabilísticas; independente do banco de dados) e avaliação do fenótipo (caracterização, avaliação; dependente do

banco de dados). Por fim, os bancos de dados observacionais fornecem um subconjunto do conjunto completo de características observáveis de um indivíduo e espera-se que os resultados de desempenho do fenótipo, reflitam isso. As informações do fenótipo podem ser capturadas em 5 seções: 1) definições, 2) caracterizações, 3) avaliações, 4) metadados e 5) disseminação. Um fenótipo para uma condição pode incluir várias definições.

5.2.3 Tipos de fenótipos: heurísticos (baseados em regras) e probabilísticos

Definição de coorte baseada em regras:

Usa regras explícitas para descrever quando um paciente está na coorte. A definição dessas regras normalmente depende muito da seleção do domínio para os critérios de inclusão de coorte.

Toda vez que vemos uma lista de critérios de inclusão e exclusão e uma pessoa pertence ou não pertence ao grupo, estamos vendo um exemplo de heurística baseada em regras.

O Atlas é um exemplo de ferramenta que implementa uma heurística baseada em regras e Criteria2Query¹⁶ é outra ferramenta que traduz critérios de elegibilidade complexos em consultas estruturadas em SQL, que podem ser aplicadas no banco de dados. Atualmente, o uso de heurística baseada em regras na pesquisa observacional representa mais de 90% dos estudos.

Definição de coorte probabilística:

No lugar de falarmos que uma pessoa pertence ou não a um determinado fenótipo, podemos usar um modelo probabilístico para calcular uma probabilidade entre 0 e 100 por cento do paciente estar na coorte e traduzir em limites de classificação de 0 e 1.

No lugar de falar se tenho diabetes ou não, podemos falar qual é a probabilidade de eu ter diabetes num determinado momento baseado no que se conhece a meu respeito. Algoritmos de aprendizagem supervisionada permitem uma melhor fenotipagem de nossas amostras de coorte e sua caracterização.

5.2.4 O fluxo da construção de fenótipos

Etapa 1: Desenvolver uma compreensão clínica compartilhada do fenótipo.

Este é definitivamente o passo mais importante, é o “alvo” que estamos tentando fenotipar. Sem uma compreensão compartilhada não podemos “avaliar” se estamos realmente identificando as pessoas que queremos estudar em nossas definições de coorte. Espera-se que seja uma descrição semiestruturada, detalhada, especificada em linguagem humana natural e não um algoritmo de computador ou código de diagnóstico.

O que deve constar na descrição clínica:

¹⁶ <https://github.com/OHDSI/Criteria2Query>

Conteúdo mínimo:

- Visão geral: é um resumo chave do fenótipo que destaca suas características
- Apresentação: é como a pessoa ou seu clínico observaria primeiro seu fenótipo
- Avaliação: O que o clínico gostaria de fazer para confirmar o diagnóstico
- Plano: É provável que a avaliação seja seguida por um plano de gestão, que se baseia nas preferências do paciente e nas evidências estabelecidas
- Prognóstico: Descreve o estado futuro esperado, ou a história natural do fenótipo, especificamente quanto tempo durará, o que se espera que melhore ou desapareça, o que se espera que piore ou se torne outra coisa?

Passo 2: Faça uma revisão da literatura.

Outros realizaram estudos epidemiológicos/observacionais baseados em pesquisas sobre o desfecho de interesse? Há literatura publicada disponível que tenha estudado o resultado de interesse em dados observacionais? Eles descreveram como desenvolveram uma definição operacional para o resultado de interesse?

Passo 3: Construindo a expressão do conjunto de conceitos.

As ideias-chave são:

- Não comece antes de ter uma descrição clínica finalizada
- Criar uma expressão de conjunto de conceitos iniciais
- Crie um critério de pesquisa lexical com base nas palavras-chave na descrição clínica
- Importe qualquer conceito inicial que foram recomendados/usados anteriormente no trabalho de outros
- Otimize a expressão do conjunto de conceitos
- Revise os conceitos incluídos
- Iterar até ficar satisfeito
- Use PHOEBE
- Finalizar

Como o diagnóstico de projeto envolve a tomada de decisões com base no sistema recomendado, uma boa prática é documentar os principais motivos/insights como notas. Estas notas devem ter detalhes suficientes para informar a um novo usuário o motivo pelo qual um determinado conceito foi incluído (ou excluído) da expressão do conjunto de conceitos.

Se foi tomada a decisão de incorporar conceitos não padronizados na expressão do conjunto de conceitos, então, é necessário relatar os insights que levaram à adição à equipe de vocabulário OMOP. Esse relatório ajuda a melhorar a lógica de mapeamento no vocabulário OMOP para uma versão futura.

É importante ressaltar que, nesta fase, a expressão do conjunto de conceitos resultante passou por um diagnóstico baseado na compreensão léxica/semântica. Pode não ter sido aplicado a uma definição de corte e, se aplicado, o diagnóstico de

desempenho de tais definições de coorte não foi avaliado, ou seja, a definição de coorte, que é construída usando a saída da expressão do conjunto de conceitos do diagnóstico de projeto não foi revisada. Portanto, não é considerada uma definição de coorte pronta para pesquisa.

Passo 4: Construindo o conjunto de conceitos que serão utilizados na construção das coortes.

Estes conjuntos de conceitos (Concept Sets) são expressões que basicamente representam a lista de conceitos do vocabulário da OHDSI que queremos utilizar para identificar eventos, e podem ser reutilizados em diferentes estudos. Estes conjuntos de conceitos são representados de maneiras que podem ser legíveis por humanos e executáveis por computador. Existe um conjunto de ferramentas desenvolvidas pela OHDSI e disponíveis através da interface do Atlas para criar e utilizar estes conjuntos de conceitos.

Passo 5: Descrever a lógica do algoritmo do fenótipo.

Contamos com as ferramentas padronizadas da OHDSI que ajudam a escrever esta lógica, tanto usando a interface gráfica quanto por programação. Resumidamente isto envolve responder uma série de questões: quais são os eventos de início da coorte, quais são os critérios de inclusão e qual é a estratégia de saída da coorte. Este é um processo iterativo no qual, após ter desenvolvido a definição da coorte e aplicado esta definição no banco de dados real, podemos reavaliar o conjunto de conceitos utilizados ou os critérios de inclusão/exclusão.

Passo 6: Avaliar a definição de coorte.

A comunidade disponibiliza ferramentas, como por exemplo, o diagnóstico de coortes, que gera estatísticas resumidas agregadas da nossa rede de bancos de dados permitindo avaliar se uma definição de coorte faz sentido e é apropriada para um estudo diferente do que foi proposto. E temos novamente um processo iterativo no qual aprendemos com os nossos dados e refinamos tanto a definição da coorte quanto a definição do conjunto de conceitos.

A medida que pensamos no fluxo de trabalho para desenvolver e avaliar fenótipos, podemos pensar em tentar avançar a ciência da fenotipagem evoluindo para um processo sistemático e disciplinado que podemos acompanhar, e que nos fornece algum tipo de evidência empírica de que as definições e fenótipos que criamos são confiáveis o suficiente para serem usados em qualquer uma das análises que formos fazer.

A Figura 5.6 apresenta o fluxo de construção e validação de fenótipos.

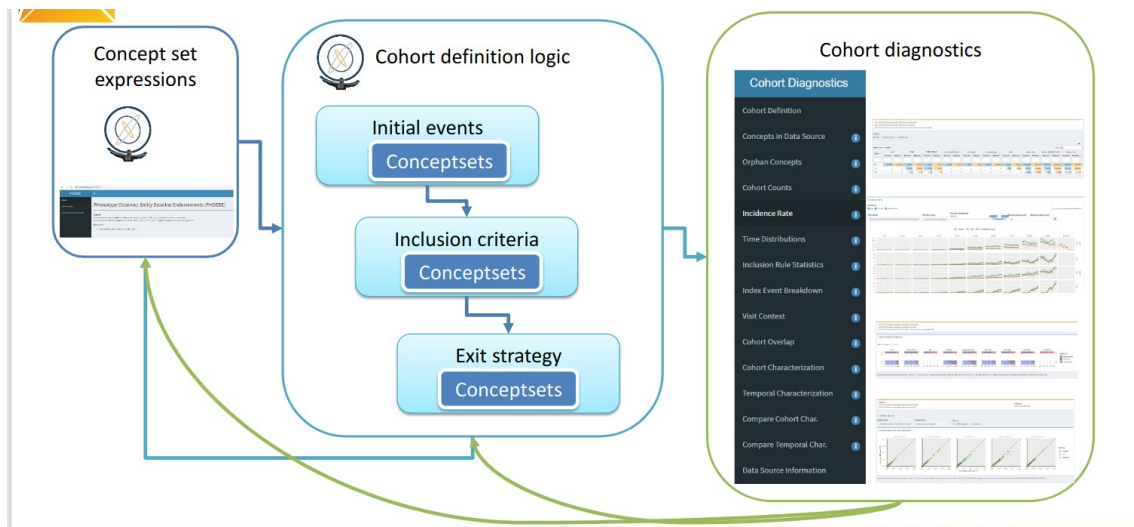


Figura 5.6 Processo de construção e avaliação dos fenótipos

5.2.5 Ferramentas de suporte ao desenvolvimento de fenótipos

A OHDSI oferece um conjunto grande de ferramentas para a dar suporte ao desenvolvimento de fenótipos, algumas das quais já foram mencionadas. A continuação, apresentamos uma descrição resumida de seis delas, as quais se integram para formar o processo de geração de fenótipos que acabamos de descrever.

Atlas é a ferramenta de interação primária do usuário para o projeto e geração de estudos. Ela possui componentes que são muito úteis no desenvolvimento de fenótipos:

- O componente de definição de coortes permite definir uma coorte baseada em regras usando uma interface gráfica
- A ferramenta de perfis (profile) permite a partir de uma coorte, identificar um determinado indivíduo e examinar os dados a nível de paciente, para validar se corresponde ou não, ele pertencer a essa coorte.

CapR: Enquanto o Atlas é uma ferramenta gráfica, este é um pacote R que permite criar programaticamente a mesma definição de coorte que seria criada pelo Atlas. Isto é particularmente útil porque queremos ter um processo totalmente reproduzível ou, por exemplo, alterar programaticamente para criar um conjunto de fenótipos e este pacote cria uma linguagem que torna isso possível.

Phoebe¹⁷¹⁸ é um sistema de recomendações que ajuda no processo de construção de conjuntos de conceitos (concept sets) de forma que podemos escolher um conceito que achamos que representa a nossa ideia e ele procura outros conceitos que podem ser sugeridos em função de semelhança lexical e o uso da hierarquia de vocabulário ou de outras bases de conhecimento.

17 <https://data.ohdsi.org/PHOEBE/>

18 <https://www.ohdsi.org/wp-content/uploads/2021/06/tutorial-phoebe.pdf>

Aphrodite é uma abordagem para desenvolver fenótipos probabilísticos onde a ideia básica é treinar um modelo de previsão baseado em rótulos ruidosos (“noisy labels”) que daria noção da probabilidade de uma pessoa pertencer ao fenótipo alvo.

Diagnóstico de coortes é uma ferramenta para avaliar algoritmos de fenótipo usando análises de caracterização em nível de população que nos permitem identificar erros na sensibilidade ou especificidade ou erros de classificação de data de indexação. A data de indexação é o momento no qual consideramos a entrada de uma pessoa na coorte, porém acontece que existem situações nas quais o critério de escolha dessa data é fonte de erros que o diagnóstico de coortes ajuda a detectar.

PheValuator é uma ferramenta que permite que dado um conjunto de fenótipos produzir uma estimativa da especificidade de sensibilidade e valor positivo treinando um modelo de previsão e criando um padrão de referência probabilístico. A Tabela 5.2 resume as ferramentas de fenotipagem.

Tabela 5.2 Ferramentas de Fenotipagem

Ferramenta	Descrição
Atlas	<ul style="list-style-type: none"> - Definição de Coortes baseada em regras - Perfis para revisão de casos individuais
CapR	Definição programática em R de coortes baseadas em R consistente com a especificação CIRCE JSON
PHOEBE	Desenvolvimento e avaliação de conjunto de conceitos (concept sets) pela exploração dos vocabulários OHDSI para recomendar conceitos candidatos
APHRODITE	Desenvolvimento de fenótipos probabilísticos por treinamento de um modelo preditivo usando rótulos ruidosos
CohortDiagnostics	Avalia algoritmos de fenotipagem usando caracterização populacional para identificar erros de sensibilidade/especificidade falhas na especificação da data de indexação
PheValuator	Avalia algoritmos de fenótipos estimando sensibilidade, especificidade e o valor preditivo positivo através do treinamento de um modelo preditivo e a criação de um padrão de referência probabilístico.

5.2.6 Biblioteca de fenótipos padrão ouro OHDSI

Para ajudar a comunidade no inventário e na avaliação geral das definições e algoritmos de coorte existentes, foi formado o Grupo de Trabalho OHDSI Gold Standard Phenotype Library (GSPL). O objetivo do grupo de trabalho GSPL é desenvolver uma biblioteca de fenótipos apoiada pela comunidade a partir de métodos probabilísticos e baseados em regras. A GSPL permite que os membros da comunidade OHDSI encontrem, avaliem e utilizem definições de coorte validadas pela comunidade para pesquisa e outras atividades. Essas definições de “padrão ouro” residirão em uma biblioteca, cujas entradas são mantidas em padrões específicos de projeto e avaliação.

Para obter informações adicionais relacionadas à GSPL, consulte a página do grupo de trabalho OHDSI¹⁹. A pesquisa dentro deste grupo de trabalho inclui AFRODITE [Banda et al. 2017] e a ferramenta PheValuator [Swedel, Hripcsak e Ryan 2019], bem como o trabalho feito para compartilhar a Biblioteca de Fenótipos eMERGE de Registros Médicos Eletrônicos e Genômica em toda a rede OHDSI [Hripcsak et al. 2019]. A Figura 5.7 apresenta a arquitetura da biblioteca de fenótipos padrão ouro da OHDSI.

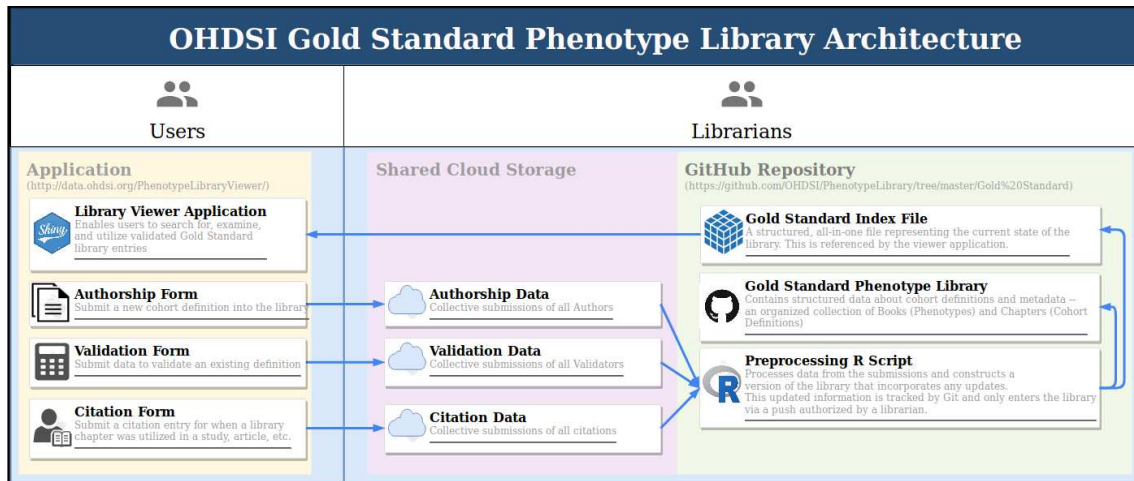


Figura 5.7 Arquitetura da Biblioteca de Fenótipos OHDSI

5.3 Definições de coortes

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. Um dos principais problemas no agrupamento de fontes de dados diversas é a procura por uma definição comum do significado das informações nelas armazenadas. Geralmente precisamos criar uma coorte (um agrupamento de pessoas com características semelhantes) usando alguma definição de como um evento clínico se manifesta. Considerações semelhantes podem ser aplicadas aos dados capturados durante as interações de rotina de assistência médica registradas em um sistema de registros eletrônicos. Como os dados estão sendo usados para uma finalidade secundária, devemos estar cientes do que cada banco de dados foi originalmente projetado para fazer. Cada vez que projetamos um estudo, precisamos pensar nas nuances de como nossa coorte existe em uma variedade de ambientes de assistência médica.

19 <https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:gold-library-wg>

5.3.1 O que é uma coorte?

Uma coorte é um conjunto de pessoas que satisfazem um ou mais critérios de inclusão por um período. É a aplicação da definição de um fenótipo instanciado sobre um banco de dados específico.

Embora os conjuntos de códigos sejam uma peça importante na montagem de uma coorte, uma coorte não é definida por um conjunto de códigos. Uma coorte requer uma lógica específica de como usar o conjunto de códigos para os critérios (por exemplo, é a primeira ocorrência do código CID-10? alguma outra ocorrência?). Uma coorte bem definida específica como um paciente entra em uma coorte e como um paciente sai de uma coorte.

Existem nuances únicas para utilizar a definição de coorte na OHDSI, incluindo:

- Uma pessoa pode pertencer a várias coortes
- Uma pessoa pode pertencer à mesma coorte por vários períodos diferentes
- Uma pessoa pode não pertencer à mesma coorte várias vezes durante o mesmo período
- Uma coorte pode ter zero ou mais membros

5.3.2 Definições de coortes

Existem duas abordagens principais para construir uma coorte:

As definições de **coorte baseadas em regras** usam regras explícitas para descrever quando um paciente está na coorte. A definição dessas regras normalmente depende muito da experiência de domínio do indivíduo que projeta a coorte para usar seu conhecimento da área terapêutica de interesse para construir regras para os critérios de inclusão da coorte.

As definições de **coorte probabilística** usam um modelo probabilístico para calcular uma probabilidade entre 0 e 100% do paciente estar na coorte. Essa probabilidade pode ser transformada em uma classificação sim-não usando algum limite, ou em alguns projetos de estudo pode ser usada como está. O modelo probabilístico é normalmente treinado usando aprendizado de máquina (por exemplo, regressão logística) em alguns dados de exemplo para identificar automaticamente as características relevantes do paciente que são preditivas.

5.3.3 Coorte baseadas em Regras

Uma definição de coorte baseada em regras começa com a declaração explícita de um ou mais critérios de inclusão (por exemplo, “pessoas com angioedema”) em um período específico (por exemplo, “que desenvolveu esta condição nos últimos 6 meses”).

Os componentes padrão que usamos para montar esses critérios são:

- Domínio: O(s) domínio(s) do CDM onde os dados são armazenados (por exemplo, “Ocorrência de Procedimento”, “Exposição a Medicamentos”) definem o tipo de informação clínica e os conceitos permitidos que podem ser

representados dentro dessa tabela do CDM. A Figura 5.7 apresenta os blocos de construção de uma coorte.

- **Conjunto de Conceitos:** Uma expressão independente de dados que define um ou mais Conceitos Padrão que abrangem a entidade clínica de interesse. Esses conjuntos de conceitos são interoperáveis em diferentes dados de saúde observacionais, pois representam os termos padrão para os quais a entidade clínica mapeia no vocabulário.
- **Atributo específico do domínio:** atributos adicionais relacionados à entidade clínica de interesse (por exemplo, `DAYS_SUPPLY` para `DRUG_EXPOSURE` ou `VALUE_AS_NUMBER` ou `RANGE_HIGH` para `MEASUREMENT`.)
- **Lógica temporal:** São os intervalos de tempo entre os quais a relação entre um critério de inclusão e um evento são avaliados (por exemplo, a condição indicada deve ocorrer durante 365 dias antes ou no início da exposição).

À medida que você está construindo sua definição de coorte, pode ser útil pensar em Domínios análogos a blocos de construção que representam atributos de coorte. A Figura 5.8 apresenta esses blocos definidos no CDM.



Figura 5.8 Blocos de construção das definições de uma coorte

Eventos de entrada na coorte:

O evento de entrada de coorte (evento inicial) define a hora em que as pessoas entram na coorte, chamada de data do índice de coorte. Um evento de entrada de coorte pode ser qualquer evento registrado no CDM, como exposições a medicamentos, condições, procedimentos, medições e visitas. Os eventos iniciais são definidos pelo domínio do CDM onde os dados são armazenados (por exemplo, `PROCEDURE_OCCURRENCE`, `DRUG_EXPOSURE`, etc), os conjuntos de conceitos construídos para identificar a atividade clínica (por exemplo, códigos SNOMED para condições, códigos RxNorm para medicamentos), bem como quaisquer outros atributos específicos (por exemplo, idade na ocorrência, primeiro diagnóstico/procedimento/etc, especificando a data de início e término, especificando o tipo ou critério de visita,

fornecimento de dias, etc). O conjunto de pessoas que têm um evento de entrada é chamado de coorte inicial do evento.

Critérios de inclusão:

Os critérios de inclusão são aplicados à coorte inicial do evento para restringir ainda mais o conjunto de pessoas. Cada critério de inclusão é definido pelo(s) domínio(s) do CDM onde os dados são armazenados, conjuntos de conceitos que representam a atividade clínica, atributos específicos do domínio (por exemplo, fornecimento de dias, tipo de visita, etc.) e a lógica temporal relativa ao data do índice de coorte. Cada critério de inclusão pode ser avaliado para determinar o impacto na perda de pessoas da coorte inicial do evento. A coorte qualificada é definida como todas as pessoas na coorte inicial do evento que atendem a todos os critérios de inclusão.

Critérios de saída da coorte:

O evento de saída da coorte significa quando uma pessoa não se qualifica mais para ser membro da coorte. A saída da coorte pode ser definida de várias maneiras, como o final do período de observação, um intervalo de tempo fixo relativo ao evento de entrada inicial, o último evento em uma sequência de observações relacionadas (por exemplo, exposição persistente a drogas) ou por meio de outra censura do período de observação. A estratégia de saída da coorte afetará se uma pessoa pode pertencer à coorte várias vezes durante diferentes intervalos de tempo.

Nas ferramentas OHDSI não há distinção entre critérios de inclusão e exclusão. Todos os critérios são formulados como critérios de inclusão. Por exemplo, o critério de exclusão “Excluir pessoas com hipertensão prévia” pode ser formulado como o critério de inclusão “Incluir pessoas com 0 ocorrências de hipertensão prévia”.

5.3.4 Conjunto de conceitos

Um conjunto de conceitos é uma expressão que representa uma lista de conceitos que podem ser usados como um componente reutilizável em várias análises. Pode ser pensado como um equivalente padronizado e executável por computador das listas de códigos frequentemente usadas em estudos observacionais. Uma expressão de conjunto de conceitos consiste em uma lista de conceitos com os seguintes atributos:

- **Excluir:** Excluir este conceito (e qualquer um de seus descendentes, se selecionado) do conjunto de conceitos.
- **Descendentes:** Considere não apenas este conceito, mas também todos os seus descendentes.
- **Mapeado:** Permite pesquisar conceitos não padronizados.

Por exemplo, uma expressão de conjunto de conceitos pode conter dois conceitos, conforme descrito na Tabela 5.3. Aqui incluímos o conceito 4329847 (“Infarto do miocárdio”) e todos os seus descendentes, mas excluimos o conceito 314666 (“Infarto do miocárdio antigo”) e todos os seus descendentes.

Tabela 5.3 Um exemplo de expressão de conjunto de conceitos

ID do conceito	Nome do conceito	Excluído	Descendentes	Mapeado
4329847	Infarto do miocárdio	NÃO	SIM	NÃO
314666	Infarto do miocárdio antigo	SIM	SIM	NÃO

A Figura 5.9 inclui “Infarto do miocárdio” e todos os seus descendentes, exceto “Infarto do miocárdio antigo” e seus descendentes. No total, esta expressão de conjunto de conceitos implica quase uma centena de Conceitos Padrão. Esses Conceitos Padrão, por sua vez, refletem centenas de códigos-fonte (por exemplo, códigos CID-9 e CID-10) que podem aparecer em vários bancos de dados.

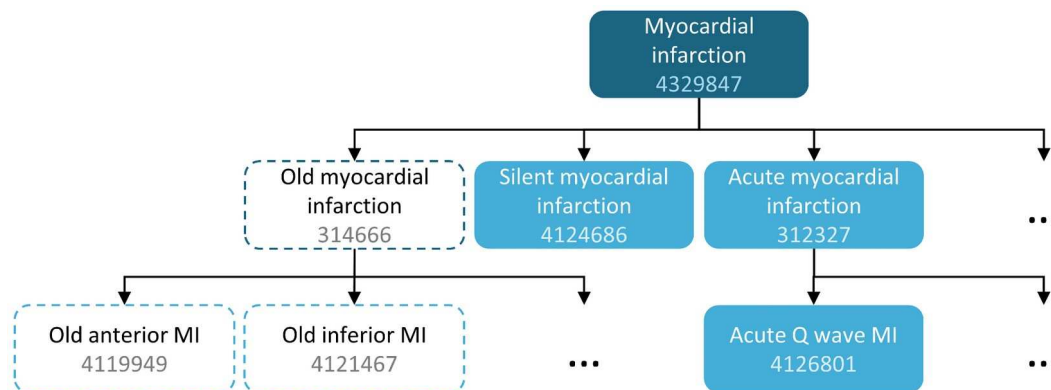


Figura 5.9 Um conjunto de conceitos incluindo “Infarto do miocárdio” (com descendentes), mas excluindo “Infarto do miocárdio antigo” (com descendentes)

5.3.5 Coorte Probabilística

As definições de coorte baseadas em regras são um método popular para montar definições de coorte. No entanto, reunir o consenso de especialistas necessário para criar uma coorte de estudo, pode consumir muito tempo. O projeto de coorte probabilístico é um método alternativo, controlado por máquina, para agilizar a seleção de atributos de coorte. Nessa abordagem, o aprendizado de máquina supervisionado permite que um algoritmo de fenotipagem aprenda com um conjunto de exemplos rotulados (casos) de quais atributos contribuem para a associação à coorte. Esse algoritmo pode então ser usado para determinar melhor as características definidoras de um fenótipo e quais compensações ocorrem na precisão geral do estudo ao optar por modificar os critérios do fenótipo.

Um exemplo de aplicação dessa abordagem em dados no CDM é o pacote R AFRODITE (Rotina de Fenótipo Automatizado para Definição, Identificação,

Treinamento e Avaliação Observacional)²⁰. Este pacote fornece uma estrutura de construção de coorte que combina a capacidade de aprender a partir de dados rotulados de forma imperfeita.

5.4 Criando coortes na ferramenta Atlas

O ATLAS é uma ferramenta gratuita, publicamente disponível e baseada na web, desenvolvida pela comunidade OHDSI que facilita o projeto e execução de análises em dados observacionais em nível de paciente, padronizados no modelo CDM-OMOP. É a ferramenta usada para fenotipagem baseada em regras ou definição de coorte baseada em regras.

O ATLAS é implantado como um aplicativo da web em combinação com o OHDSI WebAPI. Para o desempenho de análises em tempo real requerer acesso a dados de nível de paciente no CDM. Para executar o Atlas, você deve estar atrás do Firewall de sua instituição.

As coortes são usadas (e reutilizadas) nas ferramentas de análise OHDSI para definir, por exemplo, as exposições e os resultados de interesse. Ao construir uma coorte, deve-se considerar o que é mais importante: encontrar todos os pacientes elegíveis ou obter apenas aqueles sobre os quais está confiante?

A estratégia para construir uma coorte dependerá do rigor clínico de como seu consenso de especialistas define a doença. Isso quer dizer que o design de coorte certo dependerá da pergunta que está tentando responder. Pode-se optar por criar uma definição de coorte que use tudo o que puder obter, use o menor denominador comum para compartilhá-la com outras instituições. Em última análise, fica a critério do pesquisador qual limite de rigor é necessário para estudar adequadamente a coorte de interesse.

Uma definição de coorte é uma tentativa de inferir algo que gostaríamos de observar a partir dos dados registrados. Isso levanta a questão de quão bem conseguimos nessa tentativa. Em geral, a validação de uma definição de coorte baseada em regras ou algoritmo probabilístico, pode ser pensada como um teste da coorte proposta em comparação com alguma forma de referência “padrão ouro” (por exemplo, revisão manual de gráficos de casos).

5.4.1 Definindo uma coorte

A interface do Atlas para criar coortes, disponibiliza quatro seções: eventos de entrada de coorte, critérios de inclusão, saída de coorte e eras das coortes. A Figura 5.10 apresenta as quatro sessões.

²⁰ <https://github.com/OHDSI/Aphrodite>

Cohort Entry Events

Eventos de Entrada de Coorte: registros do CDM selecionados a partir de dados do paciente que qualificam o mesmo para entrar na coorte

+ Add Initial Event

Restrict initial events

Inclusion Criteria

New inclusion criteria

Critérios de Inclusão: Utilizar outros dados do paciente relativos ao evento inicial para qualificar o evento como válido para inclusão. Representa o fluxo da "Qualificação da Coorte"

Limit qualifying events to: earliest event per person.

Cohort Exit

Event Persistence: Saída de Coorte: Definir quanto tempo cada evento mantém a presença do paciente na coorte, ou identificar registros que censuram o evento.

Event will persist until: end of continuous ob

Censoring Events:

Exit Cohort based on the following criteria:

- Fim do período de observação
- Tempo fixo em relação à data de início do evento
- Observações de censura, p. ex: procedimentos ou fim da terapia

No censoring events selected.

Cohort Eras

Specify era collapse gap size: 0 days

add trimming options

Eras da Coorte: Encadeia as datas restantes dos eventos permitindo lacunas com censura opcional à esquerda e/ou à direita nas eras finais (recorte)

Figura 5.10 Critérios de seleção de uma coorte

Definindo uma coorte para hipertensão:

Começamos criando uma definição de coorte usando uma abordagem baseada em regras. Neste exemplo, queremos encontrar pacientes que iniciam a monoterapia com inibidores da ECA (enzima conversora de angiotensina) como tratamentos de primeira linha para hipertensão. A Figura 5.11 mostra a estrutura lógica de como queremos construir essa coorte.

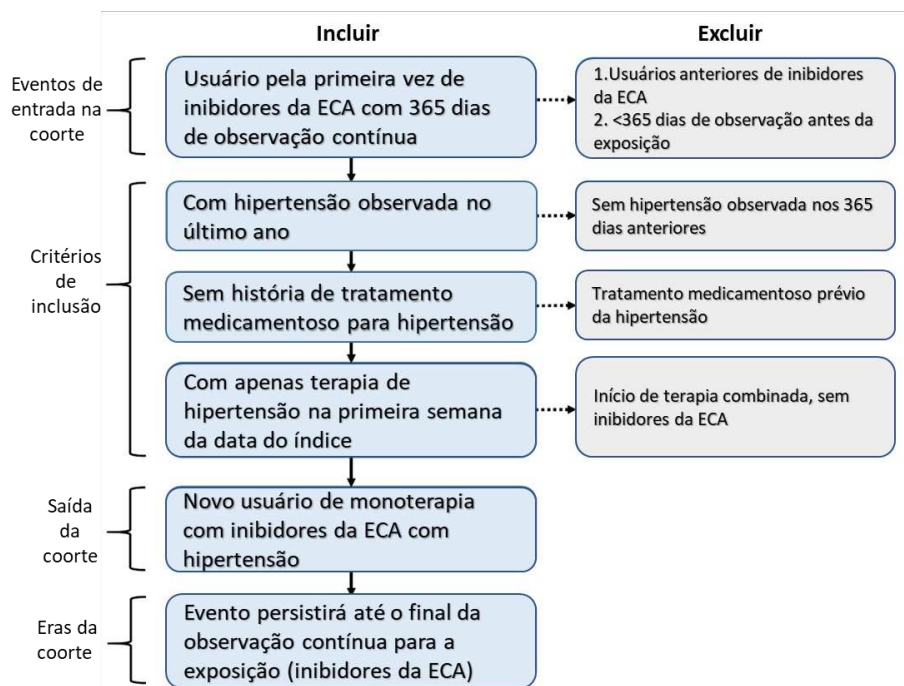


Figura 5.11 Diagrama lógico da coorte pretendida

Passo 1: Criando uma nova coorte



Para iniciar no ATLAS, clique na aba  Cohort Definitions. Quando o módulo carregar, clique em “Nova coorte”. A próxima tela que você verá será uma definição de coorte vazia. A Figura 5.12 apresenta a tela inicial do módulo para criação de coorte.

Figura 5.12 Nova definição de coorte

Antes de fazer qualquer outra coisa, altere o nome da coorte de “Nova definição de coorte” para um nome exclusivo para esta coorte. O ATLAS não permitirá que duas coortes tenham os mesmos nomes exatos. Apresenta uma mensagem de erro se escolher um nome já usado por outra coorte.

Depois de escolher um nome, você pode salvar a coorte clicando em .

Passo 2: Definição do evento inicial da coorte.


Selecione “Adicionar evento inicial”. Agora precisa escolher em qual domínio está construindo um critério. Como vemos na Figura 5.13, o ATLAS fornece descrições abaixo de cada critério. Se estivéssemos construindo um critério baseado em `CONDITION_OCCURRENCE`, a pergunta seria procurar pacientes com um diagnóstico específico. Se estivéssemos construindo um critério baseado em `DRUG_EXPOSURE`, a pergunta seria procurar pacientes com um medicamento específico ou classe de medicamentos. Como queremos encontrar pacientes que iniciam a monoterapia com inibidores da ECA como tratamento de primeira linha para hipertensão, queremos escolher um critério `DRUG_EXPOSURE`. A data de início da coorte é definida pelo início do tratamento com inibidor da ECA, que é, portanto, o evento inicial. Selecione “Adicionar exposição a drogas”.

Figura 5.13 Adicionando um evento inicial


A tela será atualizada com os critérios selecionados e é preciso informar ao ATLAS qual conjunto de conceitos está associado aos inibidores da ECA.

Figura 5.14 Definindo uma exposição a um medicamento

Passo 3: Definindo o Conjunto de Conceitos

Selecionar  para abrir a caixa de diálogo que permitirá recuperar um conjunto de conceitos para definir os Inibidores da ECA.

Cenário 1: não construiu um conjunto de conceitos

Construir um conjunto de conceitos dentro da definição de coorte navegando até a guia “Conjunto de conceitos” e clicando em “Novo conjunto de conceitos”. Renomear o conjunto de conceitos. A partir daí, usar o  Search para procurar conceitos clínicos que representam os inibidores da ECA (Figura 5.15).

← EXAMPLE: new users of ACE inhibitors as first-line mono-therapy for hypertension → ACE Inhibitors

Search

Search Import

ace inhibitors

Advanced Options

Column visibility Copy CSV Show 15 entries Filter:

Showing 1 to 9 of 9 entries Previous 1 Next

Vocabulary	Id	Code	Name	Class	RC	DRC	Domain	Vocabulary
ATC (6)	21601784	C09AA	ACE inhibitors, plain	ATC 4th	0	507,772	Drug	ATC
Multilex (1)	21601783	C09A	ACE INHIBITORS, PLAIN	ATC 3rd	0	507,772	Drug	ATC
VA Class (1)	21601802	C09BA	ACE inhibitors and diuretics	ATC 4th	0	10,982	Drug	ATC
LOINC (1)								
Class	21601801	C09B	ACE INHIBITORS, COMBINATIONS	ATC 3rd	0	10,982	Drug	ATC
ATC 4th (4)								

Figura 5.15 Pesquisando o vocabulário - Inibidores ECA

Selecionar o conceito clicando em  e retornar à sua definição de coorte.

A Figura 5.16 apresenta a expressão do conjunto de conceitos e os seus descendentes que foram selecionados.

Concept Set Expression Included Concepts 21538 Included Source Codes Export Import

Name: ACE Inhibitors

Show 25 entries Search:

Showing 1 to 15 of 15 entries Previous 1 Next

Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption	Exclude	Descendants	Mapped
1335471	18867	benazepril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1340128	1998	Captopril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
19050216	21102	Cilazapril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1341927	3827	Enalapril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1342001	3829	Enalaprilat	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1363749	50166	Fosinopril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
19122327	60245	imidapril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1308216	29046	Lisinopril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1310756	30131	moexipril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1373225	54552	Perindopril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1331235	35208	quinapril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1334456	35296	Ramipril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
19040051	36908	spirapril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1342439	38454	trandolapril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
19102107	39990	zofenopril	Drug	Standard	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>




 Classification  Non-Standard  Standard

Figura 5.16 Um conjunto de conceitos contendo drogas inibidoras da ECA

Cenário 2: já construiu um conjunto de conceitos

Selecione “Importar Conjunto de Conceitos”. Uma caixa de diálogo será aberta, para a seleção dos conceitos de interesse no repositório de conjunto de conceitos do ATLAS, conforme mostrado na Figura 5.17. Selecione a linha do conjunto de conceitos.

Import Concept Set From Repository...

[New Concept Set](#)

Show entries Filter Repository Concept Sets:

Id	Title	Created	Modified	Author
1794480	[OHDSI EU 2019] Excluded concepts of ACE inhibitors or Thiazide diuretics	03/28/2019 11:04 AM	03/28/2019 11:04 AM	anonymous
963	ACE Inhibitors			anonymous
3268	COPY OF: ACE Inhibitors			anonymous
99283	Ace Inhibitors			anonymous
142965	PheKB ACE-I ACE inhibitors			anonymous

Showing 1 to 5 of 5 entries (filtered from 11,667 total entries) Previous Next

Figura 5.17 Importando um Conjunto de Conceitos do Repositório ATLAS

Passo 4: Critérios de eventos iniciais adicionais

Sua pergunta é sobre a procura de novos usuários ou a primeira vez na história de alguém que é exposto a inibidores da ECA. Isso se traduz na primeira exposição de inibidores da ECA no prontuário do paciente. Para especificar isso, selecione “+ Adicionar atributo” e “Adicionar primeiro critério de exposição”. É possível especificar outros atributos de um critério que foi construído, como: um atributo de idade na ocorrência, a data de ocorrência, sexo ou outros atributos relacionados ao medicamento. Os critérios disponíveis para seleção são diferentes para cada domínio.

Selecionamos 365 dias de observação contínua antes do evento inicial. Este critério refere-se ao histórico anterior e não envolve tempo após o evento do índice. Portanto, exigimos 0 dias após o evento de índice. Nosso evento de qualificação é o primeiro uso de inibidores da ECA. Assim, limitamos os eventos iniciais ao “evento mais antigo” por pessoa, conforme apresenta a Figura 5.18.

Cohort Entry Events ?

Events having any of the following criteria: [+ Add Initial Event](#)

a drug exposure of ACE inhibitors [+ Add attribute...](#)

✘ for the first time in the person's history [Delete Criteria](#)

with continuous observation of at least days before and days after event index date

Limit initial events to: earliest event per person.

[Restrict initial events](#)

Figura 5.18 Configurando a observação contínua necessária antes da data do índice

Passo 5: Critérios de Inclusão

A funcionalidade nesta seção é idêntica à maneira como discutimos a construção dos critérios de coorte acima. Especifica os critérios e adiciona atributos específicos. Nosso primeiro critério adicional é subdividir a coorte para apenas pacientes com pelo menos 1 ocorrência de distúrbio de hipertensão entre 365 e 0 dias após a data de índice (primeira iniciação de um inibidor da ECA). Selecionar “Novos critérios de inclusão” para adicionar um novo critério, nomear os critérios e colocar uma pequena descrição do que está procurando.

Depois de ter anotado esses novos critérios, selecione o botão “+ Adicionar critérios ao grupo” para construir seus critérios reais para esta regra. Um exemplo seria se houvesse várias maneiras de encontrar uma doença (por exemplo, para uma condição, uma exposição a drogas, uma medição ou exame). Esses seriam domínios separados e exigiram critérios diferentes, mas podem ser agrupados em um critério procurando essa condição. Neste caso, queremos encontrar um diagnóstico de hipertensão, então “Adicionar ocorrência da condição”. Seguiremos etapas semelhantes às que fizemos com o evento inicial, anexando um conjunto de conceitos a este registro. Também queremos especificar se o evento começa entre 365 dias antes e 0 dias após a data do índice (a ocorrência do primeiro uso de inibidor da ECA). Agora verifique a lógica na Figura 5.19.

Figura 5.19 Critérios de inclusão adicionais

A Figura 5.20 ilustra os critérios de inclusão necessários para a seleção desta coorte.

The screenshot shows the 'Inclusion Criteria' interface. On the left, there is a list of criteria:

- has hypertension diagnosis in 1 yr prior to treatment
- Has no prior antihypertensive drug exposures in medical history
- Is only taking ACE as monotherapy, with no concomitant combination treatments

 The third criterion is selected. The main area shows a rule description: 'Is only taking ACE as monotherapy, with no concomitant combination treatments'. Below this, there is a section for 'having all of the following criteria:'. A detailed configuration box is open, showing:

- with exactly 1 using distinct occurrences of:
- a drug era of Hypertension drugs
- where event starts between 0 days Before and 7 days After
- index start date add additional constraint
- allow events from outside observation period (unchecked)

 At the bottom, there is a limit: 'Limit qualifying events to: earliest event per person.'

Figura 5.20 Critérios de inclusão adicionais


Passo 6: Critérios de Saída da Coorte

Especificar os critérios de saída da coorte, definindo quando as pessoas não mais se qualificam para serem incluídas nesta coorte. Nesta coorte, estamos acompanhando novos usuários de uma exposição a drogas. Queremos olhar para o período de observação contínua no que se refere à exposição ao medicamento. Como tal, o critério de saída é especificado a seguir para a totalidade da exposição contínua ao medicamento. Se houver uma interrupção subsequente na exposição ao medicamento, o paciente sairá da coorte neste momento. Fazemos isso porque não podemos determinar o que aconteceu com a pessoa durante o intervalo da exposição ao medicamento. Também podemos definir um critério na janela de persistência para especificar uma lacuna permitida entre as exposições ao medicamento. Nesse caso, adicionaremos nossa janela de persistência para “permitir um máximo de 30 dias” e selecionamos o conjunto de conceitos para “inibidores da ECA”. A Figura 5.21 apresenta essa seleção.

The screenshot shows the 'Cohort Exit' interface. It is divided into three main sections:

- Event Persistence:** 'Event will persist until: end of a continuous drug exposure'.
- Continuous Exposure Persistence:** A detailed text explaining the concept set and persistence window. Below this, 'Concept set containing the drug(s) of interest: ACE Inhibitors'. Two bullet points specify:
 - Persistence window: allow for a maximum of 30 days between exposure records when inferring the era of persistence exposure
 - Surveillance window: add 0 days to the end of the era of persistence exposure as an additional period of surveillance prior to cohort exit.
- Censoring Events:** 'Exit Cohort based on the following criteria:'. Below this, it says 'No censoring events selected.' and there is a '+ Add Censoring Event' button.

Figura 5.21 Critérios de saída da coorte

No caso desta coorte, não há outros eventos de censura. No entanto, pode ser necessário criar outras coortes em que precise especificar esses critérios. Salve a coorte . Construir uma coorte é o bloco mais importante para responder a uma pergunta nas ferramentas OHDSI. Agora, use a guia “Exportar” para compartilhar sua definição de coorte com outros colaboradores na forma de código SQL ou arquivos JSON para carregar no ATLAS.

5.5 Avaliação de fenótipos: PheValuator

A principal abordagem para definir a doença em bancos de dados observacionais de saúde é construir algoritmos de fenótipo, heurísticas baseadas em regras, que se baseiam na presença, ausência e lógica temporal de observações clínicas para identificar um conjunto de pessoas que satisfazem um determinado critério durante um período específico de tempo. No entanto, uma avaliação completa dos algoritmos, determinando sensibilidade, especificidade e valor preditivo positivo (VPP), raramente é realizada.

Não existe uma maneira correta de definir qualquer doença em particular ou qualquer droga em particular, existem múltiplas formas alternativas de definir doenças e drogas, nem certas nem erradas, apenas diferentes, que tentam representar a mesma ideia clínica.

No fim, o que precisamos saber é como se comportam estas definições na hora de escolher as pessoas que nos interessam e não escolher os que não nos interessam e quanto bem conseguem descobrir a data correta de indexação para a entrada na coorte e a data correta de saída para deixar a coorte. Mesmo com todos os cuidados, as definições que criamos são imperfeitas, seja por causa que não temos certeza que a lógica escolhida está certa, seja porque os códigos que selecionamos do vocabulário mudam, seja porque os médicos não ingressam todos os dados no prontuário eletrônico, seja porque os pacientes não vão no médico toda vez que estão doentes, seja por qualquer outro motivo. O fato é que qualquer seja a definição de coorte que você crie, ela será uma representação imperfeita da ideia clínica que queremos representar.

Medir o nível real de imperfeição, o que se traduz em quantificar o erro de medição na definição da coorte, não é uma fraqueza, senão uma qualidade que nos permite criar ferramentas para comparar diferentes definições de coortes para a mesma ideia clínica. Agora podemos testar empiricamente estas alternativas e descobrir qual delas tem um melhor desempenho operacional, definido aqui pela sensibilidade, especificidade e valor preditivo positivo. Como sabemos se nossa definição de coorte é boa, se ela realmente executou corretamente a tarefa de fazer a tradução da ideia clínica em uma lista de pacientes com algum nível de confiança?

Também nos interessa saber quanto generalizável é a nossa definição de coorte. Se num estudo, após aplicar a definição de coorte com seus critérios de inclusão e exclusão, apenas 15% dos candidatos são incorporados na coorte, será que ao aplicar essa mesma definição em bases de dados diferentes ao redor do mundo, continuaremos vendo essa porcentagem de 15%? Ou será diferente segundo a base?

Até agora não conseguimos fazer estas validações ao redor do mundo por causa das diferenças entre bases, codificações e modelos de dados. A OHDSI, através da padronização e disseminação no mundo, permitiu que isto fosse possível, e assim podemos testar a nossa definição de coortes e verificar que mais fontes observaram a mesma perda.

Agora podemos validar diferentes definições de coortes que pretendem descrever a mesma ideia, e para isso vamos fazer uso de métricas de avaliação:

- Sensibilidade: A proporção de casos verdadeiros que consegui identificar. Na base de dados, quais são todos os casos reais da doença? No caso, quanto maior a sensibilidade, maior a quantidade de pessoas que serão escolhidas para fazer parte da coorte.
- Valor Preditivo Positivo: Entre as pessoas selecionadas para a coorte, que porcentagem realmente tem a doença?

Não existe uma definição correta de coorte, o que vai existir é um compromisso na escolha de qual a definição que melhor se comporta para o nosso estudo. Podemos escolher uma maior sensibilidade porque queremos recuperar a maior quantidade de pacientes possível, ou, podemos querer escolher apenas os pacientes nos quais temos mais certeza que possuem a doença e para isso procuramos aquelas definições que possuem um valor preditivo maior. Ao conhecer o desempenho de uma definição de coorte, posso fazer escolhas razoáveis sobre como usar, quando usar, e se devo usar esta definição.

Mas como fazer para obter estes números? Para isso precisamos ter uma certa verdade absoluta para poder avaliar o desempenho, uma base na qual magicamente todos os indivíduos são rotulados como sendo ou não um caso da ideia clínica que estamos querendo descrever. Nesse caso, fazendo uma matriz de confusão, podemos simplesmente contar os casos corretamente identificados e os falsos positivos, e com isso levantar a sensibilidade e o valor preditivo positivo. O problema é que não existe mágica para rotular casos. Apenas podemos tomar uma amostra dos casos identificados pela nossa definição e calcular o valor preditivo positivo. Nunca vamos conseguir saber quantos casos que ficaram de fora da definição, perdemos. Para descrever corretamente o desempenho de uma definição de coorte precisamos dos três valores: sensibilidade, especificidade e valor preditivo positivo.

A comunidade OHDSI desenvolveu uma abordagem diferente para construir um padrão-ouro usando modelos preditivos de diagnóstico [Swerdel, Hripcsak e Ryan 2019].

Vamos criar duas coortes e treinar um modelo. A primeira coorte, chamada de xSpec, vai ser uma coorte extremamente específica, e isto significa uma definição de coorte na qual temos certeza que as pessoas que fazem parte dela são casos verdadeiros. Esta definição de coorte vai selecionar um grupo pequeno de pessoas nas quais temos a maior certeza de que realmente possuem a doença. A outra coorte, chamada de xSense,

vai ser uma coorte extremamente sensível, quer dizer, que identifique qualquer um que possa ter a mínima chance de ter a doença para poder excluí-lo. Como não temos um classificador humano que marque as pessoas definitivamente como sendo ou não um caso, vamos substituí-lo por uma probabilidade e vamos rotular os casos com rótulos incertos: este caso provavelmente tenha a doença porque pertence a coorte xSpec e este outro, provavelmente não tenha a doença porque não faz parte da coorte xSense. Usamos a coorte de alta sensibilidade para selecionar os casos com maior probabilidade de não ter a doença, esses são os que não fazem parte da coorte xSense. Vamos tomar uma amostra aleatória da nossa base, retirar as pessoas que fazem parte da coorte xSense, e o que sobrar serão as pessoas que não tem nenhum indicador da doença. Não posso dizer que ninguém não tem a doença, mas as chances são grandes de não ter.

Agora então tenho dois grupos, um com pessoas que provavelmente têm a doença e outro com pessoas que provavelmente não têm a doença. Posso converter isto em um problema de aprendizado de máquina, onde treino um modelo com os dados disponíveis dos grupos para que ele descubra o que os diferencia. Uma vez o modelo treinado, aplico ele a uma nova amostra aleatória da base e obtenho um grande conjunto de dados onde tenho pessoas nas quais consigo conhecer a chance delas de ter a doença ou não, transformando este grupo em um padrão-ouro probabilístico. Podemos assim, aplicar nossa definição de coorte original a ser avaliada, para as pessoas neste padrão-ouro probabilístico e conseguimos calcular uma estimativa da especificidade, da sensibilidade e do valor preditivo positivo porque podemos preencher a tabela 2x2 de nossa matriz de confusão e produzir diretamente esses números como se tivéssemos realmente um padrão-ouro totalmente determinístico.

Este método produz estimativas dos três parâmetros que pelo menos geram um limite inferior dos mesmos para a definição testada, o que significa que a definição de coorte pode ter um desempenho melhor do que o estimado por este método, e mais importante, quando comparo diversas definições de coortes, serve como fator de decisão para avaliar qual definição vai ter um melhor desempenho.

Existem limitações para este método. É possível que essas pessoas da coorte xSpec possam ter características diferentes das outras com a doença. Também pode ser que essas pessoas tenham tido mais tempo de observação após o diagnóstico inicial do que o paciente médio. Esse algoritmo produz um modelo parcimonioso e normalmente remove muitas das covariáveis colineares que podem estar presentes no conjunto de dados.

5.5.1 Pacote R PheValuator

O PheValuator é um algoritmo que aceita uma ou várias definições de coortes e produz estimativas das três medidas: sensibilidade, especificidade e valor preditivo positivo, usando um modelo probabilístico sem precisar de intervenção humana revisando manualmente cada caso.

A ideia do PheValuator é construir um padrão ouro de referência através de modelos probabilísticos que nos permita rotular os casos sem precisar de intervenção humana. A Figura 5.22 apresenta a descrição do processo do PheValuator.

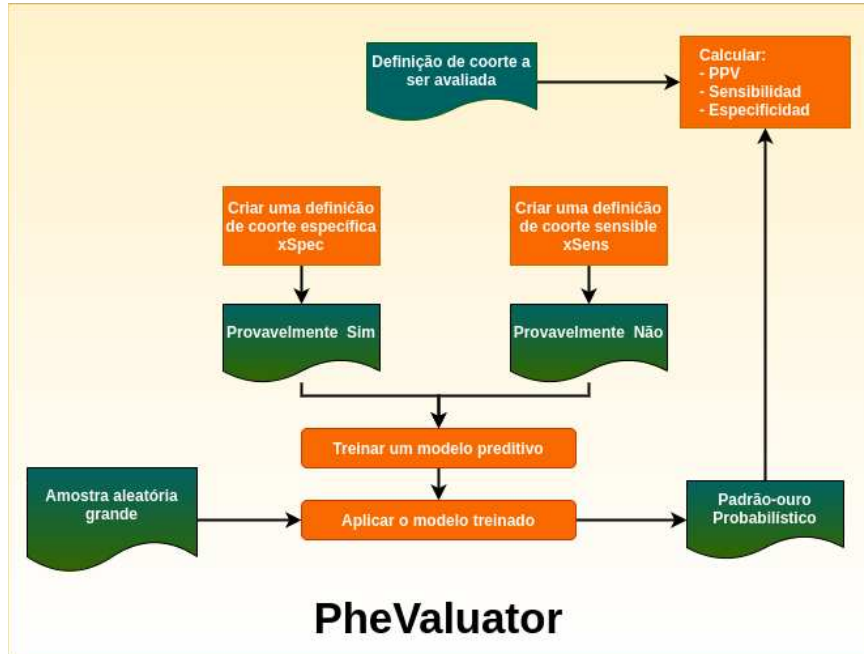


Figura 5.22 Descrição conceitual do PheValuator

A ferramenta foi desenvolvida como um pacote R de código aberto chamado PheValuator²¹. O PheValuator usa funções do pacote Patient Level Prediction. Na versão atual do software PheValuator, o status do resultado (sim/não) é avaliado com base em todos os dados de uma pessoa (todo o tempo de observação) e não avalia a precisão da data de início da coorte.

Em resumo:

- A validade clínica pode ser estabelecida entendendo as características da fonte de dados subjacente, avaliando as características de desempenho das coortes dentro de uma análise e avaliando a generalização do estudo para a população-alvo de interesse.
- Uma definição de coorte pode ser avaliada na medida em que as pessoas identificadas na coorte com base na definição de coorte e os dados observacionais disponíveis refletem com precisão as pessoas que realmente pertencem ao fenótipo.
- A validação da definição de coorte requer a estimativa de várias características de desempenho, incluindo sensibilidade, especificidade e valor preditivo positivo, para resumir completamente e permitir o ajuste do erro de medição.

²¹<https://github.com/OHDSI/PheValuator>;

<https://github.com/OHDSI/PheValuator/blob/main/inst/doc/EvaluatingPhenotypeAlgorithms.pdf>

- A adjudicação clínica através da verificação do registro de origem e o PheValuator representam duas abordagens alternativas para estimar a validação da definição de coorte.
- Os estudos da rede OHDSI fornecem um mecanismo para examinar a heterogeneidade da fonte de dados e expandir a generalização dos achados para melhorar a validade clínica das evidências do mundo real.

5.6 Diagnósticos de coortes: Ferramenta de Análise de Coorte

Ferramentas como o diagnóstico de coortes que gera estatísticas agregadas das coortes, nos permitem avaliar a definição escolhida da coorte e se a mesma faz sentido para o tipo de estudo que queremos realizar num processo iterativo no qual aprendemos dos próprios dados e refinamos a definição do fenótipo ou o conjunto de conceitos incluídos.

Podemos imaginar então, um fluxo na definição e avaliação dos fenótipos que nos permita avançar a ciência da construção de fenótipos para incorporar um processo sistemático e disciplinado que possamos seguir para obter algum tipo de evidência empírica a respeito da definição de fenótipos que criamos é confiável o suficiente para ser utilizada para qualquer uma das análises que estamos tentando efetuar.

5.6.1 Pacote R CohortDiagnostics

O Cohort Diagnostics²² [OHDSI 2021] é um pacote OHDSI HADES [OHDSI 2021a]. O seu uso é considerado uma etapa de melhor prática recomendada antes de realizar um estudo de rede OHDSI. O software permite a tomada de decisão interativa, permitindo a comparação de uma ou mais definições de coorte para ideias clínicas semelhantes. Os usuários podem inferir variações introduzidas por escolhas de sensibilidade, especificidade e consistência em uma rede de fontes de dados.

O Cohort Diagnostics permite decisões como, a viabilidade de desenvolver definições de coorte para uma ideia clínica, melhoria das definições comparando o desempenho diagnóstico entre si e as expectativas a priori. Além disso, pode ser usado para gerar evidências com base descritiva.

Usando o Cohort Diagnostics, as inferências são feitas no nível da fonte de dados, nível de definição de coorte e nível de ID do conceito:

- O diagnóstico no nível da fonte de dados pode ser usado para inferir sobre a heterogeneidade da fonte de dados. Se os dados de heterogeneidade da fonte são observados, então os pesquisadores devem entender os dados de origem e determinar se esse entendimento pode introduzir limitações na interpretação dos resultados do estudo proposto.

²² <https://data.ohdsi.org/phenotypePhebruary/>

- Diagnósticos de nível de coorte: Estes são os principais diagnósticos e nos ajudam a inferir sobre as coortes, revisando e comparando características descritivas das coortes entre as fontes de dados.
- Diagnóstico de nível de ID de conceito: como conceitos órfãos, são quaisquer conceitos que parecem ter semelhança de substring com os conceitos nas coortes sendo diagnosticadas, mas não estão presentes na definição de coorte, mas parecem capturar a ideia clínica por trás da definição de coorte e têm contagens suficientes em uma ou mais fontes de dados. Se sim, este código foi perdido na coorte original, ou seja, a definição de coorte deve incluir esses conceitos.

A Figura 5.23 apresenta a interface Pacote R CohortDiagnostics.

The screenshot displays the CohortDiagnostics interface. On the left is a navigation menu with options like Cohort Definition, Concepts in Data Source, Orphan Concepts, Cohort Counts, Incidence Rate, Time Distributions, Index Event Breakdown, Visit Context, Cohort Overlap, Cohort Characterization, Temporal Characterization, Compare Cohort Char., Compare Temporal Char., and Data Source Information. The main area is titled 'Cohort Definition' and shows a table of cohort definitions. Below this, there are tabs for 'Details', 'Cohort Count', 'Cohort definition', 'Concept Sets', 'JSON', and 'SQL'. The 'Cohort Count' tab is active, showing a table with columns for Database Id, Cohort Subjects, and Cohort Entries. The table lists various database IDs and their corresponding subject and entry counts.

Database Id	Cohort Subjects	Cohort Entries
cpird_v1958_20220219	149,908	149,908
ims_australia_bpd_v1945_20220208	4,083	4,083
ims_france_v1793_202211005	1,622	1,622
ims_germany_v1894_20220208	52,051	52,051
jmdc_v1893_20220102	18,612	18,612
ims_us_open_claims_202109	8,129,829	8,129,829
truven_ccae_v1917_20220122	434,524	434,524
truven_mdcc_v1734_20210824	231,751	231,751
truven_mdcc_v1911_20220122	426,765	426,765

Figura 5.23 Interface do Pacote R CohortDiagnostics

Conseguimos melhorar a confiança em nossa definição de coorte usando o Cohort Diagnostics. A ferramenta reúne e explora evidências empíricas que são as principais métricas para entender diferentes elementos de implementação do fenótipo em várias fontes de dados. Os diagnósticos descritivos incluídos em níveis de dados, códigos e coortes podem ser usados para avaliar as compensações de erros de sensibilidade/especificidade. Esses insights empíricos podem ajudar os pesquisadores a identificar o impacto das escolhas que fazem para qualquer fenótipo. O uso do Cohort Diagnostics em estudos de rede OHDSI é uma prática recomendada.

5.7 Phenotype Phebruary 2022: Resumo do evento

Dentro da OHDSI, criamos a maior comunidade de ciência aberta do mundo para pesquisa observacional em saúde, com uma rede de dados distribuídos de mais de 300 bancos de dados harmonizados com um modelo de dados comum, representando coletivamente mais de 10% da população mundial.

Realizamos pesquisas metodológicas para avaliar e estabelecer as melhores práticas científicas para análise observacional. Desenvolvemos ferramentas analíticas de código aberto que possibilitam a realização de análises em larga escala para caracterização clínica, estimativa em nível populacional e previsão em nível de paciente em toda a nossa rede. Aplicamos nossas melhores práticas e ferramentas para gerar evidências confiáveis que tiveram impacto na vida de pacientes em todo o mundo. E, podemos expandir ainda mais nosso impacto desenvolvendo um sistema para gerar evidências que caracterizem a utilização de doenças e tratamentos, estimem os efeitos de intervenções médicas e prevejam os resultados dos pacientes em nossa rede de bancos de dados observacionais de saúde.

Mas para alcançar essa aspiração, temos que começar construindo uma base sólida. Precisamos desenvolver fenótipos para todos os resultados de saúde que desejamos investigar e avaliar esses algoritmos de fenótipo em nossa rede, para que possamos construir a confiança de que podemos interpretar os resultados de análises usando esses algoritmos adequadamente.

"Phenotype Phebruary" foi uma iniciativa de toda a comunidade para desenvolver e avaliar fenótipos para resultados de saúde que poderiam ser investigados pela comunidade.

Temos uma ampla gama de ferramentas de código aberto desenvolvidas pela comunidade para apoiar aspectos do processo de desenvolvimento e avaliação do fenótipo (ATLAS, CapR, PHOEBE, AFRODITE, CohortDiagnostics, PheValuator...). Agora é a hora de juntar tudo com um esforço da comunidade para construir um recurso da comunidade que possa apoiar todas as nossas atividades de análise na geração de evidências confiáveis.

Mas 'fenotipar todos os resultados de saúde' pode parecer um pouco avassalador. Por onde começamos? Durante a chamada da comunidade OHDSI, foi pedido à comunidade que compartilhasse sobre 'Quais fenótipos você gostaria de desenvolver e avaliar juntos?', e foi uma participação ativa de todos. Foram selecionados uma lista de alvos de fenótipo que receberam pelo menos 5 votos.

Proposta do evento Phenotype Phebruary: 28 dias, 28 fenótipos.

No mês de fevereiro foram realizadas discussões diárias sobre os esforços de validação do padrão-ouro para realmente aprofundar as definições dos fenótipos selecionados. Com certeza, a heterogeneidade significativa do banco de dados pode ser explicada devido à coleta dos dados e devemos manter a proveniência dos dados próximos às equipes de pesquisa.

Dado um alvo de fenótipo, criar uma descrição clínica, revisar trabalhos anteriores, desenvolver uma definição de coorte usando ferramentas OHDSI (como PHOEBE, ATLAS, AFRODITE), avaliar usando ferramentas OHDSI (como CohortDiagnostics, PheValuator), escrever um resumo das constatações.

Em resumo:

- Construir as definições de coorte e disponibilizá-las no OHDSI Phenotype Library (começar com a descrição clínica, revisar o trabalho anterior, desenvolver definições de coorte)
- Executar diagnósticos de coorte em uma rede de fontes de dados <https://github.com/ohdsi-studies/PhenotypePhebruary>
- Publicar os resultados em um site disponível ao público <https://data.ohdsi.org/phenotypePhebruary/>
- Revisar os resultados e avaliá-los, escrever resumos nos Fóruns OHDSI

Recursos Phenotype Phebruary

- <https://www.ohdsi.org/phenotype-phebruary/>
- <https://atlas-phenotype.ohdsi.org/>
 - ATLAS instance for OHDSI Phenotype Development WG to share cohort definitions
 - Want to get read access to this ATLAS instance?
Fill out form here: <https://forms.gle/6fxcZFyufhL39pLj7>
- <https://data.ohdsi.org/phenotypePhebruary/>
 - CohortDiagnostics instance that we'll be results each day from Phenotype Phebruary evaluations <https://data.ohdsi.org/phenotypePhebruary/>
- <https://github.com/ohdsi-studies/PhenotypePhebruary>
 - Git repository where we can share code to run CohortDiagnostics

5.8 Considerações finais e conclusões

A missão da OHDSI é gerar evidências de alta qualidade por meio de pesquisa observacional. Uma maneira de conseguir isso é através de estudos de pesquisa colaborativa. A comunidade OHDSI criou padrões e ferramentas para facilitar a pesquisa reprodutível e de alta qualidade, incluindo os vocabulários padrões, Modelo Comum de Dados (CDM), pacotes de métodos analíticos, e o ATLAS para executar estudos retrospectivos nos bancos de dados. Os estudos de rede OHDSI representam uma maneira transparente, consistente e reproduzível de conduzir pesquisas em muitos dados geograficamente dispersos.

Embora uma coorte possa ser bem definida e totalmente avaliada no contexto de um determinado banco de dados observacional, a validade clínica é limitada pela extensão em que os resultados são considerados generalizáveis para a população-alvo de interesse. Múltiplos estudos observacionais sobre o mesmo tópico podem produzir resultados diferentes, que podem ser causados não apenas por seus desenhos e métodos analíticos, mas também pela escolha da fonte de dados. Madigan et al. [Madigan 2013] demonstraram que a escolha do banco de dados afeta o resultado do estudo observacional. Eles investigaram sistematicamente a heterogeneidade nos resultados para 53 pares de drogas-desfecho e dois delineamentos de estudo (estudos de coorte e séries de casos autocontrolados) em dez bancos de dados observacionais. Mesmo mantendo o desenho do estudo constante, observou-se uma heterogeneidade substancial nas estimativas de efeito.

Em toda a rede OHDSI, os bancos de dados observacionais variam consideravelmente nas populações que representam (por exemplo, pediátricos versus idosos, funcionários com seguro privado versus desempregados com seguro público), os ambientes de atendimento onde os dados são capturados (por exemplo, pacientes internados versus ambulatoriais), os processos de captura de dados (por exemplo, registros clínicos eletrônicos de atendimentos) e o sistema de saúde nacional e regional no qual os cuidados são baseados. Essas diferenças podem se manifestar como heterogeneidade observada ao estudar a doença e os efeitos das intervenções médicas e podem influenciar a confiança que temos na qualidade de cada fonte de dados que pode contribuir com evidências em um estudo em rede. Embora todos os bancos de dados dentro da rede OHDSI sejam padronizados para o CDM, é importante reforçar que a padronização não reduz a verdadeira heterogeneidade inerente que está presente nas populações, mas simplesmente fornece uma estrutura consistente para investigar e entender melhor a heterogeneidade em toda a rede. A rede de pesquisa OHDSI fornece o ambiente para aplicar o mesmo processo analítico em vários bancos de dados em todo o mundo, para que os pesquisadores possam interpretar os resultados em várias fontes de dados, mantendo os aspectos metodológicos constantes.

Através destes estudos foi levantada a importância de um tema que tinha ficado relegado a um segundo plano, a definição dos fenótipos. A confiabilidade das evidências que geramos muitas vezes vive e morre pela qualidade dos fenótipos que usamos como entradas de indicações, exposições, resultados e outras características que colocamos em nossas análises. E, no entanto, em todo o empreendimento de pesquisa mais amplo, a ciência do desenvolvimento e avaliação do fenótipo é relativamente imatura.

Para superar a deficiência neste assunto, a comunidade OHDSI fez um esforço concentrado em um mês para atacar o desafio de implementar 28 fenótipos em 28 dias.

O mundo ainda não possui melhores práticas consensuais para projetar fenótipos, não possui ferramentas padronizadas acordadas para construir fenótipos, não possui métodos consistentes e reprodutíveis para avaliar fenótipos. Nossos fenótipos estão repletos de erros de medição substanciais; sabemos que provavelmente temos sensibilidade, especificidade e valor preditivo positivo abaixo do ideal, mas não estimamos consistentemente o erro de medição e ainda mais raramente integramos o erro de medição em nossas análises. Nenhuma regressão em larga escala ou modelo sofisticado de aprendizado profundo poderá compensar o fato que as nossas análises se baseiam em fenótipos suspeitos.

Esta é uma tremenda lição com a qual todos podemos aprender: às vezes você tem que considerar que uma de suas opções é decidir que os dados não são apropriados para a questão de interesse e o melhor caminho a seguir pode ser parar.

Uma boa demonstração do importante trabalho que a OHDSI está fazendo para fornecer ferramentas e métodos confiáveis que nos permitam aproveitar da melhor forma possível as descobertas que se escondem na selva de informações que hoje compõem o nosso mundo.

5.9 Referências bibliográficas

Abrahão M T; Nobre, M R C; Madril, P J; O estado da arte em pesquisa observacional

- de dados de saúde: A iniciativa OHDSI. In: Artur Ziviani; Natalia Castro Fernandes; Débora Christina Muchaluat Saade. (Org.). Livro de Minicursos do 19o Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2019). 19o ed. Porto Alegre: Sociedade Brasileira de Computação-SBC, (2019), I SBN-13 (15) 978-85-7669-472-4, v. 1, p. 141-189.
- Banda, J. M., Y. Halpern, D. Sontag, and N. H. Shah. 2017. “Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network.” *AMIA Jt Summits Transl Sci Proc* (2017): 48–57.
- Hripcsak George, Albers David J, High-fidelity phenotyping: richness and freedom from bias, *Journal of the American Medical Informatics Association*, Volume 25, Issue 3, March (2018), Pages 289–294, <https://doi.org/10.1093/jamia/ocx110>
- Hripcsak, G., N. Shang, P. L. Peissig, L. V. Rasmussen, C. Liu, B. Benoit, R. J. Carroll, et al. (2019). “Facilitating phenotype transfer using a common data model.” *J Biomed Inform*, July, 103253.
- Johannsen, W. (1909). *Elemente der exakten Erblchkeitslehre [Elements of the exact theory of heredity]* (in German). Jena, Germany: Gustav Fischer
- Lane, Jennifer C E, James Weaver, Kristin Kostka, Talita Duarte-Salles, Maria Tereza F Abrahao, Heba Alghoul, Osaid Alser, et al. (2020). “Risk of Hydroxychloroquine Alone and in Combination with Azithromycin in the Treatment of Rheumatoid Arthritis: A Multinational, Retrospective Study.” *The Lancet Rheumatology* 2 (11): e698–711. [https://doi.org/10.1016/S2665-9913\(20\)30276-9](https://doi.org/10.1016/S2665-9913(20)30276-9).
- Li, Xintong, Anna Ostropelets, Rupa Makadia, Azza Shoaibi, Gowtham Rao, Anthony G. Sena, Eugenia Martinez-Hernandez, et al. (2021). “Characterising the Background Incidence Rates of Adverse Events of Special Interest for Covid-19 Vaccines in Eight Countries: Multinational Network Cohort Study.” *BMJ* 373 (June): n1435. <https://doi.org/10.1136/bmj.n1435>.
- Madigan, D., P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, and J. A. Berlin. (2013). “Evaluating the impact of database heterogeneity on observational study results.” *Am. J. Epidemiol.* 178 (4): 645–51.
- OHDSI Observational Health Data Sciences and Informatics. (2021). Cohort Diagnostics. Cohort Diagnostics: An R package for performing various cohort diagnostics. Retrieved June 15, 2021, from <https://ohdsi.github.io/CohortDiagnostics/>
- OHDSI Observational Health Data Sciences and Informatics. (2021 a). HADES. Health Analytics Data-toEvidence Suite (HADES): A collection of R packages for performing analytics against the Common Data Model. Retrieved June 15, 2021, from <https://ohdsi.github.io/Hades/>

OHDSI <https://www.ohdsi.org/>

OHDSI. The book of OHDSI <https://github.com/OHDSI/TheBookOfOhdsi> Este livro está licenciado sob Creative Commons Zero v1.0 Universal.

Olsen LA, Aisner D, McGinnis JM, editors. Institute of Medicine (US) Roundtable on Evidence-Based Medicine; The Learning Healthcare System: Workshop Summary. Washington (DC): National Academies Press (US); (2007). Institute of Medicine Roundtable on Evidence-Based Medicine. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK53483/>

Suchard, Marc A., Martijn J. Schuemie, Harlan M. Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G. Reich, et al. (2019). “Comprehensive Comparative Effectiveness and Safety of First-Line Antihypertensive Drug Classes: A Systematic, Multinational, Large-Scale Analysis.” *The Lancet* 394 (10211): 1816–26. [https://doi.org/10.1016/S0140-6736\(19\)32317-7](https://doi.org/10.1016/S0140-6736(19)32317-7).

Swerdel, Joel N et al. “PheValuator: Development and evaluation of a phenotype algorithm evaluator.” *Journal of biomedical informatics* vol. 97 (2019): 103258. [doi:10.1016/j.jbi.2019.103258](https://doi.org/10.1016/j.jbi.2019.103258)

Unesco (2021), Draft recommendation on Open Science on its way to final adoption <https://en.unesco.org/news/draft-recommendation-open-science-its-way-final-adoption>

Williams, R.D., Markus, A.F., Yang, C. et al. “Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network.” *BMC Med Res Methodol* 22, 35 (2022). <https://doi.org/10.1186/s12874-022-01505-z>