Maria Tereza Fernandes Abrahão

Método de extração de coortes em bases de dados assistenciais para estudos da doença cardiovascular

Tese apresentada à Faculdade de Medicina da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Programa de Cardiologia

Orientador: Prof. Dr. Marco Antônio Gutierrez

São Paulo

2016

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

Abrahão, Maria Tereza Fernandes
Método de extração de coortes em bases de dados assistenciais para estudos da doença cardiovascular / Maria Tereza Fernandes Abrahão. -- São Paulo, 2016.
Tese(doutorado)--Faculdade de Medicina da Universidade de São Paulo.

Programa de Cardiologia.
Orientador: Marco Antônio Gutierrez.

Descritores: 1.Informática médica 2.Sistemas de informação hospitalar 3.Registros eletrônicos de saúde 4.Estudos de coortes 5.Estudos retrospectivos 6.Mineração de dados

USP/FM/DBD-116/16

Dedico esta tese a Pablo Jorge Madril, meu marido, que fez o sonho possível, trilhando comigo esse longo caminho com muito amor e compreensão.

Agradecimentos

Ao meu orientador Prof. Dr. Marco Antônio Gutierrez, pelos desafios propostos que contribuíram para o engrandecimento desse trabalho, pela competência acadêmica que conduziu essa orientação e pelas diversas horas dedicadas a realização desse trabalho.

Ao meu "sempre orientador" Prof. Dr. Moacyr Roberto Cuce Nobre, principal responsável e incentivador do meu ingresso no programa de pós-graduação em cardiologia da FMUSP.

Ao Dr. Fabio Antero Pires, responsável pela minha inclusão na era do "big data", pela confiança, auxílio e disponibilização de dados fundamentais para a realização desse trabalho.

Aos professores José Carlos Nicolau, Antônio Mansur e Raul Dias dos Santos Filho, pelas importantes sugestões apresentadas a este trabalho.

Aos professores e funcionárias do programa de pós-graduação em cardiologia da FMUSP e a comissão científica, sempre eficientes e disponíveis, facilitando os processos burocráticos.

À Rachel Zanetta, pela amizade, presença constante e ajuda competente em todos os momentos.

Ao grupo de Pesquisa e Desenvolvimento do InCor, Admar Longo, Ramon Moreno e Marina Rebelo, pelo espaço cedido e apoio para a realização deste trabalho.

À amiga Teresa Vieira, meu muito obrigada por todos os "problemas" resolvidos, espaços liberados, "matança" de processos, dicas e tudo mais...

Ao amigo Fabiano Matos, competente "garoto de programa", que muito me ajudou na busca das tabelas e dados do SI^3 e nas dúvidas nos comandos *SQL* utilizados para a implementação do método descrito nessa tese.

Aos meus grandes amigos Maysa Macedo e John Sims, pelas longas horas de discussão que contribuíram para enriquecer o conteúdo desta tese, pela leitura crítica, sugestões, correções, traduções, conselhos, apoio, ombro amigo...

Aos amigos Alice Bacici e Anderson Santiago, que juntos com John e Maysa fizeram do "café da tarde" sempre a melhor hora do dia. Obrigada a todos pelas tardes de papos "coxinha", risadas, apoio e contribuições.

Aos colegas de doutorado, ao grupo de pesquisa do Dr. Moacyr, ao amigo Valter Garcia, a todos, pelas longas horas de discussão, engrandecimento e conhecimentos compartilhados.

Agradeço aos 1.116.848 pacientes atendidos no InCor, os quais debulhei e vasculhei anos a fio seus dados. Aos vivos, que Deus lhes deem conforto de seus males e os que foram a óbito, o descanso eterno.

Aos médicos com quem tive a honra de conviver por estes anos, pelos ensinamentos e lições de conduta e vida, bem mais que conceitos de doenças e tratamentos, mas modelos de humanidade e compaixão pela dor humana.

Enfim, agradeço aos amigos do Serviço de Informática e outros setores do Instituto do Coração que me acolheram, apoiaram e incentivaram na realização desse trabalho, tornando os anos passados aqui inesquecíveis... esse texto é para cada um de vocês.

A Tese no coração!

Um final e um recomeço…

A Tese está aí chegando ao fim, carregada de tabelas, gráficos e textos. Mas o seu significado é o que realmente importa. Em suas páginas ficaram registrados momentos inesquecíveis e pessoas especiais que vão estar para sempre dentro do meu coração. Muito caminhei e muito aprendi. Bem mais que sobre base de dados, doenças do coração e medicamentos. Aprendi que amar cura e nos permite limpar a alma e transformar as relações humanas. O coração é um órgão imenso, onde o amor corre e pulsa, com válvulas que se abrem e fecham ficando com o que de melhor passar e chegar… E nesses anos no InCor, chegaram… novos e grandes amigos se juntaram em meu caminho e para sempre vão estar no meu coração…

Além de um texto, o desafio maior é deixar lembranças e marcas pelo caminho que possam servir a quem precise por ele passar…

Agradecimentos Especiais

À Deus, força maior sem a qual nada na vida seria possível.

A todos da minha família e meus "amigos família", os que estão por perto e os que longe estão, mas seguramente juntos no coração, pelo amor, confiança e grande torcida para que mais essa fase da minha vida fosse concluída.

Em especial, agradeço a minha irmã Vera Cristina e meu cunhado Jorge Luiz, sempre presentes, participando ativamente e compartilhando todos os momentos.

A minha querida tia Nair, no topo dos seus 86 anos, que mesmo de coração remendado, continua sua jornada com jovialidade, alegria de viver, amor e muitas e muitas "lições de vida".

Meu mais sincero agradecimento e amor à Pablo, autor maior na realização de todos os meus sonhos, que lutou de capa e espada com dragões e ogros pra me ter ao seu lado. E aqui eu estou e aqui quero ficar. Com você todo lugar é o melhor lugar do mundo e todo dia é dia do "porque te amo"...

Aos "nossos" filhos, frutos da união dos nossos corações, nosso melhor da vida e nossas maiores realizações: Érika, Juliana, Ruth, Ariel, Marilia e ainda os genros filhos: Mario Augusto e Vilmar, agradeço pela imensa paciência e tempos difíceis de espera, mas sempre na torcida. Curitiba, Manaus e Toronto, lá vamos nós...

Enfim, aos "nossos" netos: Gabriel, José Renato, Maria Eduarda, Guilherme Augusto, Leonardo, Ana Júlia, Natalie e Caio. Meus amores e fontes inesgotáveis de força no meu viver... Agradeço a recepção a cada chegada, sempre com um grande e apertado abraço e muito amor, sem nunca terem me esquecido, a vovó que vai e vem... Agora a vovó fica e antes que perguntem, o Pablo também.

A vocês, desejo que tenham tempo para brincar, sonhar, recordar, ler, ouvir estórias e em essência, sempre serem crianças. Tempo para que sejam o melhor que houver em cada um de vocês na busca e realização de seus próprios sonhos...

Vovó Tetê esta aqui e vocês sempre vão estar no meu coração.

Esta tese está de acordo com as seguintes normas, em vigor no momento desta publicação:

Referências: adaptado de *International Committee of Medical Journals Editors* (Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação. Guia de apresentação de dissertações, teses e monografias.
Elaborado por Anneliese Carneiro da Cunha, Maria Julia de A. L. Freddi, Maria F. Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3a ed. São Paulo: Divisão de Biblioteca e Documentação; 2011.

Abreviaturas dos títulos dos periódicos *List of Journals Indexed in Index Medicus.*

SUMÁRIO

Lista de tabelas

Lista de figuras

Lista de gráficos

Lista de siglas

CFM	Conselho Federal de Medicina
CID-10	Classificação Internacional de Doenças - 10º revisão
DATASUS	Departamento de Informática do SUS
DCbV	Doença Cerebrovascular
DCV	Doença Cardiovascular
DIC	Doença Isquêmica do Coração
DM	Diabetes Mellitus
FMUSP	Faculdade de Medicina da Universidade de São Paulo
HAS	Hipertensão arterial sistêmica
HC FMUSP	Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo
HDL	Lipoproteína de alta densidade (*High Density Lipoproteins*)
IAM	Infarto Agudo do Miocárdio
IC	Intervalo de confiança
InCor	Instituto do Coração
ISPOR	*International Society for Pharmacoeconomics and Outcomes Research*
LDL	Lipoproteína de baixa densidade (*Low Density Lipoproteins*)
MS	Ministério da Saúde
OMOP	*Observational Medical Outcomes Partnership*
OMS	Organização Mundial de Saúde
RES	Registro Eletrônico de Saúde
SBIS	Sociedade Brasileira de Informática em Saúde
SI^3	Sistema Integrado de Informações Institucionais
SIGFar	Sistema de Dispensação de Medicamentos da Farmácia
SQL	*Structured Query Language*
SUS	Sistema Único de Saúde

RESUMO

Abrahão MT. *Método de extração de coortes em bases de dados assistenciais para estudos da doença cardiovascular* [tese]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2016.

A informação coletada de prontuários manuais ou eletrônicos, quando usada para propósitos não diretamente relacionados ao atendimento do paciente, é chamado de uso secundário de dados. A adoção de um sistema de registro eletrônico em saúde (RES) pode facilitar a coleta de dados para uso secundário em pesquisa, aproveitando as melhorias na estruturação e recuperação da informação do paciente, recursos não disponíveis nos tradicionais prontuários em papel. Estudos observacionais baseados no uso secundário de dados têm o potencial de prover evidências para a construção de políticas em saúde. No entanto, a pesquisa através desses dados apresenta problemas característicos a essa fonte de dados. Ao longo do tempo, os sistemas e seus métodos de armazenar dados se tornam obsoletos ou são reestruturados, existem questões de privacidade para o compartilhamento dos dados dos indivíduos e questões relacionadas ao uso desses dados em um contexto diferente do seu propósito original. É necessária uma abordagem sistemática para contornar esses problemas, onde o processamento dos dados é efetuado antes do seu compartilhamento. O objetivo desta Tese é propor um método de extração de coortes de pacientes para estudos observacionais contemplando quatro etapas: (1) mapeamento: a reorganização de dados a partir de um esquema lógico existente em um esquema externo comum sobre o qual é aplicado o método; (2) limpeza: preparação dos dados, levantamento do perfil da base de dados e cálculo dos indicadores de qualidade; (3) seleção da coorte: aplicação dos parâmetros do estudo para seleção de dados longitudinais dos pacientes para a formação da coorte; (4) transformação: derivação de variáveis de estudo que não estão presentes nos dados originais e transformação dos dados longitudinais em dados anonimizados prontos para análise estatística e compartilhamento. O mapeamento é uma etapa específica para cada RES e não é objeto desse trabalho, mas foi realizada para a aplicação do método. As etapas de limpeza, seleção de coorte e transformação são comuns para qualquer RES. A utilização de um esquema externo possibilita o uso parâmetros que facilitam a extração de diferentes coortes para diferentes estudos sem a necessidade de alterações nos algoritmos e garante que a extração seja efetuada sem perda de informações por um

processo idempotente. A geração de indicadores e a análise estatística fazem parte do processo e permitem descrever o perfil e qualidade da base de dados e os resultados do estudo. Os algoritmos computacionais e os dados são disponibilizados em um repositório versionado e podem ser usados a qualquer momento para reproduzir os resultados, permitindo a verificação, alterações e correções de erros. Este método foi aplicado no RES utilizado no Instituto do Coração – HC FMUSP, considerando uma base de dados de 1.116.848 pacientes cadastrados no período de 1999 até 2013, resultando em 312.469 registros de pacientes após o processo de limpeza. Para efetuar uma análise da doença cardiovascular em relação ao uso de estatinas na prevenção secundária de eventos evolutivos, foi constituída uma coorte de 27.915 pacientes, segundo os seguintes critérios: período de 2003 a 2013, pacientes do gênero masculino e feminino, maiores de 18 anos, com um diagnóstico no padrão CID-10 (códigos I20 a I25, I64 a I70 e G45) e com registro de no mínimo duas consultas ambulatoriais. Como resultados, cerca de 80% dos pacientes tiveram registro de estatinas, sendo que, 30% tiveram registro de estatinas por mais de 5 anos, 42% não tiveram registro de nenhum evento evolutivo e 9,7% tiveram registro de dois ou mais eventos. O tempo médio de sobrevida calculado pelo método *Kaplan-Meier* foi de 115 meses (intervalo de confiança 95% 114-116) e os pacientes sem registro de estatinas apresentaram uma maior probabilidade de óbito pelo teste *log-rank* $p<0,001$. Conclui-se que a adoção de métodos sistematizados para a extração de coortes de pacientes a partir do RES pode ser uma abordagem viável para a condução de estudos epidemiológicos.

Descritores: 1.Informática Médica 2.Sistemas de Informação Hospitalar 3.Registros Eletrônicos de Saúde 4.Estudos de Coortes 5.Estudos Retrospectivos 6.Mineração de Dados

ABSTRACT

Abrahão MT. *A method for the cohort selection of cardiovascular disease records from an electronic health record system* [thesis]. São Paulo: "Faculdade de Medicina, Universidade de São Paulo"; 2016.

Information collected from manual or electronic health records can also be used for purposes not directly related to patient care delivery, in which case it is termed secondary use. The adoption of electronic health record (EHR) systems can facilitate the collection of this secondary use data, which can be used for research purposes such as observational studies. These studies have the power to provide necessary evidence for the formation of healthcare policies. However, several problems arise when conducting research using this kind of data. For example, over time, systems and their methods of storing data become obsolete, data concerns arise since the data is being used in a different context to where it originated and privacy concerns arise when sharing data about individual subjects. To overcome these problems a systematic approach is required where local data processing is performed prior to data sharing. The objective of this thesis is to propose a method to extract patient cohorts for observational studies in four steps: (1) data mapping from an existing local logical schema into a common external schema over which information can be extracted; (2) cleaning of data, generation of the database profile and retrieval of indicators; (3) computation of derived variables from original variables; (4) application of study design parameters to transform longitudinal data into anonymized data sets ready for statistical analysis and sharing. Mapping is a specific stage for each EHR and although it is not the focus of this work, a detail of the mapping is included. The stages of cleaning, selection of cohort and transformation are common to all EHRs and form the main objective. The use of an external schema allows the use of parameters that facilitate the extraction of different cohorts for different studies without the need for changes to the extraction algorithms. This ensures that, given an immutable dataset, the extraction can be done by the idempotent process. The generation of indicators and statistical analysis form part of the process and allow profiling and qualitative description of the database. The set extraction / statistical processing is available in a version controlled repository and can be used at any time to reproduce results, allowing the verification of alterations and error corrections. The method was applied to EHR from the Heart Institute - HC

FMUSP, with a dataset containing 1,116,848 patients' records from 1999 up to 2013, resulting in 312,469 patients records after the cleaning process. An analysis of cardiovascular disease in relation to statin use in the prevention of secondary events was defined using a cohort selection of 27,915 patients with the following criteria: study period: 2003-2013, gender: Male, Female, age: ≥ 18 years old, at least 2 outpatient visits, diagnosis of CVD (ICD-10 codes: I20-I25, I64-I70 and G45). Results showed that around 80% of patients had a prescription for statins, of which 30% had a prescription for statins for more than 5 years. 42% had no record of a future event and 9,7% had two or more future events. Survival time was measured using a univariate Kaplan-Meier method resulting in 115 months (CI 95% 114-116) and patients without statin prescription showed a higher probability of death when measured by log-rank ($p<0.001$) tests. The conclusion is that the adoption of systematised methods for cohort extraction of patients from EHRs can be a viable approach for conducting epidemiological studies.

Descriptors: 1.Medical Informatics 2.Hospital Information Systems 3.Electronic Health Records 4.Cohort Studies 5.Retrospective Studies 6. Data Mining

1 INTRODUÇÃO

Os estudos epidemiológicos são desenhados considerando as teorias de prevalência, as medidas de fatores de risco e os custos para obtenção de dados relevantes. Até os anos 1950, pesquisadores consideravam as estatísticas vitais para conduzir estudos transversais e séries temporais em doenças não infecciosas. Consequentemente, a ausência de dados longitudinais limitavam as inferências causais nesses estudos.

Na segunda metade do Século XX, o financiamento de pesquisas permitiu aos pesquisadores o desenvolvimento de coortes de indivíduos, os quais passaram a ser seguidos ao longo de um período de tempo. Nesse contexto, o estudo de *Framingham* (*Framingham Heart Study*) foi um dos pioneiros no campo da epidemiologia cardiovascular. Um estudo de várias gerações, ao longo de mais de cinco décadas de acompanhamento de um grupo de participantes que no início do estudo não haviam desenvolvido sintomas evidentes de doença cardiovascular (DCV), fez uma contribuição significativa para a prevenção da doença nos Estados Unidos e indiretamente, tem influenciado estratégias globais de prevenção e tratamento. O Estudo de *Framingham* forneceu conhecimento sobre a prevalência, incidência, prognóstico, fatores predisponentes, e determinantes da DCV. O conceito fator de risco, fundamental para a prevenção de doenças cardiovasculares, originou-se do estudo de *Framingham*. Isso gerou descobertas seminais, tais como os efeitos do uso do tabaco, dieta pouco saudável, sedentarismo, obesidade, altas taxas de colesterol no sangue, pressão arterial elevada e diabetes na DCV (Oppenheimer, 2005; Mendis, 2010).

Entretanto, na última década, observa-se um declínio no financiamento de pesquisas e na taxa de participação de indivíduos nas coortes (Galea e Tracy, 2007), o que têm limitado a condução dos tradicionais longos e custosos estudos prospectivos.

Por outro lado, observa-se também nesse período o crescimento na adoção e uso secundário dos dados (o uso de dados para outros fins diferente do seu propósito original) de registros eletrônicos em saúde (RES), os quais oferecem uma nova alternativa na condução de estudos epidemiológicos. Os bancos de dados produzidos a

partir do RES, denominados bases de dados assistenciais, têm como vantagens o baixo custo e a possibilidade de utilizar amostras maiores, proporcionando maior força para detectar pequenas diferenças ou eventos raros, dispensando o contato direto com o paciente. Possibilita também, maior diversidade metodológica, permitindo que modificações no estudo possam ser implementadas diferentemente dos projetos de pesquisa de maior rigor de protocolos, como os ensaios randomizados. As desvantagens estão relacionadas à falta da padronização na coleta dos dados, que afeta a qualidade dos dados registrados, a cobertura que pode variar no tempo e a falta de informações que podem ser importantes para as análises de interesse, incluindo variáveis de desfecho, explicativas, mediadoras, de confusão ou modificadoras de efeito (Carrilho et al., 2010).

Além desses aspectos, a adoção do RES não envolve apenas uma versão digital do tradicional registro em papel (Wager et al., 2009), tais registros podem ser combinados com dados contextuais, por exemplo, através de sistemas de georeferenciamento para o estabelecimento de redes de casualidade (Goncalves Sa et al., 2012). Tais recursos têm possibilitado o surgimento de uma nova abordagem e uma evolução para os estudos epidemiológicos, utilizando as tecnologias disponíveis no Século XXI (Lauer, 2012; Mooney et al., 2015).

Na Europa, para diminuir a distância entre a prática farmacêutica e a pesquisa de desfechos clínicos, instituições como a *International Society for Pharmacoeconomics and Outcomes Research* (ISPOR) foram criadas para definir guias de recomendações para estudos comparativos de efetividade, e consequentemente, acabaram criando critérios de validação para bases de dados assistenciais nos quais estes estudos se fundamentam (Motheral et al., 2003; Cox et al., 2009; Johnson et al., 2009; Berger et al., 2014).

Ao mesmo tempo, nos Estados Unidos, a necessidade de vigilância farmacológica adequada tem impulsionado a definição de modelos comuns que permitam coletar dados de diversas fontes de informação e aplicar análises estatísticas padronizadas (Madigan e Ryan, 2011; Madigan et al., 2013; Hripcsak et al., 2015). Em particular, a *Pharmaceutical Research and Manufacturers of America* (PhRMA), o *Food and Drug Administration* (FDA) e a *Foundation for the National Institutes of Health* (FNIH) criaram a *Observational Medical Outcomes Partnership* (OMOP), que consiste em uma parceria público-privada. Este grupo interdisciplinar tem como objetivo identificar os

métodos mais confiáveis para analisar grandes volumes de dados coletados de fontes heterogêneas resultando na criação de um modelo comum de dados (OMOP, 2016).

A origem das informações dos grupos de pesquisas da ISPOR e OMOP, foram bases de dados assistenciais e ambos têm enfrentado os mesmos tipos de problemas, sendo necessário criar um modelo comum para extração de dados juntamente com indicadores e critérios de validação da qualidade dos mesmos.

No Brasil, a crescente preocupação com a qualidade da informação médica tem levado à definição e adoção de padrões para os sistemas de informação em saúde (DATASUS, 2016). Os atendimentos à saúde no Brasil são realizados por entidades públicas e privadas. No entanto, de um modo geral, a população brasileira é atendida pelo Sistema Único de Saúde (SUS) em sua rede própria e em serviços privados contratados pelo governo. No ano de 2013 foram realizadas cerca de 11 milhões de internações (TabNet - Internações) e 4 bilhões de atendimentos ambulatoriais (TabNet - Ambulatorial). Isto torna o SUS um setor predominante no cenário de informações médicas e muitas vezes as suas ações definem diretrizes que acabam influenciando todos os setores.

Entre alguns dos padrões definidos e adotados pelo SUS, podemos citar o Cadastro Nacional de Estabelecimentos de Saúde (CNES) e o Cartão Nacional de Saúde (CNS) que visa identificar univocamente o cidadão brasileiro no SUS.

A Agência Nacional de Saúde Suplementar (ANS) elaborou uma norma nacional relacionada ao intercâmbio eletrônico de informação, conhecida como padrão TISS (Troca de Informação em Saúde Suplementar), que define o padrão para a troca de informação sobre o atendimento prestado aos beneficiários, entre operadoras de plano privado e prestadores de serviços em saúde.

Dentro das mais recentes ações na área da saúde, a Sociedade Brasileira de Informática em Saúde (SBIS), em parceria com o Conselho Federal de Medicina (CFM), elaborou o processo de certificação para sistemas de Registro Eletrônico de Saúde (RES) no qual, com o estabelecimento dos requisitos do manual da certificação, garante o armazenamento seguro das informações dos pacientes nos prontuários eletrônicos (Silva, 2011).

Podemos assim constatar, que os sistemas informatizados têm evoluído e as novas bases de dados contêm estruturas mais adequadas para pesquisa clínica. As bases assistenciais são fonte de dados para uso secundário que pela sua natureza favorecem as

pesquisas reprodutíveis. Em particular, diversos estudos prospectivos e retrospectivos já são elaborados usando as informações nelas contidas. As bases de dados históricas se constituem como uma fonte de informação importante para pesquisa de desfechos, porém apresentam uma série de desafios metodológicos peculiares a este tipo de fonte de dados (Berger et al., 2014).

Apresenta-se nesta Tese uma proposta de metodologia para construção de uma coorte, a partir de uma base de dados assistencial, para uso em estudos observacionais retrospectivos. Para tanto, utiliza-se o mapeamento das informações da base de origem para um modelo externo, sobre o qual são aplicados métodos sistemáticos de limpeza e tratamento de dados.

O método de extração de coorte se fundamenta nos critérios de reprodutibilidade e é validado pela aplicação na base de dados assistenciais do Instituto do Coração (InCor) do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC FMUSP), um hospital terciário de referência em doença cardiovascular, na cidade de São Paulo, através do estudo da doença cardiovascular (DCV) e a medicação estatina.

1.1 Objetivos

1.1.1 Objetivo principal

Propor uma metodologia para extração de coorte em uma base de dados assistencial, que possibilite estudos observacionais da DCV.

1.1.2 Objetivos secundários

1. Atender os critérios de uma pesquisa reprodutível, fornecendo dados anonimizados para serem disponibilizados em um repositório;

2. Selecionar e padronizar, a partir do conjunto de dados assistenciais, um conjunto de variáveis que atendam as necessidades de uma pesquisa retrospectiva da DCV, e que possam ser estendidas para o estudo de outras doenças correlacionadas;

3. Criar indicadores do perfil da base assistencial em relação aos diagnósticos mais frequentes, pacientes/ano de registro, entre outros, para caracterizá-la e possibilitar comparações com outras bases assistenciais;

4. Descrever as características clínico-demográfica da coorte resultante da aplicação do método e os resultados da análise estatística;

5. Disponibilizar uma base de dados assistencial, limpa e formatada que possibilite, a partir da aplicação de critérios de seleção de coortes, realizar estudos observacionais retrospectivos sobre doenças.

1.2 Organização do texto

Este texto está organizado da seguinte forma:

- No capítulo 2 é apresentada uma revisão abordando, os tipos de pesquisas, a importância dos estudos observacionais, os estudos observacionais realizados no Brasil, os estudos observacionais na base de dados do InCor, as vantagens e desvantagens dos estudos observacionais. Apresentamos os critérios para a reprodutibilidade da pesquisa clínica, para a preparação da base de dados, os critérios para estudos observacionais retrospectivos e uma proposta de um modelo comum de dados. Por fim, serão apresentados os esforços para definição e adesão a padrões nacionais em sistemas de saúde;
- No capítulo 3 são apresentadas as estruturas das bases de dados assistenciais utilizadas no InCor, especialmente a gerada pelo Sistema Integrado de Informações Institucionais (SI^3), um breve histórico do desenvolvimento do SI^3, a incorporação de funcionalidades no decorrer do tempo, os modelos e arquitetura desse sistema;
- No capítulo 4 são apresentados os critérios que foram seguidos na definição e preparação do método e uma contextualização sobre o modelo relacional e os níveis de abstração do banco de dados relacional;
- No capítulo 5 apresentamos uma descrição da metodologia utilizada para a elaboração do método de extração de coortes, descrevendo as fases de preparo (mapeamento, limpeza e transformação), a definição dos parâmetros e extração da coorte;
- No capítulo 6 apresentamos a aplicação da metodologia de extração de coortes na base de dados assistenciais do InCor. Como resultados, foram gerados indicadores para a caracterização dessa base e um estudo a respeito da DCV e o uso de estatina;
- No capítulo 7 discute-se o uso de base de dados assistenciais como fonte de dados de uso secundário para estudos observacionais e as inconsistências detectadas na aplicação da metodologia proposta;
- No capítulo 8 são apresentadas as conclusões desta Tese.

2 REVISÃO BIBLIOGRÁFICA

O objetivo em geral das pesquisas na área da saúde é aprender mais sobre um problema clínico de uma forma controlada. É organizar a observação clínica. A maneira de compreender um problema clínico é estudar as pessoas que o têm. Esses estudos auxiliam no melhor entendimento sobre a causa do problema e servem para testar e validar novos tratamentos para determinar se eles são melhores do que os estão disponíveis para utilização naquele momento. Estudar um grupo de pessoas com o mesmo problema podem ajudar cientistas a prever melhor se o tratamento irá funcionar para outras pessoas com a mesma doença. Existem diferentes tipos de estudos na área de saúde, e podemos citar: os ensaios clínicos, estudos que avaliam os resultados de novos tratamentos para determinar se eles são melhores do que os tratamentos já classicamente utilizados e os estudos observacionais, estudos que seguem as pessoas portadoras de uma doença por anos para aprender mais sobre a evolução da doença (Hulley et al., 2003).

2.1 Tipos de Pesquisas

A pesquisa clínica que estuda o valor de intervenções terapêuticas pode ser classificada em dois grupos. O primeiro grupo, chamado ensaio de eficácia, testa um tratamento sob condições ideais, em que a amostra de pacientes, a intervenção e as medidas de resultado estão definidas de forma minuciosa. A pergunta de pesquisa do tratamento busca a eficácia. Porém, os resultados de tais estudos poderão não ser reproduzidos quando o tratamento for aplicado a uma população de serviços de saúde em geral. O segundo grupo, chamado ensaios pragmáticos ou estudos de efetividade, são estudos observacionais que visam aproximar os resultados de eficácia encontrados nos ensaios clínicos à realidade da prática clínica. Efetividade refere-se ao desempenho de um tratamento em circunstâncias usuais ou na "vida real". Os ensaios pragmáticos foram definidos como "aqueles em que a hipótese e o desenho do estudo são formulados buscando produzir informações necessárias para a tomada de decisão". Em geral, esse desenho de estudo complementa os estudos de eficácia ao testar a utilidade

de um tratamento na vida real após a sua eficácia ter sido comprovada em estudos controlados (Hulley et al., 2003).

O estudo de coorte é um tipo de estudo observacional que se refere a um grupo de pessoas que tem alguma característica comum, constituindo uma amostra a ser acompanhada por certo período de tempo, para se observar e analisar o que acontece com elas. Podem ser conduzidos de dois modos distintos: prospectivos (concorrentes) ou retrospectivos (não-concorrentes ou históricos). O estudo de coorte prospectivo é elaborado no presente, com previsão de acompanhamento determinado, segundo o objeto do estudo. Propicia um planejamento criterioso que confere um rigor científico que o aproxima dos estudos de delineamento experimental. O estudo retrospectivo é elaborado com base em registros do passado com seguimento até o presente. Os estudos de coorte apresentam limitações, como a não utilização do critério de aleatoriedade na formação dos grupos e a necessidade de uma amostra significativamente grande (Gordis, 2014).

A Figura 2.1 ilustra a diferença entre um estudo hipotético de coorte prospectivo e retrospectivo em relação ao tempo de seleção e separando os grupos de pacientes que foram expostos dos que não foram expostos a alguma intervenção, assumindo-se que o ano de início do estudo seja 2013.

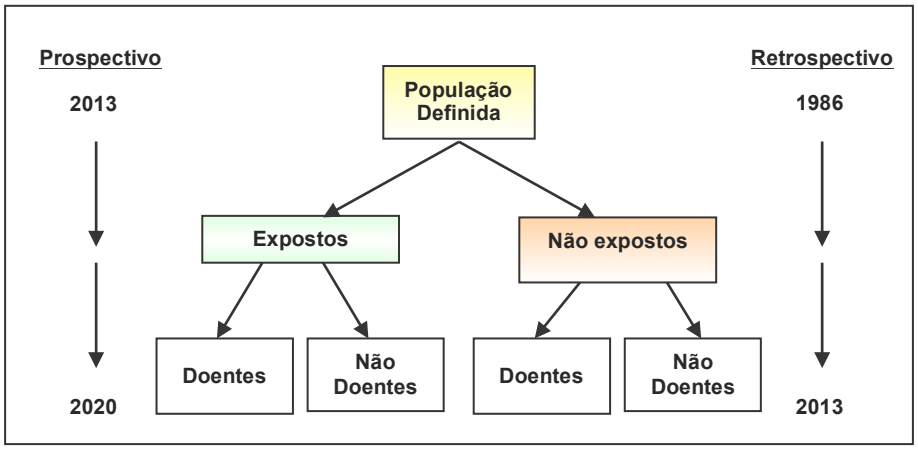


Figura 2.1 Tempo de seleção para estudos prospectivos e retrospectivos, considerando-se o ano de início em 2013: adaptado (Gordis, 2014)

Nos estudos de coorte prospectivos, identifica-se e seleciona-se os grupos expostos e não expostos no momento do início da investigação e os acompanha por um determinado período de tempo. Nos estudos de coorte retrospectivos, a identificação dos grupos é feita em algum momento do passado e estes grupos são acompanhados até o

presente, com objetivo de se identificar um evento ocorrido. Ambos conservam o princípio básico dos estudos de coorte: exposição a alguma intervenção, verificando o que ocorreu até a ocorrência do desfecho (exposição em direção ao evento). O desfecho é o evento que marca o término do tempo de observação do paciente. Em estudos de coorte retrospectivos, principalmente quando as informações sobre exposição e acompanhamento são de longo tempo, o maior problema é o da qualidade das informações. Métodos diagnósticos, identificação, mensuração podem ter mudado através do tempo, tornando as comparações impossíveis de serem realizadas.

Nos estudos observacionais retrospectivos os dados já foram obtidos. Utilizam dados do dia-a-dia do atendimento ao paciente, e tornam possível investigar grandes populações por longos períodos de tempo. Permitem estudos de grupos ou subgrupos que sofreram alguma intervenção e a avaliação de desfechos em saúde. No entanto, os dados não foram obtidos e anotados com o objetivo de investigação (uso secundário diferente do seu propósito original) e por serem produzidos por diferentes profissionais, necessitam de tratamento nas suas variações, uniformizando termos, medições, codificações, procedimentos e posologias.

2.1.1 A importância dos estudos observacionais

Um dos estudos prospectivos pioneiros na área da saúde nos Estados Unidos foi o *Framingham Heart Study* (Oppenheimer, 2005; Mendis, 2010; Framingham Heart Study, 2016), iniciado em 1948, sob a direção do *National Heart Institute*, agora conhecido como o *National Heart, Lung, and Blood Institute* (NHLBI), um projeto de pesquisa em saúde. Na época, pouco se sabia sobre as causas gerais de doença cardíaca e do acidente vascular cerebral, mas as taxas de mortalidade para doenças cardiovasculares (DCV) vinham aumentando de forma constante desde o início do século, tronando-se uma epidemia americana. O *Framingham Heart Study*, é um projeto conjunto do NHLBI e da Universidade de Boston. O objetivo do estudo foi identificar os fatores comuns ou características que contribuem para a DCV, seguindo o seu desenvolvimento ao longo de um período de tempo em um grande grupo de participantes que ainda não haviam desenvolvido sintomas evidentes de DCV ou sofrido um ataque cardíaco ou acidente vascular cerebral.

Ao longo dos anos, os integrantes do estudo *Framingham* eram acompanhados com uma monitoração cuidadosa que levou à identificação dos principais fatores de

risco para DCV. Esses fatores incluem, fatores que podem ser modificados, como a hipertensão arterial sistêmica (HAS), obesidade, diabetes mellitus (DM), tabagismo, sedentarismo, dislipidemias e outros fatores não modificáveis, como a idade, gênero e hereditariedade. No último meio século, o estudo produziu cerca de 1.200 artigos em importantes revistas médicas. A base de dados do estudo já acompanha e disponibiliza dados da terceira geração da coorte inicial. O conceito de fatores de risco para DCV tornou-se parte integrante do currículo médico e levou ao desenvolvimento de um tratamento eficaz e estratégias preventivas na prática clínica (Mendis, 2010).

O Instituto PHARMO (PHARMO), fundado na Holanda em 1999, uma organização de pesquisa científica independente, dedicada ao estudo da utilização, segurança de medicamentos e recursos de saúde. O Instituto mantém uma rede de grandes bases de dados, de alta qualidade e trabalha em estreita colaboração com as universidades e outras bases de dados europeias. O sistema PHARMO inclui os registros de dispensação de medicamentos de farmácias comunitárias e de prescrição de alta hospitalar de mais de 2 milhões de habitantes, incluindo todos os residentes de 45 áreas demograficamente e geograficamente definidas espalhadas por toda a Holanda. Todos os registros que compõe o sistema PHARMO são captados de fontes heterogêneas e estão integrados por meio da aplicação de algoritmos de *linkage* (vinculação de registros).

As dispensações informatizadas contém dados relativos à droga dispensada, o prescritor, data de distribuição, quantidade dispensada, os regimes de dosagem prescritos, custos reembolsados e a duração estimada de uso. Os nomes dos medicamentos são codificados de acordo com a classificação do *Anatomical Therapeutic Chemical* (ATC) (ATC - Guidelines, 2015). Os registros hospitalares contém informações sobre diagnósticos (Classificação Internacional de Doenças, 9ª Revisão, Modificação Clínica, ICD-9-CM), motivo de admissão e alta, datas e achados secundários. Através de estudos longitudinais com os dados dos pacientes, o Instituto PHARMO contribui para a gestão de riscos e fornecem evidências para os tomadores de decisão em economia da saúde. Desde a sua criação, mais de 750 estudos retrospectivos farmacoepidemiológicos têm sido realizados, em conjunto com universidades, organizações não governamentais e a indústria farmacêutica (PHARMO). Uma publicação de 2006, apresenta os resultados de um estudo de coorte histórica sobre o

uso da rosuvastatina em 45.000 holandeses, um estudo que utiliza a base de dados PHARMO (Goettsch et al., 2006).

Outros estudos em base populacional foram publicados recentemente e buscam verificar se os resultados de eficácia de tratamentos ou medicações, encontrados em estudos randomizados, estão refletidos pela efetividade no mundo real (Sicras-Mainar et al., 2012; Fortuna et al., 2013; Wijeysundera et al., 2014).

De Vera et al. (De Vera et al., 2014), realizou uma revisão sistemática de estudos observacionais em bases de dados assistenciais, com objetivo de sintetizar a evidência atual sobre os impactos da eficácia e adesão das estatinas no mundo real em comparação ao observado em ensaios clínicos. Nesse contexto, observa-se o crescimento na adoção e uso de registros eletrônicos em saúde (RES), os quais oferecem uma nova alternativa na condução de estudos epidemiológicos.

A partir dos ensaios clínicos, dos estudos observacionais, das meta-análises e revisões sistemáticas, são gerados o conjunto de evidências científicas que são utilizadas na preparação das diretrizes médicas com intuito de padronizar condutas e auxiliar o médico na decisão clínica de diagnóstico e tratamento de várias doenças (Sackett et al., 1996; Chambers, 1997; Greenhalgh, 2004).

A cada ano, a *American Heart Association* (AHA), em conjunto com os *Centers for Disease Control and Prevention*, o *National Institutes of Health* e outros órgãos governamentais, reúnem estatísticas sobre doenças do coração, acidente vascular cerebral, outras doenças cardiovasculares e seus fatores de risco e apresentam o *Heart Disease and Stroke Statistical Update* (Go et al., 2013). A atualização estatística é um recurso crítico para pesquisadores, clínicos, os formuladores de políticas de saúde, profissionais da área e muitos pacientes que buscam o melhor dos dados nacionais disponíveis sobre doenças cardíacas, qualidade do atendimento, utilização de procedimentos médicos, operações e os custos associados com a gestão destas doenças em um único documento.

A agência de pesquisa e qualidade em saúde americana, *Agency for Healthcare Research and Quality* (AHRQ), tem como missão produzir evidências para tornar os cuidados de saúde mais seguros, de maior qualidade, mais acessíveis e equitativos, e trabalhar em conjunto com o departamento de saúde e serviços humanos dos Estados Unidos (*U.S. Department of Health and Human Services)* (HHS) e outros parceiros, garantindo que a evidência gerada seja compreendida e utilizada. As revisões de

efetividade comparada oferecem um método sistemático para avaliar criticamente a pesquisa existente e identificar áreas clínicas na doença cardiovascular que estão sem cobertura e onde persistem variações significativas no uso de recursos, na morbidade e na mortalidade (CER-Guide).

O *Effective Health Care Program* da *Agency for Healthcare Research and Quality* realizou uma avaliação de eficácia comparativa em doenças cardiovasculares (DCV), onde delineou e selecionou áreas de interesse que representam uma ampla gama de perspectivas clínicas, políticas e do paciente. Para auxiliar na priorização de temas para a síntese de evidências criaram um quadro avaliando 12 subcategorias das DCV que refletem as diretrizes da *American College of Cardiology* (ACC) *e American Heart Association* (AHA). Através de um processo formalizado, quatro subcategorias da doença foram priorizadas: doença crônica da artéria coronária, arritmia ventricular, insuficiência cardíaca e doença cerebrovascular. Foi criada uma lista de temas prioritários em síntese de evidências, na geração de informações de segurança e eficácia de uma série de tratamentos, relevantes para o estado de doença prevalente, no auxílio aos tomadores de decisão, médicos e pacientes. Entre os temas identificados que necessitam de futuras revisões sistemáticas em DCV estão: estatinas para diminuir o risco de DCbV; tratar com as metas de colesterol contra tratando todos com estatina em doses moderadas; e em pacientes com DCV, as estatinas em doses moderadas devem ser utilizadas exclusivamente para a redução do colesterol LDL, ou para a redução de outros subtipos de lipídios (Eapen et al., 2013).

2.1.2 Estudos observacionais utilizando bases de dados assistenciais

A construção de coortes e grupos de pacientes com doenças específicas a partir de base de dados assistenciais geradas de um ou mais RES vem se tornando cada vez mais comum. A Kaiser Permanente nos Estados Unidos possui várias coortes constituídas a partir do seu RES (Croen et al., 2005; Armstrong-Wells et al., 2009; Musser et al., 2014), incluindo o *Diabetes Study of Northern California* (DISTANCE) (Moffet et al., 2009). A coorte DISTANCE envolve mais de 20.000 pacientes com diabetes e tem possibilitado uma variedade de estudos como a avaliação dessa doença em asiáticos (Kanaya et al., 2011), o impacto de privações socioeconômicas e indicadores de saúde cardiometabólicos (Laraia et al., 2012) e a relação entre a classe social e o risco de hipoglicemia (Berkowitz et al., 2014).

Outras experiências semelhantes nos Estados Unidos envolvem a colaboração entre coortes de múltiplos sistemas de Saúde. A rede de organizações sociais (*Health Maintenance Organizations*) tem se tornado líder nesse tipo de pesquisa desde 1994 (Selby, 1997). Outros exemplos ainda nos Estados Unidos são o Consórcio *Safe and Labor* (Hibbard et al., 2010; Männistö et al., 2015), que utiliza bases de dados de nascidos-vivos oriundas do RES em 19 hospitais; As coortes envolvendo dados clínicos dos hospitais que compõem a rede dos *Veterans Administration* (VA) (Vigen, 2013); A coorte para estudo de hepatite B e C crônicas, a qual combina dados obtidos de 4 sistemas de saúde e mais de 1,6 milhão de indivíduos (Moorman et al., 2013). A partir dos dados dessa coorte, Mahajan et al. (Mahajan et al., 2014) observaram que, apenas 30% dos pacientes com hepatite C, os quais tiveram como causa do óbito doença no fígado, apresentavam o registro de hepatite C em seus atestados de óbito, revelando assim uma enorme subnotificação de hepatite C como causa de mortalidade nos Estados Unidos.

Estudos envolvendo coortes a partir de repositórios centrais de dados anonimizados incluem o *Clinical Data Analysis Report System* em Hong Kong (Cheuk et al., 2004) e o *Clinical Practice Research Datalink* (CPRD) (Alonso et al., 2006), a rede *Health Improvement Network* (THIN) (Zhou et al., 2013; Dhalwani et al., 2014) e o *QResearch* no Reino Unido (Thomas et al., 2014). A coorte CPRD, que reuniu dados de mais de 500 clínicas contempla informações de mais de 5 milhões de indivíduos. Tais exemplos de coortes baseadas em base de dados assistenciais oferecem dados normalizados e longitudinais e enorme oportunidades para pesquisa em epidemiologia, que pode ser evidenciado, por exemplo, por mais de 1.000 artigos científicos originados a partir da coorte CPDR.

2.1.3 Estudos observacionais realizados nas bases de dados do DATASUS

O Departamento de Informática do SUS (DATASUS) é o responsável por fornecer os sistemas de informação e suporte de informática, necessários ao processo de planejamento, operação e controle do SUS, por meio da manutenção de bases de dados nacionais, apoio e consultoria na implantação de sistemas e coordenação das atividades de informática inerentes ao funcionamento integrado dos mesmos. O DATASUS possui vários sistemas na produção de informações necessárias a gestão do SUS. Dentre eles tem-se o Sistema de Informações Ambulatoriais (SIA), o Sistema de Informações

Hospitalares (SIH), o Cadastro Nacional de Estabelecimentos de Saúde (CNES), o Sistema de informações de Mortalidade (SIM) e o Sistema de Informações de Nascidos Vivos (SINASC). Cada um desses sistemas tem seus dados armazenados e disponibilizados para consulta através de um tabulador de informações, o TABNET, no portal do DATASUS ou através de arquivos disponíveis para *download*. Esses sistemas já produzem uma grande quantidade de informações, porém, essas bases são anonimizadas, não apresentando a identificação do paciente. A ausência de um identificador comum impede o seguimento longitudinal dos pacientes atendidos pelo SUS (DATASUS, 2016).

Em 2007, foi desenvolvido e implantado no InCor, um protótipo inicial de um *Data Warehouse* (DW), integrando dados provenientes dos diferentes sistemas de informação do DATASUS, tendo por objetivo a disponibilização de informações de modo integrado e não por meio de bases isoladas (Santos, 2007). A partir desse protótipo, foi desenvolvido um sistema mais amplo que permitiu a integração de um maior volume de dados e a extração de informações gerenciais, utilizando o modelo de DW proposto, denominado MinerSUS (Pires, 2011). Os dados de origem utilizados foram selecionados das bases de dados do SIA, SIH, SIM e SINASC, no período de 2000 a 2007, para o Estado de São Paulo. A Secretaria de Estado da Saúde de São Paulo (SES-SP), a partir de um acordo de cooperação entre o Serviço de Informática do InCor e a SES-SP, forneceu dados identificados dos sistemas SIH e SIA, com os quais foi realizado um processo de *linkage* entre essas duas bases, sendo possível a realização de seguimento por paciente e uma comparação de populações com registros nos dois sistemas (Pires et al., 2011).

Foram realizados estudos retrospectivos por Abrahão et al. na base do MinerSUS, como uma avaliação do uso de medicação de alto custo no tratamento da artrite reumatoide e sobre a medicação estatina no tratamento da doença cardiovascular (Abrahão et al., 2008, 2010a, 2010b, 2012b).

2.1.4 Estudos observacionais utilizando as bases de dados do InCor

Vários estudos foram publicados utilizando os dados da base do sistema de informações assistencial do InCor. Um estudo realizado em 2003 por Caramelli et al. com o objetivo de avaliar as características clínico demográficas da população com doença isquêmica do coração, analisou, de forma retrospectiva, 15.347 pacientes que

tiveram registro de internação no InCor, no período de 1982 a 1994, com forma aguda e crônica da doença. Nesse estudo, foram analisadas retrospectivamente as variáveis idade, gênero, letalidade intra-hospitalar e a presença de diabetes mellitus, como doença associada (Caramelli et al., 2003a).

Em outro estudo, realizado por Nicolau et al., com objetivo de avaliar a evolução de pacientes atendidos com infarto agudo do miocárdio, usuários do SUS e de outros convênios, foram analisados 1.588 pacientes, incluídos de forma prospectiva em banco de dados específico entre 1998 e 2005. Como resultado desse estudo, os usuários SUS apresentam uma taxa de mortalidade similar a dos pacientes de convênios durante a fase hospitalar, porém, têm pior prognóstico em longo prazo, reforçando a necessidade de esforços adicionais no sentido de melhorar o nível de atendimento desses pacientes após a alta hospitalar (Nicolau et al., 2008).

Caramelli et al. realizou um estudo retrospectivo com objetivo de obter informações sobre o perfil e o comportamento de uma população com doença cardíaca isquêmica submetida a cinecoronariografia para determinar a gravidade da doença. Foram selecionados 18.221 pacientes internados no período de 1986 a 1995 e analisados em relação as variáveis idade, gênero e número de artérias principais com grau de obstrução maior ou igual a 40%. Como conclusão, este estudo sugere que houve uma mudança no perfil da população com doença cardíaca isquêmica tratadas na Instituição. Os pacientes são mais velhos, possuem um número maior de grandes artérias coronárias afetadas e há um aumento da incidência da doença entre as mulheres, os quais são índices sugestivos de um pior prognóstico (Caramelli et al., 2003b).

Em uma publicação de Lisboa et al. sobre a evolução da cirurgia cardiovascular foi apresentada uma análise das 71.305 operações realizadas entre 1984 e 2007. Esta análise mostrou que o InCor realiza em média 2.971 cirurgias por ano e a cirurgia de revascularização miocárdica é a mais realizada. Mostrou que as taxas de mortalidade hospitalar apresentadas são superiores, quando comparadas aos índices internacionais (Lisboa et al., 2010).

Estudos realizados por Abrahão et al. analisaram uma coorte de cerca de 65.000 pacientes com doença cardiovascular, descrevendo o perfil dos pacientes e a relação da doença com a prescrição da medicação estatina na prevenção secundária de novos eventos (Abrahão et al., 2012a, 2013a, 2013b, 2014). Outro estudo, utilizando os dados da mesma coorte, realizado por Luque et al., analisou a relação custo-efetividade das

estatinas em uso secundário no tratamento de pacientes com doença cardiovascular (Luque et al., 2015).

2.1.5 Vantagens e desvantagens dos estudos em bases assistenciais

O uso de bases de dados assistenciais (ou secundárias) para pesquisa tem como vantagens o baixo custo e a possibilidade de se usar amostras maiores, proporcionando maior força para detectar pequenas diferenças ou eventos raros (Carrilho et al., 2010). Possibilita maior diversidade metodológica, permitindo que modificações no desenho do estudo possam ser implementadas diferentemente dos desenhos de pesquisa de maior rigor de protocolos, como os ensaios randomizados. Além disso, o uso de bases de dados assistenciais dispensa o contato direto com pacientes. Como limitação, as bases de dados assistenciais não foram desenhados para pesquisa, mas sim para atender processos assistenciais e administrativos. Como consequência, dependendo do tipo de análise ou pesquisa desejada pode haver limitação no conjunto de dados disponíveis. A qualidade das informações pode, em alguns casos estar comprometida, podendo haver erros de codificação, que podem causar resultados tendenciosos. A validade da base de dados também deve ser considerada, pois pode ser necessário comparar informações provenientes das bases assistenciais com o prontuário médico tradicional (físico em papel), para avaliar a qualidade das informações. Por outro lado, o custo e o tempo associados aos estudos que utilizam bases de dados assistenciais podem ser menores, quando comparado a outros tipos de estudo. Outra limitação é a necessidade de conhecimento de sua estrutura, que possibilite a correta leitura e captura das variáveis de interesse (Hulley et al., 2003; Carrilho et al., 2010).

O estudo da utilização de medicações a partir de dados do mundo real pode auxiliar os gestores de saúde na avaliação dos custos quando comparados aos resultados de eficácia dos ensaios randomizados, além disso, possibilita novas análises que contribuem para a efetividade na tomada de decisão e na incorporação de novas técnicas e medicações.

2.2 Reprodutibilidade

A reprodutibilidade é a possibilidade de um experimento ou estudo poder ser repetido, tanto pelo mesmo pesquisador quanto por um pesquisador atuando

independentemente. Reproduzir um experimento é chamado de replicar o mesmo. A reprodutibilidade é um dos princípios fundamentais do método científico.

Segundo Peng (Peng, 2009), a replicação de achados científicos utilizando investigadores, métodos, dados, equipamentos e protocolos independentes tem sido, e continuará a ser, o padrão pelo qual as afirmações científicas são avaliadas. No entanto, em muitos campos da ciência existem exemplos de investigações científicas que não podem ser totalmente replicadas por causa de falta de tempo ou recursos. Em tal situação, existe a necessidade de um padrão mínimo que possa preencher o vazio entre a replicação completa e nada. Um candidato para este padrão mínimo é a pesquisa denominada reprodutível, que exige que os conjuntos de dados e códigos de um experimento sejam disponibilizados a terceiros, de modo a possibilitar a verificação dos resultados e a realização de análises alternativas.

A Figura 2.2 apresenta o espectro de reprodutibilidade proposto por Peng (Peng, 2011).

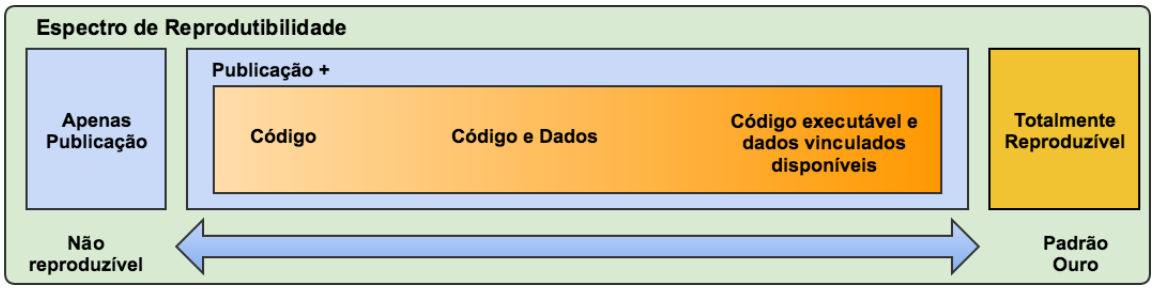


Figura 2.2 Espectro de reprodutibilidade: adaptado (Peng, 2011)

A partir das considerações de Peng (Peng, 2009), o *Biostatistics Oxford Online Journal*, estabeleceu as políticas da reprodutibilidade de seus artigos:

- Dados: Os dados analíticos dos quais provem os principais resultados são disponibilizados no site da revista. Os autores são responsáveis por garantir que as permissões necessárias serão obtidas antes dos dados serem distribuídos;
- Código: Qualquer código de computador, *software* ou outros tipos de instruções de computador que foram usados para obtenção dos resultados publicados devem ser fornecidos. Para o *software* que está amplamente disponível em repositórios públicos (por exemplo *CRAN, Statlib*), uma referência ao local onde eles podem ser obtidos será suficiente;

- Reprodutível: Um artigo é designado como reprodutível se o revisor consegue executar o código nos dados fornecidos e produz resultados semelhantes aos que os autores reivindicaram serem reprodutíveis. Ao reproduzir esses resultados serão considerados limites razoáveis de tolerância numérica.

A habilidade de replicar resultados computacionais publicados depende da disponibilidade dos dados e do código usado para gerar os mesmos (Donoho, 2010). Uma análise de 170 publicações entre 2011 e 2012 mostrou um aumento de 16% no número de jornais que definem políticas de compartilhamento de dados, de 30% de jornais que definem políticas de compartilhamento de código e de 7% dos que definem políticas para materiais suplementares (Stodden et al., 2013).

No editorial (Nature, 2015) da edição especial da *Nature* (Nature News, 2015) a respeito de reprodutibilidade, destaca-se o problema chave que afeta a pesquisa irreproduzível:

> "O hábito do cérebro humano de encontrar o que ele quer encontrar é um problema chave para a pesquisa. Estabelecer métodos robustos para evitar tal viés irá tornar os resultados mais reprodutíveis.
>
> ...
>
> Um inimigo da ciência robusta é a nossa própria humanidade - o nosso apetite por estar certo, nossa tendência para encontrar padrões no ruído, para ver provas para o que já acreditamos ser verdade, e para ignorar os fatos que não se encaixam" (Nature, 2015).

Os artigos que compõem a edição especial (Nature News, 2015) descrevem os problemas mais frequentes e as políticas adotadas pela *Nature* para evitá-los.

Dentre eles, ressaltamos o editorial no qual a revista *Nature* introduziu o *Scientific Data*, um novo jornal para a publicação de dados que permite aos autores publicar conjuntos de dados e receber o devido crédito (Nature Editorial, 2014).

A revista *Nature* tem políticas específicas para publicação de artigos (Nature Policies, 2016), em particular para a disponibilidade de dados e materiais, e uma lista de repositórios recomendados (Scientific Data):

> "A condição da publicação em uma revista *Nature* é que os autores devem deixar materiais, dados, códigos e protocolos associados disponíveis para os leitores sem empecilhos indevidos" (Nature Policies, 2014).

A revista *Science* e a revista *Nature*, num editorial conjunto (McNutt, 2014; Nature Journals, 2014), *"Journals unite for reproducibility"*, ressaltam que:

> "Reprodutibilidade, rigor, transparência e verificação independente são pedras angulares do método científico. Claro, só porque um resultado é reprodutível não necessariamente o torna certo, e só porque ele não é reprodutível não necessariamente o torna errado. Uma abordagem transparente e rigorosa, no entanto, quase sempre pode lançar uma luz sobre questões da reprodutibilidade. Esta luz assegura que a ciência avança, por meio de verificações independentes, bem como das correções de curso que vêm de refutações e do exame objetivo dos dados resultantes" (McNutt, 2014; Nature Journals, 2014).

O trabalho da *Nature*, da *Science* e de outras editoras deu origem ao *"Principles and Guidelines for Reporting Preclinical Research"* (NHI - Guidelines, 2015), elaborado durante o workshop organizado pelo *National Institutes of Health* (NIH) em Junho de 2014 com a participação de editores representando cerca de 30 dos principais jornais de ciência básica e pré-clínica. Nesse guia se delineiam as recomendações para que as editoras definam suas políticas de publicação em relação aos seguintes tópicos:

- Análise estatística rigorosa;
- Transparência no relato;
- Compartilhamento dos dados e materiais utilizados;
- Consideração das refutações;
- Estabelecimento de guias de melhores práticas para dados baseados em imagens e descrição de material biológico.

Estas recomendações têm sido endossadas por mais de 70 associações, jornais e sociedades (NHI-Associations). Nos Estados Unidos, estas políticas também estão sendo disseminadas para áreas governamentais que fornecem verbas e incentivos para pesquisa (Stodden, 2014).

Em relação à pesquisa observacional, precisamos salientar dois dos mais importantes problemas:

- Existe evidência da influência da escolha da base de dados nos resultados de estudos clínicos que usam bases de dados observacionais (Ioannidis, 2005a; Madigan et al., 2013). Estes estudos procuram utilizar dados administrativos ou

registros eletrônicos de saúde para abordar questões importantes a respeito dos efeitos de produtos médicos utilizando estudos observacionais. Muitos vieses potenciais desafiam a validade deste tipo de estudos e a literatura aborda amplamente estas preocupações (Mayes et al., 1988). Mesmo quando os estudos publicados discutem as várias limitações, raramente mencionam bases alternativas que poderiam ter sido usadas e como a escolha da base afetaria o resultado do estudo. Na verdade, na maioria das vezes a fonte dos dados é apenas identificada sem prover nenhuma discussão a respeito do processo utilizado na seleção dela. Estudos a respeito de um determinado assunto, quando aplicados em bases diferentes, podem gerar resultados contraditórios apenas mudando a base (Ioannidis, 2005b).

- Toda afirmação vinda de um estudo observacional está provavelmente errada (Young e Karr, 2011). Aqui a proposta é discutir uma forma de evitar o viés de publicação, onde apenas artigos que apresentam resultados estatisticamente significativos ($p < 0.05$) são publicados e que estimulam o pesquisador a procurar modelos de estudos e métodos estatísticos que favoreçam o aparecimento do resultado desejado (Nature, 2015). É proposta a separação de uma parte dos dados para uso como controle e a aplicação dos métodos citados no estudo sobre o conjunto de controle, publicando os achados independentemente do resultado.

O princípio para resolução destes problemas, segundo Madigan et al., passa por um método sistemático de definição e análise de métodos de estudos observacionais (Madigan et al., 2014). Nessa publicação, foram discutidos alguns dos desafios encontrados em estudos observacionais e avaliaram uma abordagem alternativa, para o desenho do estudo observacional, execução e análise. Foi feita uma replicação de 4 estudos, sobre: (i) insuficiência renal aguda; (ii) insuficiência hepática aguda; (iii) infarto agudo do miocárdio; (iv) sangramento do segmento superior gastrointestinal. Para cada um dos estudos foram aplicados 7 métodos analíticos diferentes:

- *SCC - Self-controlled cohort;*
- *SCCS - Self-controlled case series;*
- *Case control;*
- *ICTPD - Information component temporal pattern Discovery;*

- *New-user cohort;*
- *DP - Disproportionality analysis;*
- *LGPS – Longitudinal gamma Poisson shrinker;*

Sobre 5 diferentes bases de dados assistenciais disponíveis no OMOP:

- *CCAE – MarketScan Comercial Claims and Encounters (total: 46,5 milhões, 49% gênero masculino e média de idade de 31,4 anos (18,1); tempo observacional (t.o.): 97,6 milhões de anos-paciente, 2003-2009);*
- *MDCD – MarketScan Multi-State Medicaid (total: 10,8 milhões, 42% gênero masculino e média de idade de 21,3 anos (21,5); t.o: 20,7 milhões de anos-paciente, 2002-2007);*
- *MDCR – Medicare Supplemental Beneficiaries (total: 4,6 milhões, 44% gênero masculino e média de idade de 73,5 anos (8,0); t.o: 13,4 milhões anos-paciente, 2003-2009);*
- *MSLR – MarketScan Lab Supplemental (total: 1,2 milhão, 35% gênero masculino e média de idade de 37,6 anos (17,7); t.o: 2,2 milhões anos-paciente, 2003-2007);*
- *GE – GE Centricity (total: 11,2 milhões, 42% gênero masculino e média de idade de 39,6 anos (22,0); t.o: 22,4 milhões anos-paciente, 1996-2008)*;

Foi concluído que, mesmo que nenhum dos métodos resultassem numa discriminação perfeita, muitos deles foram substancialmente melhores (área sob a curva ROC entre 0.76 a 0.94) que a escolha aleatória e demonstraram o resultado do desempenho dos vários métodos analíticos para estudos observacionais em uma série de combinações de banco de dados assistenciais.

2.3 Preparação da base de dados

A tecnologia da informação tem oferecido modelos e arquiteturas apropriadas para as diversas áreas de negócio relacionadas com a busca de informações para apoiar as tomadas de decisões. São tecnologias que oferecem recursos e ferramentas para a extração, tratamento, limpeza, integração e análise de dados em evolução histórica. Os dados assim preparados servem para alimentar as bases de dados destinados as pesquisas clínicas.

Na Europa, para diminuir a distância entre a prática farmacêutica e a pesquisa de desfechos clínicos, instituições como a *International Society for Pharmacoeconomics*

and Outcomes Research (ISPOR) foram criadas para definir guias de recomendações para estudos comparativos de efetividade, e consequentemente, acabaram criando critérios de validação para bases de dados assistenciais nos quais estes estudos se fundamentam. A ISPOR tem como missão, aumentar a eficiência, efetividade e equidade dos cuidados ao paciente e melhorar a saúde. A ISPOR promove a ciência da farmacoeconomia e pesquisa de desfechos, e facilita a transição desta pesquisa em informação útil para as decisões na área da saúde (Berger et al., 2014).

Na preparação da base de pesquisa foram seguidos os critérios da ISPOR para coleta retrospectiva de dados (Motheral et al., 2003; Berger et al., 2009; Cox et al., 2009; Johnson et al., 2009). A orientação da ISPOR abrange 27 itens de utilidade para realizar, avaliar ou aproveitar os estudos retrospectivos, para aplicação em bases de dados de origem médica. A aplicação destes itens permite avaliar a metodologia empregada nestes estudos e a validade de suas conclusões. Os itens aplicados na preparação da base de pesquisa foram:

1. Relevância das fontes de dados: os dados estão suficientemente detalhados que possam ser empregados, generalizados, interpretados e permitam uma conclusão que atenda o estudo em questão? Possui atributos que permita traçar um perfil sócio-demográfico, dos serviços, procedimentos e medicações disponíveis?
2. Confiabilidade e validade: é necessário checar a qualidade dos dados, considerando-se que os atributos não são estáticos, podem depender do tempo e do tipo de análise. Devem-se descrever as medidas e recursos usados para normalizar e eliminar os dados não confiáveis. Devem-se apontar as mudanças no relato e na codificação no decorrer do tempo, as possíveis repercussões dessas mudanças nos resultados, a verificação de dados muito distantes da faixa habitual (*outliers*) e se houver grandes inconsistências.
3. Conexões (*Linkage*): quando se usam dados procedentes de diferentes bases, deve-se relatar como foram recuperados e ligados ao mesmo paciente. Em bases distintas, um mesmo evento de saúde pode ser relatado de diferentes formas.
4. Delineamento da pesquisa (*research design*): é necessária a sistematização da busca, definir um plano de análise dentro de uma linha de testar hipóteses.
5. Escolha do delineamento ou desenho (*design*): diferentes tipos de desenho podem ser usados, conforme o ambiente, o tipo de questões levantadas e até os

dados. O investigador deve declarar as vantagens e fraquezas do tipo de desenho escolhido.

6. Limitações do tipo de projeto: há discussão sobre os possíveis desvios ou tendenciosidades deste tipo de desenho?
7. Efeitos do tratamento: se a pesquisa busca inferências sobre certa intervenção deverá haver um grupo de comparação ou grupo controle (pessoas expostas e não expostas a intervenção).
8. Seleção da amostra: definição dos critérios de inclusão e de exclusão para se definir a população potencial do estudo. A aplicação destes critérios deve ser detalhada e discorrida sobre sua influência nos resultados. Devem ser destacados o número total de sujeitos na população e quais os números após a aplicação destes critérios.
9. Definições operacionais: devem-se definir diagnósticos, códigos de procedimentos, medidas, nomenclaturas e critérios.
10. Definição de validade: relatar os casos de incertezas e inconsistências encontrados.
11. Tempo da exposição até o desfecho (desfecho é considerado como o evento que marca o término do tempo de observação do paciente): há uma relação temporal clara entre a exposição e o evento definido como desfecho? Levou em conta a proximidade entre intervenções importantes, sua duração e a ocorrência do desfecho?
12. Captação incompleta dos eventos: alguns procedimentos podem escapar da avaliação por não terem sido considerados na busca.
13. História natural da doença: doenças com história natural longa não se prestam bem a estudos transversais que não levem em consideração esta característica.
14. Avaliação de recursos: para estudos que examinam custos, deve ser definida a lista de recursos que são afetados pela intervenção.
15. Definição da estatística: todas as decisões de estatísticas devem ser definidas e descritas antes do início do estudo. Dar um maior enfoque a:
    a. Controle das variáveis: nos estudos retrospectivos devem-se considerar todas as variáveis para evitar conclusões equivocadas (fatores de confusão);

b. Modelo estatístico: devem ser declarados e bem detalhadas as razões de adoção de certo modelo estatístico;
c. Verificação dos pontos fora da curva (*outliers*): identificar e justificar os pontos que não estão no padrão;
d. Variáveis relevantes: identificar todas as variáveis que possam influenciar os desfechos relevantes e incluir no modelo.

16. Discussão e conclusões: examinar a relação causal nos estudos retrospectivos, pois neste tipo de estudo não existe randomização do tratamento. Geralmente as amostras contêm um grande número de indivíduos e podem mostrar diferenças que não são reais ou não tem significado estatístico. Na generalização das conclusões é preciso ficar claro a qual população e em quais circunstâncias seus dados podem ser generalizados.

Além das orientações da ISPOR para coleta dos dados, em relação as publicações de estudos observacionais, a iniciativa denominada *Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE), formulou uma lista de verificação que contém 22 itens, denominada *STROBE Statement* "Declaração STROBE", com recomendações sobre o que deveria ser incluído em uma descrição mais precisa e completa de estudos observacionais. Essas recomendações foram traduzidas e adaptadas para o português, apresentando o contexto de utilização, as potencialidades e limitações da Iniciativa STROBE. Essas recomendações oferecem um modelo que poderá ser seguido por autores de estudos observacionais e que objetiva contribuir para um relato mais adequado desses estudos e, consequentemente, facilitar a leitura crítica dessas publicações por parte de editores, revisores e leitores em geral (Ferreira e Passos, 2010).

2.4 Modelo comum de dados

A finalidade do modelo comum de dados é padronizar o formato e o conteúdo dos dados observacionais, oriundos de sistemas heterogêneos, para que aplicativos, ferramentas e métodos padronizados possam ser aplicados.

Nos Estados Unidos, a necessidade de vigilância farmacológica adequada tem impulsionado a definição de modelos comuns que permitam coletar dados de diversas fontes de informação e aplicar análises estatísticas padronizadas (Madigan e Ryan, 2011; Madigan et al., 2013; Hripcsak et al., 2015).

Em 2007, reconhecendo o aumento da utilização dos RES em pesquisas epidemiológicas, o Congresso Americano requisitou ao *Food and Drug Administration* (FDA) criar um novo programa de vigilância farmacológica para identificar de forma mais agressiva, potenciais problemas de segurança. O FDA lançou várias iniciativas para alcançar esse objetivo, incluindo o programa "Sentinela", para criar uma rede de dados a nível nacional para o monitoramento de drogas, utilizando dados eletrônicos de detentores de informações de saúde (FDA, 2016).

Em particular, a *Pharmaceutical Research and Manufacturers of America* (PhRMA), o FDA e a *Foundation for the National Institutes of Health* (FNIH) criaram a *Observational Medical Outcomes Partnership* – OMOP, uma parceria público-privada (OMOP, 2016). Este grupo de pesquisa interdisciplinar abordou uma tarefa que é fundamental para os objetivos mais amplos da comunidade de investigação: identificar os métodos mais confiáveis para a análise de grandes volumes de dados extraídos de fontes heterogêneas.

Empregando uma variedade de abordagens das áreas de epidemiologia, estatística e ciências da computação, OMOP procura responder a um desafio crítico: o que podem pesquisadores médicos aprender com a avaliação dessas novas bases de dados de saúde? Poderia uma abordagem única ser aplicada a várias doenças e poderiam as suas conclusões ser provadas? Sucesso significa a oportunidade para a comunidade de pesquisa médica fazer mais estudos em menos tempo, utilizando menos recursos e obtendo resultados mais consistentes. No final, isso significaria um melhor sistema de monitoramento de drogas, dispositivos e procedimentos para a comunidade de saúde poder identificar com segurança os riscos e oportunidades para melhorar o atendimento ao paciente (Madigan e Ryan, 2011).

Para conseguir estes objetivos, o OMOP reuniu num modelo comum de dados diversas fontes de dados heterogêneas (Overhage et al., 2012). A Figura 2.3 apresenta o modelo comum de dados (*Common Data Model - CDM*) proposto pelo OMOP.

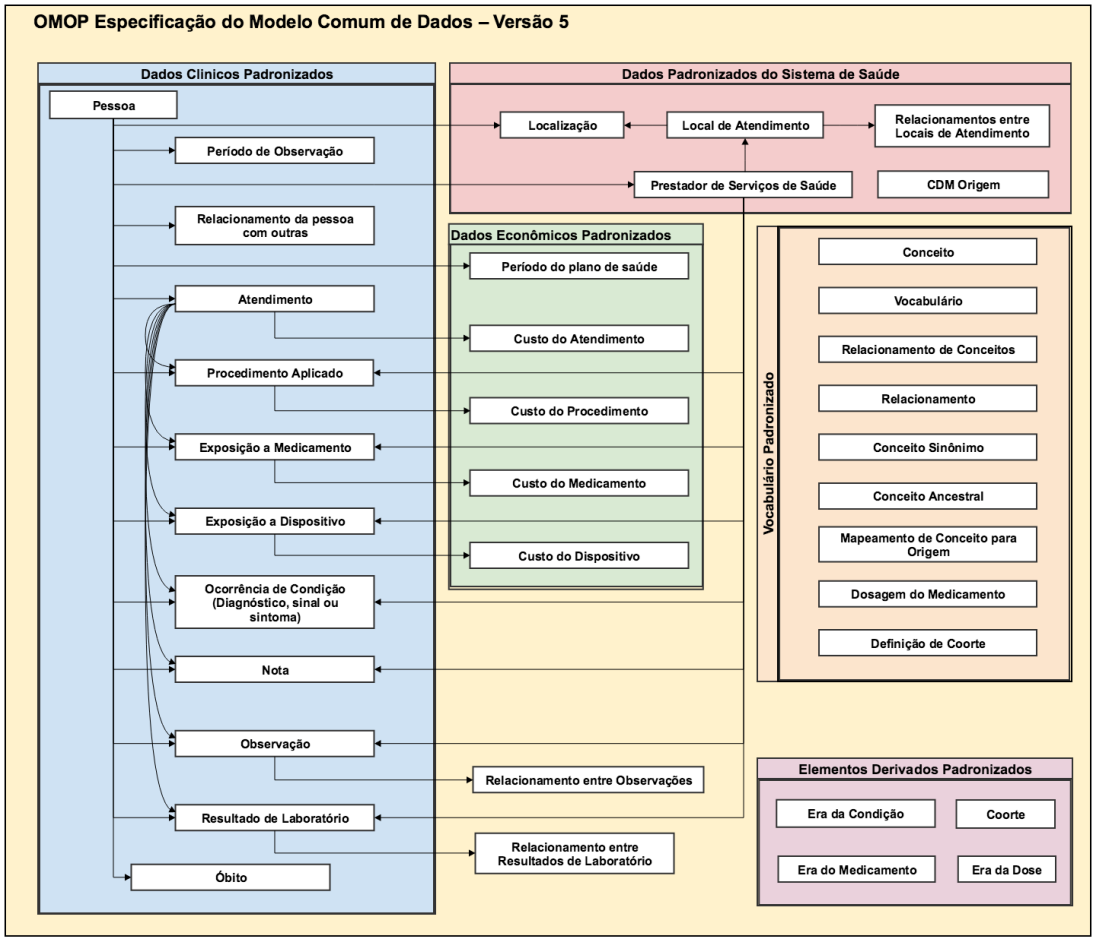


Figura 2.3 Modelo comum de dados – OMOP: Adaptado (OMOP - CDM, 2016)

Além dos dados da pessoa, da condição, droga, procedimento e informações de visitas, o modelo prove informações de custo e do provedor do atendimento. Isto irá apoiar a economia da saúde e estudos de casos de uso de resultados de tratamento médico, incluindo a segurança de dispositivos médicos, eficácia comparativa e qualidade de saúde. O mapeamento é a parte mais demorada da criação de um banco de dados em formato OMOP. Além do mapeamento e transformação dos dados para o CDM, o conteúdo também tem de estar em conformidade com os vocabulários padrão, descrito nas especificações OMOP. O vocabulário padrão é uma ferramenta fundamental, desenvolvida pela equipe OMOP, para padronizar o conteúdo em todos os bancos de dados observacionais díspares e serve para apoiar a comunidade de pesquisa OMOP na realização de pesquisas observacionais eficientes e reprodutíveis.

O vocabulário padrão contém todos os conjuntos de códigos, terminologias, vocabulários, nomenclaturas, léxicos, dicionários, ontologias, taxonomias, classificações, abstrações e outros dados que são necessários para: (i) criar dados

padronizados a partir da transformação do conjunto de dados brutos; (ii) pesquisar, consultar os dados e navegar nas hierarquias de classes e abstrações inerentes aos dados transformados; (iii) interpretar os significados dos dados.

Todo material do modelo OMOP está disponível para o público como *Open Source* com restrições mínimas (OMOP, 2016).

2.5 Esforços para definição e adesão a padrões nacionais em sistemas de saúde

Segundo a *International Organization for Standardization* (ISO, 2016), padrão é um documento estabelecido por consenso e aprovado por um grupo reconhecido, que estabelece um conjunto de regras, protocolos, requisitos ou características de processos com o objetivo de ordenar e organizar atividades em contextos específicos, para o benefício de todos.

Segundo a *Healthcare Information and Management Systems Society* (HIMSS, 2016):

> "Na área da saúde, a interoperabilidade é a capacidade de diferentes sistemas de informação e aplicações de software para se comunicar, trocar dados, e utilizar as informações trocadas. Padrões e normas devem permitir que os dados sejam compartilhados entre médicos, laboratórios, hospitais, farmácias, e pacientes, independentemente das aplicações ou dos fabricantes. A interoperabilidade significa a capacidade dos sistemas de informação de saúde para trabalhar em harmonia dentro e através das fronteiras organizacionais, a fim de conseguir a prestação eficaz de cuidados de saúde para os indivíduos e as comunidades. Existem três níveis de interoperabilidade de tecnologia da informação da saúde: 1) fundacional; 2) estrutural; e 3) semântica" (HIMSS - Definition).

A interoperabilidade semântica representa a possibilidade dos sistemas interpretarem, automaticamente e sem ambiguidade, a informação intercambiada tanto em significado quanto em acurácia, para poder produzir resultados úteis para usuários finais que necessitam da informação. Para poder aplicar os mesmos estudos a bases diferentes, é necessário que os dados extraídos sejam semanticamente interoperáveis.

Os padrões, tanto de estrutura de dados quanto de vocabulários, são o caminho para tornar isso possível.

O Ministério da Saúde (MS) vem propondo nos últimos anos a construção de uma "Estratégia de e-Saúde" para o Brasil, buscando a qualificação dos processos de atenção à saúde para a população, tendo como base a publicação do "Pacote de Ferramentas da Estratégia Nacional de eSaúde" (National eHealth, 2012), pela Organização Mundial da Saúde (OMS) e a União Internacional das Telecomunicações (UIT). Esta publicação é um trabalho partilhado que reflete um objetivo: dar resposta às necessidades dos países, em todos os níveis de desenvolvimento, que procuram adaptar e empregar as mais recentes tecnologias da informação e da comunicação na saúde, para benefício dos seus cidadãos. O Pacote de Ferramentas é um marco de referência quanto ao que a eSaúde é, o que pode fazer, e por que e como deve ser aplicada aos cuidados de saúde de hoje. É um guia prático e abrangente que todos os governos e respectivos ministérios, departamentos e organismos podem adaptar às suas próprias circunstâncias, bem como às suas visões e metas.

Na proposta do MS, um dos elementos fundamentais é a conformação de um Registro Eletrônico de Saúde (RES) Nacional, pois é por meio dele que as informações podem ser reorganizadas, agregando valor para se tornarem um componente estratégico para tomada de decisão clínica e de gestão do sistema de saúde (BVSMS). A visão de eSaúde pode ser sintetizada:

> "Até 2020, a e-Saúde estará incorporada ao SUS como uma dimensão fundamental, sendo reconhecida como estratégia de melhoria consistente dos serviços de saúde por meio da disponibilização e uso de informação abrangente, precisa e segura que agilize e melhore a qualidade da atenção e dos processos de saúde, nas três esferas de governo e no setor privado, beneficiando pacientes, cidadãos, profissionais, gestores e organizações de saúde".

A equipe de desenvolvimento do DATASUS atua na remodelagem das bases de dados do SUS e na perspectiva de definir um conjunto mínimo de dados, com foco no indivíduo, visando diminuir a fragmentação das bases atuais de atenção a saúde e de modernização tecnológica.

O Conjunto de Dados Mínimos da Atenção à Saúde (CDMS) é uma tradução literal do conceito de *Minimum Dataset Healthcare* (MDH) existente atualmente em diversos países, contendo um conjunto de variáveis objetivando processos administrativos, médicos-administrativos e clínicos (DATASUS - CDM, 2016). O CDMS substituirá gradativamente os modelos de informação atuais do SUS, entre eles, o BPA-C, BPA-I, APAC, SIA, AIH e outros. O modelo de informação do CDMS está dividido em: (i) informações básicas: foram definidas 34 variáveis. É a parte comum que deverá estar presente em qualquer registro de processo assistencial; (ii) informações complementares: necessidade informacional específica de algum processo assistencial particular (especialidades, atenção domiciliar, etc.).

A Associação Brasileira de Normas Técnicas (ABNT, 2016) é o órgão brasileiro de normatização, que representa o Brasil na ISO. Na Informática em Saúde, a representação é feita por meio da Comissão de Estudo Especial em Informática em Saúde - ABNT/CEE-78, fundado em dezembro de 2006, que tem como escopo a normalização no campo de informação para a saúde, tecnologias da informação e comunicação da saúde para adquirir compatibilidade e interação operacional entre sistemas independentes. Visa também assegurar a compatibilidade de dados para propósitos de comparações estatísticas (por exemplo: classificações), e reduzir a duplicação de esforços e redundâncias (ABNT/CEE-78). Essa comissão é espelho do comitê técnico ISO/TC-215 – *Health Informatics* (ISO/TC-215, 2016).

O DATASUS participa da ABNT/CEE-78 em grupos de trabalhos que atuam na padronização dos dados de saúde do paciente traduzindo normas da ISO, as adequando e discutindo seus conteúdos para a realidade brasileira (DATASUS, 2016).

O MS a partir da publicação da Portaria 2.073/2011 (BVSMS) regulamenta o uso de padrões de interoperabilidade e informação para sistemas de informação em saúde no âmbito do SUS, dos sistemas privados e do setor de saúde suplementar. O capítulo II do anexo dessa portaria especifica um catálogo com os padrões de informação para adoção na área de saúde, entre eles podemos citar: HL7, para estabelecer a interoperabilidade entre sistemas; DICOM, para a representação da informação relativa a exames de imagem; LOINC, para a codificação de exames laboratoriais; CID, para a codificação das doenças nos atendimentos em saúde; TISS, para a troca de informações em saúde suplementar (BVSMS; UAB, 2013).

O TISS é uma norma nacional relacionada ao intercâmbio eletrônico de informação, elaborado pela Agência Nacional de Saúde Suplementar - ANS (TISS, 2016). O TISS define o padrão para a troca de informação sobre o atendimento prestado aos beneficiários, entre operadoras de plano privado e prestadores. O objetivo do padrão TISS é atingir a compatibilidade e interoperabilidade funcional e semântica entre os diversos sistemas independentes para fins de avaliação da assistência à saúde (caráter clínico, epidemiológico ou administrativo) e seus resultados, orientando o planejamento do setor.

Entre alguns dos padrões definidos e adotados pelo SUS, podemos citar o Cadastro Nacional de Estabelecimentos de Saúde (CNES, 2016), com o cadastro de todos que realizam qualquer tipo de serviço de atenção à saúde e o Cartão Nacional de Saúde (CNS, 2016), que visa identificar univocamente o cidadão brasileiro no SUS.

A SBIS tem como objetivo promover o desenvolvimento de todos os aspectos da tecnologia da informação aplicada à saúde. Dentro das mais recentes ações, a SBIS, em parceria com o CFM, elaboraram o processo de certificação para Sistemas de Registro Eletrônico de Saúde (S-RES), que se baseia em conceitos e padrões nacionais e internacionais da área de Informática em Saúde, garantindo o armazenamento seguro das informações dos pacientes nos prontuários eletrônicos (Certificação SBIS-CFM, 2013). A definição do que é um S-RES é bastante ampla e abrangendo todos os subsistemas e componentes (SGBDs, servidores, bibliotecas, etc.). O processo de certificação avalia o conjunto completo que compõem o S-RES, devidamente configurados de forma a atender aos requisitos especificados no manual. Dessa forma, qualquer sistema que capture, armazene, apresente, transmita ou imprima informação identificada em saúde pode ser considerado como sendo um S-RES.

A Certificação SBIS-CFM é um processo voluntário e pode ser entendida como "uma opinião técnica, qualificada e imparcial" de duas instituições dispostas a garantir a privacidade e confidencialidade da informação de saúde dos cidadãos, atender a legislação brasileira sobre documentos eletrônicos e melhorar a qualidade dos sistemas de informação em saúde.

Estas iniciativas representam alguns dos esforços de padronização do conteúdo das bases de dados dos sistemas de saúde o que permitirá, uma vez atingido, a comparação direta de resultados de estudos efetuados nas mesmas.

3 BASE DE DADOS ASSISTENCIAIS DO INCOR

O ambiente utilizado no desenvolvimento desta Tese, foi o Instituto do Coração (InCor), um dos institutos vinculado ao Hospital das Clínicas da Faculdade de Medicina da Universidade São Paulo (HC FMUSP). O InCor é um hospital público universitário de alta complexidade, especializado em cardiologia, pneumologia e cirurgias cardíaca e torácica. Utiliza um sistema de informação, denominado Sistema Integrado de Informações Institucionais (SI^3) (Furuie et al., 2003, 2007; Pires et al., 2003; Tess et al., 2009), desenvolvido pelo InCor ao longo de mais de duas décadas com funcionalidades para a assistência ao paciente e administração hospitalar. Sua versão inicial foi implantada em novembro do ano de 2002 e com o passar dos anos, novas funcionalidades foram sendo incorporadas.

O SI^3 encontra-se atualmente em rotina com mais de 20 milhões de acesso por ano e disponibiliza uma base de dados assistencial com informações de mais de 1 milhão de pacientes que foram atendidos nos últimos 20 anos. O SI^3 contempla os requisitos exigidos para o S-RES definidos pela SBIS e CFM com base nas resoluções:

- 1.638/2002 (publicada no D.O.U. de 9 de agosto de 2002, Seção I, p.184-5), que define prontuário médico eletrônico como um conjunto de informações assistenciais, oriundas de atendimentos de saúde, em diferentes âmbitos (ambulatorial ou internação), registradas de forma eletrônica. E que torna obrigatória a criação da Comissão de Revisão de Prontuários nas instituições de saúde;
- 1.639/2002 que aprova as "Normas Técnicas para o Uso de Sistemas Informatizados para a Guarda e Manuseio do Prontuário Médico";
- 1.331/1989 que define a guarda permanente do prontuário do paciente, revogada pela Resolução CFM 1821/2007, que define a guarda do prontuário papel em 20 anos após a data do último atendimento e ainda definiu guarda permanente para os prontuários microfilmados, digitalizados e nascidos eletronicamente.

O SI^3 contempla padrões internacionalmente aceitos para interoperabilidade funcional e semântica, tais como: classificação internacional de doenças, padrão CID-10

(CBCD, 2008; DATASUS CID-10, 2008), transferência de dados e imagens nos padrões HL7 (HL7, 2016) e DICOM (DICOM, 2016). O SI^3 é disponibilizado em um modelo de três camadas, com acesso via interface padrão para a Internet e capacidade de se integrar a outros sistemas de informação existentes no complexo do HCFMUSP, permitindo a incorporação de dados pré-existentes e informações geradas em outros setores ou locais de atendimento.

O Centro de Distribuição de Medicamentos (CDM) do HC FMUSP, onde ocorre a dispensação das medicações prescritas para pacientes ambulatoriais, possui um sistema próprio denominado Sistema de Dispensação de Medicamentos da Farmácia (SIGHFar), desenvolvido pela Companhia de Processamento de Dados do Estado de São Paulo (PRODESP). O SIGHFar está integrado ao SI^3 a partir do módulo de prescrição eletrônica. Quando é elaborada uma prescrição de medicamentos no nível da assistência ambulatorial, é gerada uma receita, que é registrada eletronicamente no SIGHFar e essa receita é a base do controle da dispensação dos medicamentos para o paciente. Os medicamentos disponíveis no CDM e em todas as Unidades de Farmácia do HC FMUSP são padronizados por uma Comissão de Farmacologia vinculada à Diretoria Clínica da Instituição.

Os dados que compõem a base do SI^3 foram incorporados ao sistema em vários momentos de tempo. De 1986 até 1999, os dados foram recuperados de tabelas e de sistemas legados, como o do centro cirúrgico. Em 2005 foi feita uma nova carga de dados em complementação a carga realizada em 1999. No processo de importação, os pacientes receberam como data de cadastro a data da importação.

A base de dados do SI^3 contempla informações do paciente, sua admissão em qualquer processo de atendimento, seja ele do tipo ambulatorial, pronto socorro ou internação no hospital, até a sua alta ou término de atendimento, incluindo os diagnósticos, cirurgias, procedimentos e exames realizados, bem como a medicação prescrita. Uma vez que essa base de dados foi desenhada para registrar a assistência ao paciente, foi necessário o conhecimento prévio de sua estrutura, tabelas, campos de dados, fluxos e relacionamentos, além de ser necessário um processo de seleção, limpeza, transformação e a formatação desses dados para atender as necessidades de um estudo observacional.

As medicações dispensadas pela farmácia ambulatorial começaram a ser registradas no SI^3 em 2003 e as medicações dispensadas para os pacientes internados,

tiveram registro a partir do ano de 2004. No ano de 2006, os resultados dos exames laboratoriais foram integrados e em 2008 foi disponibilizado o módulo de registro das consultas ambulatoriais semiestruturado. Tabelas históricas de cirurgias (1986 a 2007) que continham informações do registro das cirurgias realizadas e diagnósticos dos pacientes, foram incorporadas. A Tabela 3.1 apresenta alguns registros de diagnóstico e quantidade em que foram cadastrados recuperados da tabela de cirurgia na sua forma textual.

Tabela 3.1 Diagnósticos e quantidades de registros na tabela de cirurgia

Diagnóstico	Quantidade
990000000~OUTROS - ICO	11984
990000000~OUTROS	4440
010401200~INSUFICIÊNCIA CORONÁRIA (ICO) CRÔNICA	3903
~	2794
I44.2~Bloqueio atrioventricular total	2053
040200300~BLOQUEIO ATRIO VENTRICULAR (BAV) DO TERCEIRO GRAU	1507
I25~Doenca isquemica cronica do coracao	740
062300000~HEMORRAGIA PÓS-OPERATÓRIA	736
I44.1~Bloqueio atrioventricular de segundo grau	711
040200900~EXAUSTÃO DE GERADOR DE MARCA-PASSO	608
990000000~OUTROS - CIA	569
990000000~OUTROS - ESTENOSE MITRAL	521
I35.0~Estenose (da valva) aortica	519

Esta evolução temporal se reflete na seleção de uma amostra de pacientes, dado que o SI^3 apresenta registros de diagnósticos desde o ano de 1986 e os registros complementares só foram incorporados a partir de 2003 e muitos pacientes tiveram registro de atendimento antes do registro de um diagnóstico.

4 CRITÉRIOS PROPOSTOS PARA A DEFINIÇÃO DO MÉTODO DE EXTRAÇÃO DE COORTE

A aplicação dos critérios de reprodutibilidade em uma pesquisa permite que os resultados de uma análise ou experimento, possam ser repetidos e que se obtenham os mesmos resultados descritos no estudo. Os códigos de computador, software ou outros tipos de instruções que foram usados para processar os dados, o conjunto dos dados e os resultados devem ser disponibilizados em repositórios públicos. Uma pesquisa é designada como reprodutível se for possível executar o código nos dados fornecidos e produzir resultados semelhantes aos que os autores reivindicaram serem reprodutíveis (Peng, 2009).

No domínio da pesquisa de dados em bases assistenciais isto se traduz em uma série de critérios que são propostos e seguidos na definição do método de extração de coorte e estão detalhados a seguir:

1. Uso de padrões em linguagens de programação e ferramentas: O uso de padrões permite que tanto o processo de seleção dos dados quanto os algoritmos de análise estatística, possam ser não apenas reproduzidos de forma independente da base de origem dos dados, mas também, analisados e criticados por um conjunto de revisores focados na qualidade da pesquisa sem precisar depender de uma expertise específica na ferramenta utilizada.

2. Processamento sem perda de informação, preservação do dado não processado (*raw data*) (Green et al., 2009): Os dados originais são aqueles obtidos diretamente como seleção primária da base, sem nenhum tipo de tratamento, e com a seleção mais abrangente possível. Para manter esse conjunto inalterado se faz necessário que a limpeza e transformação se de por meio de um processo que preserve o valor original. No caso de bases de dados estamos falando do uso de visões (consultas) para seleção, limpeza e transformação dos dados, no lugar de inserções, atualizações ou deleções de registros na base.

3. Definição de um conjunto de dados: A seleção de um conjunto de variáveis que serão utilizadas para o mapeamento do esquema lógico da base de origem para o esquema externo;

4. Vocabulários padronizados: Normalmente numa base de dados assistencial são encontrados vocabulários específicos à prática local, aos quais são atribuídos significados variados para melhor implementar o sistema de informação. Estes vocabulários não devem fazer parte das consultas de seleção dos dados na base de origem. Se necessário, para os vocabulários onde existam padrões, os valores originais devem ser processados para gerar traduções que os mapeiem a estes vocabulários. Isto permite uma melhor leitura e crítica aos resultados da pesquisa. O próprio processo de tradução, que deve fazer parte do estudo, pode ser analisado em relação à acurácia do processo.

5. Estatística descritiva da base: Avaliar a qualidade dos dados originais é imprescindível para poder formar com o tempo, critérios de comparação entre bases de estudos semelhantes. Para isto são levantadas estatísticas como, por exemplo, porcentagem de registros repetidos, campos em branco, campos sem identificação, dados inválidos ou incoerentes e dados faltantes. Esta estatística faz parte de relatórios que ficam disponíveis e que contextualizam a base utilizada para a pesquisa.

6. Geração de indicadores da base: Da mesma forma, dentro dos registros que compõem a base de estudo, é possível encontrar coeficientes e indicadores, que uma vez calculados possam ser referenciados a valores conhecidos da literatura. Isto permite qualificar os desvios e vieses que a base possa ter em relação a parâmetros semelhantes. Exemplo: prevalência da ocorrência de infarto agudo do miocárdio em homens na faixa etária acima de 60 anos.

7. Uso de tecnologias abertas: O desenvolvimento de ferramentas estatísticas de código aberto tem crescido ao ponto de se transformar em referência para as análises estatísticas. Isto permite duas coisas: (i) possibilidade de reproduzir os resultados do estudo por qualquer um sem limitação da disponibilidade de uma

determinada ferramenta; (ii) possibilidade de correção de eventuais erros que possam existir na implementação dos algoritmos utilizados no estudo.

8. Geração do estudo como conjunto de códigos executável: O estudo pode ser escrito de forma a incorporar no texto do mesmo, o código necessário para a execução dos algoritmos que geram os resultados da análise.

9. Controle de versão de todos os componentes do estudo: Componentes como conjuntos de códigos de extração e análise, parâmetros, ferramentas estatísticas e qualquer outro elemento que faça parte do estudo, devem estar armazenados em repositórios acessíveis com controle de versão.

O método de extração de coorte proposto se baseia na junção do processo de seleção e depuração dos dados crus, com o processo de análise estatística dos mesmos.

Teremos assim, o alvo da pesquisa reprodutível, que consiste na possibilidade de aplicação constante do estudo a novos conjuntos de dados com a consequente revalidação dos resultados.

4.1 Fundamentos do processamento sem perda de informação, preservação dos dados na extração da coorte

As informações dos sistemas assistenciais que são armazenadas em bancos de dados relacionais se baseiam em uma teoria formal chamada modelo relacional que é baseado na lógica de primeira ordem (ou cálculo de predicados de primeira ordem) (Date, 2000).

O modelo relacional foi introduzido por Edgar F. Codd em 1970 (Codd, 1970). Nesse modelo, todos os dados estão representados como tuplas (listas finitas ordenadas de elementos ou registros) agrupadas em tabelas ou relações.

Uma tabela ou relação é a principal construção usada para representação de dados em um modelo relacional. É uma estrutura composta de linhas e colunas usada para armazenamento de informações. Colunas ou atributos, correspondem aos itens (campos) que deverão ser armazenados. Linhas ou tuplas são um conjunto de campos/atributos, ou seja, o conjunto dos valores que representam uma ocorrência. A

Tabela 4.1 apresenta um exemplo de uma tabela, com linhas e colunas. A Tabela 4.2 apresenta um modelo de uma tabela denominada 'PACIENTE' com as colunas <Id> identificador do paciente, <Nome> nome do paciente e <Idade> idade do paciente, com 5 linhas preenchidas com dados fictícios.

Tabela 4.1 Modelo de uma tabela

	Nome da Tabela		
	Coluna 1	Coluna 2	Coluna n
Linha 1			
Linha 2			
Linha 3			
Linha 4			
Linha n			

Tabela 4.2 Tabela de exemplo: PACIENTE

PACIENTE		
Id	**Nome**	**Idade**
1	MARIA	2
2	LUIZA	10
3	PEDRO	22
4	JOAO	20
5	LUCAS	15

Uma base relacional apresenta os dados em três níveis diferentes de abstração, cada um deles sendo representado por um esquema (*Schema*):

- Esquema Físico: descreve como os dados são armazenados no meio físico: arquivos, índices, etc;
- Esquema Lógico ou Conceitual: apresenta os dados como conjunto de tabelas;
- Esquema Externo: especifica uma reordenação dos dados derivada do esquema lógico, apresentando os dados agrupados de maneiras diferentes de acordo com o utilizador. São realizadas em consultas que oferecem diferentes modos de visualização dos dados das tabelas e não ocupam espaço físico na base de dados.

A Figura 4.1 representa uma abstração do modelo conceitual do banco de dados relacional.

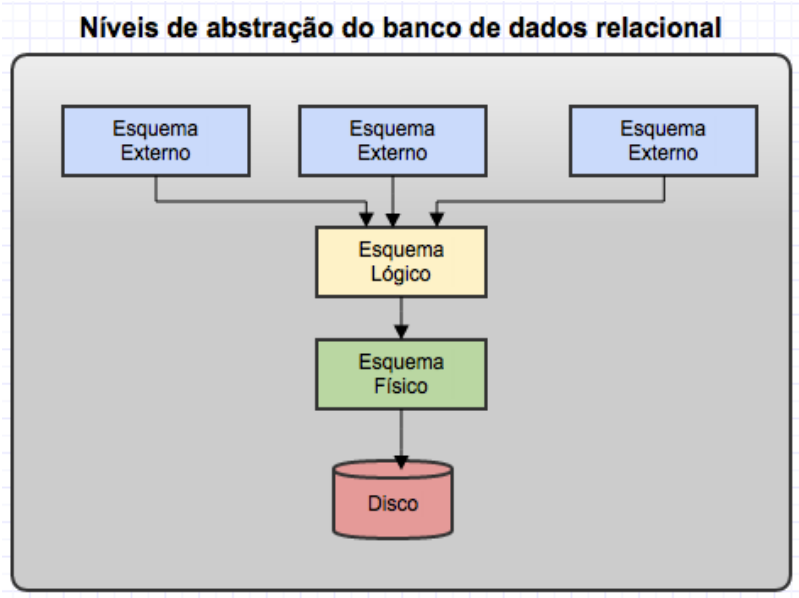


Figura 4.1 Abstração do modelo conceitual

A Figura 4.2 apresenta um exemplo do esquema externo das tabelas com dados de paciente (EM_PATIENT) e da admissão (EM_ADMISSION_DISCHARGE), onde as tabelas do esquema lógico são visualizadas mostrando somente os campos que se quer ver.

EM_PATIENT	
PATIENT_ID	NUMBER(8)
PATIENT_REG_DATE	DATE
PATIENT_NAME	VARCHAR2(60)
GENDER	VARCHAR2(1)
BIRTH_DATE	DATE
PATIENT_MOTHER_NAME	VARCHAR2(60)
COUNTRY	NUMBER(3)
STATE_CODE	VARCHAR2(2)
STATE	VARCHAR2(60)
COUNTY_CODE	NUMBER(8)
COUNTY	VARCHAR2(60)
MARITAL_STATUS	NUMBER(4)
EDUCATION_LEVEL	NUMBER(4)
RELIGION	NUMBER(4)
OCUPATION	NUMBER(5)
RACE	VARCHAR2(3)
MEDICAL_RECORD_CODE	VARCHAR2(12)
MEDICAL_RECORD_DIGIT	VARCHAR2(1)
MEDICAL_RECORD_TYPE	VARCHAR2(3)
MEDICAL_RECORD_DATE	DATE

EM_ADMISSION_DISCHARGE	
PATIENT_ID	NUMBER(8)
ADMISSION_YEAR	NUMBER(4)
ADMISSION_ID	NUMBER(8)
ADMISSION_DATE	DATE
ADMISSION_SECTOR	VARCHAR2(2)
PROVIDER	NUMBER(3)
WEIGHT	NUMBER
HEIGHT	NUMBER
TREATMENT_TYPE	VARCHAR2(3)
SERVICE_TYPE	VARCHAR2(3)
DISCHARGE_DATE	DATE
DISCHARGE_TYPE	NUMBER(2)

Figura 4.2 Exemplo do esquema externo

As linguagens de consulta permitem a manipulação e extração de dados de um banco de dados. Uma consulta é aplicada a instâncias de relações e tem como resultado outras instâncias de relações (Codd, 1970).

O modelo relacional suporta linguagens de consulta como o SQL (*Structured Query Language*) que é uma linguagem padronizada ANSI/ISO (Date e Darwen, 1996).

Esta linguagem tem como fundamentos matemáticos (Ramakrishnan e Gehrke, 2000; Ullman et al., 2009) a Álgebra Relacional e o Cálculo Relacional, duas linguagens formais equivalentes de consulta. A álgebra relacional é uma coleção de operações sobre relações. As operações definidas originalmente por Codd (Codd, 1972) são união, interseção, diferença, produto, restrição, projeção, junção e divisão. Esses conjuntos de operadores originais se estendem a linguagem SQL, que inclui operações de definição e de manipulação de dados. Como exemplos de instruções em SQL, podemos citar:

- Combinações de operações permitem aplicar uma seleção a partir do resultado de outra seleção (subconsulta aninhada).
    - Com base na Tabela 4.2, em SQL, podemos escrever:
        - `select NOME, IDADE from (select * from PACIENTE where IDADE > 18) where NOME='JOAO';`
- As consultas de seleção (*select*) são idempotentes, ou seja, podem ser aplicadas várias vezes sem que o valor do resultado se altere após a aplicação inicial.
    - Em SQL, o resultado da aplicação de uma segunda consulta não altera em nada o resultado da primeira aplicação:
        - `select * from (select * from PACIENTE where IDADE > 18) where IDADE > 18;`
    - A primeira consulta já descartou os pacientes menores de 18 anos, portanto da segunda vez que aplicarmos essa consulta, não haverá novo resultado.
- As operações de seleção são comutativas, a ordem dos fatores não altera o resultado.
    - Em SQL, as seguintes consultas retornam o mesmo resultado:
        - `select NOME, IDADE from (select * from PACIENTE where IDADE > 18) where NOME='JOAO';`
        - `select * from (select NOME, IDADE where NOME='JOAO') where IDADE > 18;`

Estas operações são o fundamento do método para extração de coortes, pois se utiliza apenas de operações de seleção. Garantem que a aplicação do método sobre os dados permite a extração das informações sem alterar o estado original da base e

possibilita a reprodução dos mesmos resultados. Segue um exemplo do processo de descarte de duplicados e valores nulos.

Vamos supor que uma tabela de pacientes proveniente do sistema assistencial pode ser descrita por três campos: identificador do registro do paciente, nome do paciente e data de nascimento, representados pela tupla <id, nome, dataNasc> e os valores apresentados na Tabela 4.3.

Tabela 4.3 Tabela PACIENTE de uma base assistencial

id	nome	dataNasc
1	maria	12/01/1950
2	maria	12/01/1950
3	pedro	12/01/1950
4	null	12/01/1950
5	null	null
6	ana	25/01/1940
7	Jose	01/01/1980
8	jose	01/01/1980
9	jose	01/01/1980
10	joao	null

Podemos notar que na tabela existem registros com valores nulos (inexistentes) e registros repetidos. A seguinte consulta seleciona os valores não nulos:

```
SELECT id, nome, dataNasc FROM PACIENTE
WHERE nome IS NOT NULL AND dataNasc IS NOT NULL;
```

Com esta consulta podemos criar uma visão que fará parte do esquema externo, nomeando a consulta. Uma visão ou vista (*view*) é um conjunto resultado de uma consulta armazenada sobre os dados, em que os usuários do banco de dados podem consultar simplesmente como eles fariam em um objeto de coleção de banco de dados persistente. Este comando de consulta pré-estabelecido é armazenado no dicionário de banco de dados. Diferente das tabelas de base ordinárias em um banco de dados relacional, uma visão não forma parte do esquema físico: como um conjunto de resultado, ele é uma tabela virtual computada ou coletada dinamicamente dos dados no banco de dados quando o acesso àquela visão é solicitado. Alterações aplicadas aos dados em uma tabela subjacente relevante são refletidos nos dados mostrados em invocações subsequentes da visão. O exemplo a seguir apresenta a criação de uma visão dos dados da tabela PACIENTE.

```
CREATE OR REPLACE VIEW PACIENTE_SEM_NULO AS (
   SELECT id, nome, dataNasc
   FROM PACIENTE
   WHERE nome IS NOT NULL AND dataNasc IS NOT NULL );
```

Ao aplicar esta seleção obtemos o resultado apresentado na Tabela 4.4.

Tabela 4.4 Visão da tabela PACIENTE sem registros nulos, denominada PACIENTE_SEM_NULO

id	nome	dataNasc
1	maria	12/01/1950
2	maria	12/01/1950
3	pedro	12/01/1950
6	ana	25/01/1940
7	jose	01/01/1980
8	jose	01/01/1980
9	jose	01/01/1980

A partir deste momento, podemos acessar a visão PACIENTE_SEM_NULO como se fosse outra tabela do banco e utilizá-la em outras seleções. Porém, se verificarmos o banco, notaremos que não existe nenhuma tabela com esse nome, nem o banco aumentou o seu espaço físico. Os dados provenientes do resultado da seleção não estão armazenados no banco de dados, sendo calculada toda vez que acessamos a visão PACIENTE_SEM_NULO.

Agora precisamos achar os registros duplicados na visão PACIENTE_SEM_NULO, o que é expresso pela seguinte seleção:

```
CREATE OR REPLACE VIEW DUPLICADO AS (
   SELECT DISTINCT  A.id, A.nome, A.dataNasc
   FROM PACIENTE_SEM_NULO A
   INNER JOIN  PACIENTE_SEM_NULO B
   ON A.nome = B.nome
   WHERE A.id  != B.id );
```

Ao aplicar o conjunto de instruções acima obtemos uma visão dos registros duplicados, denominada DUPLICADOS apresentados na Tabela 4.5.

Tabela 4.5 Visão dos registros duplicados denominada DUPLICADOS

id	name	dataNasc
1	maria	12/01/1950
2	maria	12/01/1950
7	jose	01/01/1980
8	jose	01/01/1980
9	jose	01/01/1980

Por último criamos uma visão dos dados sem registros duplicados e sem valores nulos:

```
CREATE OR REPLACE VIEW PACIENTE_LIMPO AS (
   SELECT * FROM PACIENTE_SEM_NULO
   MINUS
   SELECT * FROM DUPLICADOS );
```

Onde obteremos como resultado uma visão dos registros de pacientes sem campos nulo e sem duplicados, denominada PACIENTE_LIMPO apresentados na Tabela 4.6.

Tabela 4.6 Visão de pacientes sem dados nulos e duplicados

id	nome	dataNasc
3	pedro	12/01/1950
6	ana	25/01/1940

Podemos observar que:

- Não foi necessário intervir manualmente ou alterar os dados originais apagando registros;
- Todos os registros nulos e/ou duplicados foram descartados.

A segunda observação é uma condição desejável porque não temos como saber qual dos duplicados é o registro real e consequentemente, a real história clínica do paciente. Como nos dados originais existem registros de identificação do paciente duplicados, a história do mesmo paciente pode ter sido fragmentada em registros diferentes.

Evoluções e complementações do método de extração de coorte podem introduzir algoritmos para recuperar os casos de pacientes duplicados. Como os dados originais não foram alterados, eles ainda estão lá.

No Anexo 1 é relatado uma descrição mais formal e detalhada deste processo descrito em SQL, com o auxílio da álgebra relacional.

5 MÉTODO

O método de extração de coortes em base de dados assistencial, proposto nesta Tese, consiste na aplicação de instruções em linguagem padrão SQL, que seleciona e mapeia as informações da base de origem para um esquema externo, sobre o qual possam ser aplicados métodos sistemáticos de limpeza, tratamento e extração, de forma coesa e consistente, de um conjunto de dados para uso em estudos observacionais retrospectivos (Gini et al., 2016).

5.1 Premissas da base assistencial

- As informações do sistema assistencial devem estar armazenadas em um banco de dados relacional;
    - Nem todos os sistemas hospitalares utilizam bancos relacionais. Exemplos: VistA (MUMPS) (Brown et al., 2003), NoSQL (Yu et al., 2013);
- Deve existir um identificador unívoco das informações para o registro do paciente;
- Processamento sem perda de informação: Os dados originais devem ser preservados em relação com a data de cadastro inicial.
    - O sistema deve garantir rastreabilidade das alterações dos dados. O principio é que o método não altera o estado da base, portanto, se existem diferenças, elas são devidas a alterações na base de origem e não consequência da aplicação do método.

5.2 Regras do método de extração de coorte

- É fundamentado na lógica relacional e representado como um **esquema externo** que através de sucessivas aplicações de relações (visões) permite a extração das coortes de forma **sistemática,** isto é, sem a intervenção manual de qualquer usuário da base;
- É idempotente, ou seja, pode ser executado várias vezes gerando sempre os mesmos resultados, sem alterar o estado da base ou os dados originais;

- Propõe um conjunto de dados para ser utilizado no mapeamento do esquema lógico da base de origem para o esquema externo onde o método será aplicado;
- Os dados originais são selecionados e mapeados para um esquema externo composto de visões. Sobre os dados mapeados são aplicadas outras visões nos processos de limpeza e transformação dos dados;
- Possui parâmetros que definem a coorte (data de início e fim de estudo, diagnósticos, intervenção, desfecho, etc.);
- Inclui a geração de análise estatística a partir da aplicação de um conjunto de funções e algoritmos em uma linguagem estatística diretamente nos dados resultantes da extração da coorte;
- Inclui a geração de indicadores que permitem traçar o perfil da base de dados assistencial como: qualidade da base (porcentagem de: duplicados, valores nulos, dados fora de especificação, atendimentos sem diagnósticos), perfil das doenças predominantes no atendimento do hospital (distribuição de diagnósticos), que permitam balizar o resultado das análises;
- O conjunto dos parâmetros utilizados, as análises estatísticas realizadas e os indicadores obtidos compõem o resultado da aplicação do método;
- Para uma determinada coorte extraída pelo método, o valor dos parâmetros, os códigos em SQL e os códigos da análise estatística, devem estar armazenados em um repositório com controle de versão;

5.3 Descrição do método de extração de coorte

Após a definição das fontes de dados é necessário o acesso e conhecimento da estrutura dos dados armazenados na base de origem. A preparação dos dados está relacionada com as tarefas de obtenção e tratamento inicial dos dados.

Os dados selecionados das fontes de origem devem ser mapeados para o esquema externo. A partir do mapeamento os dados devem ser limpos, transformados e consolidados. Após a conclusão da preparação dos dados, são aplicados os critérios do modelo do estudo para a extração da coorte e os códigos de análise estatística para a geração dos resultados. Cada uma dessas operações envolve considerações que serão detalhadas a seguir.

A Figura 5.1 apresenta o esquema geral do método para extração e análise de coortes, aplicado sobre os dados de origem previamente mapeados para o esquema externo.

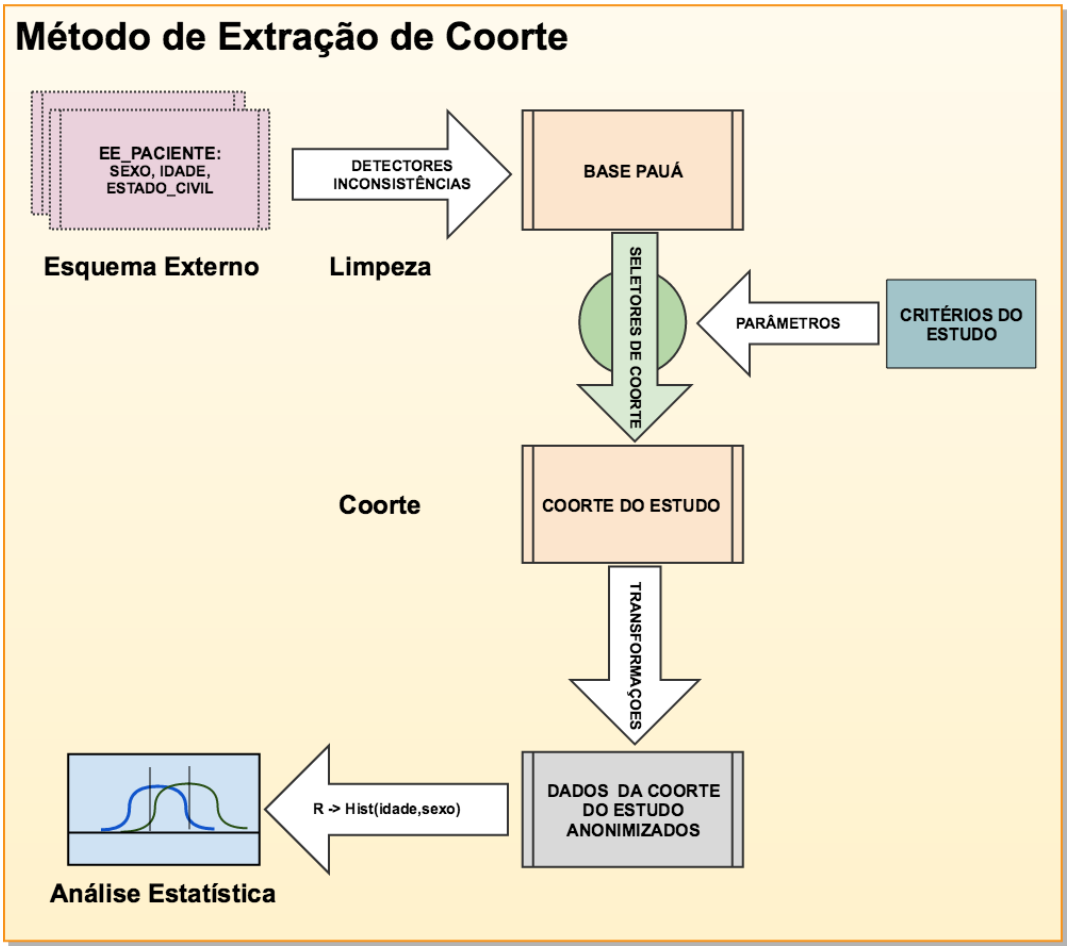


Figura 5.1 Diagrama geral do método de extração de coorte

5.3.1 Mapeamento dos dados de origem para o esquema externo

O método de extração é aplicado sobre o esquema externo. Um conjunto básico de relações mapeia inicialmente os dados do sistema de origem para o esquema externo. Este conjunto básico de relações precisa ser criado uma única vez para cada base de dados de origem. Uma vez mapeadas as informações, o resto do processo pode ser aplicado sem alterações, pois ele depende apenas das informações do esquema externo.

Para o mapeamento, é necessário um levantamento prévio nas fontes de origem, na busca das variáveis com as informações necessárias para a realização do método proposto. Nas visões para mapeamento dos dados, as variáveis estão distribuídas e organizadas de forma a atender as necessidades do estudo. A visão criada da tabela

PACIENTE se relaciona com as outras visões a partir da relação de um para muitos registros (por exemplo: um paciente pode ter registro de uma ou várias consultas, internações, receitas ou exames). No entanto, a relação da visão PACIENTE com a visão ÓBITO se restringe a relação de um para um registro (um paciente pode ter um ou nenhum registro de óbito).

A Figura 5.2 apresenta uma abstração do grupo de variáveis resultantes das visões que foram preparadas para o mapeamento dos dados originais para o esquema externo.

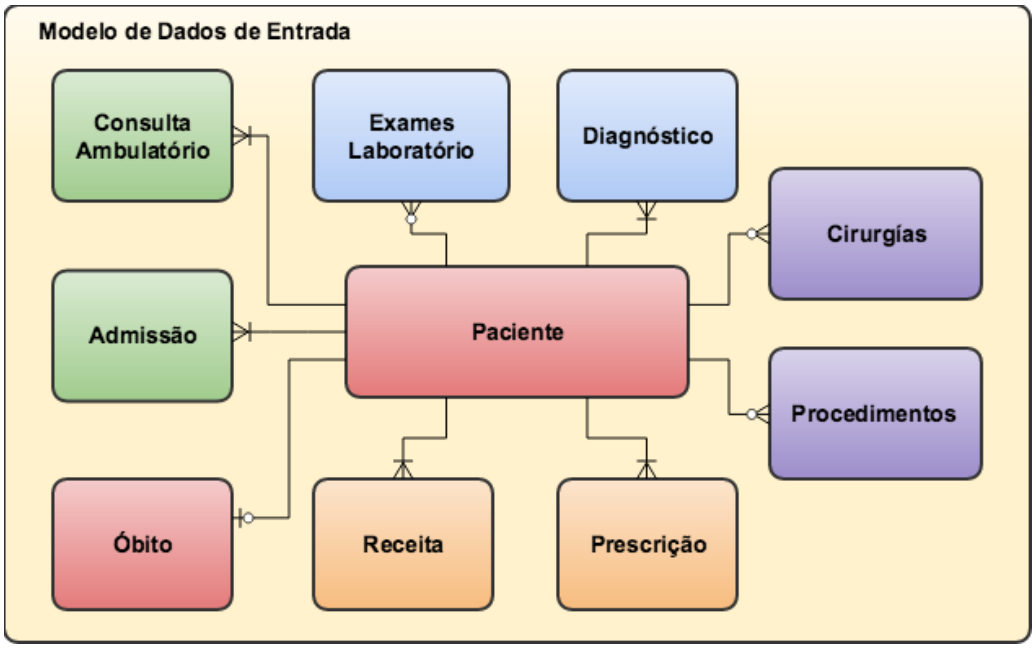


Figura 5.2 Diagrama do esquema externo para o mapeamento da base de origem

As Tabelas 5.1 a 5.10 apresentam a lista completa dos campos e variáveis do mapeamento. A coluna <Campo de mapeamento> corresponde a informação que deve ser buscada nos dados da fonte de origem. A coluna <Variável> corresponde ao campo para qual o dado de origem vai ser mapeado. Como exemplo: a partir do mapeamento a variável que contém a informação do identificador do paciente na base de origem passa a ter como nome de variável <PATIENT_ID> e o conteúdo do identificador. A visão mapeada do paciente é denominada: EE_PATIENT.

Tabela 5.1 Mapeamento das variáveis relativas ao paciente

PACIENTE (EE_PATIENT)	
Campo de mapeamento	**Variável**
Identificação do paciente	PATIENT_ID
Data de registro do paciente	PATIENT_REG_DATE
Nome completo do paciente	PATIENT_NAME
Gênero	GENDER
Data de nascimento	BIRTH_DATE
Nome completo da mãe	MOTHER_NAME
Nacionalidade	COUNTRY
Código do estado	STATE_CODE
Unidade Federação do Estado (UF)	STATE
Código do Município de residência	COUNTY_CODE
Município de residência	COUNTY
Estado Civil	MARITAL_STATUS
Nível educacional	EDUCATION_LEVEL
Religião	RELIGION
Ocupação	OCUPATION
Raça	RACE
Número da matrícula	MEDICAL_RECORD_CODE
Digito da matrícula	MEDICAL_RECORD_DIGIT
Tipo de matrícula	MEDICAL_RECORD_TYPE
Data da matrícula	MEDICAL_RECORD_DATE

Tabela 5.2 Mapeamento das variáveis relativas a admissão

ADMISSÃO (EE_ADMISSION_DISCHARGE)	
Campo de mapeamento	**Variável**
Identificação do paciente	PATIENT_ID
Ano da Admissão	ADMISSION_YEAR
Identificador da admissão (ID)	ADMISSION_ID
Data da admissão	ADMISSION_DATE
Setor de admissão (AMB, INT, PS, SADT)	ADMISSION_SECTOR
Provedor (SUS, CONV, PART)	PROVIDER
Peso	WEIGHT
Altura	HEIGHT
Tipo de tratamento	TREATMENT_TYPE
Tipo de atendimento (URG, ELE)	SERVICE_TYPE
Data da alta ou término da admissão	DISCHARGE_DATE
Tipo de alta	DISCHARGE_TYPE
Identificador de alta	DISCHARGE_ID

Tabela 5.3 Mapeamento das variáveis relativas a exames laboratoriais

LABORATÓRIO (EE_LABORATORY)	
Campo de mapeamento	**Variável**
Identificador do paciente	PATIENT_ID
Data do exame	LAB_TEST_DATE
Nome do exame	LAB_TEST_NAME
Valor do resultado	LAB_TEST_VALUE
Unidade do valor	LAB_TEST_UNIT

Tabela 5.4 Mapeamento das variáveis relativas ao óbito

ÓBITO (EE_DEATH)	
Campo de mapeamento	**Variável**
Identificação do paciente	PATIENT_ID
Data de óbito	DEATH_DATE
CID-10 do óbito (causa básica)	DEATH_ICD

Tabela 5.5 Mapeamento das variáveis relativas ao diagnóstico

DIAGNÓSTICO (EE_DIAGNOSIS)	
Campo de mapeamento	**Variável**
Identificação do paciente	PATIENT_ID
Sequencial do diagnóstico	DIAGNOSIS_ID
Ano da admissão	ADMISSION_YEAR
Identificador da admissão	ADMISSION_ID
Indica se evento ativo, resolvido ou cancelado	STATUS
CID-10 Categoria do diagnóstico	ICD10_CATEGORY
CID-10 do diagnóstico	ICD10
Diagnóstico Principal	PRIMARY_DIAGNOSIS
Data do diagnóstico	DIAGNOSIS_DATE
Data do registro do diagnóstico	DIAGNOSIS_INCLUSION_DATE

Tabela 5.6 Mapeamento das variáveis relativas a cirurgia

CIRURGIA (EE_SURGERY)	
Campo de mapeamento	**Variável**
Identificador do paciente	PATIENT_ID
Ano da admissão	ADMISSION_YEAR
Identificador da admissão (ID)	ADMISSION_ID
Ano da cirurgia	SURGERY_YEAR
Data da cirurgia	SURGERY_DATE
Peso do paciente	WEIGHT_CABG
Altura do paciente	HEIGHT_CABG
Data de óbito	DEATH_DATE
Diagnóstico	DIAGNOSIS
Especialidade Cirúrgica	SURGICAL_SPECIALITY
Procedimento	PROCEDURE

Tabela 5.7 Mapeamento das variáveis relativas a consulta ambulatorial

CONSULTA AMBULATÓRIO (EE_OUTPATIENT_VISIT)	
Campo de mapeamento	**Variável**
Identificação do paciente	PATIENT_ID
Data de registro	REGISTRATION_DATE
Descrição da queixa	SYMPTOM
Duração da queixa (DIAS)	SYMPTOM_DAYS_SINCE
Interrogatório	QUESTIONING
Exame geral físico	PHYSICAL_EXAM
Orientação e conduta	PROCEDURE

Tabela 5.8 Mapeamento das variáveis relativas aos medicamentos

MEDICAMENTOS (EE_PHARMACY)	
Campo de mapeamento	**Variável**
Identificador do paciente	PATIENT_ID
Identificador da receita	PRESCRIPTION_NUMBER
Data da receita	PRESCRIPTION_DATE
Identificador do medicamento	DRUG_ID
Descrição do medicamento	DRUG_NAME
Apresentação do medicamento	DRUG_APRE
Indicador se é intervenção ou não	INTERVENTION_DRUG
Quantidade de medicamentos prescritos	DRUG_QTD
Data de dispensação	DISPENSATION_DATE
Data de retorno	RETURN_DATE
Quantidade solicitada	ORDER_QUANTITY
Quantidade dispensada	DISPENSATION_QUANTITY

Tabela 5.9 Mapeamento das variáveis relativas a angioplastias

PROCEDIMENTO - ANGIOPLASTIA (EE_PCI)	
Campo de mapeamento	**Variável**
Identificador do paciente	PATIENT_ID
Data do exame	DATE_PCI
Procedimento realizado	PROCED_ID

Tabela 5.10 Mapeamento das variáveis relativas a fator de risco cardiovascular

FATOR DE RISCO CARDIO (EE_HEART_RISK_FACTOR)	
Campo De Mapeamento	**Variável**
Identificação Do Paciente	PATIENT_ID
Data Da Consulta	VISIT_DATE
Hipercolesterolemia	HIGH_BLOOD_CHOESTEROL
Diabetes	DIABETES
Obesidade	OBESITY
Tabagismo	SMOKING
Hipertrigliceridemia	HYPERTRIGLYCERIDEMIA
Hiperuricemia	HYPERURICEMIA
Alcoolismo	ALCOHOLISM
Drogas	DRUGS
Contraceptivo Oral	ORAL_CONTRACEPTION
Hipertensão	HIGH_BLOOD_PRESSURE
Estresse	STRESS
Sedentarismo	PHYSICAL_INACTIVITY
Menopausa	MENOPAUSE
Endocardite	ENDOCARDITIS
Febre Reumática	RHEUMATIC_FEVER

A partir desse ponto é necessária a definição do período de inclusão dos registros. Esse intervalo de datas será utilizado como um filtro dos dados que, a partir do mapeamento, os registros serão selecionados a partir de uma data de início até uma

data final determinada. Deverão ser selecionados os dados dos pacientes, diagnósticos, admissões e outros eventos que apresentarem registros nesse intervalo de datas. Para aplicação desse filtro são definidos os parâmetros:

- **Período de inclusão dos pacientes**: Define o período disponível dos dados. Limita o tempo disponível para seleção dos dados na base de origem;
    - Data mínima de inclusão (data de início para seleção dos dados);
    - Data máxima de inclusão (data final para a seleção dos dados;

5.3.2 Limpeza dos dados (detectores de inconsistências)

A limpeza dos dados é o processo de descarte de registros que apresentam variáveis que são fundamentais no preparo do processo, que apresentam alguma inconsistência, como exemplo: campos nulos, datas inválidas, campos de identificação duplicados, que caracterizam registros em duplicidade. Normalmente, as operações de limpeza incluem o preenchimento de valores faltantes, correção de erros de digitação, estabelecimento de abreviações e formatos padrão, substituição de valores por valores padrões, e outros mais. Na aplicação do método, os dados reconhecidos como errados ou com alguma inconsistência, foram separados em visões auxiliares, que na aplicação de novas visões com descartes desses dados, resultou um conjunto de registros consistentes para cada paciente. Como exemplo: o campo <dataNasc>, data de nascimento do paciente, é uma variável necessária para a identificação e os cálculos de idade nos eventos relativos a um paciente, portanto, esse campo deverá estar preenchido e no formato de data válida. Os registros dos pacientes com inconsistência nesse campo foram descartados.

Para realizar a limpeza dos dados foram criadas visões denominadas: detectores de inconsistências. Foi preparada uma visão para cada conjunto de dados mapeados. Como exemplo dos detectores de inconsistências, no caso do paciente, temos: registros duplicados, nome inválido e data de nascimento inválida e no futuro. Estes detectores geram a visão dos registros inválidos, no nosso exemplo, a visão dos pacientes inválidos. Com base nessa visão é gerada a visão dos registros válidos como uma operação de subtração dos registros inválidos do total de registros:

Pacientes válidos = total de pacientes – pacientes inválidos

Cada conjunto de dados mapeados foram validados em questão das datas de registro e as relações com os dados dos pacientes. Por exemplo: um paciente com registro de óbito não pode ter registro de mais nenhum evento a partir desta data, um registro de uma admissão tem de ter um registro de alta associado. Os resultados da aplicação dos detectores são utilizados para criar os indicadores de qualidade da base.

5.3.3 Visão intermediária dos registros assistenciais – Base PAUÁ

O conjunto de visões formadas por todos os registros validados após a limpeza dos dados é denominado base **PAUÁ –** base virtual de dados de registros eletrônicos assistenciais. (Pauá, palavra em tupi-guarani que quer dizer: tudo, muito, no sentido de grande extensão). Os pacientes que compõem a base Pauá possuem no mínimo um diagnóstico no padrão CID-10 e os dados dos eventos complementares registrados.

5.3.4 Definição dos parâmetros

Para a realização de um estudo específico é necessária a definição de uma série de fatores e critérios que possibilitem realizar a seleção dos dados de interesse.

Para tal, foi definido um conjunto de parâmetros relativos a uma questão de pesquisa em especial, ou seja, a criação de uma coleção de dados orientada ao assunto, integrada e consolidada, que permita escolher de formas diversas os dados que serão analisados (seleção da coorte). Foi criada uma tabela de parâmetros para variáveis de seleção, como: período para a seleção dos dados, CID-10 de referência para a doença em estudo, idade mínima e máxima dos pacientes na ocorrência da doença, data de início e fim do estudo e um conjunto de medicamentos, exames e cirurgias. A utilização de parâmetros permite configurar os seletores de coorte e reproduzir a partir do mesmo conjunto de dados, novos estudos, por exemplo: mudando o conjunto de CID, medicamentos ou o período do estudo.

A Figura 5.3 representa o conjunto de pacientes de uma determinada coorte e as diversas situações na janela temporal do estudo e serve como guia dos parâmetros que foram definidos para a configuração dos seletores de coorte (Carvalho et al., 2011; Lui, 2012).

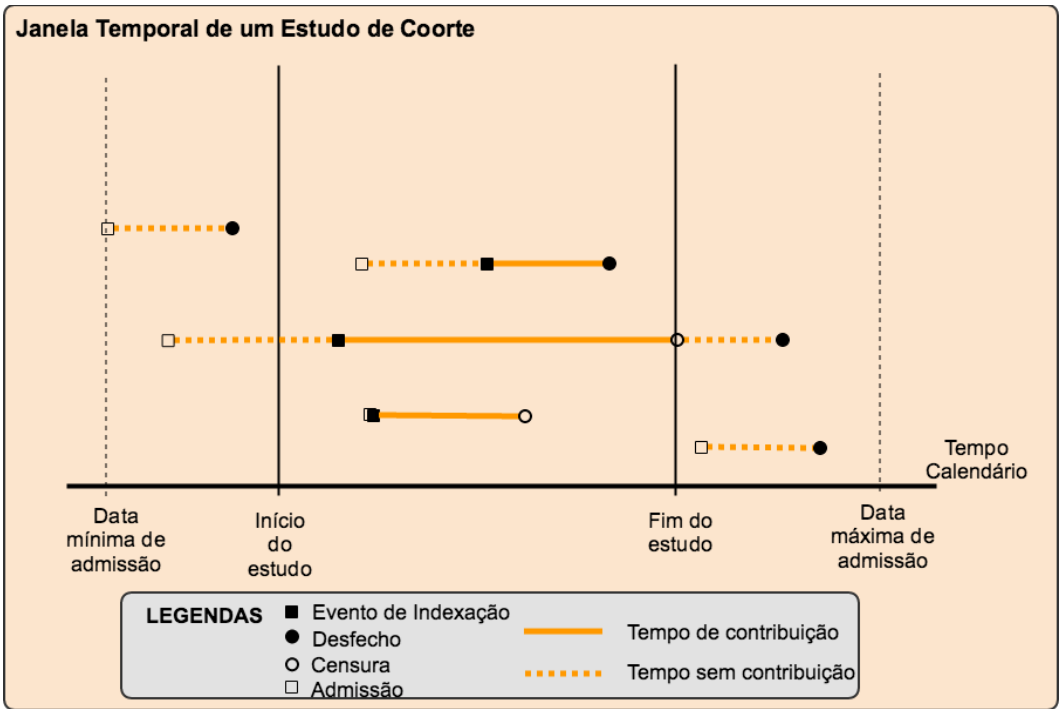


Figura 5.3 Janela temporal de um estudo

Com base na Figura 5.3 foram definidos os seguintes parâmetros:

- **Evento de indexação**: O evento pelo qual se dá o início do acompanhamento do paciente. Por exemplo: Evento = diagnóstico.
  - Evento;
- **Data de indexação**: A data que o evento de indexação foi cadastrado para o paciente.
  - Data de cadastro do evento de indexação;
- **Definição do estudo**: Representa as datas do intervalo para a ocorrência do evento de indexação do paciente na coorte (primeiro diagnóstico);
  - Data inicial;
  - Data final;
- **Características dos pacientes**: Idade mínima e máxima do paciente quando da ocorrência do evento de indexação e os gêneros de interesse para estudo;
  - Idade mínima do paciente na indexação;
  - Idade máxima do paciente na indexação;
  - Gêneros;
- **Seleção da coorte:** Representa o conjunto de diagnósticos possíveis para seleção dos pacientes para compor a coorte. Foi estabelecido o uso do padrão da

Classificação Internacional de Doenças, na sua versão 10 (padrão CID-10), como codificação dos diagnósticos;

- Diagnósticos de inclusão; Códigos do padrão CID-10 para a seleção dos diagnósticos de interesse. Podem ser incluídos grupos ou códigos específicos. Como exemplo: Diagnóstico incluído: I21 – Infarto agudo do miocárdio, I25.2 – Infarto antigo do miocárdio;
- Diagnósticos de exclusão; Códigos do padrão CID-10 que fazem parte de um grupo, mas não serão usados para a seleção dos diagnósticos.
  Como exemplo: Diagnóstico incluído: I63 - Infarto cerebral e diagnóstico excluído: I63.9 – Infarto cerebral não especificado. Inclui todo grupo do I63 excluindo o I63.9.

- **Eventos de interesse**: Ocorrências cadastradas para o paciente que podem ser:
  - Diagnóstico subsequente; Códigos do padrão CID-10 para a seleção dos diagnósticos subsequentes, ou seja, códigos que selecionam o próximo diagnóstico cadastrado após o diagnóstico de indexação. Podem ser incluídos grupos ou códigos específicos.
    - Data do cadastro do diagnóstico subsequente;
    - Intervalo mínimo de tempo a partir da data de indexação até a data do diagnóstico subsequente;
  - Cirurgia;
    - Grupo de cirurgia (seleção da especialidade cirúrgica);
    - Data do cadastro da cirurgia realizada;
    - Intervalo mínimo de tempo a partir da data de indexação até a data da cirurgia;
  - Exames Hemodinâmicos;
    - Grupo de exames (seleção do tipo de exame);
    - Data do cadastro do exame realizado;
    - Intervalo mínimo de tempo a partir da data de indexação até a data do exame;
- **Medicamentos**: Seleção de grupos e medicamentos. Exemplo: Estatina.
  - Conjunto de medicamentos para acompanhamento;
- **Desfecho**: Evento que marca o fim do tempo de participação do paciente na seleção;

- Evento de desfecho;
- Data do desfecho;

O Anexo 2 apresenta uma explicação detalhada dos parâmetros.

A aplicação dos parâmetros permite a detecção de uma situação particular que acontece a partir de um tempo predefinido após o primeiro diagnóstico (evento de indexação) e que chamamos de evento subsequente.

Para este método, foi estabelecido como evento subsequente a ocorrência do primeiro de algum dos seguintes eventos:

- Algum diagnóstico de um grupo pré-selecionado de diagnósticos de interesse;
- Um procedimento de angioplastia percutânea;
- Uma cirurgia de revascularização do miocárdio.

Para cada um destes eventos, pode ser parametrizado um tempo específico para ocorrência após o primeiro diagnóstico. Só serão considerados eventos que aconteçam depois desse tempo específico (parâmetro de intervalo mínimo de tempo após a data de indexação).

Na ocorrência do evento subsequente, são calculadas as seguintes informações: idade do paciente no evento (dias desde o nascimento até a data de ocorrência do evento), intervalo de tempo em dias entre o primeiro diagnóstico e o evento, e um campo identificador do evento subsequente ao primeiro diagnóstico (se ocorreu um diagnóstico do grupo de interesse ou uma angioplastia, ou uma revascularização).

Em relação aos medicamentos, foi preparada uma tabela com os códigos e descrição dos medicamentos de interesse que filtra quais medicamentos serão observados. Para cada paciente, é calculado um campo que indica se o paciente recebeu ou não algum dos medicamentos de interesse durante o período de estudo. Para cada um dos medicamentos é contada a quantidade total de receitas e de comprimidos que foram dispensados para o paciente e calculada uma variável com o tempo entre a primeira receita e a última dispensação.

Por exemplo: no caso de um estudo com foco na doença cardiovascular, foi selecionada uma tabela com um grupo de medicamentos utilizados com maior frequência no tratamento e prevenção da doença. A tabela possui os seguintes campos:

- Nome do grupo: campo texto, exemplo: hipolipemiantes, antiagregantes, outros;
- Código do medicamento no sistema de origem: Campo texto, exemplo: 1141263X

- Nome do medicamento no sistema de origem: Campo texto, exemplo: atorvastatina cálcica 10mg;
- Indicador de intervenção: Indica se o medicamento deve ser considerado como intervenção. Campo texto (0)Não e (1)Sim;
- Pesquisa: Campo texto que indica o nome da pesquisa, exemplo: pesquisa DCV;

O desfecho é definido como o evento que marca o final do período de observação do paciente. Foi considerado como desfecho o registro de óbito. Foi incluído um campo calculado da idade do paciente no momento da ocorrência e um campo com a causa básica do óbito.

5.3.5 Transformação dos dados

Os sistemas hospitalares por característica se modelam no apoio à assistência, onde todas as ocorrências de um paciente são registradas a cada momento, e a disposição da estrutura das tabelas não é a ideal para efetuar análises estatísticas. Mesmo após serem limpos, os dados ainda não estarão na forma apropriada a permitir comparações e análises. Como resultado, a transformação dos dados pode envolver uma divisão ou combinação dos dados de origem, e a geração de novos dados a partir dos dados originais. Nessa fase também são encontradas inconsistências necessitando descarte de dados. A transformação permite mesclar dados de várias origens, consolidando-os e validando datas e horas associadas com o registro dos eventos, correlacionando os eventos de interesse e criando uma sincronização de tempo.

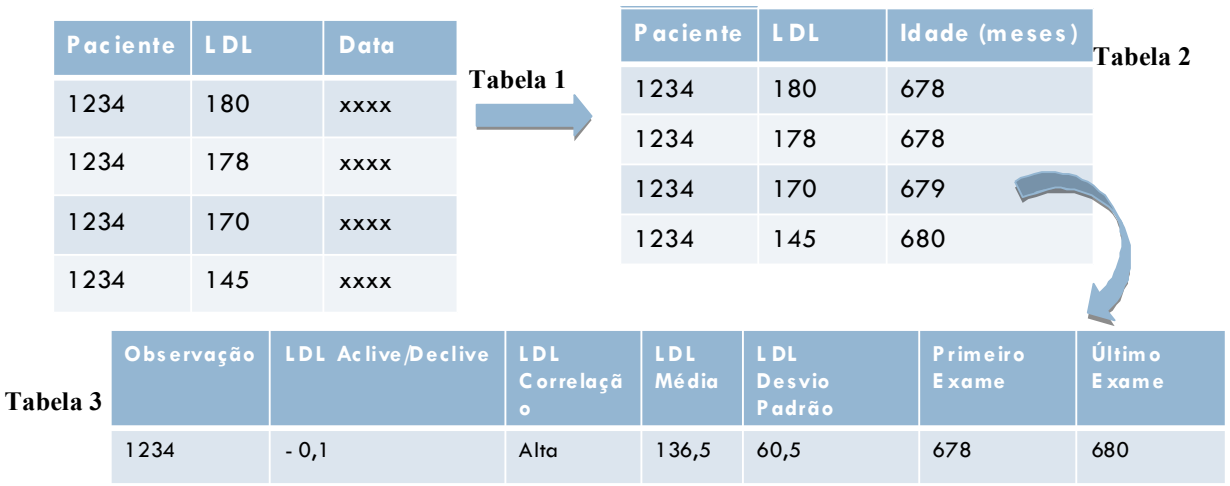
Na preparação desta etapa, os dados de um indivíduo precisam estar organizados em uma única linha e seus valores ajustados para facilitar comparações estatísticas entre pacientes. Foram utilizadas visões que permitiram tratar os dados e formatar o conjunto de variáveis no padrão necessário para a extração da coorte.

Os dados foram separados da seguinte forma:

- Eventos, onde o valor do campo é a data de ocorrência do acontecimento (por exemplo: data de realização da cirurgia, data do óbito, etc.);
- Resultados de exames laboratoriais, onde cada indivíduo pode ter um ou vários resultados para cada tipo de exame. Exemplo: um paciente pode ter registro de vários resultados em várias datas, para o exame de LDL Colesterol.

No caso dos eventos, as datas foram convertidas em dias a partir da data de nascimento do paciente. Foram criadas outras variáveis com os intervalos de tempo em

dias entre o primeiro diagnóstico e esses eventos. Para cada exame selecionado foi feita a regressão linear dos resultados, sendo criadas colunas com os parâmetros da correlação para cada tipo de exame laboratorial. A Figura 5.4 apresenta uma ilustração da transformação dos exames de LDL Colesterol para um paciente com quatro resultados registrados e o resultado da aplicação da regressão linear. Na Figura 5.4 a Tabela 1 apresenta os resultados dos quatro exames de LDL com as datas em que foram realizados, para o paciente fictício 1234. A Tabela 2 apresenta os resultados com as datas de realização transformados em idade do paciente em dias a partir do cálculo de uma nova variável <Idade>. A Tabela 3 apresenta os resultados da regressão linear que foi aplicada nos resultados dos exames de LDL do paciente 1234, com as colunas calculadas do valor da correlação, a média, o desvio padrão e acréscimo das colunas <Primeiro exame> e <Último exame> com valor da idade no primeiro e no último exame.



Paciente	LDL	Data
1234	180	xxxx
1234	178	xxxx
1234	170	xxxx
1234	145	xxxx

Paciente	LDL	Idade (meses)
1234	180	678
1234	178	678
1234	170	679
1234	145	680

Observação	LDL Aclive/Declive	LDL Correlação	LDL Média	LDL Desvio Padrão	Primeiro Exame	Último Exame
1234	- 0,1	Alta	136,5	60,5	678	680

Figura 5.4 Transformação dos resultados de exame de LDL Colesterol

Ainda no processo de transformação, foram acrescentadas outras variáveis com dados que precisaram ser calculados, como por exemplo:

- Número de visitas que o paciente fez no hospital por setor de atendimento (por exemplo: visitas ao pronto socorro, ambulatório e internações) e calculado o intervalo de tempo entre o primeiro e último registro de atendimento no hospital;
- Total de receitas por medicamento e por grupo de medicações;
- Campo indicador de registro de medicação (0)Não e (1)Sim;
- Campo indicador de registro de medicação de interesse como intervenção (0)Não e (1)Sim;

- Criados campos dicotômicos com valor (0)Não e (1)Sim, para indicar a ocorrência de um evento. Por exemplo: óbito, cirurgia e fatores de risco.
- Para anonimizar a identificação do paciente foi gerada um *hash*[1] concatenando o número de registro com o nome do paciente utilizando o algoritmo MD5.

5.3.6 Extração da coorte (seletores de coorte)

Na base formada são aplicados os seletores de coorte, que são visões que filtram registros de acordo com os parâmetros do estudo. Por exemplo, para paciente, temos o seletor de idade no início e fim do estudo (idade mínima e máxima do paciente) e seletor de gênero (masculino, feminino, desconhecido, outros). Cada um dos seletores pode ser independentemente ativado, no caso filtram os registros, ou desativado, para não influenciarem na seleção. O resultado após aplicação dos seletores na base é a coorte do estudo, selecionada segundo os seus critérios de inclusão e exclusão. A coorte anonimizada é disponibilizada para o pesquisador requisitante.

Os cálculos podem ser realizados em uma ferramenta estatística, sendo que a extração disponibiliza todas as informações necessárias no formato adequado. O conjunto das variáveis disponíveis estão resumidas na Tabela 5.11. A descrição completa das variáveis disponíveis e o dicionário dos dados resultante da aplicação dos seletores de coorte estão no Anexo 3.

Tabela 5.11 Variáveis de saída da coorte

Variáveis	Descrição dos Valores
Diagnóstico de indexação	Primeiro diagnóstico registrado para o paciente no padrão CID-10 do grupo definido nos parâmetros de entrada;
Ano do diagnóstico de indexação	Ano do registro do primeiro diagnóstico;
Idade do paciente no diagnóstico de indexação	Idade do paciente em dias no diagnóstico de indexação;
Gênero	Feminino, Masculino;
Faixa etária	Estratificação das idades dos pacientes em anos, no diagnóstico de indexação, por períodos de 10 em 10 anos;
Estado civil	Casado, solteiro, viúvo, divorciado, amasiado, outros;
Escolaridade	Analfabeto, 1 grau incompleto, 1 grau completo, 2 grau incompleto, 2 grau completo, superior;
Indicador de dispensa de medicações de intervenção	1-Sim, 0-Não;

[1] Uma função *hash* é um algoritmo que mapeia dados de comprimento variável para dados de comprimento fixo.

Continuação

Variáveis	Descrição dos Valores
Indicador de dispensa de medicações por grupos de medicamentos	1-Sim, 0-Não;
Quantidade de receitas e comprimidos	Quantidades das medicações dispensadas;
Tempo de dispensação de medicação	Quantidade de dias entre a primeira e a última dispensação;
Intervenção	Data e idade do paciente em dias na ocorrência da intervenção;
Evento	Ano de registro e a idade do paciente em dias na ocorrência do evento (cirurgia, angioplastia);
Desfecho	Óbito (ano de registro e idade do paciente em dias na ocorrência do óbito);
Intervalo entre óbito	Tempo em dias entre o diagnóstico de indexação e o óbito;
Diagnóstico subsequente	CID-10, ano de registro e idade do paciente em dias na ocorrência do diagnóstico subsequente;
Evento evolutivo*	Evento subsequente ao diagnóstico inicial; idade do paciente em dias quando da ocorrência do evento;
Indicador de evento evolutivo**	Indica qual evento foi marcado como evolução da doença inicial (primeiro evento que ocorreu após o diagnóstico de indexação);
Quantidade de eventos***	Indicador de quantos eventos do grupo de eventos evolutivos, o paciente recebeu no período de estudo;
Intervalo entre eventos	Tempo em dias entre o diagnóstico de indexação e o evento evolutivo;
Intervalo entre atendimento	Tempo em dias entre o diagnóstico de indexação e o último registro de atendimento do paciente no hospital - entende-se por atendimento: consulta ambulatorial, exames, pronto socorro ou internações;
Quantidade de atendimentos do paciente no hospital	Total de registros de atendimentos prestados ao paciente no hospital;
Valores dos exames laboratoriais	Parâmetros da regressão linear, média, desvio padrão e variância dos resultados de todas as amostras dos exames para cada paciente;
Quantidade de resultados de exame laboratorial	Quantidade de resultados para cada exame laboratorial;
Fatores de risco	HAS, DM, dislipidemia, tabagismo, inatividade física;

*A variável <Evento evolutivo> representa a idade do paciente em dias na ocorrência do primeiro evento subsequente que foi registrado após o diagnóstico de indexação.

**A variável <Indicador de evento evolutivo> indica qual foi o evento subsequente ocorrido seguinte ao diagnóstico de indexação. Os eventos podem ser: hospitalização por código do padrão CID-10 do grupo dos diagnósticos subsequentes, cirurgia ou angioplastia.

***A variável <Quantidade de eventos> indica quantos eventos do grupo de eventos evolutivos foram registrados para o paciente. O paciente pode ter recebido um novo diagnóstico, uma angioplastia e realizado uma cirurgia. Seriam contados três eventos.

5.3.7 Ambiente computacional

Para o desenvolvimento do método de extração de coorte, foram utilizadas ferramentas e soluções disponibilizadas em código aberto (*open-source*), denominados *software* livre e soluções gratuitas. O seguinte conjunto ferramentas foram utilizadas:

- *SQL Developer* versão 4.1.2: ferramenta gráfica de acesso a bancos de dados, para a seleção, extração e limpeza dos dados e criação da base de dados (Oracle SQL, 2016);
- R versão 3.1.2: ferramenta de mineração de dados e estatística, utilizada na análise dos dados das amostras (R - Project, 2016);
- *RStudio* -Versão 0.98.1091 – © 2009-2014 RStudio, Inc. (RStudio, 2016);
- *RCommander* - Versão 2.1-7 (Karp, 2010);
- Repositório público *GitHub* (GitHub - MTFA).

Em todo o processo foram utilizadas rotinas em linguagem ANSI-SQL-99 (Date e Darwen, 1996).

5.3.8 Considerações éticas

A realização deste estudo foi aprovada pela Comissão de Ética para Análise de Projetos de Pesquisa do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, CAPPesq – HC FMUSP, tendo o projeto sido registrado com o número CEP-6426/2012 (Apêndice 1).

6 RESULTADOS

O método de extração de coortes em base de dados assistenciais, apresentado no Capítulo 5, foi aplicado na base de dados assistenciais do InCor-HC FMUSP.

6.1 Mapeamento dos dados da base assistencial para o esquema externo

No esquema lógico da base de dados assistenciais proveniente do sistema SI^3 foi selecionado um conjunto de tabelas para o preenchimento das visões do esquema externo. Por exemplo: para a visão do paciente denominada EE_PATIENT, foram localizadas e recuperadas na base de origem as tabelas que continham as informações necessárias para preenchimento dos campos. A Tabela 6.1 (uma amostra da tabela 5.1) ilustra o exemplo.

Tabela 6.1 Mapeamento da visão EE_PACIENTE

PACIENTE (EE_PATIENT)	
Campo de mapeamento	**Variável**
Identificação do paciente	PATIENT_ID
Data de registro do paciente	PATIENT_REG_DATE
Nome completo do paciente	PATIENT_NAME
Gênero	GENDER
Data de nascimento	BIRTH_DATE

O campo <PATIENT_ID> foi preenchido pelo campo correspondente a identificação do paciente na base do SI^3. O campo <PATIENT_REG_DATA> com a data em que foi feito o cadastro do paciente no sistema. Todas as variáveis das visões mapeadas foram preenchidas da mesma forma.

A Figura 6.1 apresenta um exemplo do mapeamento dos dados da base do SI^3 para pacientes e diagnósticos, apresentando a seleção das tabelas do esquema lógico e o seu mapeamento para o conjunto de dados do esquema externo.

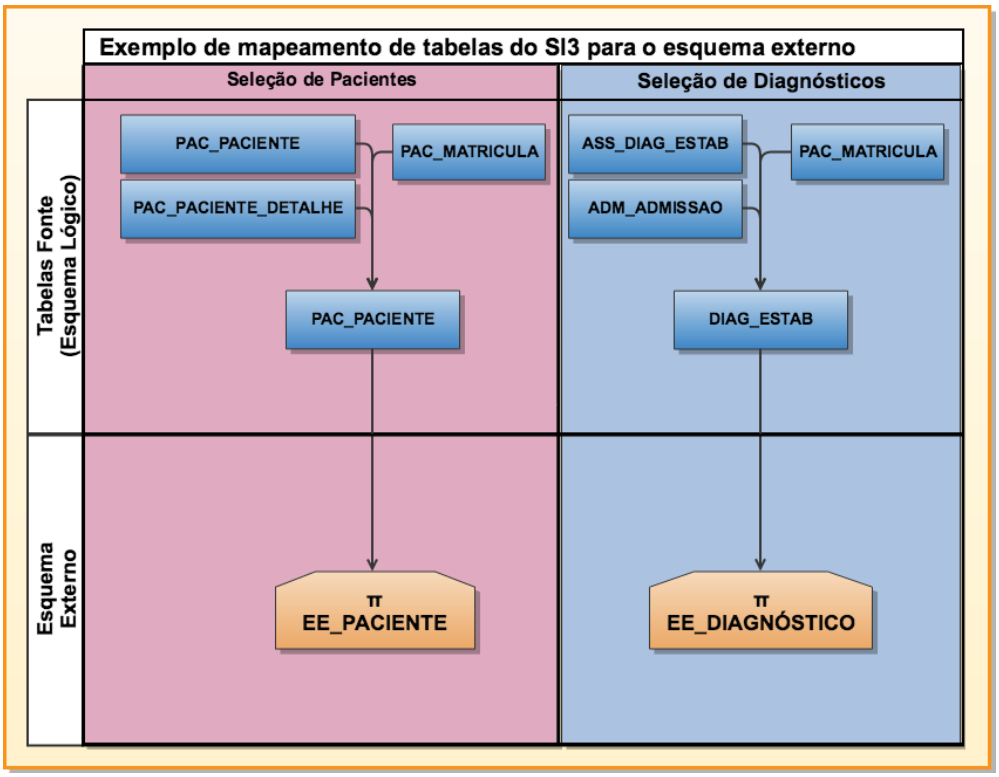


Figura 6.1 Exemplo de mapeamento do esquema lógico do SI^3 para o esquema externo

A Figura 6.2 apresenta o modelo relacional das operações algébricas que descrevem o método. A figura apresenta o fluxo seguido para a seleção dos pacientes e dos diagnósticos, do mapeamento até a extração da coorte. Os operadores básicos da álgebra relacional representados na figura 6.2 estão descritos no Anexo 1. Como exemplo, (σ) seleção, seleciona um subconjunto de registros de uma relação; (π) projeção, descarta colunas indesejadas de uma relação.

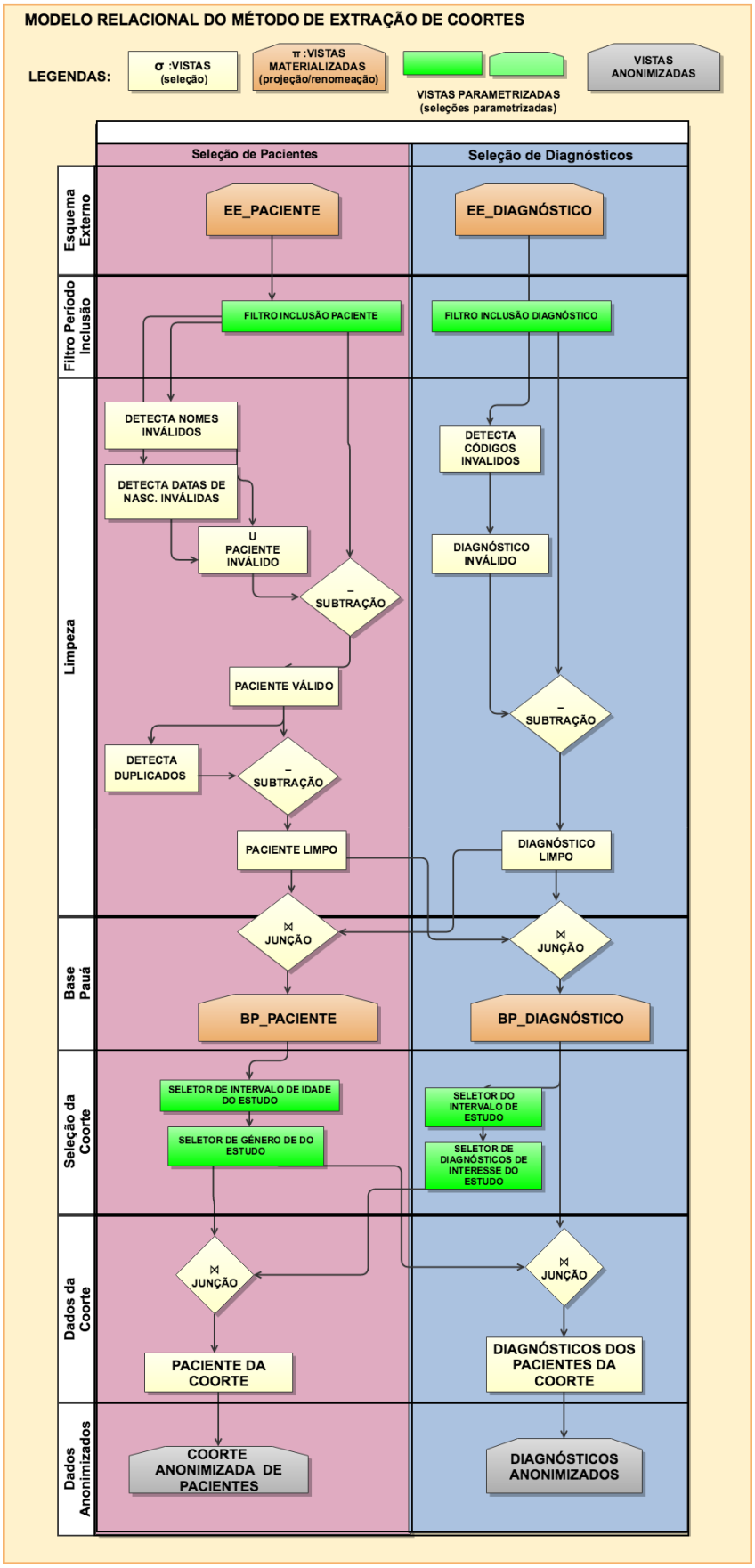


Figura 6.2 Modelo relacional do método de seleção de coorte

6.2 Limpeza dos dados

A partir do mapeamento são aplicados os detectores de inconsistências, para a limpeza e descarte dos dados inconsistentes. O resultado são outras visões que apresentam somente os dados válidos. Os indicadores de qualidade da base são gerados a partir da compilação das inconsistências encontradas e o perfil da base, a partir dos dados válidos.

6.3 Base Pauá de registros assistenciais

A base Pauá é formada pelo conjunto de visões dos registros válidos para a base de dados do SI[3].

6.4 Definição do estudo e dos parâmetros

Foi realizado um estudo observacional retrospectivo, para análise de pacientes portadores de doença cardiovascular (DCV), com dispensação da medicação estatina na prevenção secundária de novos eventos da doença.

Como critérios de inclusão foram definidos: período de 2003 a 2013, pacientes maiores de 18 anos, dos gêneros masculino e feminino, com um diagnóstico no padrão CID-10 (I20, I21, I22, I23, I24, I25, I64, I65, I67, I69, I70 e G45), com registro de no mínimo duas consultas ambulatoriais e um mês de período mínimo de acompanhamento do paciente (tempo entre eventos).

A coorte é composta de pacientes com e sem registro da medicação estatina. A análise compara os grupos, verifica a ocorrência de óbito, eventos evolutivos, o tempo entre esses eventos e o tempo de dispensa de estatinas. A Figura 6.3 apresenta o modelo do estudo.

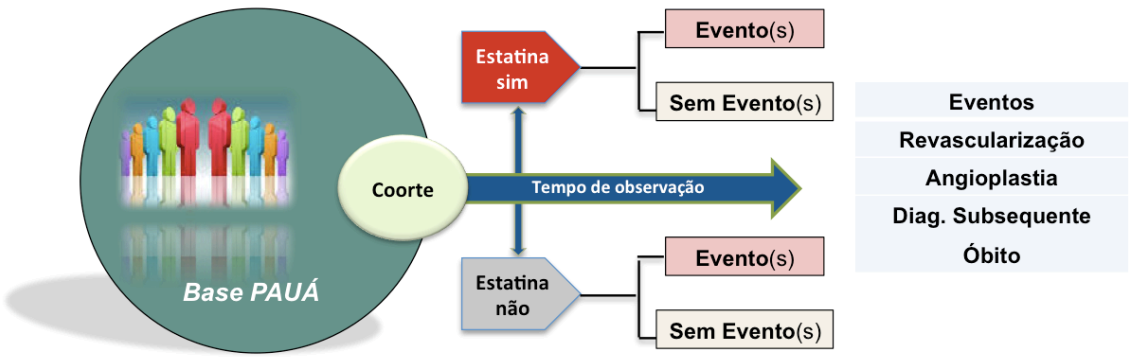


Figura 6.3 Modelo do estudo retrospectivo da DCV

Foram definidos os seguintes parâmetros para aplicação dos seletores de coorte:

- **Definição do período do estudo**:
  - Data inicial = 01/01/2003;
  - Data final = 31/12/2013;
- **Características dos pacientes**:
  - Idade mínima => igual ou maior que 18 anos na data do diagnóstico de indexação;
  - Idade máxima => sem limite de idade;
  - Gêneros dos pacientes => Masculino e Feminino;
- **Diagnósticos para seleção da coorte**:
  - Diagnósticos de inclusão => diagnósticos no padrão CID-10: I20, I21, I22, I23, I24, I25, I64, I65, I67, I69, I70 e G45;
  - Diagnósticos de exclusão => não se aplicam;
- **Evento de indexação**:
  - Diagnósticos de inclusão; (evento que define da data de indexação pelo qual se dá o início do acompanhamento do paciente na coorte).
- **Evento subsequente/interesse**:
  - Diagnósticos para evento subsequente: Diagnósticos no padrão CID-10: I21, I22, I23, I50, I63, I64, I65, I66, I69 e G45; (diagnóstico de agravo que ocorreu após o diagnóstico de inclusão).
  - Tempo mínimo para ocorrência do evento subsequente após a data de indexação => um mês;
  - Eventos de interesse: angioplastia percutânea e cirurgias de revascularização do miocárdio;
- **Desfecho**:
  - Óbito;
- **Exames laboratoriais**:
  - Selecionado os registros com resultados dos exames: Colesterol total, LDL colesterol, HDL colesterol, Glicemia, hemoglobina glicada;
- **Medicamentos**:
  - Conjunto de medicamentos para acompanhamento, agrupados em categorias: hipolipemiantes, hipotensores, hipoglicemiantes, antiagregantes descritos no Anexo 4;

- **Intervenção**: medicações estatinas.

6.5 Transformação dos dados

Na transformação dos dados, as datas foram convertidas em idade do paciente em dias na ocorrência do evento e foram criadas colunas para cada ocorrência. Para os exames selecionados (HDL, LDL, Colesterol Total, Glicemia e Hemoglobina Glicada) foi feita a regressão linear dos resultados de cada tipo de exame, sendo criadas colunas com os parâmetros da correlação. Ainda no processo de transformação, foram acrescentadas as variáveis com dados calculados, como por exemplo: quantas visitas (atendimentos do paciente no hospital), intervalo de tempo entre o primeiro e último registro de atendimento no hospital, total de receitas por medicamento e por grupo de medicações, intervalo de tempo em dias, entre o primeiro diagnóstico e o óbito e outros eventos e um campo indicador de registro de estatina e dos fatores de risco (1-sim e 0-não).

6.6 Extração da coorte

Na base Pauá foram aplicados os seletores de coorte a partir dos parâmetros definidos para o estudo. Foi extraída a coorte DCV, com um conjunto de variáveis formatadas para importação na ferramenta estatística. As variáveis e suas descrições são apresentadas no Anexo 3. A Figura 6.4 apresenta um diagrama de representação da coorte DCV, mostrando uma tupla para cada paciente.

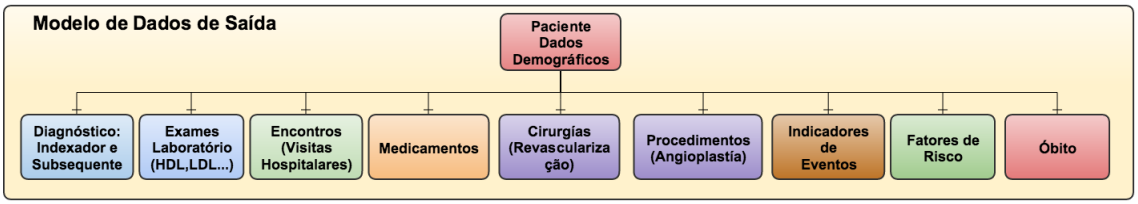


Figura 6.4 Diagrama de representação dos dados de saída

6.7 Caracterização da base de dados SI^3

Os indicadores do perfil da base apresentam os dados selecionados, que foram mapeados do SI^3 para o esquema externo, sem nenhum tratamento, somente com uma

contagem dos dados. Os indicadores de qualidade da base foram gerados na aplicação da limpeza dos dados do esquema externo para formar a visão dos dados validos.

6.7.1 Indicadores de perfil da base

- *Pacientes cadastrados e pacientes com admissões:*

A quantidade de pacientes cadastrados na base de dados do SI^3 no período de 01/01/1985 a 01/01/2014 (filtro de inclusão) é de 1.116.848, sendo 45,4% de pacientes do gênero feminino, 43,2% do gênero masculino e 11% desconhecido e outros. A média de idade para o gênero feminino é de 65 anos (IC 95% 62-68 anos) e para o gênero masculino é de 63 anos (IC 95% 60-66 anos). Foram registradas 4.782.708 admissões, distribuídas entre internações, atendimento ambulatorial e pronto socorro, sendo em média 6 atendimentos por paciente. A Tabela 6.2 apresenta a distribuição dos pacientes registrados e a distribuição dos pacientes que tiveram registro de admissões por gêneros.

Tabela 6.2 Distribuição de pacientes/admissões por gêneros

Gêneros	Pacientes n (%)	Pacientes com Admissões n (%)
Feminino	507.335 (45,4)	374.673 (49,6)
Masculino	482.957 (43,2)	343.007 (45,4)
Desconhecido	85.392 (7,6)	2.252 (0,3)
Outros	41.164 (3,7)	35.382 (4,7)
Total	**1.116.848 (100)**	**755.314 (100)**

- *Pacientes por ano de cadastro:*

No ano de 1999 foi efetuada a importação dos dados de sistemas legados para a base do SI^3. Em novembro de 2002 teve início o sistema SI^3 a partir da implantação de seus primeiros módulos. Em abril de 2005 foi realizada uma nova carga de complementação dos dados para a base do SI^3. O Gráfico 6.1 apresenta a distribuição dos pacientes cadastrados no SI^3 por ano. O Gráfico 6.2 apresenta a quantidade total de pacientes no período de 1999 a 2013.

Gráfico 6.1 Quantidade de pacientes no SI[3] por ano de cadastro

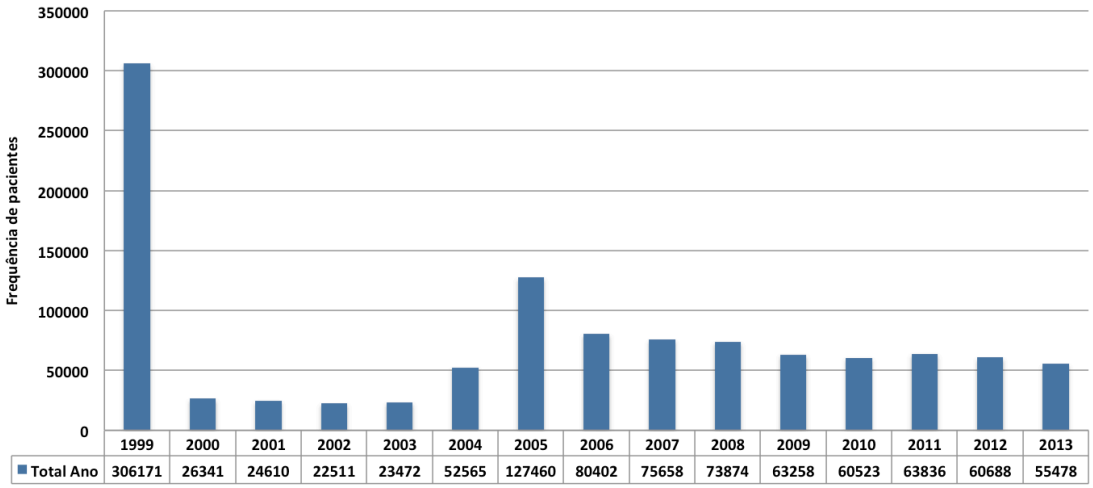
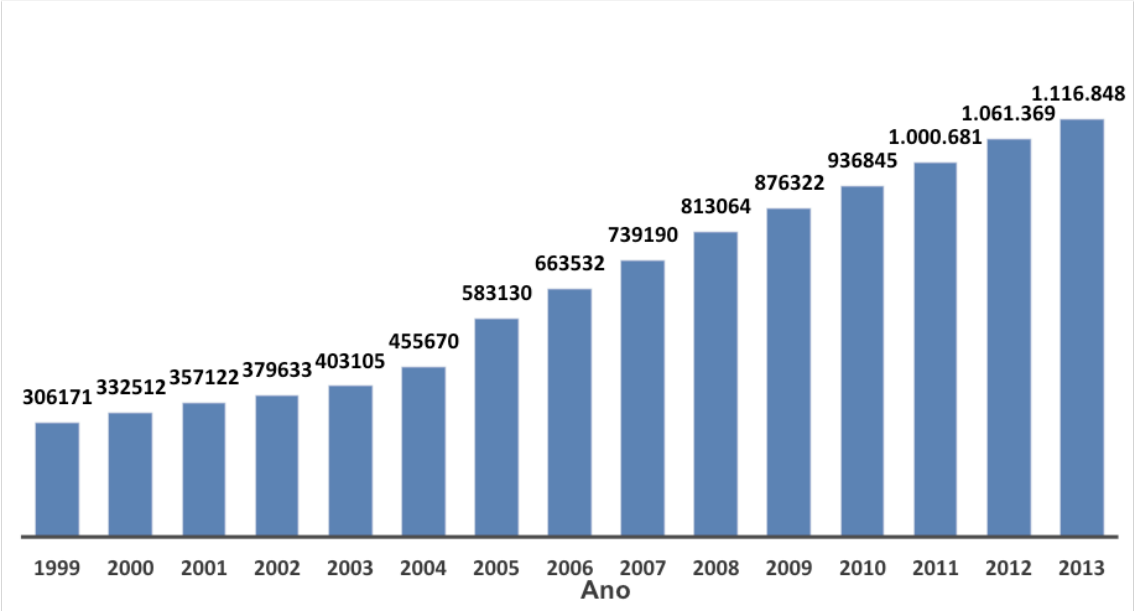


Gráfico 6.2 Total acumulado de pacientes por ano (1999-2013)



- Distribuição de diagnósticos:

O total de diagnósticos cadastrados na base do SI[3] para o período de seleção foi de 949.669 diagnósticos para 361.662 pacientes (32,4% do total de pacientes possuíam um ou mais diagnósticos registrados). Do total, 10.911 diagnósticos (1,1%) não estavam especificados no padrão CID-10. Os diagnósticos codificados no CID-10 totalizaram em 938.758 diagnósticos, referentes a 360.564 pacientes (32,3% do total de pacientes), sendo em média 2,6 diagnósticos por paciente. A Tabela 6.3 apresenta as quantidades e percentuais de diagnósticos encontrados na base do SI[3].

Tabela 6.3 Perfil da codificação dos diagnósticos da base do SI[3]

Diagnósticos	Quantidade	%
Total de diagnósticos	949.669	100
Codificados no CID-10	938.758	98,9
Outras Codificações	10.911	1,1
*Total de pacientes com diagnóstico CID-10	360.564	32,3

* Quantidade total de pacientes com registro de algum diagnóstico CID-10

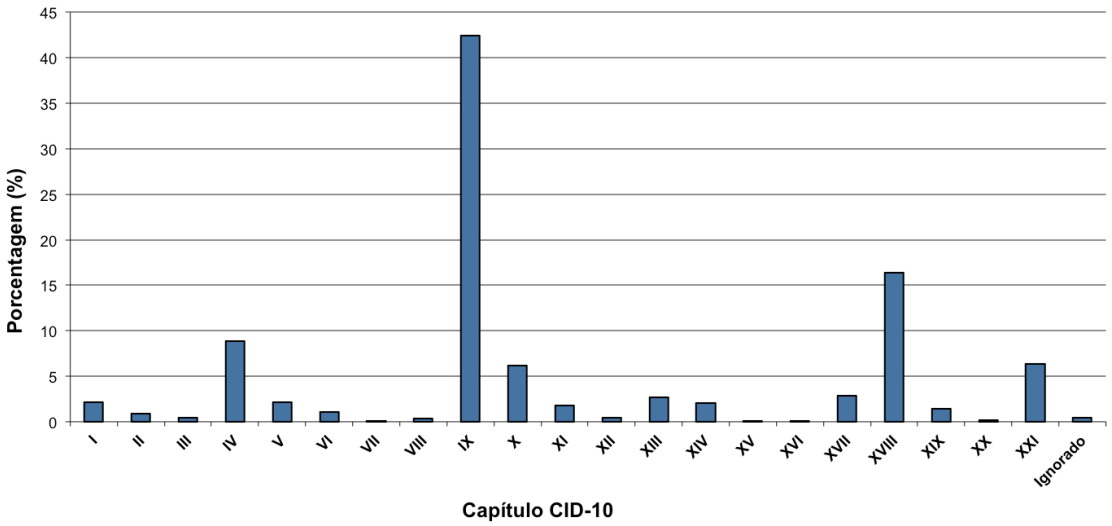
O perfil dos diagnósticos no padrão CID-10 encontrado nos dados do SI[3], mostra que 42,5% dos diagnósticos são do capítulo IX – Doenças do aparelho circulatório, 16,4% para o capítulo XVIII – Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte, e 8,8% para o capítulo IV – Doenças endócrinas, nutricionais e metabólicas. A Tabela 6.4 apresenta o total dos diagnósticos registrados no SI[3] por capítulo do padrão CID-10.

Tabela 6.4 Diagnósticos por capítulos CID-10 registrados na base SI[3]

Capítulo CID	Descrição	Quantidade	(%)
I	Algumas doenças infecciosas e parasitárias	20.356	2,1
II	Neoplasias [Tumores]	8.342	0,9
III	Doenças do sangue e dos órgãos hematopoiéticos e alguns transtornos imunitários	4.267	0,4
IV	Doenças endócrinas, nutricionais e metabólicas	83.879	8,8
V	Transtornos mentais e comportamentais	20.265	2,1
VI	Doenças do sistema nervoso	9.916	1,0
VII	Doenças do olho e anexos	1.022	0,1
VIII	Doenças do ouvido e da apófise mastoide	3.135	0,3
IX	Doenças do aparelho circulatório	403.158	42,5
X	Doenças do aparelho respiratório	58.322	6,1
XI	Doenças do aparelho digestivo	17.195	1,8
XII	Doenças da pele e do tecido celular subcutâneo	4.408	0,5
XIII	Doenças do sistema osteomuscular e do tecido conjuntivo	25.287	2,7
XIV	Doenças do aparelho geniturinário	19.771	2,1
XV	Gravidez, parto e puerpério	400	0,0
XVI	Algumas afecções originadas no período perinatal	144	0,0
XVII	Malformações congênitas, deformidades e anomalias cromossômicas	27.101	2,9
XVIII	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	155.801	16,4
XIX	Lesões, envenenamentos e algumas outras consequências de causas externas	13.990	1,5
XX	Causas externas de morbidade e de mortalidade	1.260	0,1
XXI	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	60.739	6,4
Outros	Diagnóstico não especificado no padrão CID-10	10.910	1,1
Total	Total dos diagnósticos	949.669	100

O Gráfico 6.3 apresenta a distribuição gráfica do perfil dos diagnósticos do InCor em porcentagem por Capítulo do CID-10.

Gráfico 6.3 Distribuição dos diagnósticos por capítulo do CID-10



O perfil encontrado se justifica por ser o InCor um hospital de referência em doenças cardiovasculares.

- *Distribuição das cirurgias:*

O InCor é um hospital de referência em cirurgias cardíacas e torácicas e tem registrado no SI[3] 112.104 cirurgias realizadas de novembro de 1986 até dezembro de 2013. O Gráfico 6.4 apresenta a quantidade de cirurgias realizadas no seu centro cirúrgico por especialidade. A especialidade "Coronária" corresponde as cirurgias de revascularização do miocárdio e foram realizados 27.885 procedimentos, que representa cerca de 24,9% do total de cirurgias realizadas no InCor. Na legenda "Outras" foram acumuladas especialidades com menos de 3.000 cirurgias/período, somando 7.624 (6,8%) das cirurgias. O Gráfico 6.5 apresenta a distribuição das cirurgias de revascularização do miocárdio para a especialidade médica de coronária, realizadas no InCor por ano de seu registro. O ano de 1986 não foi considerado no gráfico, pois no ano em questão, só foram registradas 133 cirurgias. As cirurgias realizadas antes de

1999 não apresentam informação do registro de uma internação associada. Esses registros foram recuperados diretamente de sistema legado do centro cirúrgico e foram utilizados para a complementação dos eventos evolutivos.

Gráfico 6.4 Quantidade de cirurgias por especialidade (1986 a 2013)

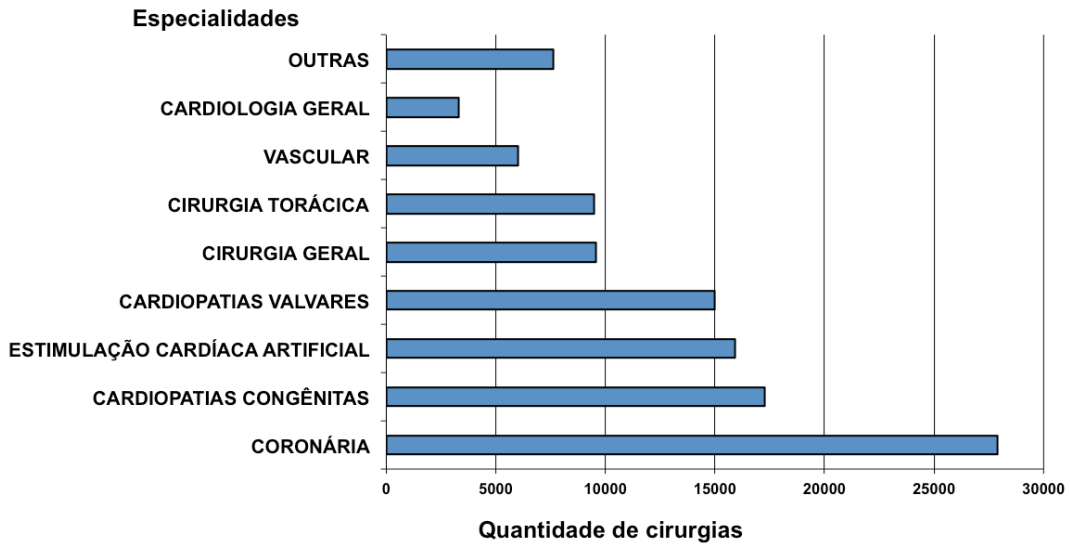
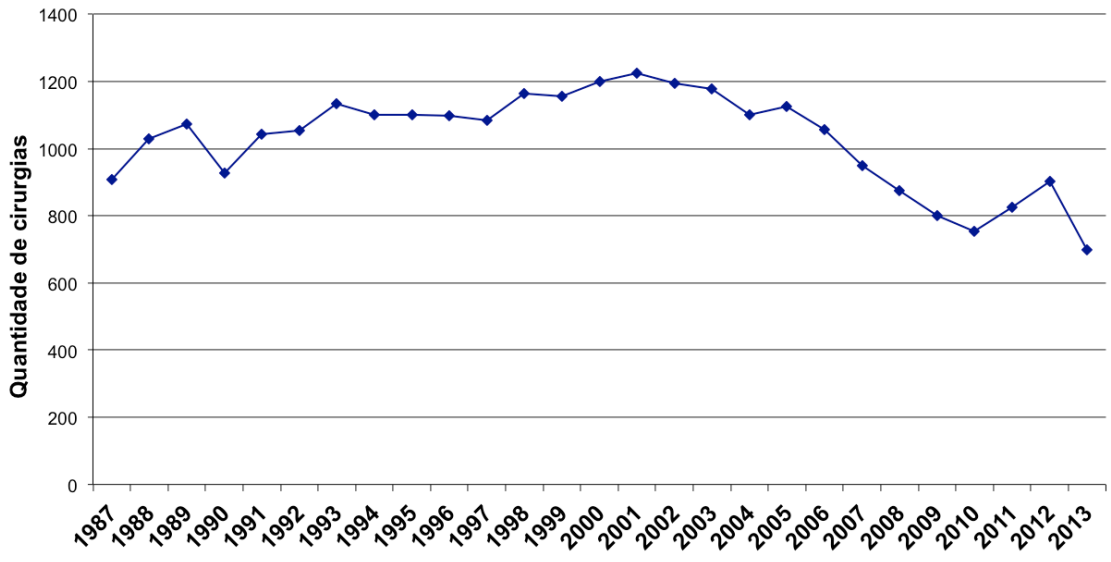


Gráfico 6.5 Distribuição das cirurgias de revascularização do miocárdio (1987 a 2013)



6.7.2 Indicadores de qualidade

No processo de limpeza dos campos <nome do paciente> e <data de nascimento> sem preenchimento, foram descartados da seleção 20.997 registros de pacientes sem data de nascimento preenchida e não foram encontrados registros sem nome de paciente cadastrado. Foram encontrados 14 pacientes com data de nascimento no futuro (data de nascimento posterior ao cadastro do paciente). No processo de retirada de nomes de pacientes em duplicidade (foram comparados nomes e datas de nascimento igual) foram descartados 114.441 pacientes. Após a limpeza destes campos, foi verificado o campo <diagnostico> e foram descartados 669.017 pacientes que não estavam com esse campo preenchido ou com diagnóstico não codificado no padrão CID-10. Restaram na seleção 312.469 pacientes, sendo esse o total dos pacientes validos, com diagnósticos codificados no padrão CID-10. A Figura 6.5 apresenta o fluxo da limpeza dos dados.

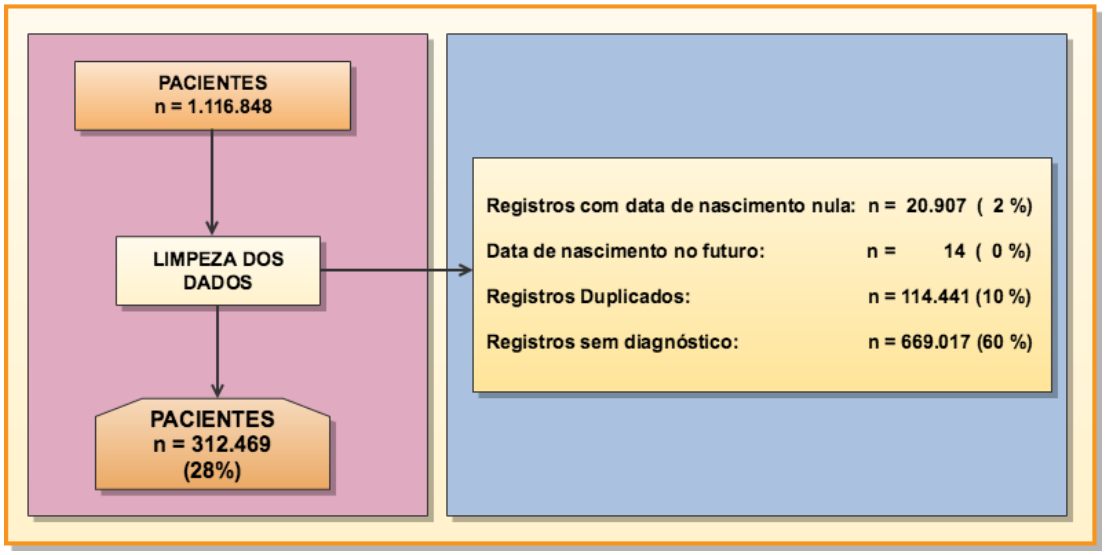


Figura 6.5 Fluxo da aplicação da limpeza dos dados

A Tabela 6.5 apresenta a compilação das inconsistências encontradas na preparação da base por ano do cadastro dos pacientes. A coluna <Ano do cadastro> informa o ano de cadastro do paciente no SI[3], a coluna <Paciente cadastro/ano> a quantidade de pacientes cadastrados no ano, a coluna <Data Nasc. nula> a quantidade de registros com data de nascimento do paciente sem preenchimento, a coluna <Data Nasc. Futura> registros de data de nascimento posterior a data de cadastro do paciente, a coluna <Nome da mãe nulo> com o total dos registros sem preenchimento dessa

informação, a coluna <Registros duplicados> com o total dos registros de pacientes com nome em duplicidade e a coluna <Sem diagnóstico> o total de pacientes sem nenhum registro de diagnóstico. Cada coluna apresenta a quantidade para cada detector de inconsistências aplicado. Um registro de paciente pode ser contato em cada um dos detectores.

Tabela 6.5 Resumo das inconsistências dos dados da base SI[3]

Ano do cadastro	Pacientes cadastrado/ano	Data Nasc. nula	Data Nasc. Futura	Nome da mãe nulo	Registros duplicados	Sem diagnóstico
1999*	306.171	1.526	4	1.742	9.402	237.459
2000	26.341	212	0	192	3.268	10.466
2001	24.610	108	1	238	2.906	8.458
2002	22.511	51	0	171	1.866	7.082
2003	23.472	176	0	379	2.015	8.079
2004	52.565	2.426	1	7.344	7.405	22.622
2005*	127.460	3.095	1	13.939	14.960	81.225
2006	80.402	5.713	0	26.051	15.317	41.965
2007	75.658	2.477	0	24.984	13.837	44.247
2008	73.874	1.623	0	31.584	15.956	42.068
2009	63.258	950	0	19.262	8.069	36.035
2010	60.523	792	0	15.377	5.631	33.914
2011	63.836	614	0	15.392	5.882	35.443
2012	60.688	688	6	10.344	4.618	30.394
2013	55.479	456	1	8.287	3.309	29.560
Total	**1.116.848**	**20.907**	**14**	**175.286**	**114.441**	**669.017**

* Ano de realização de carga dos dados para a base do SI[3].

A Tabela 6.6 apresenta a quantidade e percentual de pacientes após o descarte dos dados inconsistentes, para os pacientes e os diagnósticos.

Tabela 6.6 Limpeza dos dados: percentual de dado de pacientes validados

Descrição	Quantidade Pacientes	Percentual Total (%)
Total de pacientes cadastrados	1.116.848	100
Pacientes com data nascimento válidos	1.094.927	98
Sem pacientes em duplicidade	981.486	88
Com diagnóstico no padrão CID-10	312.469	28

A Tabela 6.7 apresenta os totais de registros dos pacientes por ano de sua inclusão no SI[3], após a limpeza. A coluna <Paciente registro/ano> apresenta a quantidade de pacientes que tiveram registro no ano. A coluna <Registros não selecionados> apresenta o total dos registros de pacientes com alguma inconsistência

(data de nascimento nulo, data de nascimento no futuro, registros duplicados e registros sem diagnóstico cadastrado). A coluna <Registros válidos> apresenta o total dos registros limpos disponíveis para a seleção da coorte a partir da aplicação dos critérios pré-definidos para o estudo da doença.

Tabela 6.7 Totais de registros de pacientes por ano após limpeza

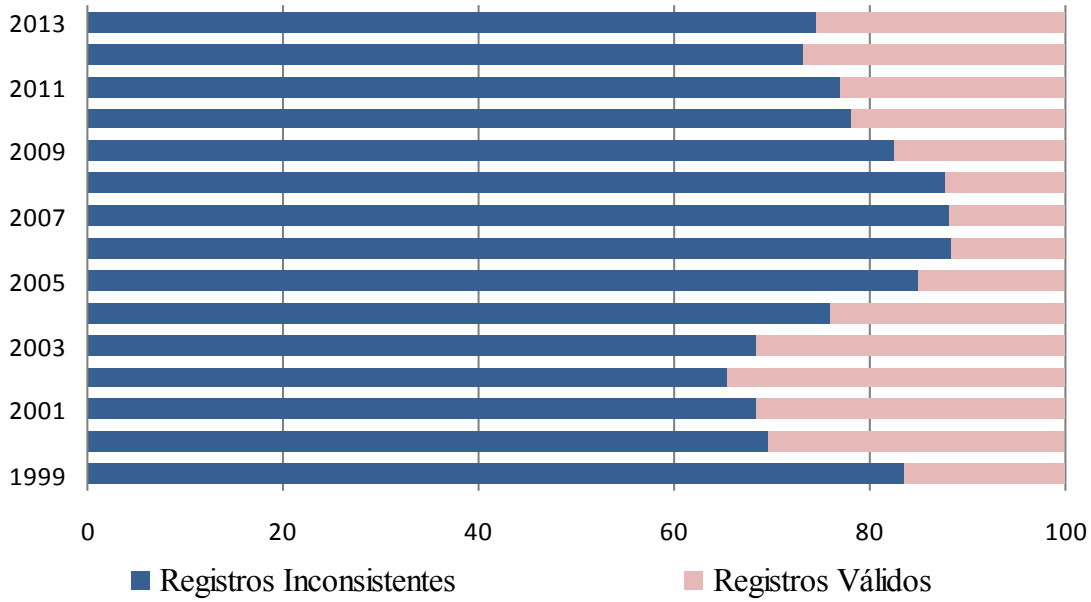
Ano do registro	Paciente registrado/ano	Registros não selecionados	Registros Válidos
1999*	306.171	248.391	57.780
2000	26.341	13.946	12.395
2001	24.610	11.473	13.137
2002	22.511	8.999	13.512
2003	23.472	10.270	13.202
2004	52.565	32.454	20.111
2005*	127.460	99.281	28.179
2006	80.402	62.995	17.407
2007	75.658	60.561	15.097
2008	73.874	59.647	14.227
2009	63.258	45.054	18.204
2010	60.523	40.337	20.186
2011	63.836	41.939	21.897
2012	60.688	35.706	24.982
2013	55.479	33.326	22.153
Total	**1.116.848**	**804.379**	**312.469**

*Ano de realização de carga dos dados para a base do SI[3]

A etapa de limpeza resultou em uma visão de 312.469 pacientes com diagnósticos codificados no padrão CID-10. Para os pacientes válidos foram selecionados todos os eventos vinculados no mesmo período de inclusão, para complementação das informações, como: todas as admissões, cirurgias, angioplastias, exames laboratoriais, medicamentos, fatores de risco e óbito. Foi aplicada uma limpeza nos dados complementares, descartando registros com datas inválidas e o tratamento dos dados complementares. Foi constituída a visão intermediária dos dados limpos denominada de base **Pauá,** com os registros válidos de 312.469 pacientes com diagnósticos codificados no padrão CID-10 e os dados complementares, para o estudo de doenças.

O Gráfico 6.6 apresenta a distribuição percentual dos registros totais dos pacientes, os que não foram selecionados e os que possuíam registros válidos para a seleção de coorte, resultantes da preparação da base, por ano do cadastro dos pacientes.

Gráfico 6.6 Distribuição dos registros da base (%pacientes por ano/cadastro)



6.8 Coorte DCV

A coorte dos pacientes DCV foi selecionada a partir da base Pauá, com a aplicação dos critérios de inclusão, resultando um conjunto de 27.915 pacientes (9%) com diagnóstico DCV, com segmento de atendimento ambulatorial e internações no InCor, no período de 2003 a 2013. A Figura 6.6 apresenta o fluxo de seleção da coorte a partir da aplicação dos critérios de inclusão, destacando os quantitativos de pacientes que foram descartados.

O ponto de partida para a seleção da coorte foi o registro do primeiro diagnóstico de DCV. O CID I25-Doença isquêmica crônica, foi encontrado como primeiro diagnóstico em 40% dos pacientes, seguido do CID I20-Angina pectóris com 23%, o I21-Infarto agudo com 14% e o I24-Doença isquêmica crônica com 13%. O grupo da DIC representou 91% dos primeiros diagnósticos. O Gráfico 6.7 apresenta a distribuição dos 27.915 pacientes, por diagnósticos (categoria CID-10) utilizados na composição da coorte.

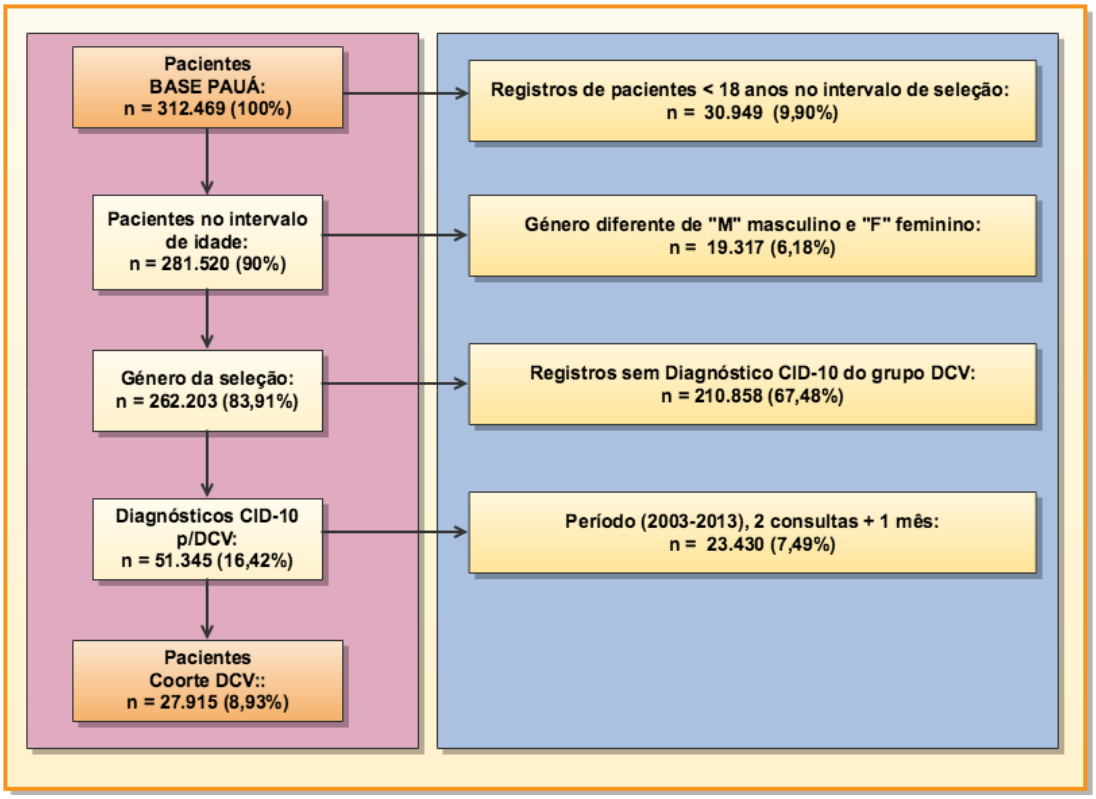
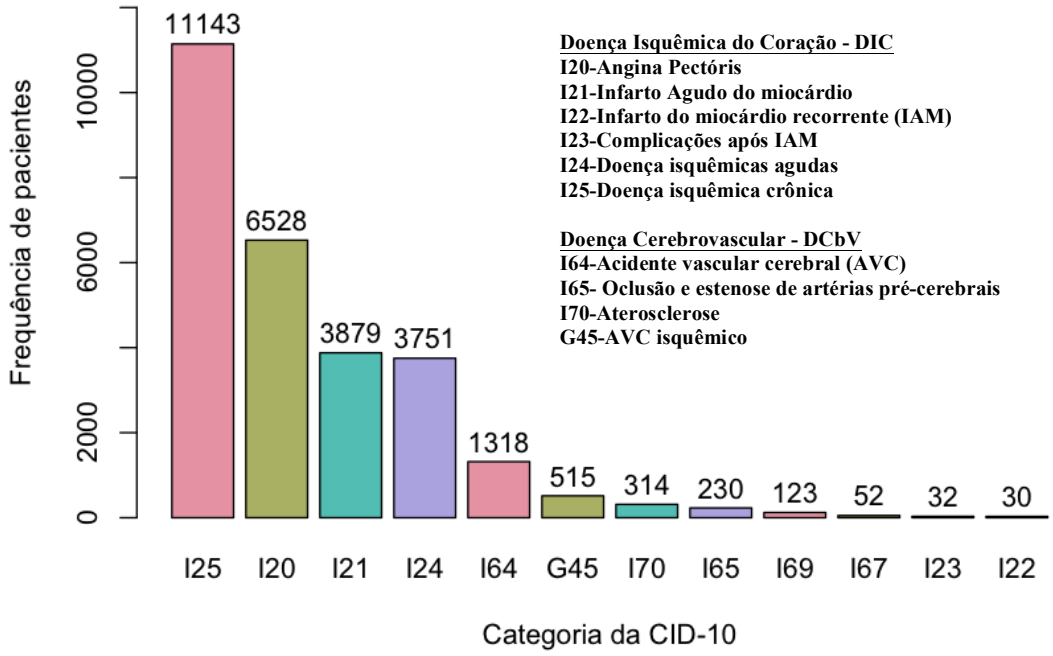


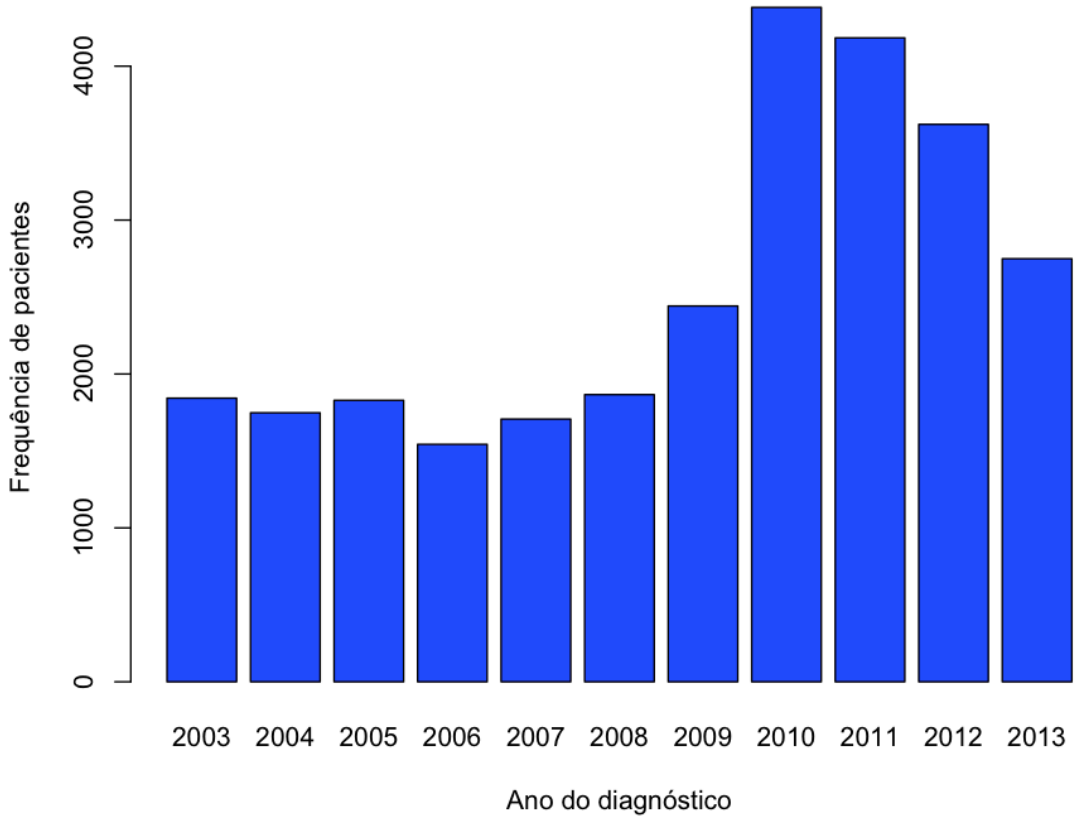
Figura 6.6 Fluxo de seleção da coorte DCV

Gráfico 6.7 Distribuição dos pacientes da coorte DCV por diagnóstico



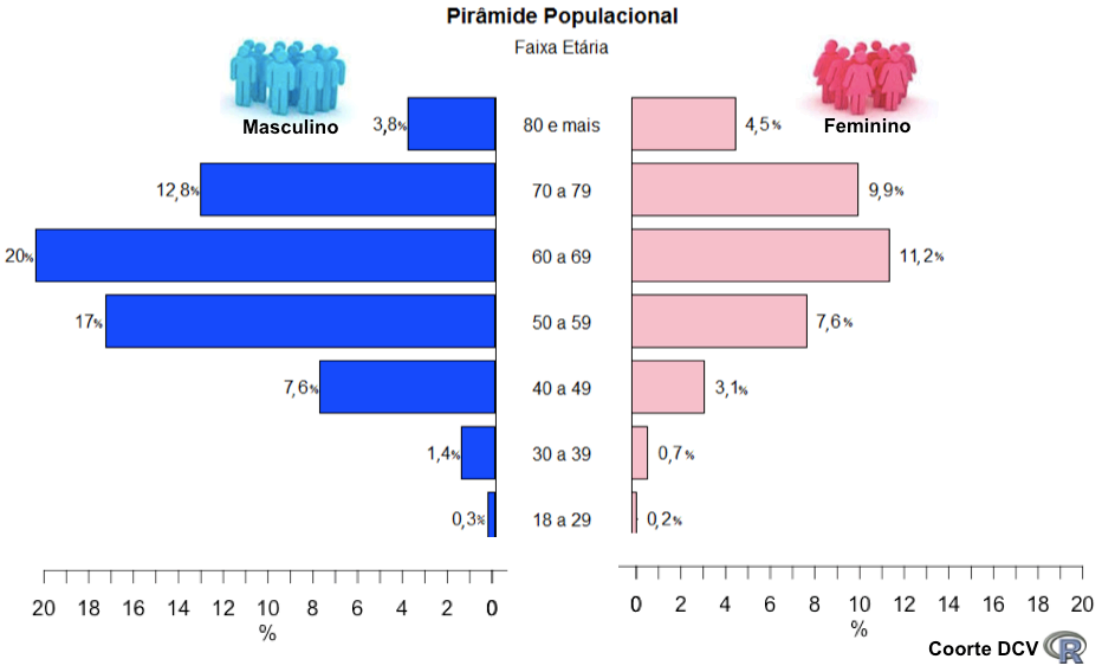
O Gráfico 6.8 apresenta a distribuição dos pacientes por ano do registro do primeiro diagnóstico de DCV.

Gráfico 6.8 Distribuição dos pacientes da coorte DCV por ano do diagnóstico



A população masculina apresentou um percentual total de 61% sendo 22% a mais que a população feminina na coorte DCV. Os pacientes do gênero masculino foram predominantes em quase todas as faixas de idades, sendo inferior em cerca de 0,3% na faixa acima de 80 anos. Os pacientes acima de 50 anos representaram 88% do total da coorte DCV. O Gráfico 6.9 apresenta a distribuição percentual por gênero dos pacientes da coorte DCV por faixa etária no primeiro diagnóstico.

Gráfico 6.9 Distribuição populacional da coorte DCV



A coorte DCV apresentou 45% dos pacientes com a variável "município de origem" preenchida. Os pacientes atendidos no InCor são oriundos de todos os estados do Brasil, sendo aproximadamente 63% do estado de São Paulo, 10% do estado de Minas Gerais, 8% da Bahia e 19% de outros estados.

Os pacientes da coorte foram separados em dois grupos: 1- estatina sim (pacientes com registro da medicação) e 2 - estatina não (pacientes sem registro da medicação). As variáveis foram analisadas em relação à demografia, os atendimentos, cirurgias de revascularização do miocárdio, angioplastias, exames laboratoriais, medicamentos, fatores de risco e óbito. O grupo com estatina correspondeu a 80% dos pacientes da coorte DCV, sendo 49% desse percentual do gênero masculino, com média de idade de 63 anos (60-66), sendo 89% com mais de 50 anos, 51% casados e 46% com escolaridade até o primeiro grau. A Tabela 6.8 apresenta as variáveis demográficas dos pacientes da coorte DCV por grupos.

Tabela 6.8 Demografia dos pacientes DCV quanto ao registro de estatina

Variáveis	Total n (%)		Estatina Não		Estatina Sim		Valor *p*
Gênero	**27.915 (100)**		**5.561 (20)**		**22.354 (80)**		<0,001*
Feminino	11.001	39,4	2.338	8,4	8.663	31,0	
Masculino	16.914	60,6	3.223	11,5	13.691	49,0	
Faixa Etária no diagnóstico							<0,001*
18 a 49	3.575	22,1	985	9,3	2.590	12,8	
50 a 59	7.118	25,5	1.269	4,5	5.849	21,0	
60 a 69	8.547	30,6	1.491	5,3	7.056	25,3	
70 a 79	6.302	22,6	1.226	4,4	5.076	18,2	
80 e mais	2.373	8,5	590	2,1	1.783	6,4	
Idade (anos), média (dp)	64		63		64		<0,001**
≥ 50 anos, n (%)	24.340	87,2	4.576	16,4	19.764	70,8	<0,001**
Estado Civil							<0,001*
Solteiro	2.866	10,3	635	2,3	2.231	8,0	
Casado	17.736	63,5	3.377	12,1	14.359	51,4	
Separado	869	3,1	140	0,5	729	2,6	
Divorciado	1.506	5,4	275	1,0	1.231	4,4	
Viúvo	3.955	14,1	677	2,4	3.278	11,7	
Amasiado	154	0,6	22	0,1	132	0,5	
Não informado	829	3,0	435	1,6	394	1,4	
Nível de Escolaridade							<0,001*
Analfabeto	776	2,4	111	0,4	665	2,0	
1º grau - incompleto	10.114	37,5	1.330	5,0	8.784	32,5	
1º grau completo	3.738	14,8	468	2,1	3.270	12,7	
2º grau completo	3.994	16,0	726	3,3	3.268	12,7	
Superior completo	4.379	17,3	1.107	4,6	3.272	12,7	
Não informado	4.914	19,1	1.819	7,0	3.095	12,1	

* *Pearson's Chi-squared test*; ** Teste t; dp=desvio padrão; n= quantidade de pacientes; %= Percentual do total de pacientes;

Dos pacientes com registro de estatina 40% tiveram como primeiro diagnóstico a doença isquêmica crônica, seguido de 16% com Infarto agudo; 31% tiveram registro de diabetes, 62% de hipertensão arterial, 25% tabagistas, 58% com dislipidemia, 40% com hipertriglicemia e 52% registraram inatividade física. Para esses pacientes, 42% não tiveram registro de nenhum evento subsequente, sendo que 1,6% tiveram registro de mais de 3 eventos (angioplastias, revascularizações e óbito), os procedimentos de angioplastia representaram 16% e as cirurgias de revascularização 17% e 7% dos pacientes foram a óbito. A Tabela 6.9 apresenta a distribuição dos CID-10 de DCV, fatores de risco para a doença cardiovascular, angioplastias, cirurgias de revascularização do miocárdio e óbito para os grupos de pacientes com e sem estatinas.

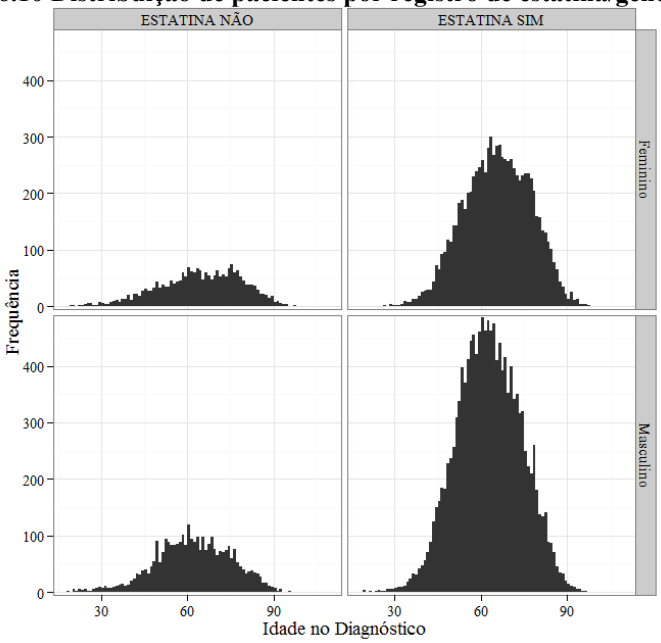
Tabela 6.9 Distribuição dos pacientes DCV quanto ao registro de estatina

Variáveis	Total n (%)		Estatina Não		Estatina Sim		Valor *p*
Diagnóstico Inicial	**27.915 (100)**		**5.561 (20)**		**22.354 (80)**		<0,001*
I20 Angina pectoris	6.528	23,8	1.197	4,3	5.331	19,1	
I21 Infarto agudo do miocárdio-IAM	3.879	14,8	304	1,1	3.575	12,8	
I22 Infarto recorrente	30	0,1	7	0,0	23	0,1	
I23 Complicações após IAM	32	0,1	9	0,0	23	0,1	
I24 Doença isquêmica aguda	3.751	13,9	622	2,2	3.129	11,2	
I25 Doença isquêmica crônica	11.143	38,9	2.284	8,2	8.859	31,7	
I64 AVC	1.318	4,8	526	1,9	792	2,8	
I65 Oclusão e estenose pré-cerebral	230	0,6	108	0,4	122	0,4	
I67 Aterosclerose cerebral	52	0,2	14	0,1	38	0,1	
I69 Sequelas do AVC	123	0,4	57	0,2	66	0,2	
I70 Aterosclerose	314	0,5	213	0,8	101	0,4	
G45 AVC isquêmico	515	1,8	220	0,8	295	1,1	
Diabetes mellitus	7.376	26,4	434	1,6	6.942	24,9	<0,001*
Hipertensão arterial sistêmica	15.021	53,8	1.092	3,9	13.929	49,9	<0,001*
Tabagismo	6.075	21,8	462	1,7	5.613	20,1	<0,001*
Dislipidemia	13.687	49,0	747	2,7	12.940	46,4	<0,001*
Hipertriglicemia	9.387	33,6	537	1,9	8850	31,7	<0,001*
Inatividade Física	12.659	49,3	984	3,5	11.675	46,1	<0,001*
Não tiveram evento evolutivo	16.061	57,6	4.369	15,7	11.692	41,9	<0,001*
Diagnóstico subsequente	3.527	12,6	341	1,2	3.186	11,4	<0,001*
Angioplastia	4.261	15,3	178	0,6	4.083	14,6	<0,001*
Revascularizações	4.770	17,1	314	1,1	4.456	16,0	<0,001*
Óbito	2.383	8,5	617	2,2	1.766	6,3	<0,001*

* *Pearson's Chi-squared test*; n=quantidade de pacientes; %=percentual do total de pacientes;

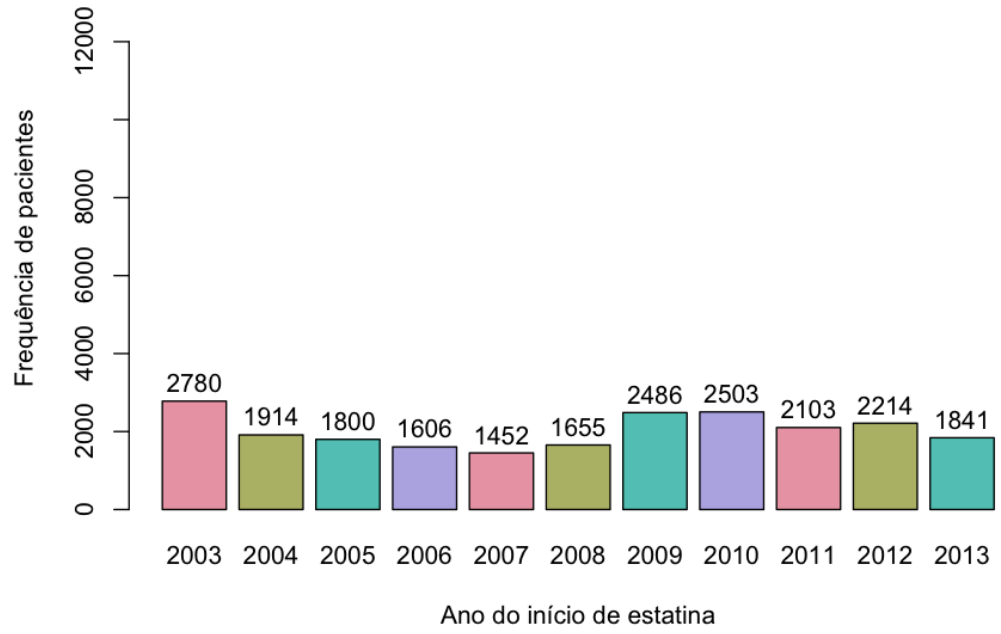
O Gráfico 6.10 apresenta um gráfico da distribuição dos pacientes com registro ou não de estatina, por gênero e idade no diagnóstico. Os pacientes do gênero masculino foram em maior frequência e tiveram um maior período de registro de estatina.

Gráfico 6.10 Distribuição de pacientes por registro de estatina/gênero/idade



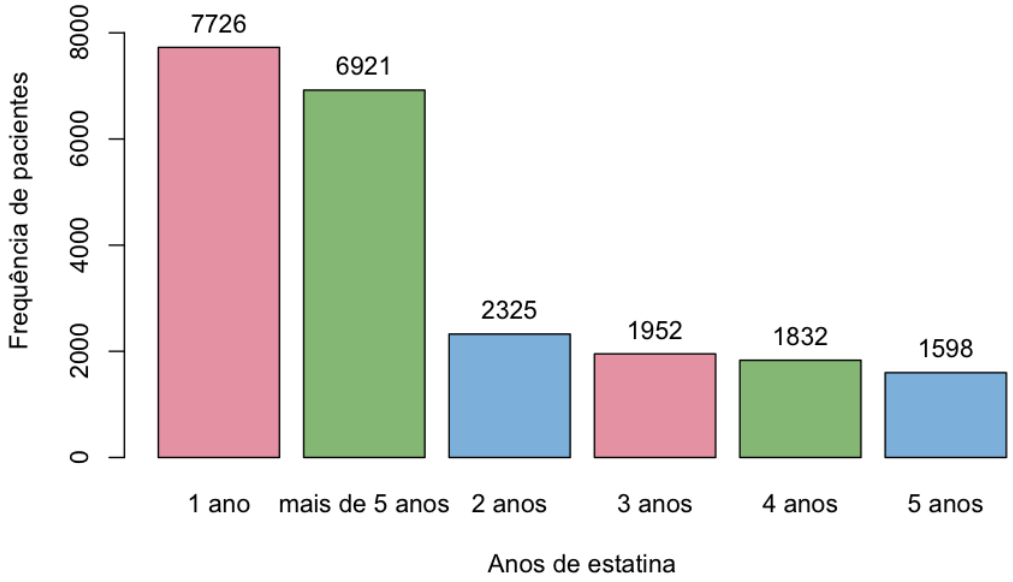
O Gráfico 6.11 apresenta o gráfico da distribuição dos pacientes por ano do início de registro de dispensa de estatina.

Gráfico 6.11 Distribuição dos pacientes por ano de inicio de dispensa de estatinas



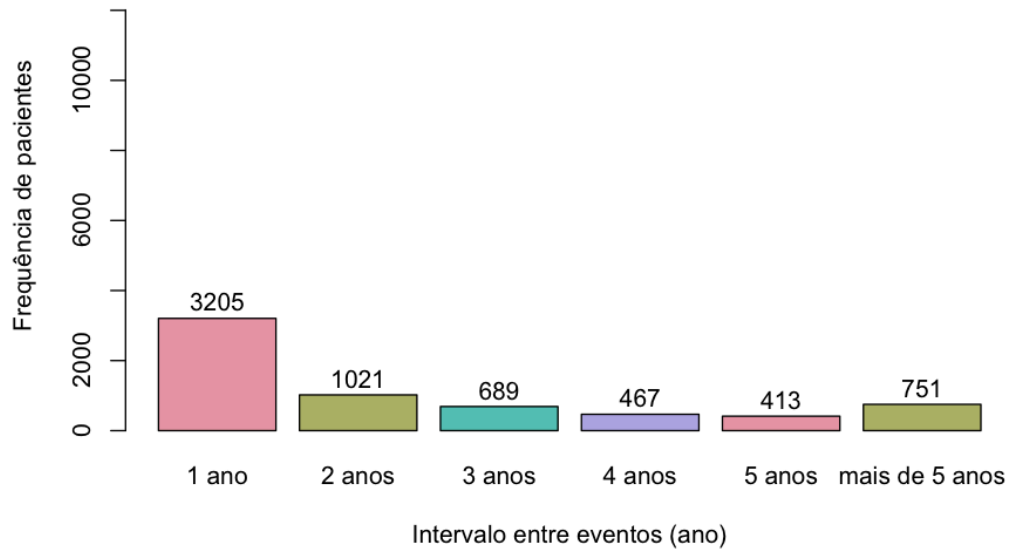
O Gráfico 6.12 apresenta o gráfico da distribuição dos pacientes por tempo (ano) de registro de dispensação da estatina. Os pacientes com até um ano de dispensa representaram 34% do total e 30% dos pacientes com mais de cinco anos.

Gráfico 6.12 Distribuição dos pacientes tempo de dispensa de estatinas (anos)



O Gráfico 6.13 apresenta o gráfico da distribuição dos pacientes por intervalo entre primeiro diagnóstico e evento evolutivo. O total de pacientes com registro de um evento evolutivo foi de 6.546 pacientes, sendo que 49% tiveram registro de um evento em até um ano e 11% dos pacientes com mais de cinco anos.

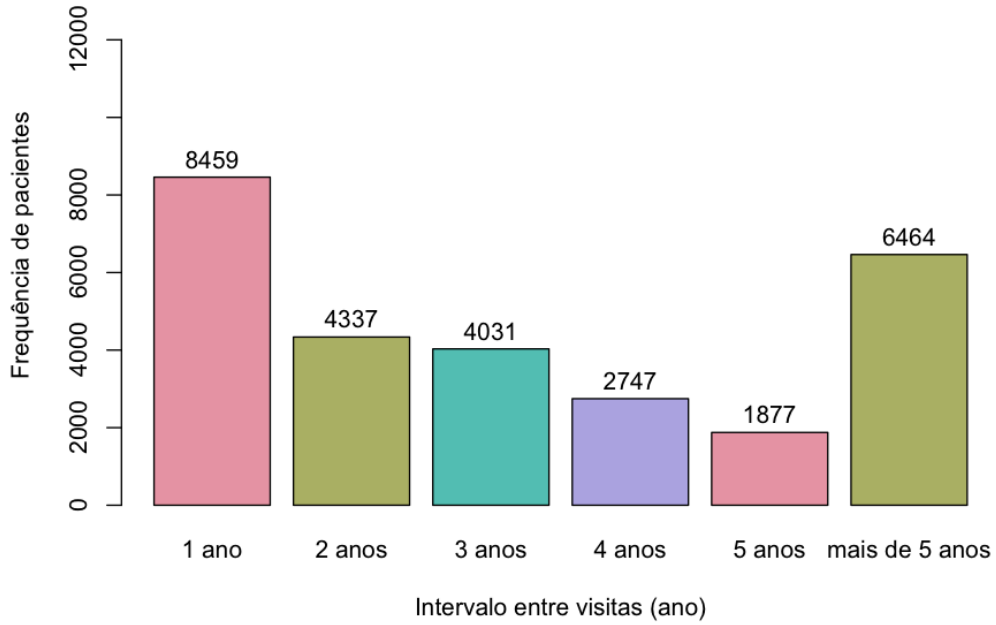
Gráfico 6.13 Distribuição dos pacientes por intervalo entre o primeiro diagnóstico e evento evolutivo (ano)



Os pacientes tiveram o primeiro registro de um diagnóstico do grupo DCV durante todo o período, sendo que cada paciente contribuiu com um tempo de segmento, tempo entre o primeiro diagnóstico e a última visita do paciente ao InCor.

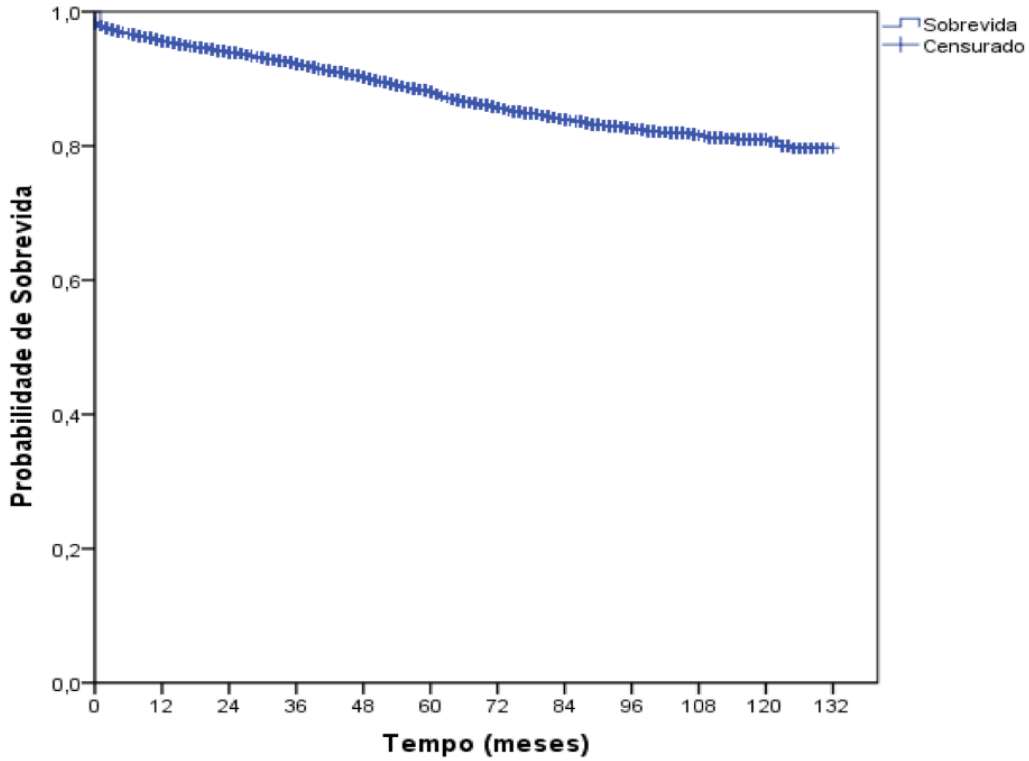
O Gráfico 6.14 apresenta o gráfico da distribuição dos pacientes por intervalo entre primeiro diagnóstico e a última visita do paciente ao hospital. Os pacientes com registro de até um ano de intervalo foi de 30%, sendo que 23% dos pacientes tiveram mais de cinco anos de intervalo.

Gráfico 6.14 Distribuição dos pacientes por intervalo entre primeiro diagnóstico e última visita (ano)



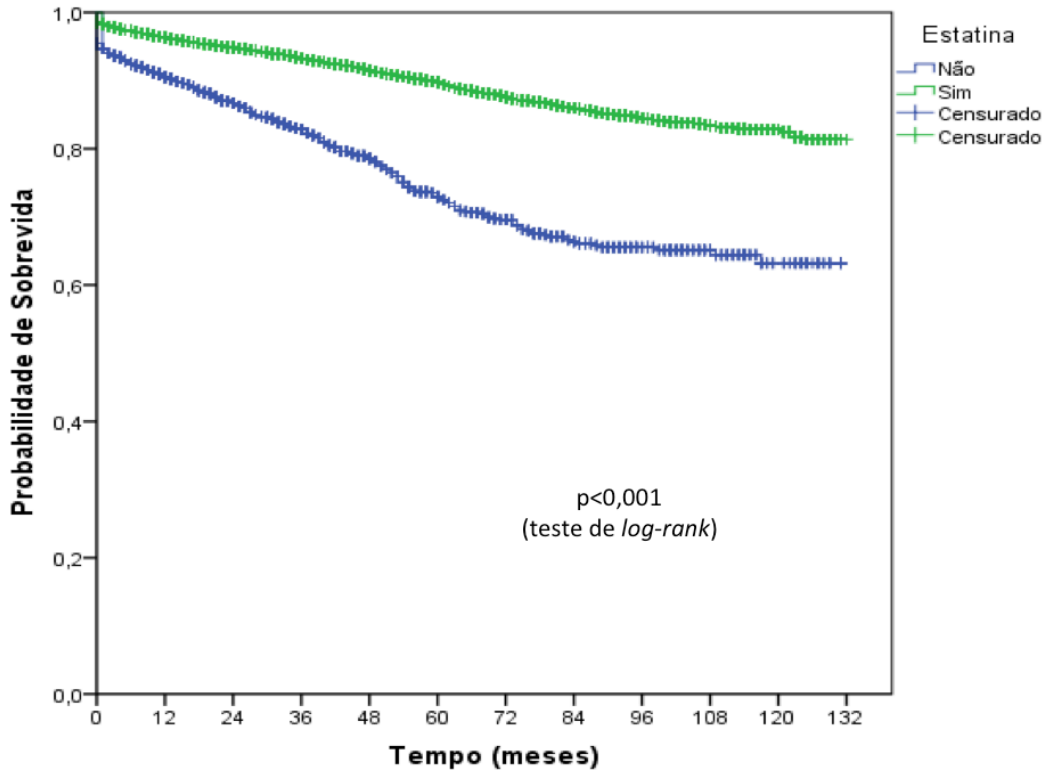
O Gráfico 6.15 apresenta a curva de sobrevida da coorte DCV (n=27.915 pacientes, sendo os pacientes com óbitos n=2.383), utilizando o método *Kaplan-Meier*, que apresentou um tempo médio de sobrevida de 115 meses (IC 95% 114-116).

Gráfico 6.15 Estimativa de sobrevida (tempo em meses)



O Gráfico 6.16 apresenta a curva de sobrevida da coorte DCV, para os grupos de pacientes com registro de estatina (n=23.354 pacientes) e sem registro de estatina (n=5.561 pacientes), utilizando o método *Kaplan-Meier*.

Gráfico 6.16 Estimativa de sobrevida em relação a estatina (tempo em meses)



O método *Kaplan-Meier* apresentou um tempo médio de sobrevida de 117 meses (IC 95% 116-118) para os pacientes com estatinas e um tempo médio de 97 meses (IC 95% 95-100) para os pacientes sem registro de estatina. O teste de *log-rank* foi significativo com um valor de $p<0,001$, mostrando uma maior probabilidade de óbito para o grupo de pacientes sem registro de estatinas.

7 DISCUSSÃO

Essa Tese apresentou um método de extração de coorte em base de dados assistencial, tendo utilizado como campo para sua implementação a base de dados do SI^3, e como fundamento a álgebra relacional aplicada num gerenciador de bases de dados relacional. Reúne a definição de um conjunto de dados organizado em um esquema externo, o mapeamento dos dados da base de origem no mesmo, e um processo sistemático de limpeza e tratamento expresso em instruções da linguagem SQL que possuem seu equivalente em álgebra relacional.

Neste escopo, as bases de dados assistenciais surgem como uma fonte de informações adequada para estudos observacionais, mas ao mesmo tempo oferecem desafios notórios. A idade e consequentemente a evolução dos sistemas de informação, a maturidade do uso de terminologias na prática médica diária, a importância do processo de coleta das informações ganha uma nova ótica na hora que percebemos que o sistema se transforma em elemento chave da pesquisa clínica.

A partir da base de dados do SI^3, que contempla informações de assistência de mais de um milhão de pacientes ao longo dos últimos 20 anos, a preparação da base foi um processo complexo e demorado. Os dados passaram por várias migrações e no decorrer do período de mais de 10 anos da implantação do SI^3, novos dados foram incorporados e complementados, o que justifica a descontinuidade de algumas informações. No entanto, as informações do conjunto de dados do esquema externo proposto foram preenchidas e o método foi aplicado nos dados da base do SI^3.

Após o mapeamento dos dados da base de origem para o conjunto de variáveis definidas no esquema externo, tem inicio a aplicação do método. É feita a limpeza, o tratamento e derivação dos dados, a geração das visões de dados limpos e formatados, denominado base Pauá, que disponibiliza um conjunto de pacientes com diagnósticos no padrão CID-10 e a seleção da coorte final anonimizada, que foi ser analisada na ferramenta estatística R. O método pode ser aplicado a qualquer momento ou quando for necessária a alteração de algum critério de seleção da coorte.

No desenvolvimento do método foram seguidos os critérios propostos pela ISPOR para coleta retrospectiva de dados. A Tabela 7.1 apresenta um descritivo da aplicação de cada critério.

Tabela 7.1 Resumo da aplicação dos critérios da ISPOR

Critérios da ISPOR	Aplicação dos Critérios
1. Relevância das fontes de dados	A base de dados do SI^3 após o processo de limpeza, resultou em mais de 200 mil pacientes. Possui atributos que permitiram traçar o perfil demográfico dos pacientes e gerar indicadores de qualidade da base;
2. Confiabilidade e validade	Os dados inconsistentes (campos nulos, datas inválidas, registros duplicados) foram descartados e outros dados foram tratados (diagnóstico de cirurgia);
3. Conexões (*linkage)*	Os dados do SI^3 e do SIGFar foram recuperados por um identificador comum;
4. Delineamento da pesquisa	Foi realizado um estudo de caso retrospectivo sobre DCV e estatinas;
5. Escolha do delineamento	Estudo de coorte observacional retrospectivo em base de dados assistencial;
6. Limitações do tipo de delineamento	Foi definido o primeiro diagnóstico de DCV para a seleção da coorte. Aproximadamente 18% do total dos pacientes possuíam o registro de algum diagnóstico;
7. Efeitos do tratamento	Foi avaliado o tempo de dispensação da medicação estatina e o intervalo de tempo entre o primeiro diagnóstico e o óbito;
8. Seleção da amostra	A partir de mais de 1 milhão de registros, foram selecionados os pacientes com DCV, maiores de 18 anos, com no mínimo 2 consultas ambulatoriais, no período de 2003 a 2013;
9. Definições operacionais	Foram considerados os diagnósticos classificados no padrão CID-10, o procedimento de angioplastia, a cirurgia de revascularização, selecionado um grupo de medicamentos e cinco exames de laboratório;
10. Definição de validade	As inconsistências encontradas foram descartadas. Não foram selecionados registros de pacientes que apresentaram alguma inconsistência;
11. Tempo entre a exposição e o desfecho	O desfecho alisado foi o óbito e foi considerado um tempo mínimo de um mês entre o primeiro diagnóstico e o óbito;
12. Captação incompleta dos eventos	Os óbitos e outros eventos registrados no SI^3 são ocorridos somente em atendimentos hospitalares. As informações extra hospitalares não foram consideradas;
13. História natural da doença	A doença em estudo foi a cardiovascular aguda e crônica que pode ser avaliada em relação ao tempo entre o primeiro diagnóstico e o óbito;
14. Avaliação de recursos	Não foram avaliados custos e outros recursos adicionais;
15. Definição da estatística	Foram preparadas mais de 200 variáveis, mas foi selecionado um conjunto de 80 variáveis para compor o estudo da DCV. As variáveis foram analisadas descritivamente e realizados testes estatísticos em relação ao desfecho de óbito;
16. Discussão e conclusão	A partir da aplicação do método foi possível gerar indicadores do perfil e qualidade da base assistencial do SI^3, sintetizar as inconsistências que foram descartadas e as que foram tratadas e extrair uma coorte de 27.915 pacientes com DCV realizando uma análise descritiva;

A aplicação dos detectores de inconsistências apontou vários problemas na etapa de preparação dos dados, o que já era esperado, por se tratar de dados de registro assistencial não direcionado para pesquisa e por incorporar um longo período de tempo, durante o qual o processo de registro eletrônico foi sendo complementado com novos módulos e informações. Esses problemas foram divididos em dois grupos, as inconsistências que foram descartadas e as que foram tratadas, sendo:

Grupo 1: Inconsistências em registros que foram descartados:

- ✓ Registros de pacientes com campo da data de nascimento nulo;
- ✓ Pacientes sem registro de diagnóstico, com diagnósticos não codificados no padrão CID-10 ou que os diagnósticos não estavam associados a uma admissão;
- ✓ Registros com nome de paciente e data de nascimento em duplicidade. Seria necessário realizar uma verificação minuciosa, interpolação e unificação desses registros. Optou-se por não selecionar nenhum registro de pacientes em duplicidade;
- ✓ Data de nascimento do paciente anterior ao ano de 1900 e posterior ao ano de 2013;
- ✓ Foram descartados todos os registros que estavam com data de registro no futuro, acima do ano de inclusão do registro;
- ✓ Requisições de medicamentos com registro anterior a 2003;
- ✓ As admissões que estavam como canceladas e as admissões que não tinham uma alta associada (todas as admissões devem ter uma alta associada).
- ✓ Vários sequenciais de alta válidos (que não estavam indicados como cancelados), associados a mesma admissão. Foram descartados os registros e só foi considerado o último sequencial de alta válido;
- ✓ Registro de admissão ou evento para o paciente com data posterior ao registro de óbito;

Grupo 2: Inconsistências em registros que foram tratados:

- ✓ Na busca do primeiro diagnóstico, foram encontrados vários CIDs do grupo DCV cadastrados na mesma data/hora ou para o mesmo evento de um único

paciente. Foi estabelecido o seguinte critério de prioridade na seleção do primeiro diagnóstico, observando-se em sequência, a doença aguda (CIDs I21, I22, I23 e I24), crônica (CID I25), angina (CID I20), doença cerebrovascular (CIDs I63 a I69 e G45) e a aterosclerose (CID I70).

- ✓ Na etapa do mapeamento na seleção do campo para preenchimento do diagnóstico no esquema externo, foram encontradas várias codificações fora do padrão CID-10. A base de dados do SI[3] possui uma tabela que possibilita o relacionamento desses valores com o CID-10. Foi necessário o uso dessa tabela para a conversão e unificação dos códigos para o padrão CID-10. Os códigos que não puderam ser convertidos foram descartados;
- ✓ As descrições dos medicamentos precisaram ser tratadas e padronizadas, com filtragem de caracteres não textuais ou espaços indevidos (exemplo: '.', '( )', '*', espaços antes da descrição, etc.);
- ✓ No campo de diagnóstico da tabela do centro cirúrgico foi encontrado informação sem codificação no CID-10, descrita como texto livre ou utilizando codificações não padronizadas. Foi aplicado um algoritmo de processamento de texto com utilização de expressões regulares, para a transformação desses valores em códigos CID-10. Na sequência se apresenta uma descrição mais detalhada do tratamento aplicado na recuperação dos diagnósticos na tabela de cirurgia.

As tabelas do centro cirúrgico continham informações de diagnóstico com cadastro a partir do ano de 1986. Esses registros de cirurgias foram recuperados para a base de dados do InCor e pertencem ao histórico de informações. Optou-se por considerar também esses diagnósticos. Entretanto, nestas tabelas o campo de <diagnóstico> estava em formato descritivo (campo texto). Foi aplicado um algoritmo de processamento de texto, utilizando expressões regulares, para a transformação desses valores em códigos no padrão CID-10. A Tabela 7.2 apresenta a relação das expressões regulares e os códigos CID-10 utilizados para transformação dos textos em diagnósticos codificados, na sequência de aplicação. Quando a sequência de caracteres da primeira coluna era encontrada dentro do campo diagnóstico, era computado o código CID-10 indicados na segunda coluna. Nesse processo foi também aplicada a tabela de conversão

das codificações internas em CID-10. A transformação era interrompida na primeira sequência encontrada, considerando-se a ordem apresentada na tabela a seguir:

Tabela 7.2 Relação das expressões regulares/códigos CID-10

Expressão	Substituição (CID-10)
ICO	I25.8
IAM	I21.9
Infarto agudo	I21.9
Angina instavel	I20
Insuficiencia cardiaca	I50
Doença aterosclerotica	I25
010300100*	I21.9
010400900*	I22
010401200*	I25
010200000*	I24.8
I35	I35
Infarto recorrente	I22
I[2][0-9][^\.]	I[0-9][0-9]
^I[2,6,7][0-9]\.[0-9]	^I[0-9][0-9]\.[0-9]
I[2,6,7][0-9]\.[0-9]'	I[2,6,7][0-9]\.[0-9]
I[6][0-9][^\.]	I[0-9][0-9]
I[7][0-9][^\.]	I[0-9][0-9]
I[5][0-9][\.]	I[0-9][0-9]\.[0-9]
I[5][0-9][^\.]	I[0-9][0-9]
^I[0-9][0-9][^\.]	^I[0-9][0-9]
^I[0-9][0-9][\.][0-9]	^I[0-9][0-9][\.][0-9]

* Para esses códigos o CID-10 foram recuperados da tabela de relacionamento de diagnósticos com o padrão CID-10, disponível na base do SI[3].

Os códigos que não puderam ser convertidos foram descartados. Foram considerados os diagnósticos que estavam codificados com 3 dígitos do CID-10. Estes CIDs apesar de estarem incompletos para definir um diagnóstico, permitiram classificar um diagnóstico em um grupo especifico de CID.

As angioplastias foram recuperadas de duas tabelas: uma tabela de histórico, com registro de procedimentos realizados de 1997 a 2007 e outra tabela com os registros de 2007 a data atual. Na tabela de histórico, o registro do paciente estava com um identificador que é utilizado no HC. Na tabela atual o registro do paciente estava com seu identificador do InCor. A tabela de histórico precisou ser convertida para um

identificador comum e unificada com a tabela atual. Foi criada uma nova visão consolidada dos procedimentos de angioplastia.

As tabelas de cirurgia também precisaram de tratamento de unificação, pois estavam divididas em uma tabela com registros históricos e outra atual, porém possuíam um identificador único.

Os óbitos dos pacientes foram recuperados de três tabelas diferentes, a do cadastro do paciente, a do registro da alta referente a uma admissão e da tabela de cirurgia. Essas datas foram recuperadas e foi feita uma interpolação dos registros unificando um registro único do óbito.

Problemas de processo

A data de um diagnóstico é essencial na análise de sobrevida e em outras análises, porém, comprovamos que foi achada uma quantidade significativa de diagnósticos registrados com a mesma data e hora e para a mesma admissão. Foi criado um indicador anual para apontar esta condição. Isto aponta para algum problema de processo que precisa ser acompanhado no qual a captura da informação está condicionada por alguma interferência externa e não acompanha a realidade da evolução clínica.

Problemas de vocabulários

O único vocabulário padronizado encontrado no sistema é aquele usado para codificação dos diagnósticos (CID-10), por ser de uso obrigatório, porém o campo do diagnóstico permite outras codificações e até mesmo texto livre. A falta de um vocabulário adequado para os medicamentos faz que não seja possível acompanhar o uso de um determinado medicamento pelo paciente ao longo do tempo (também conhecido com era) devido a que o mesmo medicamento se apresenta no sistema com diversas codificações e nomenclaturas, impedindo de ser tratado de forma uniforme.

A falta de uma estrutura de metainformação, como por exemplo, *Anatomical Therapeutic Chemical* (ATC, 2013), que desde 1996, passou a ser reconhecido pela Organização Mundial de Saúde como padrão internacional para os estudos de utilização de drogas, dificulta a definição de parâmetros tanto para seleção de uma coorte quanto para definição dos tempos de acompanhamento no estudo.

No caso dos procedimentos encontramos uma situação na qual não foi possível aplicar nenhum critério de seleção padronizado, que permitisse identificá-los adequadamente. Só foi possível a recuperação dessas informações utilizando uma busca textual na descrição dos códigos dos procedimentos por termos que não definem necessariamente de forma unívoca o procedimento procurado.

Estes dois elementos, medicamentos e procedimentos, são essenciais para estudos de acompanhamento e interações de medicamentos e sobrevida.

Em ambos os casos, o sistema possui estruturas adequadas que permitiriam o armazenamento de um vocabulário consistente e consequentemente a captura correta de informações. O que falta é a definição destes vocabulários e a sua adoção na prática clínica.

Acompanhamento da evolução do paciente

O sistema acompanha o paciente na sua interação com o hospital, porém observamos que em muitos casos, esta interação é extremadamente breve o que ocasiona a censura da maior parte das informações por motivo de perda de acompanhamento no estudo. Foi elaborado um indicador para apontar o número de pacientes com uma única visita ao hospital.

Uso secundário de dados

A constituição de coortes de pacientes, utilizando base de dados assistenciais, geradas a partir de um ou mais RES, como fonte de dados para uso secundário, vêm se tornando um recurso cada vez mais utilizado em estudos epidemiológicos. A Tabela 7.3 apresenta um resumo com as características de coortes constituídas a partir do uso secundário de dados oriundos de bases de dados assistenciais, apresentando as fontes dos dados, a doença, o período, a população disponibilizada na base assistencial, o desenho do estudo, a quantidade de pacientes e a referência.

Tabela 7.3 Características das coortes constituídas a partir de bases assistenciais

Fonte: Kaiser Permanente USA					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
AVC perinatal	1993-2003	232.532	Caso-controle	20	(Armstrong-Wells et al., 2009)
Mãe com doença auto imune, alergia e asma associado a crianças com autismo (ASD)	1995-1999	2.520	Caso-controle	407 casos, 2.095 controles	(Croen et al., 2005)
Mãe com autismo (ASD) e déficit de atenção (ADHD), associado a crianças ASD e ADHD	1998-2004	35.073	Caso-controle	35.073	(Musser et al., 2014)
Diabetes com doença renal crônica (DISTANCE)	1996-2006	64.211	Coorte prospectiva	64.211	(Kanaya et al., 2011)
Diabetes com doença renal crônica (DISTANCE)	2005-2007	19.804	Coorte retrospectiva	19.804	(Laraia et al., 2012)
Fatores de risco para hipoglicemia em pacientes com diabetes	2005-2006	14.357	Análise transversal	14.357	(Berkowitz et al., 2014)
Fonte: Kaiser Permanente + 4 sistemas de saúde - USA					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Hepatite crônica B e C	2006-2010	>1.6 milhão	Coorte prospectiva	2.202 hepatite B e 8.810 hepatite C	(Moorman et al., 2013)
Fonte: Consorcio Safe Labor (12 instituições + 19 hospitais)-USA					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Prematuros com pneumonia e falência respiratória	2002-2008	233.844	Coorte retrospectiva	7.055 UTI e 2.032 pneumonia e falência respiratória	(Hibbard et al., 2010)
Fonte: CHeCS-Chronic Hepatitis Cohort Study (4 sistemas de saúde USA)					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Óbitos por hepatite C	2006-2010	2.143.369	Coorte retrospectiva	11.703 pessoas com hepatite C e 1.590 óbitos	(Mahajan et al., 2014)

Continuação

Fonte: PHARMO-Holanda					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Uso de rosuvastatina ou outra estatina	2003-2004	>2 milhões	Coorte retrospectiva	10.147 rosuvastatina, 37.396 estatinas, 99.935sem estatinas	(Goettsch et al., 2006)
Fonte: Badalona Services Assistencials (BSA) database - Barcelona-Espanha					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Estatina na prevenção secundária de AVC	2003-2008		Coorte retrospectiva	611	(Sicras-Mainar et al., 2012)
Fonte: Cardiac Care Network (CCN) - Ontário-Canadá (18 hospitais)					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Terapia medicamentosa x revascularização (pacientes com mais de 65 anos)	10/2008-09/2011	13 milhões	Coorte retrospectiva	39.131 pacientes, 15.139 medicamento, 23.992 revascularização	(Wijeysundera et al., 2014)
Fonte: Multicentro 16 hospitais - Itália					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Angioplastia x Revascularização	06/2002-12/2008	56.060 pac/revasc	Coorte retrospectiva	11.750 pacientes, 6.246 angio, 5.504 revasc.	(Fortuna et al., 2013)
Fonte: Veterans Affairs (VA) - USA					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Riscos da terapia com testosterona associada a mortalidade e eventos	2005-2011	23.175 homens com DCV	Coorte retrospectiva	8.709 homens	(Vigen, 2013)
Fonte: Health Improvement Network (THIN) Reino Unido					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Mulheres 25 a 29 anos - Infertilidade e doença celíaca	1990-2013	2.426.225	Coorte retrospectiva	6.506 mulheres	(Dhalwani et al., 2014)

Continuação

Fonte: Base de dados do SI^3 - InCor-HC FMUSP - Brasil					
Estudo (doença)	**Período do estudo**	**População**	**Desenho do estudo**	**N (seleção Pacientes)**	**Referência**
Características demográficas de pacientes com doença isquêmica do coração	1982-1994	306.171	Coorte retrospectiva	15.347	(Caramelli et al., 2003a)
Evolução de pacientes com infarto - SUS	1998-2005	583.130	Coorte retrospectiva	1.588	(Nicolau et al., 2008)
Evolução da cirurgia cardiovascular	1984-2007	739.190	Coorte retrospectiva	71.305 cirurgias	(Lisboa et al., 2010)
DCV com prescrição de estatinas na prevenção secundária	1986-2013	1.116.848	Coorte retrospectiva	65.000	(Abrahão et al., 2013a)
DCV com uso de estatinas, caso-controle pareado	2003-2013	46.757	Coorte retrospectiva Caso-controle	29.308, 14.654 estatina e 14.654 não estatina	(Abrahão et al., 2013b)
Perfil do paciente com doença isquêmica submetidos a cinecoronariografia	1986-1995		Coorte retrospectiva	18.221	(Caramelli et al., 2003b)

A quantidade de pacientes na composição das coortes dependem da doença em foco e do período do estudo. Quanto mais rara a doença menor o número de indivíduos na coorte. A coorte DCV apresentada no capítulo de resultados desta Tese, foi composta de 27.915 pacientes em um período de 11 anos e foi derivada de um outro estudo sobre DCV na mesma base de dados, para o período de 14 anos com 65.000 pacientes com algum atendimento no InCor (Abrahão et al., 2013a). Um outro estudo derivado da mesma base foi realizado pareando pacientes por faixa etária e gênero para grupos com e sem registro de estatinas. Foi possível a seleção de 29.308 pacientes pareados em 14.654 pacientes em cada grupo (Abrahão et al., 2013b).

8 CONCLUSÃO

O custo elevado de estudos clínicos randomizados fornece motivação para a busca de outras fontes de informação mais econômicas, mas que forneçam os elementos necessários às pesquisas.

Por outra parte, a necessidade de criar elementos que permitam pesquisas reprodutíveis tem sido o ponto focal dos estudos nas últimas décadas. Hoje não mais se admite um estudo que não disponibilize os dados originais e todos os métodos pelos quais foram obtidos os dados e os resultados. Isto faz com que os sistemas de informação ganhem uma nova importância como fonte de dados para uso secundário, pois eles serão a fonte de coleta de dados ou da seleção da coorte para a realização de estudos observacionais.

Esta seleção precisa ser sistemática, baseada no uso de um método de extração com mecanismos que comprovadamente não alterem os dados na sua aplicação, para poderem ser reproduzidos a qualquer momento e servir para a validação do estudo por observadores isentos.

Em relação aos objetivos propostos nesta tese, foi desenvolvido um método de extração de coortes em uma base assistencial, utilizando dados do SI^3. Foi definido um conjunto de dados para a aplicação do método, sendo que foi possível mapear os dados da base assistencial para o conjunto de dados pré-definido no esquema externo. O método formou uma base virtual intermediária de dados limpos e padronizados de pacientes com diagnósticos no padrão CID-10 e foi selecionada uma coorte de pacientes com DCV que cumprem os critérios definidos para o estudo proposto sobre essa doença. Foram gerados indicadores do perfil, da qualidade da base, e uma análise estatística da coorte DCV.

A base de dados do SI^3 do InCor-HCFMUSP, disponibiliza 1.116.848 pacientes cadastrados de 1999 a 2013, e obteve 312.469 pacientes (28%) com no mínimo um diagnóstico no padrão CID-10 registrado, depois do processo de limpeza. Foi feita a extração de uma coorte de 27.915 pacientes, selecionada segundo os seguintes critérios: período do estudo: 2003-2013, gêneros: masculino e feminino, idade: $\geq$ 18 anos, com no

mínimo dois encontros ambulatoriais e diagnósticos de DCV (CID-10 códigos: I20 a I25, I64 a I70 e G45) e feita uma descrição estatística dos dados.

Como resultado da análise estatística em relação aos pacientes com e sem registros de prescrições de estatinas, cerca de 80% dos pacientes tiveram registros de estatinas, sendo que 34% com registros de estatinas em até um ano e 30% com registros por mais de 5 anos. Dos pacientes com estatinas, 42% não tiveram registro de nenhum evento evolutivo (angioplastias, cirurgias de revascularização, diagnóstico subsequente e óbito), 29% tiveram registro de um evento e 9,7% com dois ou mais eventos. O tempo médio de sobrevida calculado pelo método *Kaplan-Meier* foi de 115 meses (intervalo de confiança 95%: 114-116) e os pacientes sem registros de estatinas apresentaram uma maior probabilidade de óbito pelo teste *log-rank* $p<0,001$.

Conclui-se que a adoção de métodos sistematizados para a extração de coortes de pacientes a partir do RES pode ser uma abordagem viável para a condução de estudos epidemiológicos.

No processo de extração, o uso exclusivo da álgebra relacional, garante a preservação dos dados originais e abre a possibilidade de estender a sua aplicação a outras bases de dados assistenciais que comportem o mapeamento das informações para o esquema externo proposto.

Trabalhos futuros

Para que se possa dar continuidade ao trabalho iniciado nesta tese, são sugeridos os seguintes trabalhos futuros:

- Desenvolvimento de uma estrutura para facilitar a entrada dos parâmetros do estudo;
- Testar outros desenhos de estudo para validar se os parâmetros e variáveis de saída resultantes da extração da coorte, atendem as necessidades diferentes das que foram propostas nesta tese;
- Estudos de outras doenças a partir de diferentes conjuntos de CIDs, gêneros, idades e períodos de estudo;
- Testar o mapeamento dos dados de outros RES para o conjunto de dados definido no modelo externo e aplicar o método de extração de coortes a outras bases assistenciais;

9 ANEXOS

ANEXO 1 - Descrição detalhada dos fundamentos do processamento sem perda de informação: preservação do dado original

O poder expressivo da álgebra relacional é usado como medida do poder de uma linguagem de consulta de banco de dados. Se a linguagem consegue representar todas as consultas que possam ser expressas na álgebra relacional, se diz que a linguagem é relacionalmente completa (Ramakrishnan e Gehrke, 2000).

Os operadores básicos da álgebra relacional são:

- Seleção (σ): seleciona um subconjunto de registros de uma relação (tabela);
- Projeção (π): descarta colunas indesejadas de uma relação;
- Produto cartesiano ($\times$): Permite combinar duas relações;
- Diferença ($-$): Tuplas presentes em r1, mas não em r2;
- União ($\cup$): Tuplas em r1 e em r2.
- Junções:
    - Junção natural ($\bowtie$)
    - Anti junção ($\triangleright$)
    - θ-junção e equijunção ($\bowtie_\theta$)
    - Semijunção ($\ltimes$, $\rtimes$)
- Renomeação (ρ): utilizado para alterar o nome das colunas de uma tabela e para relacionamentos onde possam surgir nomes iguais para as colunas, como num relacionamento da tabela com ela mesma.

A álgebra relacional e em particular, a linguagem de consulta *SQL*, tem propriedades importantes:

- A álgebra relacional é fechada: a aplicação de operações sobre relações tem como resultado outra relação e, portanto, as operações podem ser combinadas;
- As consultas de seleção são operações idempotentes, quer dizer, múltiplas aplicações da mesma seleção têm o mesmo efeito de uma única aplicação, em

particular, são nullimpotentes, quer dizer, não geram efeitos secundários e o resultado de aplicar uma seleção não altera o estado da base de dados;

- As seleções são comutativas, a ordem em que elas são aplicadas não altera o resultado;
- Independência física dos dados (Date, 2007; Voorhis, 2015): O modelo relacional oferece independência da forma da organização física dos dados no meio de armazenamento (orientados a linhas ou colunas, particionados verticalmente ou horizontalmente) o que permite as diferentes implementações buscar melhorias de desempenho na execução das consultas;
- Independência lógica dos dados: Alterações do modelo conceitual não afetam a visão externa dos dados. Os bancos de dados atuais implementam parcialmente este conceito.

Estas propriedades são o fundamento do método para extração de coortes, pois se utiliza apenas de operações de seleção que garantem que a aplicação do mesmo sobre os dados permite a extração das informações sem alterar o estado original da base e garantem a reprodutibilidade dos resultados.

A rigor, a linguagem *SQL* não é exatamente baseada na álgebra relacional (Codd, 1985a, 1985b; Morgan, 2002), principalmente porque o modelo de uma tabela *SQL* não é exatamente uma relação (conjunto ou set), mas um saco (multiconjunto ou bag), porque permite duplicados. Se realmente não quero duplicados como resultado de uma seleção, preciso explicitar isto através da clausula ***DISTINCT***, o que aumenta o custo de execução da consulta.

A álgebra relacional é estendida com várias operações tais como junções de fora (outer joins: ⟕, ⟖), funções de agregação e encerramentos transitivos (transitive closures) (Ozsu e Valduriez, 2011).

Em particular, as junções de fora assumem que existe um valor *null*(ω), não definido, que é utilizado como preenchimento dos valores ausentes.

Na prática, isto corresponde ao valor *NULL* no *SQL*, porém dado que o *NULL* não é membro de nenhum domínio de dados, ele não é considerado um valor, mas um marcador para indicar a ausência de um valor.

Com isto, nas comparações (a = b ?), temos uma lógica de três valores possíveis: verdadeiro, falso e desconhecido (ISO/IEC). O *SQL* implementa três resultados lógicos com uma lógica de três valores (Hans-Joachim, 2001; Coles, 2008).

Mesmo assim, estas limitações não invalidam o modelo porque ainda garantem a reprodutibilidade dos resultados, que é o objetivo principal para a extração sistemática de coortes a partir dos dados da base relacional.

Os desvios que possam ser ocasionados pela limitação da linguagem de implementação, são sistêmicos e podem ser reconhecidos por uma análise estatística posterior.

Exemplo de uso da álgebra relacional: Descarte de duplicados e valores nulos

Para este exemplo foi utilizada a ferramenta RelaX (RelaX) que permite trabalhar com álgebra relacional. Vamos supor que a tabela de pacientes proveniente do sistema assistencial pode ser descrita por três campos (<id>, <name>, <birthDate>) e nela achamos os seguintes valores:

ALL_PATIENT_RAW

id	name	birthDate
1	maria	1950-01-12
2	maria	1950-01-12
3	pedro	1950-01-12
4	null	1950-01-12
5	null	null
6	ana	1940-01-25
7	Jose	1980-01-01
8	jose	1980-01-01
9	jose	1980-01-01
10	joao	null

Podemos ver que nela existem registros com valores nulos e registros repetidos.
Para definir uma relação que contém todos os registros que cumprem a condição de não ter nulos nos campos do nome ou da data de nascimento, podemos usar a seguinte expressão da álgebra relacional:

```
NOT_NULL_PATIENT = σ name≠null ∧ birthDate≠null ALL_PATIENT_RAW
```

Ao avaliar esta relação obtemos:

NOT_NULL_PATIENT

id	name	birthDate
1	maria	1950-01-12
2	maria	1950-01-12
3	pedro	1950-01-12
6	ana	1940-01-25

7	jose	1980-01-01
8	jose	1980-01-01
9	jose	1980-01-01

Agora precisamos achar os registros duplicados, o que é expresso pela relação:
`DUPLICATES = π A.id, A.name, A.birthDate (σ A.id ≠ B.id ((ρ A (NOT_NULL_PATIENT)) ⋈A.name = B.name (ρ B (NOT_NULL_PATIENT))))`

Que uma vez avaliada, gera:

DUPLICATES

id	name	birthDate
1	maria	1950-01-12
2	maria	1950-01-12
7	jose	1980-01-01
8	jose	1980-01-01
9	jose	1980-01-01

Para poder descartar os valores duplicados, utilizamos o operador relacional de subtração (-) na relação:

```
ALL_PATIENTS = NOT_NULL_PATIENT - DUPLICATES
```

Que ao ser avaliado, gera como resultado:

ALL_PATIENT

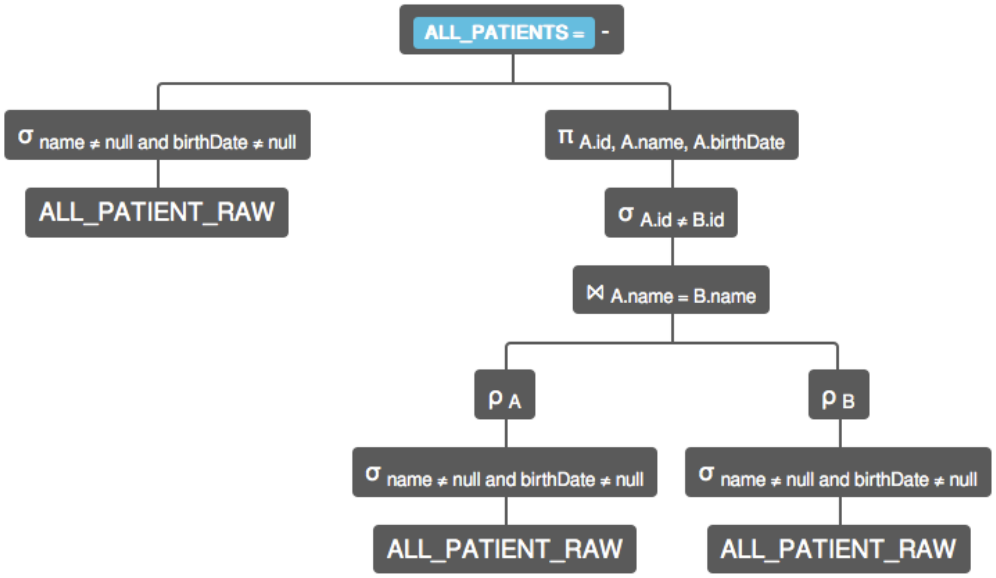
id	name	birthDate
3	pedro	1950-01-12
6	ana	1940-01-25

Podemos ver que os dados inválidos ou repetidos foram filtrados sem a necessidade de alterar o estado da base de dados (através de inserções ou deleções).

Por ser a álgebra relacional fechada, as relações podem ser combinadas numa única relação:

```
ALL_PATIENT = (σ name≠null ∧ birthDate≠null ALL_PATIENT_RAW) - (π A.id,
A.name, A.birthDate (σ A.id ≠ B.id ((ρ A ((σ name≠null ∧ birthDate≠null
ALL_PATIENT_RAW))) ⋈A.name = B.name (ρ B (σ name≠null ∧ birthDate≠null
ALL_PATIENT_RAW)))))
```

Podemos ver a árvore de análise da expressão:



A avaliação da expressão gera o mesmo resultado:

```
( σ name ≠ null and birthDate ≠ null ALL_PATIENT_RAW) – ( π A.id, A.name, A.birthDate ( σ A.id ≠ B.id (( ρ A ( σ name ≠ null and birthDate ≠ null ALL_PATIENT_RAW)) ⋈A.name = B.name ( ρ B ( σ name ≠ null and birthDate ≠ null ALL_PATIENT_RAW)))))
```

ALL_PATIENT

id	name	birthDate
3	pedro	1950-01-12
6	ana	1940-01-25

Características da Implementação

Todas as expressões *SQL* utilizadas no processo podem ser mapeadas para os operadores disponíveis na álgebra relacional como no exemplo descrito e em última instância, poderiam ser combinadas em uma única expressão que representasse uma única consulta que extrai a informação numa única execução.

Isto garante que o estado da base de dados não é alterado pelo processo de extração da coorte e que o mesmo pode ser repetido gerando sempre o mesmo resultado.

Porém, na prática a execução de uma única consulta se torna lenta e de muito difícil manutenção. Por isso duas extensões disponíveis no *Oracle SQL* foram utilizadas:

- Visões (*views*): Conceitualmente são relações, mas os registros não são armazenados na base senão calculados usando a definição da visão, a partir de relações armazenadas no banco. Permitem nomear uma consulta para poder ser combinada com outra consulta. Constituem o nível de esquema externo.
- Visões Materializadas (*materialized views*): Geram uma tabela temporária que representa a execução de uma consulta e melhora o desempenho de novas combinações de consultas que incluam a mesma.

Em qualquer dos dois casos, a consulta que deu origem a visão é preservada no banco de dados e pode ser recuperada a qualquer momento. Assim, o próprio processo pode ser recuperado através da releitura da meta-informação disponível no banco.

Por exemplo, as seguintes duas consultas recuperam todo o modelo do processo de extração de coortes diretamente do banco:

```
SELECT MVIEW_NAME, QUERY FROM USER_MVIEWS;

SELECT VIEW_NAME, TEXT FROM USER_VIEWS;
```

ANEXO 2 - Parâmetros do Modelo - Armazenamento dos parâmetros

Um dos atributos do modelo é a possibilidade de parametrização que permite realizar diferentes estudos sem precisar alterar o script de extração.
Os parâmetros são armazenados em três tabelas segundo o tipo de parâmetro:

- MODEL_STR_PARAMETERS: parâmetros de tipo *string;*
    - Onde o campo do valor é: PARAMETER_VALUE VARCHAR (80)
- MODEL_NUMBER_PARAMETERS: parâmetros de tipo numérico;
    - Onde o campo do valor é: PARAMETER_VALUE NUMBER
- MODEL_DATE_PARAMETERS: parâmetros do tipo data;
    - Onde o campo do valor é: PARAMETER_VALUE DATE

A Tabela A2.1 a seguir apresenta a estrutura comum para a inclusão dos parâmetros da pesquisa. Cada tabela têm a seguinte estrutura comum:

Tabela A2.1 Estrutura comum para inclusão dos parametros da preparação da coorte.

Campo	Tipo	Descrição do campo
RESEARCH_ID	VARCHAR2 (40) NOT NULL	Identificador da pesquisa
MODEL_ID	VARCHAR2(40) NOT NULL	Identificador do modelo de pesquisa
RESEARCHER_ID	VARCHAR2(40) NOT NULL	Identificador do pesquisador
CRITERIA_ID	VARCHAR2(40) NOT NULL	Domínio do parâmetro
PARAMETER_TYPE	VARCHAR2(40) NOT NULL	Semântica do parâmetro
PARAMETER_ID	NUMBER(4) NOT NULL UNIQUE	Identificador do parâmetro
PARAMETER_VALUE	-- Ver cada caso --	Valor do parâmetro
INCLUSION_EXCLUSION	VARCHAR2(1 BYTE) CHECK (INCLUSION_EXCLUSION IN ('I','E'))	I: parâmetro usado para inclusão do registro ou E: para exclusão
COMMENTS	VARCHAR2(80)	Disponível para comentários
GROUP	VARCHAR(40)	Campo auxiliar para agrupar parâmetros
INCLUSION_DATE	DATE	Data da inclusão do registro
MODIFICATION_DATE	DATE	Data da alteração

Os seguintes campos formam a chave da tabela:
- RESEARCH_ID,
- MODEL_ID,
- RESEARCHER_ID,
- CRITERIA_ID,
- PARAMETER_TYPE,
- PARAMETER_ID

Os valores são ingressados na tabela correspondente com um comando *SQL 'INSERT'*. O exemplo a seguir apresenta a inclusão de um parâmetro com valor de categoria CID-10 igual a 'I21', para seleção do diagnóstico de indexação:

```
INSERT INTO
  MODEL_STR_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','DIAGNOSIS_SELECTION','ICD1
0_CATEGORY_FULL',1,'I20','I','',(SELECT       TRUNC(SYSDATE)       FROM
DUAL),(SELECT TRUNC(SYSDATE) FROM DUAL));
```

O conjunto de todos os comandos *SQL* para ingressar todos os códigos dos parâmetros é incorporado em um arquivo texto e armazenado num repositório de versionamento. O sistema irá procurar por este arquivo no repositório para executar a carga dos parâmetros.

Um programa desenvolvido como uma interface *HTML* auxilia na seleção dos parâmetros e no preenchimento do arquivo com os comandos, armazenando-o diretamente no repositório. Porém, esta ferramenta não é indispensável para o processo.

Definição do parâmetro tipo diagnóstico

O parâmetro tipo diagnóstico é utilizado em várias situações. Por ele ter características especiais, se descreve em separado. Qualquer parâmetro deste tipo segue as mesmas regras.

O diagnóstico do paciente é codificado na base utilizando o padrão CID-10, que tem a estrutura apresentada na Figura 9.1.

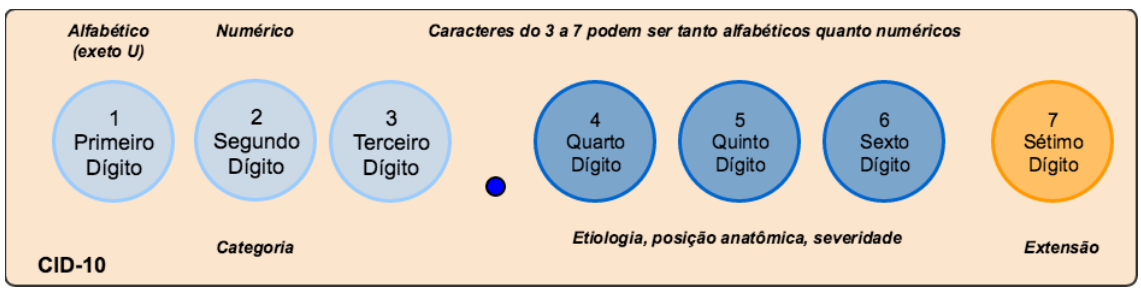


Figura 9.1 Estrutura da codificação no padrão CID-10: Adaptado (CID-10, 2008)

Porém, nem todos os dados encontrados na base seguem exatamente a codificação. O grau de afastamento do padrão CID-10 é um dos indicadores do perfil da base. Levando isto em conta, a seleção de pacientes pelo seu diagnóstico oferece flexibilidades para poder separar situações específicas. Elas estão contempladas nos valores atribuídos ao campo <PARAMETER_TYPE>, que para o parâmetro de seleção de diagnóstico pode assumir algum dos seguintes valores:

PARAMETER_TYPE:

- ICD10_CATEGORY_PARTIAL: Indica que o código entrado no campo valor (p.ex. **I21**) é para ser pesquisado exatamente por apenas 3 caracteres. Serão considerados apenas pacientes com o código de diagnóstico **I21**, sem considerar por exemplo alguém com código **I21.** ou **I21.1** .

- ICD10_CATEGORY_FULL: Pesquisa uma substring do diagnóstico dos 3 primeiros caracteres que se igualem ao ingressado no campo valor do parâmetro (p.ex. **I21**). Então, pacientes com um diagnóstico de **I21** ou **I21.** ou **I21.1** ou **I21.1???** (onde ??? representam quaisquer outros códigos) serão considerados na coorte.

- ICD10_SUBCATEGORY: Pesquisa uma substring do diagnóstico dos 5 primeiros caracteres (4 caracteres e o ponto decimal) que se igualem ao ingressado no campo valor (p.ex. **I21.1**) do parâmetro. Então, pacientes com um diagnóstico de **I21.1** ou **I21.1???** (onde ??? representam quaisquer outros códigos) serão considerados na coorte. Pacientes com diagnósticos com códigos **I21**ou **I21.** não seão incorporados na coorte.

- ICD10_ETIOLOGY: Pesquisa uma substring do diagnóstico dos 7 primeiros caracteres (6 caracteres e o ponto decimal) que se igualem ao ingressado no campo valor (p.ex. **I21.110**) do parâmetro. Então, pacientes com um diagnóstico de **I21.110** ou **I21.110?** (onde ? representa qualquer outro código) serão considerados na coorte. Pacientes com diagnósticos com códigos **I21**ou **I21.** ou **I21.1** ou **I21.11** não serão incorporados na coorte.

- ICD10_EXTENSION: Pesquisa uma substring do diagnóstico dos 8 primeiros caracteres (7 caracteres e o ponto decimal) que se igualem ao ingressado no campo valor (p.ex. **I21.110E**) do parâmetro. Apenas pacientes com o código completo serão considerados.

Exemplo:

Para definir um dos diagnósticos de seleção de paciente pelo primeiro diagnóstico, por exemplo, para pesquisar por pacientes com algum diagnóstico do grupo I21-Infarto agudo do miocárdio, os seguintes campos devem ser preenchidos:

CRITERIA_ID: DIAGNOSIS_SELECTION (fixo)

PARAMETER_TYPE: ICD10_CATEGORY_FULL (como já descrito, queremos pesquisar por I21 ou I21. Ou I21.?)

PARAMETER_VALUE: O valor escolhido para o diagnóstico, p.ex. **I21**, levando em conta as regras indicadas anteriormente.

INCLUSION_EXCLUSION: Algum dos seguintes valores: **I** ou **E**. I – os pacientes que satisfaçam a condição do diagnóstico serão incluídos na coorte. E – Os pacientes que satisfaçam a condição do diagnóstico serão excluídos da coorte.

GROUP: Gera um campo calculado que atribui um valor inteiro (1,2,3,4) para os pacientes que satisfaçam a condição do diagnóstico.

Os valores descritos são ingressados na tabela correspondente com um comando *SQL 'INSERT'*.

Para entrar o código I21 como critério de inclusão na tabela, efetuamos o seguinte comando:

```
INSERT INTO
  MODEL_STR_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','DIAGNOSIS_SELECTION','ICD1
0_CATEGORY_FULL',1,'I21','I','',(SELECT TRUNC(SYSDATE) FROM
DUAL),(SELECT TRUNC(SYSDATE) FROM DUAL));
```

Parâmetros de definição do estudo

Data de início do estudo: Campo do tipo data, intervalo fechado (inclui o primeiro dia).

CRITERIA_ID: 'STUDY_BEGIN'

PARAMETER_TYPE: 'INTERVAL_DATE'

PARAMETER_VALUE: TO_DATE('01-01-1985','DD-MM-YYYY')

Exemplo:

```
INSERT INTO
  MODEL_DATE_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','STUDY_BEGIN','INTERVAL_DAT
E',1,TO_DATE('01-01-1985','DD-MM-YYYY'),'I','',(SELECT TRUNC(SYSDATE)
FROM DUAL),(SELECT TRUNC(SYSDATE) FROM DUAL));
```

Data do fim do estudo: Campo do tipo data, intervalo aberto (exclui o último dia).

CRITERIA_ID: 'STUDY_END'

PARAMETER_TYPE: 'INTERVAL_DATE'

PARAMETER_VALUE: TO_DATE('01-01-2014','DD-MM-YYYY')

Exemplo:

```
INSERT INTO
  MODEL_DATE_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
```

```
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','STUDY_END','INTERVAL_DATE'
,1,TO_DATE('01-01-2014','DD-MM-YYYY'),'I','',(SELECT     TRUNC(SYSDATE)
FROM DUAL),(SELECT TRUNC(SYSDATE) FROM DUAL));
```

Parâmetros de seleção da coorte

Idade Mínima do Paciente no início do Estudo: Campo numérico em anos de vida do paciente.

CRITERIA_ID: 'PATIENT_AGE'

PARAMETER_TYPE: 'MIN_VALUE'

PARAMETER_VALUE: 18

Exemplo:

```
INSERT INTO
  MODEL_NUMBER_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','PATIENT_AGE','MIN_VALUE',1
,18,'E','ONLY ADULTS',(SELECT TRUNC(SYSDATE) FROM DUAL),(SELECT
TRUNC(SYSDATE) FROM DUAL));
```

Idade Máxima do Paciente no início do Estudo: Campo numérico em anos de vida do paciente.

```
CRITERIA_ID: 'PATIENT_AGE'

PARAMETER_TYPE: 'MAX_VALUE'

PARAMETER_VALUE: 100
```

Exemplo:

```
INSERT INTO
  MODEL_NUMBER_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','PATIENT_AGE','MAX_VALUE',1
,100,'E','ONLY ADULTS',(SELECT TRUNC(SYSDATE) FROM DUAL),(SELECT
TRUNC(SYSDATE) FROM DUAL));
```

Parâmetros de seleção do evento de indexação

Diagnósticos de seleção da data de indexação: Campo texto com um código CID-10 de diagnóstico, segundo descrito anteriormente, e um indicador se é um diagnóstico de inclusão ou exclusão. Podem ser adicionados vários diagnósticos, a seleção será feita pela presença de qualquer um deles (lógica OR).

CRITERIA_ID: COHORT_SELECTION_DIAGNOSIS (fixo)

PARAMETER_TYPE: Ver definição do parâmetro diagnóstico.

PARAMETER_VALUE: Ver definição do parâmetro diagnóstico.

INCLUSION_EXCLUSION: Algum dos seguintes valores: **I** ou **E**. I – os pacientes que satisfaçam a condição do diagnóstico serão incluídos na coorte. E – Os pacientes que satisfaçam a condição do diagnóstico serão excluídos da coorte.

GROUP: Gera um campo calculado que atribui um valor inteiro (1,2,3,4) para os pacientes que satisfaçam a condição do diagnóstico.

Os valores descritos são ingressados na tabela correspondente com um comando *SQL 'INSERT'*.
Exemplo:

```
INSERT INTO
  MODEL_STR_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','COHORT_SELECTION_DIAGNOSIS
','ICD10_CATEGORY_FULL',1,'I21','I','',(SELECT TRUNC(SYSDATE) FROM
DUAL),(SELECT TRUNC(SYSDATE) FROM DUAL));
```

Podem ser ingressados quantos diagnósticos sejam necessários. O paciente será selecionado (ou excluído) pela presença de qualquer um deles.

Parâmetros do Marcador de Evento Subsequente

O evento subsequente pode ser parametrizado podendo definir um grupo de diagnósticos de interesse e o tempo mínimo para considerar o evento, após o primeiro diagnóstico.

Tempo mínimo após o primeiro diagnóstico

Indica o tempo a partir do qual é registrada a ocorrência do evento subsequente. Isto permite filtrar ocorrências que aconteçam junto ou muito próximas do primeiro diagnóstico.

Para cada um dos eventos (diagnóstico do grupo de interesse, angioplastia ou revascularização) pode ser configurado um valor diferente para o tempo após o primeiro diagnóstico. O campo PARAMETER_ID (valores possíveis: PCI_DIAGNOSIS, CABG_DIAGNOSIS, DIAGNOSIS_OUTCOME) define para qual evento estamos definindo o valor do tempo.

Este é um parâmetro do tipo numérico que representa meses após o primeiro diagnóstico e deve ser incluído na tabela MODEL_NUMBER_PARAMETER.

Para incluir este parâmetro definimos os seguintes campos:

PARAMETER_TYPE: MONTHS_AFTER (fixo)

PARAMETER_ID: Algum dos seguintes valores: PCI_DIAGNOSIS, CABG_DIAGNOSIS, DIAGNOSIS_OUTCOME

PARAMETER_VALUE: Número que indica a quantidade de meses após o primeiro diagnóstico.

INCLUSION_EXCLUSION: I fixo – Não utilizado.

Exemplo, para entrar com 1 mês como tempo após o primeiro diagnóstico para a Angioplastia, efetuamos o seguinte comando:

```
INSERT INTO
  MODEL_NUMBER_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','MONTHS_AFTER','PCI_DIAGNOS
IS',1,1,'I','OUTCOME BY PCI, AT LEAST 1 MONTH AFTER FIRST
DIAGNOSIS',(SELECT TRUNC(SYSDATE) FROM DUAL),(SELECT TRUNC(SYSDATE)
FROM DUAL));
```

Diagnósticos de interesse para evento subsequente

Para definir cada um dos diagnósticos de interesse, preenchemos os seguintes campos:

CRITERIA_ID: OUTCOME_SELECTION_DIAGNOSIS (fixo)

PARAMETER_TYPE: Alguns dos seguintes valores: ICD10_CATEGORY_PARTIAL, ICD10_CATEGORY_FULL, ICD10_SUBCATEGORY, ICD10_ETIOLOGY, ICD10_EXTENSION, segundo já indicado.

PARAMETER_VALUE: O código CID do diagnóstico pesquisado, p. Ex. G45.8

INCLUSION_EXCLUSION: I ou E – Inclusão ou exclusão.

GROUP: 1 (fixo). Não utilizado

Exemplo, para entrar o código I21 como diagnóstico de interesse para evento subsequente na tabela de parâmetros, efetuamos o seguinte comando:

```
INSERT INTO
  MODEL_STR_PARAMETERS
(RESEARCH_ID,MODEL_ID,RESEARCHER_ID,CRITERIA_ID,PARAMETER_TYPE,PARAMET
ER_ID,PARAMETER_VALUE,INCLUSION_EXCLUSION,COMMENTS,INCLUSION_DATE,MODI
FICATION_DATE)
  VALUES
('STATIN_EFFECTIVITY','MODEL_ONE','TEREZA','OUTCOME_SELECTION_DIAGNOSI
S','ICD10_CATEGORY_FULL',2,'I21','I','',(SELECT TRUNC(SYSDATE) FROM
DUAL),(SELECT TRUNC(SYSDATE) FROM DUAL));
```

ANEXO 3– Grupo das variáveis de saída da extração e descrição

VARIÁVIES DE SAÍDA

SEQ	COLUNA	DESCRIÇÃO
	PACIENTE - DADOS DEMOGRÁFICOS	
1	UNIDENTIFIED_ID	IDENTIFICADOR ANONIMIZADO DO PACIENTE
2	GENDER	GÊNERO
3	COUNTY	MUNICIPIO
4	STATE	ESTADO
5	MARITAL_STATUS	ESTADO CIVIL
6	EDUCATION_LEVEL	NIVEL EDUCACIONAL
	DIAGNÓSTICO INDEX (INDEXAÇÃO DO PACIENTE)	
7	FIRST_DIAGNOSIS	IDADE EM MESES NO 1. DIAGNÓSTICO
8	FIRST_DIAGNOSIS_MONTH	MÊS DO 1. DIAGNÓSTICO
9	FIRST_DIAGNOSIS_YEAR	ANO DO 1. DIAGNÓSTICO
10	FIRST_ICD10	CID 10 DO 1. DIAGNOSTICO
11	FIRST_ICD10_CATEGORY	CATEGORIA DO CID DO 1. DIAGNÓSTICO
	DIAGNÓSTICO SUBSEQUENTE	
12	SECOND_DIAG	IDADE EM MESES NO 2. DIAGNÓSTICO
13	SECOND_DIAG_MONTH	MÊS DO 2. DIAGNÓSTICO
14	SECOND_DIAG_YEAR	ANO DO 2. DIAGNÓSTICO
15	SECOND_DIAG_ICD	CID-10 DO 2. DIAGNÓSTICO
16	SECOND_DIAG_ICD_CATEGORY	CATEGORIA DO CID-10 2. DIAGNÓSTICO
	PROCEDIMENTO - ANGIOPLASTIA	
17	PCI1	IDADE EM MESES DA 1. ANGIO
18	PCI1_MONTH	MÊS DA 1. ANGIO
19	PCI1_YEAR	ANO DA 1. ANGIO
20	PCI1_PROCED	PROCEDIMENTO DA 1. ANGIO
21	PCI2	IDADE EM MESES DA 2. ANGIO
22	PCI2_MONTH	MÊS DA 2. ANGIO
23	PCI2_YEAR	ANO DA 2. ANGIO
24	PCI2_PROCED	PROCEDIMENTO DA 2. ANGIO
25	PCI3	IDADE EM MESES DA 3. ANGIO
26	PCI3_MONTH	MÊS DA 3. ANGIO
27	PCI3_YEAR	ANO DA 3. ANGIO
28	PCI3_PROCED	PROCEDIMENTO DA 3. ANGIO
29	PCI4	IDADE EM MESES DA 4. ANGIO
30	PCI4_MONTH	MÊS DA 4. ANGIO
31	PCI4_YEAR	ANO DA 4. ANGIO
32	PCI4_PROCED	PROCEDIMENTO DA 4. ANGIO
33	PCI5	IDADE EM MESES DA 5. ANGIO
34	PCI5_MONTH	MÊS DA 5. ANGIO
35	PCI5_YEAR	ANO DA 5. ANGIO
36	PCI5_PROCED	PROCEDIMENTO DA 5. ANGIO
	CIRURGIA - REVASCULARIZAÇÃO DO MIOCÁRDIO	
37	CABG1	IDADE EM MESES DA 1. REVASCULARIZACAO
38	CABG1_MONTH	MÊS DA 1. REVASCULARIZACAO
39	CABG1_YEAR	ANO DA 1. REVASCULARIZACAO
40	CABG2	IDADE EM MESES DA 2. REVASCULARIZACAO
41	CABG2_MONTH	MÊS DA 2. REVASCULARIZACAO
42	CABG2_YEAR	ANO DA 2. REVASCULARIZACAO
43	CABG3	IDADE EM MESES DA 3. REVASCULARIZACAO
44	CABG3_MONTH	MÊS DA 3. REVASCULARIZACAO
45	CABG3_YEAR	ANO DA 3. REVASCULARIZACAO
	MEDICAÇÃO - ESTATINA	
46	STATIN_BEGIN	IDADE EM MESES DE INICIO DE ESTATINAS
47	STATIN_BEGIN_MONTH	MÊS DO INICIO DE ESTATINA (AMB OU HOSP)
48	STATIN_BEGIN_YEAR	ANO DO INICIO DE ESTATINA (AMB OU HOSP)
49	STATIN_END	IDADE EM MESES DE FIM DE DISPENSA DE ESTATINAS
50	STATIN_END_MONTH	MÊS FINAL DE ESTATINA (AMB OU HOSP)
51	STATIN_END_YEAR	ANO FINAL DE ESTATINA (AMB OU HOSP)
52	STATIN_AMOUNT_MONTHS	NUMERO DE MESES DE DISPENSA DE ESTATINAS
	ÓBITO	
53	DECEASE	IDADE EM MESES DO ÓBITO
54	DECEASE_MONTH	MÊS DO ÓBITO
55	DECEASE_YEAR	ANO DO ÓBITO
56	ICD_DECEASE	CID 10 DO ÓBITO (SIM)
	EVENTO SUBSEQUENTE	

57	OUTCOME	IDADE EM MESES DA OCORRÊNCIA DO DESFECHO
58	OUTCOME_MONTH	MÊS DO DESFECHO
59	OUTCOME_YEAR	ANO DO DESFECHO
60	OUTCOME_REASON	EVENTO RAZÃO DO DESFECHO
61	OUTCOMES_CNT	QUANTIDADE DE DESFECHOS (OUTROS EVENTOS DE DESFECHO)
62	FD_ICD_GROUP	GRUPO DO CID (1 DIC, 2 DCV)
63	**EVENTS_INTERVAL**	NÚMERO DE MESES ENTRE 1º DIAGNÓSTICO E 1º DESFECHO
	EXAMES LABORATORIAIS - HEMOGLOBINA GLICADA (HGLI)	
64	HBG_FIRST_TST_AGE	HBG IDADE DA 1º AMOSTRA
65	HBG_END_TST_AGE	HBG IDADE NA ÚLTIMA AMOSTRA
66	HBG_RESULTS_CNT	HBG QUANTIDADE DE EXAMES
67	HBG_MEAN	HBG MÉDIA DE TODAS AS AMOSTRAS
68	HBG_DEVIATION	HBG DESVIO PADRÃO DE TODAS AS AMOSTRAS
69	HBG_VARIANCE	HBG VARIÂNCIA DE TODAS AS AMOSTRAS
70	HBG_SLOPE	INCLINAÇÃO DA RETA
71	HBG_INTERCEPT	INTERCESSÃO DA RETA COM O EIXO Y
72	HBG_R2	R AO QUADRADO (% DE OBS QUE EXPLICA COM A RETA)
73	HBG_CORRELATION	HBG CORRELAÇÃO
74	HBG_AMNT_MNTH	QUANTIDADE DE MESES ENTRE 1º E ÚLTIMA AMOSTRA
75	HBG_FIRST_VALUE	HBG 1º VALOR DE AMOSTRA
76	HBG_FST_VAL_UNIT	HBG UNIDADE DA AMOSTRA
77	HBG_END_VALUE	HBG ÚLTIMO VALOR DE AMOSTRA
78	HBG_VARIATION	DIFERENÇA ENTRE ÚLTIMO VALOR MENOS O PRIMEIRO
	EXAMES LABORATORIAIS - LDL COLESTEROL	
78	LDL_FIRST_TST_AGE	LDL IDADE DA 1º AMOSTRA
79	LDL_END_TST_AGE	LDL IDADE NA ÚLTIMA AMOSTRA
80	LDL_RESULTS_CNT	LDL QUANTIDADE DE EXAMES
81	LDL_MEAN	LDL MÉDIA
82	LDL_DEVIATION	LDL DESVIO PADRÃO
83	LDL_VARIANCE	LDL VARIÂNCIA
84	LDL_SLOPE	INCLINAÇÃO DA RETA
85	LDL_INTERCEPT	INTERCESSÃO DA RETA COM O EIXO Y
86	LDL_R2	R AO QUADRADO (% DE OBS QUE EXPLICA COM A RETA)
87	LDL_CORRELATION	LDL CORRELAÇÃO
88	LDL_AMNT_MNTH	QUANTIDADE DE MESES ENTRE 1º E ÚLTIMA AMOSTRA
89	LDL_FIRST_VALUE	LDL 1º VALOR DE AMOSTRA
90	LDL_FST_VAL_UNIT	LDL UNIDADE DA AMOSTRA
91	LDL_END_VALUE	LDL ÚLTIMO VALOR DE AMOSTRA
92	LDL_VARIATION	DIFERENÇA ENTRE ÚLTIMO VALOR MENOS O PRIMEIRO
	EXAMES LABORATORIAIS - HDL COLESTEROL (HDL)	
93	HDL_FIRST_TST_AGE	HDL IDADE DA 1º AMOSTRA
94	HDL_END_TST_AGE	HDL IDADE NA ÚLTIMA AMOSTRA
95	HDL_RESULTS_CNT	HDL QUANTIDADE DE EXAMES
96	HDL_MEAN	HDL MÉDIA
97	HDL_DEVIATION	HDL DESVIO PADRÃO
98	HDL_VARIANCE	HDL VARIÂNCIA
99	HDL_CORRELATION	HDL CORRELAÇÃO
100	HDL_SLOPE	INCLINAÇÃO DA RETA
101	HDL_INTERCEPT	INTERCESSÃO DA RETA COM O EIXO Y
102	HDL_R2	R AO QUADRADO (% DE OBS QUE EXPLICA COM A RETA)
103	HDL_AMNT_MNTH	QUANTIDADE DE MESES ENTRE 1º E ÚLTIMA AMOSTRA
104	HDL_FIRST_VALUE	HDL 1º VALOR DE AMOSTRA
105	HDL_FST_VAL_UNIT	HDL UNIDADE DA AMOSTRA
106	HDL_END_VALUE	HDL ÚLTIMO VALOR DE AMOSTRA
107	HDL_VARIATION	DIFERENÇA ENTRE ÚLTIMO VALOR MENOS O PRIMEIRO
	EXAMES LABORATORIAIS - COLESTEROL TOTAL (COL-TOTAL)	
108	CHL_FIRST_TST_AGE	CHL IDADE DA 1º AMOSTRA
109	CHL_END_TST_AGE	CHL IDADE NA ÚLTIMA AMOSTRA
110	CHL_RESULTS_CNT	CHL QUANTIDADE DE EXAMES
111	CHL_MEAN	CHL MÉDIA
112	CHL_DEVIATION	CHL DESVIO PADRÃO
113	CHL_VARIANCE	CHL VARIÂNCIA
114	CHL_SLOPE	INCLINAÇÃO DA RETA
115	CHL_INTERCEPT	INTERCESSÃO DA RETA COM O EIXO Y
116	CHL_R2	R AO QUADRADO (% DE OBS QUE EXPLICA COM A RETA)
117	CHL_CORRELATION	CHL CORRELAÇÃO
118	CHL_AMNT_MNTH	QUANTIDADE DE MESES ENTRE 1º E ÚLTIMA AMOSTRA
119	CHL_FIRST_VALUE	CHL 1º VALOR DE AMOSTRA
120	CHL_FST_VAL_UNIT	CHL UNIDADE DA AMOSTRA
121	CHL_END_VALUE	CHL ÚLTIMO VALOR DE AMOSTRA
122	CHL_VARIATION	DIFERENÇA ENTRE ÚLTIMO VALOR MENOS O PRIMEIRO (CHL_END_VALUE - CHL_FIRST_VALUE)

	EXAMES LABORATORIAIS - GLICEMIA	
123	GLI_FIRST_TST_AGE	GLI IDADE DA 1º AMOSTRA
124	GLI_END_TST_AGE	GLI IDADE NA ÚLTIMA AMOSTRA
125	GLI_RESULTS_CNT	GLI QUANTIDADE DE EXAMES
126	GLI_MEAN	GLI MÉDIA
127	GLI_DEVIATION	GLI DESVIO PADRÃO
128	GLI_CORRELATION	GLI CORRELAÇÃO
129	GLI_VARIANCE	GLI VARIANCIA
130	GLI_SLOPE	INCLINAÇÃO DA RETA
131	GLI_INTERCEPT	INTERCESSÃO DA RETA COM O EIXO Y
132	GLI_R2	R AO QUADRADO (% DE OBS QUE EXPLICA COM A RETA)
133	GLI_AMNT_MNTH	QUANTIDADE DE MESES ENTRE 1º E ÚLTIMA AMOSTRA
134	GLI_FIRST_VALUE	GLI 1º VALOR DE AMOSTRA
135	GLI_FST_VAL_UNIT	GLI UNIDADE DA AMOSTRA
136	GLI_END_VALUE	GLI ÚLTIMO VALOR DE AMOSTRA
137	GLI_VARIATION	DIFERENÇA ENTRE ÚLTIMO VALOR MENOS O PRIMEIRO
	ENCONTROS - VISITAS AO HOSPITAL	
138	LAST_ENCOUNTER	IDADE EM MESES NA ÚLTIMA VISITA AO HOSPITAL
139	**ENCOUNTER_INTERVAL**	NÚMERO DE MESES ENTRE 1º DIAGNÓSTICO E A ÚLTIMA VISITA AO HOSPITAL
140	LAST_ENCOUNTER_DATE_MONTH	MES DA ÚLTIMA VISITA AO HOSPITAL
141	LAST_ENCOUNTER_DATE_YEAR	ANO DA ÚLTIMA VISITA AO HOSPITAL
142	ENCOUNTER_EMERGENCY	QUANTIDADE DE ENCONTROS - PRONTO SOCORRO
143	ENCOUNTER_OUTPATIENT	QUANTIDADE DE ENCONTROS - AMBULATÓRIO
144	ENCOUNTER_LABORATORY	QUANTIDADE DE ENCONTROS - LABORATÓRIO
145	ENCOUNTER_INPATIENT	QUANTIDADE DE ENCONTROS - INTERNAÇÃO
146	ENCOUNTER_CNT	QUANTIDADE DE ENCONTROS (VISITAS HOSPITAL)
	MEDICAMENTOS DISPENSADOS	
147	FIRST_DATE_MEDIC_MONTH	MÊS DA 1. MEDICAÇÃO (QUALQUER MEDICAÇÃO DE FARMACIA)
148	FIRST_DATE_MEDIC_YEAR	ANO DA 1. MEDICAÇÃO (QUALQUER MEDICAÇÃO DE FARMACIA)
149	END_DATE_MEDIC_MONTH	MÊS DA ULTIMA MEDICAÇÃO (QUALQUER MEDICAÇÃO DE FARMACIA)
150	END_DATE_MEDIC_YEAR	ANO DA ULTIMA MEDICAÇÃO (QUALQUER MEDICAÇÃO DE FARMACIA)
151	AMOUNT_MONTHS_MEDIC	QUANTIDADE DE MESES DE DESPENSA DE MEDICAÇÃO
	MEDICAMENTOS POR GRUPOS E QUANTIDADES DE RECEITAS	
152	MEDIC_ANTIAGGREGANTS	QUANTIDADE DE RECEITAS DE FARMACIA DE MEDICAMENTOS ANTIAGREGANTES
153	MEDIC_HYPOGLYCEMIANTS	QUANTIDADE DE RECEITAS DE FARMACIA DE MEDICAMENTOS HIPOGLICEMIANTES
154	MEDIC_HYPOLIPEMIANTS	QUANTIDADE DE RECEITAS DE FARMACIA DE MEDICAMENTOS HIPOLIPIMIANTES
155	MEDIC_HYPOTENSIVES	QUANTIDADE DE RECEITAS DE FARMACIA DE MEDICAMENTOS HIPOTENSORES
156	MEDIC_PRESCRIPTION_QTY	QUANTIDADE TOTAL DE RECEITAS DE FARMACIA (PARA MEDICAMENTOS DE SELEÇÃO)
	INDICADORES	
157	SOME_MEDIC_PHARMACY_FLAG_NUM	INDICADOR DE USO DE QUALQUER MEDICAÇÃO DE FARMACIA (1=Y e 0= N)
158	MEDIC_ANTIAGGREGANTS_FLAG	INDICADOR DE USO DE ANTIAGREGANTES (1=Y e 0= N)
159	MEDIC_HYPOGLYCEMIANTS_FLAG	INDICADOR DE USO DE HIPOGLICEMIANTES (1=Y e 0= N)
160	MEDIC_HYPOLIPEMIANTS_FLAG	INDICADOR DE USO DE HIPOLIPIMIANTES (1=Y e 0= N)
161	MEDIC_HYPOTENSIVES_FLAG	INDICADOR DE USO DE HIPOTENSORES (1=Y e 0= N)
162	**STATIN_FLAG**	INDICADOR DE USO DE ESTATINA (Y-N-NI-NULL) Y=SIM, N =NÃO ESTATINA, NI= OUTROS MEDIC E NULL=SEM REGISTRO
164	INTERVENTION_FLAG	INCDICADOR DE INTERVENÇÃO (1=Y e 0= N) 1 COM ESTATINA, 0 SEM ESTATINA
165	OUTCOME_FLAG_NUM	INDICADOR DE OUTCOME (1=Y e 0= N)
166	SECOND_DIAG_FLAG_NUM	INDICADOR DE 2. DIAGNOSTICO (1=Y e 0= N)
167	DECEASE_FLAG_NUM	INDICADOR DE OBITO (1=Y e 0= N)
168	PCI_FLAG_NUM	INDICADOR DE ANGIOPLASTIA (1=Y e 0= N)
169	CABG_FLAG_NUM	INDICADOR DE REVASCULARIZACAO (1=Y e 0= N)
170	LAB_RESULTS_FLAG_NUM	INDICADOR DE EXAMES DE LABORATORIO (1=Y e 0= N)
	INDICADORES - FATORES DE RISCO	
171	RF_SMOKE_NUM	INDICADOR DE FUMANTE (1=Y, 0= N E NULL)
172	RF_PHYSICAL_INACTIVITY_NUM	INDICADOR DE NATIVIDADE FISICA (1=Y, 0= N E NULL)
173	RF_HIGH_BLOOD_PRESSURE_NUM	INDICADOR DE PRESSAO ALTA (1=Y, 0= N E NULL)
174	RF_HIGH_BLOOD_CHOESTEROL_NUM	INDICADOR DE COLESTEROL ALT0 (1=Y, 0= N E NULL)
175	RF_OBESITY_NUM	INDICADOR DE OBESIDADE (1=Y, 0= N E NULL)
176	RF_DIABETES_NUM	INDICADOR DE DIABETES (1=Y, 0= N E NULL)
177	RF_HYPERTRIGLYCERIDEMIA_NUM	INDICADOR DE HIPERTRIGLICEMIA(1=Y, 0= N E NULL)

178	RF_HYPERURICEMIA_NUM	INDICADOR DE HYPERURICEMIA (1=Y, 0= N E NULL)
179	RF_ALCOHOLISM_NUM	INDICADOR DE ALCOLISMO (1=Y, 0= N E NULL)
180	RF_DRUGS_NUM	INDICADOR DE DROGAS ILICITAS (1=Y, 0= N E NULL)
181	RF_ORAL_CONTRACEPTION_NUM	INDICADOR DE CONTRACEPTIVO ORAL (1=Y, 0= N E NULL)
182	RF_STRESS_NUM	INDICADOR DE STRESS (1=Y, 0= N E NULL)
183	RF_MENOPAUSE_NUM	INDICADOR DE MENOPAUSA (1=Y, 0= N E NULL)
184	RF_ENDOCARDITIS_NUM	INDICADOR DE ENDOCARDITE (1=Y, 0= N E NULL)
185	RF_RHEUMATIC_FEVER_NUM	INDICADOR DE FEBRE REUMÁTICA (1=Y, 0= N E NULL)
	MEDICAÇÕES HIPOLIPIEMIANTES	
186	ATORVASTATINA_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
187	ATORVASTATINA_CALCICA_10MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
188	ATORVASTATINA_20MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
189	ATORVASTATINA_40MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
190	ATORVASTATINA_40MG_COMP_REVES	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
191	SINVASTATINA_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
192	SINVASTATINA_5MG_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
193	SINVASTATINA_10MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
194	SINVASTATINA_20MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
195	SINVASTATINA_40MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
196	PRAVASTATINA_10MG_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
197	PRAVASTATINA_20MG_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
198	PRAVASTATINA_40MG_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
199	ROSUVASTATINA_10MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
200	LOVASTATINA_10MG_COMP	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
201	CIPROFIBRATO_100MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
202	EZETIMIBE_10MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)
203	ACIDO NICOTINICO_500MG	QUANTIDADE DE RECEITAS (0 - SEM REGISTRO DE RECEITAS)

A variável **STATIN_FLAG** foi resultante da seleção de medicamentos juntando prescrições de pacientes internados e dispensações de receitas de farmácia. Todas as variáveis com prefixo (STATIN_) foram resultantes dessa seleção (farmácia ambulatorial e prescrição hospitalar).

*** Statin_flag:**

A variável STATIN_FLAG tem 4 opções:

Y = SIM (tiveram dispensação de estatina),

N= NAO (tiveram dispensação de outros medicamentos de interesse e NÃO tiveram dispensação de estatinas)

NI = NAO INTERESSE (tiveram dispensação de outros medicamentos, mas não tiveram dispensação de medicações de interesse ou estatinas),

null (vazio) = NAO TIVERAM DISPENSAÇÃO DE MEDICAMENTOS.

Regressão linear (aproximação com a reta)

Inclinação	
0	horizontal
infinito	vertical
(+) mais	cima
(-) menos	baixo

ANEXO 4 – Grupo de Medicamentos de seleção para estudo DCV

Grupo	Descrição dos Medicamentos
HIPOLIPEMIANTES	**ATORVASTATINA 20MG**
	ATORVASTATINA 40MG
	ATORVASTATINA 40 MG COMP. REVES.
	ATORVASTATINA (87) COMP.
	ATORVASTATINA CALCICA 10MG
	SINVASTATINA 5 MG
	SINVASTATINA 10MG
	SINVASTATINA 20MG
	SINVASTATINA 40MG
	PRAVASTATINA 10 MG COMP.
	PRAVASTATINA 20 MG COMP.
	PRAVASTATINA 40 MG COMP.
	LOVASTATINA 10 MG COMP.
	ROSUVASTATINA 10 MG
	EZETIMIBE 10MG
	CIPROFIBRATO 100MG
	ACIDO NICOTINICO 500MG
HIPOTENSORES	AMLODIPINA (BESILATO) 5MG
	ATENOLOL 100MG
	ATENOLOL 50MG
	CAPTOPRIL 12,5 MG
	CAPTOPRIL 25MG
	CINARIZINA 75MG
	CLONIDINA CLORIDRATO 0,100MG
	CLORTALIDONA 12,5MG
	CLORTALIDONA 50MG
	DOXAZOSINA 4 MG COMP.
	ESPIRONOLACTONA 100MG CP
	ESPIRONOLACTONA 25MG CP
	FUROSEMIDA 40MG
	HIDRALAZINA, CLORIDRATO 25MG
	HIDROCLOROTIAZIDA 25MG COMPRIMIDO
	LISINOPRIL 20MG
	LOSARTAN 50MG
	MALEATO DE ENALAPRIL 20MG
	MALEATO DE ENALAPRIL 5MG
	METILDOPA 250MG
	METOPROLOL TARTARATO 100MG
	MINOXIDIL 10MG
	NIFEDIPINA 20MG RETARD
	PROPRANOLOL 10MG
	PROPRANOLOL 40MG
	PROPRANOLOL 80MG
	VERAPAMIL 80MG
	CARVEDILOL 25MG
	CARVEDILOL 6,25MG
	DILTIAZEM 30MG
HIPOGLICEMIANTES	ACARBOSE 50 MG

	GLIBENCLAMIDA 5MG
	GLICAZIDA 30MG COMP. LIBERACAO MODIFICADA
	INSULINA HUMANA ACAO INTERMED.100UI/ML
	INSULINA HUMANA ACAO RAPIDA REGULAR 100 UI/ML - FR/AMP
	PIOGLITAZONA 30MG
ANTIAGREGANTES	ACIDO ACETILSALICILICO 100MG
	CLOPIDOGREL 75MG

10 REFERÊNCIAS

ABNT. ABNT - Associação Brasileira de Normas Técnicas [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://www.abnt.org.br/

ABNT/CEE-78. ABNT - ABNT/CEE-078 - Comissão de Estudo Especial de Informática em Saúde [Internet]. [citado 16 de janeiro de 2016]. Recuperado de: http://www.abnt.org.br/cb-78

Abrahão MT, Nobre MRC, Gutierrez MA. Descriptive Statistics of 65000 Patients Treated for Myocardial Ischemia: data from routine electronic health records. Evidence Live 2013, Oxford UK; 2013a.

Abrahão MT, Nobre MRC, Gutierrez MA. Estudos retrospectivos em base de dados assistencial em um hospital de referência em cardiologia. XIV Congresso Brasileiro de Informática em saúde - CBIS; 2014.

Abrahão MT, Nobre MR, Gutierrez MA. Estatística Descritiva de uma população de pacientes atendidos no InCor com Doença Cardiovascular Aterosclerótica. XIII Congresso Brasileiro de Informática em saúde - CBIS; 2012a.

Abrahão MT, Nobre MR, Gutierrez MA. The effectiveness of statins in the treatment of cardiovascular disease: Cross-sectional study with paired groups from electronic patient records. Value Health [Internet]. 2013b [citado 24 de janeiro de 2016];16(7):A518–A518. Recuperado de: http://linkinghub.elsevier.com/retrieve/pii/S1098301513031380

Abrahão MT, Santos RS, Almeida AL, Pires FA, Gutierrez MA. Monitoramento de Intervenções de Alta Complexidade Utilizando Técnicas de Mineração de Dados em Saúde Pública. XI Congresso brasileiro de Informática em Saúde, Campos do Jordão; 2008.

Abrahão MT, Soares TJ, Nobre MRC, Rebelo MFS, Pires FA, Gutierrez MA. Avaliação do uso de medicação de alto custo no tratamento da artrite reumatóide no Estado de São Paulo, a partir da mineração de dados do SUS. XXII Congresso Brasileiro de Engenharia Biomédica - CBEB 2010. Tiradentes - MG. p. 1089-1092; 2012b.

Abrahão MT, Soares TJ, Pires FA, Gutierrez MA. Construção de um Data Warehouse a partir de dados do SUS do Estado de São Paulo: coleta, preparação e validação da base de dados. CBIS 2010, Porto de Galinhas - Recife. XII Congresso Brasileiro de Informática em Saúde, 2010. v. 1.; 2010a.

Abrahão MT, Soares TJ, Pires FA, Gutierrez MA, Nobre MRC. Data Warehouse com dados da Saúde Pública: Estudo de Caso sobre o Tratamento de Doença Cardiovascular Aterosclerótica no Estado de São Paulo. CBIS 2010, Porto de

galinhas - Recife. XII Congresso Brasileiro de Informática em Saúde, 2010. v. 1; 2010b.

AHRQ. Agency for Healthcare Research & Quality [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://www.ahrq.gov/

Alonso A, Jick SS, Hernán MA. Allergy, histamine 1 receptor blockers, and the risk of multiple sclerosis. Neurology. 28 de fevereiro de 2006;66(4):572–5.

Armstrong-Wells J, Johnston SC, Wu YW, Sidney S, Fullerton HJ. Prevalence and predictors of perinatal hemorrhagic stroke: results from the kaiser pediatric stroke study. Pediatrics. março de 2009;123(3):823–8.

ATC. ATC- 1_2013guidelines.pdf [Internet]. 2013 [citado 24 de janeiro de 2016]. Recuperado de: http://www.whocc.no/filearchive/publications/1_2013guidelines.pdf

ATC - Guidelines. atcguidelines2015final.pdf. 2015.

Berger ML, Mamdani M, Atkins D, Johnson ML. Good Research Practices for Comparative Effectiveness Research: Defining, Reporting and Interpreting Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. Value Health [Internet]. novembro de 2009 [citado 1 de abril de 2014];12(8):1044–52. Recuperado de: http://linkinghub.elsevier.com/retrieve/pii/S1098301510603087

Berger ML, Martin BC, Husereau D, Worley K, Allen JD, Yang W, et al. A Questionnaire to Assess the Relevance and Credibility of Observational Studies to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report. Value Health [Internet]. março de 2014 [citado 14 de abril de 2014];17(2):143–56. Recuperado de: http://linkinghub.elsevier.com/retrieve/pii/S1098301514000096

Berkowitz SA, Karter AJ, Lyles CR, Liu JY, Schillinger D, Adler NE, et al. Low socioeconomic status is associated with increased risk for hypoglycemia in diabetes patients: the Diabetes Study of Northern California (DISTANCE). J Health Care Poor Underserved. maio de 2014;25(2):478–90.

Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA—U.S. Department of Veterans Affairs national-scale HIS. Int J Med Inf [Internet]. março de 2003 [citado 25 de janeiro de 2016];69(2–3):135–56. Recuperado de: http://www.sciencedirect.com/science/article/pii/S1386505602001314

BVSMS. Ministério da Saúde [Internet]. [citado 16 de janeiro de 2016]. Recuperado de: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_31_08_2011.html

Caramelli B, Fornari LS, Monachini M, Ballas D, Fachini NR, de Pádua Mansur A, et al. Tendências Seculares da População com Doença Isquêmica do Coração Internada no Instituto do Coração-São Paulo. Arq Bras Cardiol [Internet]. 2003a

[citado 1 de abril de 2014];81(4):363–8. Recuperado de: http://www.scielo.br/pdf/abc/v81n4/17720.pdf

Caramelli B, Fornari LS, Monachini M, Ballas D, Frachini R, Mansur AP, et al. Características demográficas da população submetida à cinecoronariografia no Instituto do Coração da Faculdade de Medicina da USP de 1986 a 1995. . Arq. bras. cardiol 81, no. 3: 303-308; 2003b.

Carrilho JF, Nita ME, Nobre MRC, Secili S. Avaliação de tecnologias em saúde: Evidência Clínica, Análise Econômica e Análise de Decisão. Artmed; 2010.

Carvalho MS, Andreozzi VL, Codeço CT, Campos DP, Barbosa MTS, Shimakura SE. Análise de sobrevivência: teoria e aplicações em saúde. SciELO - Editora FIOCRUZ; 2011.

CBCD. Centro Brasileiro de Classificação de Doenças - CBCD [Internet]. 2008 [citado 23 de janeiro de 2016]. Recuperado de: http://www.fsp.usp.br/cbcd/

CER-Guide. Methods Guide for Effectiveness and Comparative Effetiveness Reviews - CER-Methods-Guide-140109.pdf [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://www.effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf

Certificação SBIS-CFM. Manual de Certificacao SBIS-CFM 2013 v4_0 pre 09_docx - [Internet]. ManualCertificacaoSBIS-CFM2013v4-1pdf. 2013 [citado 24 de janeiro de 2016]. Recuperado de: http://www.sbis.org.br/certificacao/Manual_Certificacao_SBIS-CFM_2013_v4-1.pdf

Chambers LW. Evidence-Based Healthcare: How to Make Health Policy and Management Decisions. CMAJ Can Med Assoc J [Internet]. 1 de dezembro de 1997 [citado 24 de janeiro de 2016];157(11):1598–9. Recuperado de: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1228586/

Cheuk BLY, Cheung GCY, Cheng SWK. Epidemiology of venous thromboembolism in a Chinese population. Br J Surg. abril de 2004;91(4):424–8.

CID-10. Road to 10 » ICD-10 Basics [Internet]. 2008 [citado 30 de janeiro de 2016]. Recuperado de: http://www.roadto10.org/icd-10-basics/

CNES. Cadastro Nacional de Estabelecimentos de Saúde [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://cnes.datasus.gov.br/

CNS. Cartão Nacional de Saúde [Internet]. Portal Saúde – Minist. Saúde – Wwwsaudegovbr. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/sgep/cartao-nacional-de-saude

Codd EF. A relational model of data for large shared data banks. Commun ACM [Internet]. 1970 [citado 28 de janeiro de 2016];13(6):377–87. Recuperado de: http://dl.acm.org/citation.cfm?id=362685

Codd EF. This paper attempts to provide a theoretical basis which may be used to determine how complete a selection capability is provided in a proposed data sublanguage independently of any host lan-guage in which the sublanguage may be embedded. A. Data Base Syst. 1972;6:65.

Codd EF. Does Your DBMS Run By the Rules. Computerworld. Vol. 19; 21 de outubro de 1985a;:p49.

Codd EF. Is Your DBMS Really Relational? Computerworld. Vol. 19; 14 de outubro de 1985b;:pID1.

Coles M. Four Rules for NULLs - SQLServerCentral [Internet]. 2008 [citado 30 de janeiro de 2016]. Recuperado de: http://www.sqlservercentral.com/articles/Advanced+Querying/fourrulesfornulls/1915/

Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good Research Practices for Comparative Effectiveness Research: Approaches to Mitigate Bias and Confounding in the Design of Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. Value Health [Internet]. novembro de 2009 [citado 1 de abril de 2014];12(8):1053–61. Recuperado de: http://linkinghub.elsevier.com/retrieve/pii/S1098301510603099

Croen LA, Grether JK, Yoshida CK, Odouli R, Van de Water J. Maternal autoimmune diseases, asthma and allergies, and childhood autism spectrum disorders: a case-control study. Arch Pediatr Adolesc Med. fevereiro de 2005;159(2):151–7.

DATASUS. DATASUS [Internet]. 2016 [citado 16 de janeiro de 2016]. Recuperado de: http://www.datasus.gov.br/DATASUS/index.php?acao=11&id=28333

DATASUS - CDM. Arquivo(s) do CMD para download. [Internet]. 2016 [citado 20 de janeiro de 2016]. Recuperado de: http://sia.datasus.gov.br/documentos/listar_ftp_cmd.php

DATASUS CID-10. CID-10 [Internet]. 2008 [citado 23 de janeiro de 2016]. Recuperado de: http://www.datasus.gov.br/cid10/V2008/cid10.htm

Date CJ. Introdução a sistemas de banco de dados. Rio de Janeiro: Campus; 2000.

Date CJ. Logic and Databases: The Roots of Relational Theory. Trafford Publishing; 2007.

Date CJ, Darwen H. A Guide to SQL Standard. 4 edition. Reading, Mass: Addison-Wesley Professional; 1996.

De Vera MA, Bhole V, Burns LC, Lacaille D. Impact of statin adherence on cardiovascular disease and mortality outcomes: a systematic review. Br J Clin Pharmacol. outubro de 2014;78(4):684–98.

Dhalwani NN, West J, Sultan AA, Ban L, Tata LJ. Women with celiac disease present with fertility problems no more often than women in the general population. Gastroenterology. dezembro de 2014;147(6):1267–74.e1; quiz e13–4.

DICOM. DICOM Homepage [Internet]. 2016 [citado 16 de fevereiro de 2016]. Recuperado de: http://dicom.nema.org/

Donoho DL. An invitation to reproducible computational research. Biostatistics [Internet]. 1 de julho de 2010 [citado 24 de janeiro de 2016];11(3):385–8. Recuperado de: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxq028

Eapen ZJ, McBroom AJ, Gray R, Musty MD, Hadley C, Hernandez AF, et al. Priorities for Comparative Effectiveness Reviews in Cardiovascular Disease. Circ Cardiovasc Qual Outcomes [Internet]. 1 de março de 2013 [citado 30 de março de 2014];6(2):139–47. Recuperado de: http://circoutcomes.ahajournals.org/content/6/2/139

FDA. U S Food and Drug Administration Home Page [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://www.fda.gov/default.htm

Ferreira MM, Passos CMF. Iniciativa STROBE: subsídios para a comunicação de estudos observacionais. SciELO Public Health [Internet]. 2010 [citado 19 de janeiro de 2016];44:559–65. Recuperado de: http://www.scielosp.org/pdf/rsp/v44n3/21.pdf

Fortuna D, Nicolini F, Guastaroba P, Palma RD, Bartolomeo SD, Saia F, et al. Coronary artery bypass grafting vs percutaneous coronary intervention in a "real-world" setting: a comparative effectiveness study based on propensity score-matched cohorts. Eur J Cardiothorac Surg [Internet]. 1 de julho de 2013 [citado 28 de março de 2014];44(1):e16–24. Recuperado de: http://ejcts.oxfordjournals.org/content/44/1/e16

Framingham Heart Study. Framingham Heart Study [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: https://www.framinghamheartstudy.org/

Furuie SS, Rebelo MFS, Gutierrez MA, Moreno RA, Nardon FB, Tachinardi U. Prontuário Eletrônico de Pacientes: integrando informações clínicas e imagens médicas. Revista Brasileira de Engenharia Biomédica. Rio de Janeiro, v. 19, p. 125-137; 2003.

Furuie SS, Rebelo MS, Moreno RA, Santos M, Bertozzo N, Motta GHMB, et al. Managing Medical Images and Clinical Information: InCor's Experience. IEEE Trans Inf Technol Biomed [Internet]. janeiro de 2007 [citado 24 de janeiro de 2016];11(1):17–24. Recuperado de: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4049796

Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol. setembro de 2007;17(9):643–53.

Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, et al. Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 8 de fevereiro de 2016 [citado 18 de fevereiro de 2016];4(1). Recuperado de: http://repository.edm-forum.org/egems/vol4/iss1/2

GitHub - MTFA. MTFA/CohortEx [Internet]. GitHub. [citado 31 de janeiro de 2016]. Recuperado de: https://github.com/MTFA/CohortEx

Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Blaha MJ, et al. Heart Disease and Stroke Statistics—2014 Update A Report From the American Heart Association. Circulation [Internet]. 18 de dezembro de 2013 [citado 27 de março de 2014];01.cir.0000441139.02102.80. Recuperado de: http://circ.ahajournals.org/content/early/2013/12/18/01.cir.0000441139.02102.80

Goettsch W g., Heintjes E m., Kastelein J j. p., Rabelink T j., Johansson S, Herings R m. c. Results from a rosuvastatin historical cohort study in more than 45 000 Dutch statin users, a PHARMO study. Pharmacoepidemiol Drug Saf [Internet]. 1 de julho de 2006 [citado 28 de março de 2014];15(7):435–43. Recuperado de: http://onlinelibrary.wiley.com/doi/10.1002/pds.1278/abstract

Goncalves Sa JH, Sa Rebelo M, Brentani A, Grisi S, Gutierrez MA. GeoHealth: A Georeferenced System for Health Data Analysis in Primary Care. IEEE Lat Am Trans [Internet]. janeiro de 2012 [citado 15 de fevereiro de 2016];10(1):1352–6. Recuperado de: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6142483

Gordis L. Epidemiology. Fifth edition. Philadelphia, PA: Elsevier/Saunders; 2014.

Green A, Macdonald S, Rice R. Policy-making for Research Data in Respositories: A Guide [Internet]. Citeseer; 2009 [citado 22 de janeiro de 2016]. Recuperado de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.218.467&rep=rep1&type=pdf

Greenhalgh T. Effectiveness and Efficiency: Random Reflections on Health Services. BMJ [Internet]. 26 de fevereiro de 2004 [citado 24 de janeiro de 2016];328(7438):529. Recuperado de: http://www.bmj.com/content/328/7438/529.1

Hans-Joachim K. Null Values in Relational Databases and Sure Information Answers. In: Bertossi L, Katona GOH, Schewe K-D, Thalheim B, organizadores. Semant Databases [Internet]. Springer Berlin Heidelberg; 2001 [citado 30 de janeiro de 2016]. p. 119–38. Recuperado de: http://link.springer.com/chapter/10.1007/3-540-36596-6_7

HHS USD of H and H. HHS.gov [Internet]. HHS.gov. [citado 24 de janeiro de 2016]. Recuperado de: http://www.hhs.gov/

Hibbard JU, Wilkins I, Sun L, Gregory K, Haberman S, Hoffman M, et al. Respiratory morbidity in late preterm births. JAMA. 28 de julho de 2010;304(4):419–25.

HIMSS. HIMSS - Healthcare Information and Management Systems Society [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://www.himss.org/

HIMSS - Definition. HIMSS Interoperability Definition FINAL.pdf [Internet]. [citado 16 de janeiro de 2016]. Recuperado de: http://s3.amazonaws.com/rdcms-himss/files/production/public/FileDownloads/HIMSS%20Interoperability%20Definition%20FINAL.pdf

HL7. Health Level Seven International - Homepage [Internet]. 2016 [citado 16 de fevereiro de 2016]. Recuperado de: http://www.hl7.org/

Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. MEDINFO [Internet]. 2015 [citado 31 de outubro de 2015];15. Recuperado de: http://www.researchgate.net/profile/Vojtech_Huser/publication/281815340_Observational_Health_Data_Sciences_and_Informatics_(OHDSI)_Opportunities_for_Observational_Researchers/links/560d4d9208ae6cf68153e89d.pdf

Hulley SB, Cummings SR, Warren S, et al. Delineando a pesquisa clínica: uma abordagem epidemiológica. 2 Edição. Porto Alegre: Artmed, 2003; 2003.

Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 13 de julho de 2005a;294(2):218–28.

Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med [Internet]. 30 de agosto de 2005b [citado 24 de janeiro de 2016];2(8):e124. Recuperado de: http://dx.plos.org/10.1371/journal.pmed.0020124

ISO. ISO Standards - ISO [Internet]. 2016 [citado 15 de janeiro de 2016]. Recuperado de: http://www.iso.org/iso/home/standards.htm

ISO/IEC. ISO/IEC 9075-1:2003, “SQL/Framework”. ISO/IEC. Section 4.4.2: The null value [Internet]. Recuperado de: http://www.iso.org/iso/catalogue_detail.htm?csnumber=34132

ISO/TC-215. ISO - Technical committees - ISO/TC 215 - Health informatics [Internet]. 2016 [citado 16 de janeiro de 2016]. Recuperado de: http://www.iso.org/iso/iso_technical_committee?commid=54960

ISPOR. International Society For Pharmacoeconomics and Outcomes Research [Internet]. [citado 17 de fevereiro de 2016]. Recuperado de: http://www.ispor.org/

Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good Research Practices for Comparative Effectiveness Research: Analytic Methods to Improve Causal Inference from Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III. Value Health [Internet]. novembro de 2009 [citado 1 de abril de 2014];12(8):1062–73. Recuperado de: http://linkinghub.elsevier.com/retrieve/pii/S1098301510603105

Kanaya AM, Adler N, Moffet HH, Liu J, Schillinger D, Adams A, et al. Heterogeneity of diabetes outcomes among asians and pacific islanders in the US: the diabetes study of northern california (DISTANCE). Diabetes Care. abril de 2011;34(4):930–7.

Karp NA. R commander an Introduction [Internet]. 2010 [citado 6 de agosto de 2014]. Recuperado de: http://cran.vinastat.com/doc/contrib/Karp-Rcommander-intro2.pdf

Laraia BA, Karter AJ, Warton EM, Schillinger D, Moffet HH, Adler N. Place matters: neighborhood deprivation and cardiometabolic risk factors in the Diabetes Study of Northern California (DISTANCE). Soc Sci Med 1982. abril de 2012;74(7):1082–90.

Lauer MS. Time for a creative transformation of epidemiology in the United States. JAMA. 7 de novembro de 2012;308(17):1804–5.

Lisboa LAF, Moreira LFP, Mejia OV, Dallan LAO, Pomerantzeff PM, Costa R, et al. Evolução da cirurgia cardiovascular no Instituto do Coração: análise de 71.305 operações. Arq Bras Cardiol [Internet]. 2010 [citado 15 de novembro de 2015];94(2):174–81. Recuperado de: http://www.scielo.br/pdf/abc/v94n2/06.pdf

Lui X. Survival Analysis: Models and Applications. 2012.

Luque A, Nobre M, Abrahão M. Cost-Utility Of Statin In Secondary Prevention: A Propensity Score Method Of Administrative Database. Value Health [Internet]. novembro de 2015 [citado 24 de janeiro de 2016];18(7):A395–6. Recuperado de: http://linkinghub.elsevier.com/retrieve/pii/S1098301515029708

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

Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the Impact of Database Heterogeneity on Observational Study Results. Am J Epidemiol [Internet]. 15 de agosto de 2013 [citado 2 de novembro de 2015];178(4):645–51. Recuperado de: http://aje.oxfordjournals.org/cgi/doi/10.1093/aje/kwt010

Madigan D, Stang PE, Berlin JA, Schuemie M, Overhage JM, Suchard MA, et al. A Systematic Statistical Approach to Evaluating Evidence from Observational Studies. Annu Rev Stat Its Appl [Internet]. 3 de janeiro de 2014 [citado 20 de janeiro de 2016];1(1):11–39. Recuperado de: http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-022513-115645

Mahajan R, Xing J, Liu SJ, Ly KN, Moorman AC, Rupp L, et al. Mortality among persons in care with hepatitis C virus infection: the Chronic Hepatitis Cohort Study (CHeCS), 2006-2010. Clin Infect Dis Off Publ Infect Dis Soc Am. abril de 2014;58(8):1055–61.

Männistö T, Mendola P, Liu D, Leishear K, Sherman S, Laughon SK. Acute air pollution exposure and blood pressure at delivery among women with and without hypertension. Am J Hypertens. janeiro de 2015;28(1):58–72.

Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. Int J Epidemiol. setembro de 1988;17(3):680–5.

McNutt M. Journals unite for reproducibility. Science [Internet]. 7 de novembro de 2014 [citado 24 de janeiro de 2016];346(6210):679–679. Recuperado de: http://www.sciencemag.org/cgi/doi/10.1126/science.aaa1724

Mendis S. The contribution of the Framingham Heart Study to the prevention of cardiovascular disease: a global perspective. Prog Cardiovasc Dis. agosto de 2010;53(1):10–4.

Moffet HH, Adler N, Schillinger D, Ahmed AT, Laraia B, Selby JV, et al. Cohort Profile: The Diabetes Study of Northern California (DISTANCE)--objectives and design of a survey follow-up study of social health disparities in a managed care population. Int J Epidemiol. fevereiro de 2009;38(1):38–47.

Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: Epidemiology in the era of big data. Epidemiol Camb Mass. maio de 2015;26(3):390–4.

Moorman AC, Gordon SC, Rupp LB, Spradling PR, Teshale EH, Lu M, et al. Baseline characteristics and mortality among people in care for chronic viral hepatitis: the chronic hepatitis cohort study. Clin Infect Dis Off Publ Infect Dis Soc Am. janeiro de 2013;56(1):40–50.

Morgan K. Relational Database Design Clearly Explained, Second Edition. 2 edition. New York: Morgan Kaufmann; 2002.

Motheral B, Brooks J, Clark MA, Crown WH, Davey P, Hutchins D, et al. A checklist for retrospective database studies—report of the ISPOR Task Force on Retrospective Databases. Value Health [Internet]. 2003 [citado 1 de abril de 2014];6(2):90–7. Recuperado de: http://onlinelibrary.wiley.com/doi/10.1046/j.1524-4733.2003.00242.x/full

Musser ED, Hawkey E, Kachan-Liu SS, Lees P, Roullet J-B, Goddard K, et al. Shared familial transmission of autism spectrum and attention-deficit/hyperactivity disorders. J Child Psychol Psychiatry. julho de 2014;55(7):819–27.

National eHealth WH. National eHealth strategy toolkit [Internet]. International Telecommunication Union; 2012 [citado 24 de janeiro de 2016]. Recuperado de: http://apps.who.int/iris/handle/10665/75211

Nature. Let's think about cognitive bias. Nature [Internet]. 7 de outubro de 2015 [citado 24 de janeiro de 2016];526(7572):163–163. Recuperado de: http://www.nature.com/doifinder/10.1038/526163a

Nature Editorial. Data-access practices strengthened. Nature [Internet]. 19 de novembro de 2014 [citado 24 de janeiro de 2016];515(7527):312–312. Recuperado de: http://www.nature.com/doifinder/10.1038/515312a

Nature Journals. Journals unite for reproducibility. Nature [Internet]. 5 de novembro de 2014 [citado 24 de janeiro de 2016];515(7525):7–7. Recuperado de: http://www.nature.com/doifinder/10.1038/515007a

Nature News. Challenges in irreproducible research : Nature News & Comment [Internet]. 2015 [citado 13 de novembro de 2015]. Recuperado de: http://www.nature.com/news/reproducibility-1.17552

Nature Policies. Availability of data & materials : authors & referees @ npg [Internet]. 2014 [citado 24 de janeiro de 2016]. Recuperado de: http://www.nature.com/authors/policies/availability.html

Nature Policies. Editorial policies : authors & referees @ npg [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://www.nature.com/authors/policies/index.html

NHI-Associations. Endorsing Associations, Journals, and Societies - rigor-reproducibility-endorsements.pdf [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://www.nih.gov/sites/default/files/research-training/initiatives/reproducibility/rigor-reproducibility-endorsements.pdf

NHI - Guidelines. Principles and Guidelines for Reporting Preclinical Research [Internet]. Natl. Inst. Health NIH. 2015 [citado 24 de janeiro de 2016]. Recuperado de: http://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

Nicolau JC, Baracioli LM, Serrano Jr CV, Giraldez RR, Kalil Filho R, Lima FG, et al. Pacientes com Infarto Agudo do Miocárdio. Arq Bras Cardiol [Internet]. 2008 [citado 15 de novembro de 2015];91(6):347–51. Recuperado de: http://www.scielo.br/pdf/abc/v91n6/a04v91n6.pdf

OMOP. Observational Medical Outcomes Partnership [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://omop.org/

OMOP - CDM. Common Data Model | Observational Medical Outcomes Partnership [Internet]. 2016 [citado 17 de fevereiro de 2016]. Recuperado de: http://omop.org/CDM

Oppenheimer GM. Becoming the Framingham Study 1947–1950. Am J Public Health [Internet]. abril de 2005 [citado 22 de fevereiro de 2016];95(4):602–10. Recuperado de: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1449227/

Oracle SQL. Oracle SQL Developer Downloads [Internet]. 2016 [citado 28 de janeiro de 2016]. Recuperado de: http://www.oracle.com/technetwork/developer-tools/sql-developer/downloads/index.html

Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc JAMIA [Internet]. 2012 [citado 17 de fevereiro de 2016];19(1):54–60. Recuperado de: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3240764/

Ozsu MT, Valduriez P. Principles of Distributed Database Systems. 3rd ed. 2011 edition. New York: Springer; 2011.

Peng RD. Reproducible research and Biostatistics. Biostatistics [Internet]. 1 de julho de 2009 [citado 19 de janeiro de 2016];10(3):405–8. Recuperado de: http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxp014

Peng RD. Reproducible Research in Computational Science. Science [Internet]. 2 de dezembro de 2011 [citado 11 de janeiro de 2016];334(6060):1226–7. Recuperado de: http://www.sciencemag.org/cgi/doi/10.1126/science.1213847

PHARMO. PHARMO Institute [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://www.pharmo.nl/

Pires FA. Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde [Internet] [text]. Universidade de São Paulo; 2011 [citado 24 de janeiro de 2016]. Recuperado de: http://www.teses.usp.br/teses/disponiveis/5/5131/tde-08122011-145701/

Pires FA, Abrahão MT, Rebelo MS, Santos RS, Nobre MC, Gutierrez MA. Ambiente para extração de informações de saúde a partir de bases de dados do SUS. BIS Bol Inst Saúde Impresso [Internet]. abril de 2011 [citado 24 de janeiro de 2016];13(1):39–45. Recuperado de: http://periodicos.ses.sp.bvs.br/scielo.php?script=sci_abstract&pid=S1518-18122011000100007&lng=pt&nrm=iso&tlng=pt

Pires FA, Furuie SS, Gutierrez MA, Tachinardi U. Prontuário Eletrônico: Aspectos Legais e Situação Atual. l. Revista da Sociedade de Cardiologia do Estado de São Paulo, São Paulo, v. 13, p. 730-735; 2003.

Ramakrishnan R, Gehrke J. Database management systems. 2000 [citado 28 de janeiro de 2016]; Recuperado de: http://dspace.utamu.ac.ug:8080/xmlui/bitstream/handle/123456789/85/%5BRam

akrishnan_R.,_Gehrke_J.%5D_Database_Management_S(BookFi.org).pdf?sequence=1&isAllowed=y

RelaX. RelaX - relational algebra calculator [Internet]. [citado 30 de janeiro de 2016]. Recuperado de: http://dbis-ue.uibk.ac.at/relalg

R - Project. R: The R Project for Statistical Computing [Internet]. 2016 [citado 28 de janeiro de 2016]. Recuperado de: https://www.r-project.org/

RStudio. RStudio | Open source and enterprise-ready professional software for R [Internet]. 2016 [citado 28 de janeiro de 2016]. Recuperado de: https://www.rstudio.com/

Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ [Internet]. 13 de janeiro de 1996 [citado 24 de janeiro de 2016];312(7023):71–2. Recuperado de: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2349778/

Santos R da S. Ambiente para extracao de informacoes atraves da mineracao das bases de dados do Sistema Unico de Saúde. A computational framework to extract analytical information through data mining of public health databases [Internet]. 2007 [citado 24 de janeiro de 2016]; Recuperado de: http://repositorio.unifesp.br/handle/11600/23647

Scientific Data. Recommended Repositories : Scientific Data [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://www.nature.com/sdata/data-policies/repositories

Selby JV. Linking automated databases for research in managed care settings. Ann Intern Med. 15 de outubro de 1997;127(8 Pt 2):719–24.

Sicras-Mainar A, Planas-Comes A, Frias-Garrido X, Navarro-Artieda R, de Salas-Cansado M, Rejas-Gutiérrez J. Statins after recent stroke reduces recurrence and improves survival in an aging Mediterranean population without known coronary heart disease. J Clin Pharm Ther [Internet]. Agosto de 2012 [citado 28 de outubro de 2015];37(4):441–7. Recuperado de: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2710.2011.01318.x/abstract

Silva ML. Manual de Certificação para Sistemas de Registro Eletrônico em Saúde (S-RES). 2011 [citado 17 de fevereiro de 2016]; Recuperado de: http://www.sbis.org.br/certificacao/Manual_Certificacao_SBIS_CFM_2011_v4_Consulta_Publica.pdf

Stodden V. The reproducible research movement in statistics. Stat J IAOS J Int Assoc Off Stat [Internet]. 2014 [citado 2 de novembro de 2015];30(2):91–3. Recuperado de: http://www.stanford.edu/~vcs/talks/ISI-Aug302013-STODDEN.pdf

Stodden V, Guo P, Ma Z. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS ONE [Internet].

21 de junho de 2013 [citado 2 de novembro de 2015];8(6):e67111. Recuperado de: http://dx.doi.org/10.1371/journal.pone.0067111

TabNet - Ambulatorial. TabNet Win32 3.0: Produção Ambulatorial do SUS - Brasil - por local de atendimento [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sia/cnv/qabr.def

TabNet - Internações. TabNet Win32 3.0: Procedimentos hospitalares do SUS - por local de internação - Brasil [Internet]. [citado 24 de janeiro de 2016]. Recuperado de: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sih/cnv/qiuf.def

Tess BH, Furuie SS, Castro RCF, Barreto M do CC, Nobre MRC. Assessing the scientific research productivity of a Brazilian healthcare institution: a case study at the heart institute of São Paulo, Brazil. Clinics [Internet]. junho de 2009 [citado 24 de janeiro de 2016];64(6):571–6. Recuperado de: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-59322009000600013&lng=en&nrm=iso&tlng=en

Thomas SL, Minassian C, Ganesan V, Langan SM, Smeeth L. Chickenpox and risk of stroke: a self-controlled case series analysis. Clin Infect Dis Off Publ Infect Dis Soc Am. janeiro de 2014;58(1):61–8.

TISS. TISS - Troca de Informação de Saúde Suplementar - ANS - Agência Nacional de Saúde Suplementar [Internet]. 2016 [citado 24 de janeiro de 2016]. Recuperado de: http://www.ans.gov.br/prestadores/tiss-troca-de-informacao-de-saude-suplementar

UAB. UAB-2013-Informática-em-Saúde-Padroes-em-IS.pdf [Internet]. 2013 [citado 20 de janeiro de 2016]. Recuperado de: http://www.cee78is.org.br/Downloads/UAB-2013-Inform%C3%A1tica-em-Sa%C3%BAde-Padroes-em-IS.pdf

Ullman J, Molina HG, Widom J. The_Complete_Book.pdf [Internet]. Pearson Prentice Hall; 2009. Recuperado de: http://people.inf.elte.hu/nikovits/DB2/Ullman_The_Complete_Book.pdf

Vigen R. Association of Testosterone Therapy With Mortality, Myocardial Infarction, and Stroke in Men With Low Testosterone Levels. JAMA [Internet]. 6 de novembro de 2013 [citado 15 de fevereiro de 2016];310(17):1829. Recuperado de: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2013.280386

Voorhis D. Codd's Twelve Rules [Internet]. Dep. Comput. Math. 2015 [citado 30 de janeiro de 2016]. Recuperado de: http://computing.derby.ac.uk/c/codds-twelve-rules/

Wager KA, Lee FW, Glaser JP, Wager KA. Health care information systems: a practical approach for health care management. 2nd ed. San Francisco, CA: Jossey-Bass; 2009.

Wijeysundera HC, Bennell MC, Qiu F, Ko DT, Tu JV, Wijeysundera DN, et al. Comparative-Effectiveness of Revascularization Versus Routine Medical

Therapy for Stable Ischemic Heart Disease: A Population-Based Study. J Gen Intern Med [Internet]. 8 de março de 2014 [citado 30 de março de 2014]; Recuperado de: http://link.springer.com/10.1007/s11606-014-2813-1

Young SS, Karr A. Deming, data and observational studies: A process out of control and needing fixing. Significance [Internet]. setembro de 2011 [citado 24 de janeiro de 2016];8(3):116–20. Recuperado de: http://doi.wiley.com/10.1111/j.1740-9713.2011.00506.x

Yu WD, Kollipara M, Penmetsa R, Elliadka S. A distributed storage solution for cloud based e-Healthcare Information System. IEEE; 2013 [citado 25 de janeiro de 2016]. p. 476–80. Recuperado de: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6720723

Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf. fevereiro de 2013;36(2):119–34.

Apêndice A

PARECER CONSUBSTANCIADO DA COMISSÃO DE ÉTICA

3721/11/139

Hospital das Clínicas da FMUSP
Comissão de Ética para Análise de Projetos de Pesquisa - CAPPesq

PROJETO DE PESQUISA

Título: SEGUIMENTO EVOLUTIVO DE PACIENTES COM DOENÇA CARDIOVASCULAR ATEROSCLERÓTICA E O USO DE ESTATINA: ESTUDO DE COORTE RETROSPECTIVO EM BASE DE REGISTRO ASSISTENCIAL.
Pesquisador Responsável: Moacyr Roberto Cuce Nobre **Versão: 1**
Pesquisador Executante: Maria Tereza Fernandes Abrahão **CAAE:** 00594512.5.0000.0068
Orientador: Marco Antonio Gutierrez
Finalidade Acadêmica: Doutorado
Instituição: HCFMUSP
Departamento: COMISSÃO CIENTÍFICA DO INCOR

PARECER CONSUBSTANCIADO DO CEP

Número do Parecer: 6426

Data da Relatoria: 07.03.12

Apresentação do Projeto: Projeto de caráter retrospectivo que deverá avaliar os registros dos Bancos de Dados de pacientes tendidos no incor com o diagnóstico de doença cardiovascular aterosclerótica, com ênfase na análise da efetividade do emprego de estatinas na prevenção secundária de novos eventos na prática clínica.
Objetivo da Pesquisa: Descrever o seguimento evolutivo de pacientes com doença cardiovascular aterosclerótica, atendidos em um hospital universitário, durante período de oito anos e avaliar a evolução desses pacientes em função do uso de estatinas.
Comparar os resultados de efetividade obtidos no presente estudo com os achados de eficácia e efetividade obtidos por revisão sistemática de ensaios clínicos e estudos de coorte realizada em outros estudos
Avaliação dos Riscos e Benefícios: O estudo é de caráter retrospectivo, não resultando em qualquer risco ou benefício imediato aos pacientes estudados. Pode ser importante para a atualização de diretrizes para a prática clínica e gestão em saúde a partir dos resultados observados.
Comentários e Considerações sobre a Pesquisa: Pesquisa de coorte retrospectiva a ser desenvolvida em Banco de Dados bem estruturado, com número relevante de pacientes para o objetivo proposto.
Considerações sobre os Termos de apresentação obrigatória:
Recomendações: Recomendada a sua aprovação em seu formato atual.
Conclusões ou Pendências e Lista de Inadequações: Recomendada a sua aprovação em seu formato atual.
Necessita de Apreciação da CONEP: NÃO
Situação: **APROVADO**
Considerações Finais a critério do CEP: Protocolo não apresenta pendências.

São Paulo, 09 de Março de 2012

COMISSÃO CIENTIFICA
RECEBIDO
02/04/12

Prof. Dr. Luiz Eugênio Garcez Leme
Coordenador
Comissão de Ética para Análise de
Projetos de Pesquisa-CAPPesq

Rua Dr. Ovídio Pires de Campos, 225 - Prédio da Administração - 5º andar CEP 05403-010 - São Paulo - SP.
55 11 2661-6442 ramais: 16, 17, 18 e 20 | cappesq@hcnet.usp.br

Apêndice B

PUBLICAÇÕES DECORRENTES DESSE PROJETO

1. Abrahão, MT; Nobre, MR; Gutierrez, MA. Estatística descritiva de uma população de pacientes atendidos no InCor com doença cardiovascular. In: XIII Congresso Brasileiro em Informática em Saúde – Anais do CBIS, 2012. v. 1 – Curitiba, PR, Brasil. 19 a 23 Nov. 2012.

2. Abrahão, MT; Nobre, MR; Gutierrez, MA. *Descriptive Statistics of 65000 Patients Treated for Myocardial Ischemia: data from routine electronic health records. Evidence Live 2013, Oxford UK*.

3. Abrahão, MT; Nobre, MR; Gutierrez, MA. *The effectiveness of Statins in the Treatment of Cardiovascular Disease: Transverse study with paired groups from an Electronic Patient Record*. ISPOR 16th Annual European Congress, Dublin, 11/2013, VALUE IN HEALT H 1 6 (2 0 1 3) A323–A636. Volume 16, Issue 7, Page A518. doi:10.1016/j.jval.2013.08.1233. Journal ISSN: 1098-3015.

4. Abrahão, MT; Nobre, MR; Gutierrez, MA. Estudos retrospectivos em base de dados assistencial em um hospital de referência em cardiologia. In: XIV Congresso Brasileiro em Informática em Saúde – Anais do CBIS, 2014. Santos, SP, Brasil. 7 a 10 Dez. 2014.

5. Luque A, Nobre MR, Abrahão MT. *Cost-utility of statin in secondary prevention: a propensity score method of administrative database*. ISPOR 18th Annual European Congress. 7 a 11 November 2015. MiCo - Milano Congressi. Milan, Italy. Value Health. 2015 Nov;18(7):A395-6. doi: 10.1016/j.jval.2015.09.894. Epub 2015 Oct 20. No abstract available. PMID: 26532229 [PubMed - in process].