Universidade Estadual de Londrina

RENATO TOSHIO KUROE

APLICAÇÃO DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MEDIDAS DO CORAÇÃO DE PACIENTES BRASILEIROS

LONDRINA
2020

RENATO TOSHIO KUROE

APLICAÇÃO DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MEDIDAS DO CORAÇÃO DE PACIENTES BRASILEIROS

LONDRINA
2020

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

394 Kuroe, Renato Toshio.
APLICAÇÃO DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MEDIDAS DO CORAÇÃO DE PACIENTES BRASILEIROS / Renato Toshio Kuroe. - Londrina, 2020.
69 f. : il.

Orientador: Jacques Duílio Brancher.
Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, 2020.
Inclui bibliografia.

1. Mineração de dados - Tese. 2. Valores de referências brasileiras para laudo de ecocardiograma - Tese. I. Brancher, Jacques Duílio. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU 519

RENATO TOSHIO KUROE

APLICAÇÃO DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MEDIDAS DO CORAÇÃO DE PACIENTES BRASILEIROS

BANCA EXAMINADORA

Orientador: Prof. Dr. Jacques Duílio Brancher
Universidade Estadual de Londrina

Prof. Dr. Rodolfo Miranda de Barros
Universidade Estadual de Londrina - UEL

Prof. Dr. Renne Rodrigues
Departamento de Saúde Coletiva (CCS)

Londrina, 23 de Julho de 2020.

Este trabalho é dedicado aos nerds que todos os dias constroem tecnologias para facilitar a vida das pessoas.

AGRADECIMENTOS

Os agradecimentos principais são direcionados aos Doutores Jacques Brancher, Willyan Nazima e Fabrício Furtado, que orientaram, apoiaram, autorizaram a utilização do sistema para obtenção de dados valiosos coletados durante anos, e passaram o conhecimento necessário para que esse trabalho fosse idealizado.

Também aos colegas do Programa de Mestrado que se tornaram amigos, e apoiaram imensamente durante toda a trajetória do curso. Agradecimentos especiais são direcionados aos médicos, que por meio do sistema, contribuíram com os dados dos pacientes de forma massiva e constante, viabilizando uma análise rica e abrangente.

KUROE, R. T.. **APLICAÇÃO DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MEDIDAS DO CORAÇÃO DE PACIENTES BRASILEIROS**. 2020. 70f. – Universidade Estadual de Londrina, Londrina, 2020.

RESUMO

Atualmente os laudos dos exames de ecocardiograma realizados no Brasil, utilizam intervalos de medidas do coração vindas de referências americanas. Ainda que sejam parâmetros de pessoas que possuem características físicas similares, as origens étnicas, diferenças climáticas em que vivem e costumes são diferentes. Isso pode tornar o laudo final impreciso. Um milímetro a mais ou a menos em qualquer parte coração pode ser um valor normal para o americano, porém para o brasileiro pode ser indício de algum problema cardíaco. Utilizando a base de dados de um sistema que centraliza as informações, abrange várias regiões do Brasil, e tem como objetivo gerar laudos de ecocardiograma, foram utilizadas técnicas de mineração de dados para obtenção de dados de pacientes brasileiros. Todo o banco de dados é relacional, centralizado e padronizado, permitindo gerar relatórios precisos de cada paciente, em cada clínica de diferentes estados, possibilitando a diferenciação e comparação entre pacientes de regiões diferentes do Brasil. Com isso, por meio da base de dados com mais de 20 mil laudos médicos, utilizando-se da mineração de dados por meio da metodologia KDD, o trabalho apresenta valores de referência para as medidas do coração do brasileiro. Por meio desse estudo, concluiu-se ser possível obter maior assertividade no diagnóstico de situações clínicas de partes do coração, ao menos em relação aos valores de referências brasileiras.

Palavras-chave: ecocardiograma, laudo, mineração de dados, valores de referência, coração.

KUROE, R. T.. . 2020. 70p. – State University of Londrina, Londrina, 2020.

ABSTRACT

Nowadays, the reports of echocardiogram exams performed in Brazil, use intervals of heart measurements from American references. Although they are parameters of people who have similar physical characteristics, ethnic origins, climatic differences in which they live and customs are different. This can render the final report inaccurate. A millimeter more or less in any part of the heart can be a normal value for the American, but for the Brazilian it can be an indication of a heart problem. Using the database of a system that centralizes the information, it covers several regions of Brazil, and aims to generate echocardiogram reports, data mining techniques were used to obtain data from Brazilian patients. The entire database is relational, centralized and standardized, allowing accurate reports to be generated for each patient, in each clinic in different states, enabling differentiation and comparison between patients from different regions of Brazil. With this, through the database with more than 20 thousand medical reports, using data mining through the KDD methodology, the work presents reference values for the measurements of the Brazilian heart. With this study, it was concluded that it is possible to obtain greater assertiveness in the diagnosis of clinical situations of parts of the heart, at least in relation to the values of Brazilian references.

Keywords: echocardiogram, report, data mining, benchmarks, heart.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EMR	Electronic Medical Record
ASE	American Society of Ecocardiography
PHT	Paratormonio
KDD	Knowledge-discovery in databases
EROA	Effective regurgitant orifice area
API	Application Programming Interface
FE	Fração de ejeção
ASC	Área de superfície corpórea
ELSA	Estudo Longitudinal de Saúde do Adulto
PCM	Paracoccidioidomicose
WEKA	Waikato Environment for Knowledge Analysis
VE	Ventrículo esquerdo
SQL	Structured Query Language
CSV	Comma-separated values

SUMÁRIO

1 INTRODUÇÃO

O uso da mineração de dados tem como um dos objetivos, ser uma solução para a recuperação da informação intangível. Com isso é possível reunir esses dados com técnicas de agrupamento para tratamento, análise e tomada de decisões, geralmente de grandes empresas ou pesquisas. Atualmente o sistema médico sofre com a falta de centralização de dados, tornando inviável muitas pesquisas que poderiam melhorar o diagnóstico médico. Isso acaba gerando informações incompatíveis entre si, e dificultando a obtenção de dados a nível estadual ou nacional [1].

A evolução da tecnologia trouxe a reflexão sobre o modelo de trabalho analógico e sua possível melhora [2], mesmo em uma época em que muitos médicos ainda utilizam métodos tradicionais para ações de rotina, como é o caso da elaboração manual de laudo ecocardiográfico [3, 4, 5]. Apesar das resistências, seja por desconfiança da tecnologia ou por medo de mudança por se tratar de vidas, estamos em uma era de migração do papel para o digital, inclusive no setor médico. As tecnologias de informação em saúde têm o objetivo de fornecer melhores serviços aos hospitais e organizações relacionadas.

A natureza deste setor, que é profundamente influenciado pela economia, fatores sociais, políticos e tecnologia mudaram com o tempo. Com a ampliação da atuação de tecnologia na área médica, os sistemas computacionais aplicados aos cuidados de saúde foram desenvolvidos com capacidade e interatividade exponencial, e seu custo reduzido em proporção inversa. O ambiente digital no setor médico vem oferecendo muitas vantagens, mais especificamente, na questão do Relatório Médico Eletrônico [6].

O RME constitui o núcleo de um sistema de saúde computacional. O armazenamento eletrônico de informações desenvolveu um potencial para ferramentas computacionais, para auxiliar a qualidade dos cuidados médicos de forma significativa, aumentando a eficiência da prática médica [7]. Consequentemente e acompanhando a evolução, sistemas baseados em computador podem gerar relatórios dos exames e notificar imediatamente os médicos quando os resultados dos testes estiverem prontos. Assim, podem aumentar o tempo disponível para planejar estratégias em um diagnóstico que indique as condições, além de outros benefícios [6, 8].

Como citado acima, muitos médicos ainda usam métodos manuais para preparar relatórios, nos quais um dos principais problemas é a descentralização de dados, em que muitos são elaborados em sistemas incapazes de armazenar informação. E, se inserida no sistema, existe também a falta de padronização [3, 4]. Paralelamente, acompanhando o avanço da digitalização, e enxergando uma possibilidade de centralizar dados, o Sistema EcoCloud surgiu como um software especializado que ajuda os médicos por meio da ge-

ração automática de diagnóstico cardiológico por meio de ecocardiograma. Permite o uso de plataformas móveis, proporcionando mobilidade e agilidade (informações detalhadas no anexo deste documento).

Compreender um exame de ecocardiograma e tirar suas conclusões podem ser tarefas complexas e demoradas, contendo partes específicas que podem ser automatizadas. Assim, este trabalho utiliza a base de dados do Sistema EcoCloud, que possui o banco em nuvem, com informações centralizadas. Isso permite que sejam extraídos as informações necessárias para uma pesquisa por meio de mineração de dados. O ecocardiograma é um teste que utiliza ondas sonoras de alta frequência para avaliar o coração, que também é chamado de ecocardiografia ou ecografia cardíaca diagnóstica [9]. Este teste é usado para examinar a estrutura e o desempenho cardíaco do coração.

Este projeto apresenta por meio de mineração de dados, métricas obtidas e padrões do coração do brasileiro, para que sejam usados parâmetros mais próximos em relação com as que são usadas atualmente no Brasil (padrões americanos). E com isso obter laudos mais precisos. Esses dados são obtidos por meio de técnicas de mineração de dados, divididos em etapas de extração, limpeza, seleção e por fim a mineração em si. Após os dados coletados, é feita uma avaliação e visualização dos resultados, gerando também um comparativo entre os valores utilizados atualmente e os novos valores de referência obtidos.

1.1 Contextualização e problemática

Com a transição demográfica ocorrida no Brasil desde a década de 1950, as doenças cardiovasculares se tornaram a principal causa de óbitos no país, o que, com o envelhecimento da população, aumentam a necessidade de cardiologistas e profissionais especializados na fisiologia cardíaca. A quantificação do risco elevado para grupos de indivíduos expostos a partir de uma série de pares risco-resultado é importante para informar a tomada de decisões sobre a saúde individual. [10] O exame de ecocardiograma é uma ferramenta muito eficaz na entrega de informações do coração.

Em 2008, as doenças cardiovasculares foram responsáveis por dois terços das mortes no mundo, entre as 36 milhões registradas. [11] Segundo Maria de Fátima Marinho de Souza et al., em 1990, fatores como dieta inadequada, tabagismo manteve no topo da lista de problemas de saúde do SUS (Sistema Único de Saúde), contribuindo para que em 2015, problemas de pressão arterial subissem do terceiro para o segundo lugar, tanto em homem como em mulheres. [12]

Em 1990, a dieta inadequada, o tabagismo, a pressão arterial sistólica elevada e a desnutrição materno-infantil foram os principais FRs para DALYs para homens e mulheres. A dieta inadequada se manteve no topo da lista em 1990 e, em 2015, e a

pressão arterial sistólica subiu do terceiro para o segundo lugar, tanto em homens como em mulheres

Métodos e técnicas por meio de imagem têm revolucionado a medicina cardiovascular se tratando de prevenção de doenças, diagnóstico e tratamento das doenças cardíacas. Por conta de seus resultados eficientes, as solicitações de exames de imagem na área de cardiologia têm crescido exponencialmente. [13] Em paralelo, o número de solicitações de exames de ecocardiografia também cresce exponencialmente como resposta ao aumento do número de exames diagnósticos na área de cardiologia. [14]

Existem pelo menos 429 serviços de ecocardiografia no Brasil, dentre esses, 55 possuíam estágio de formação de ecocardiografistas, e 141 atendiam na área pediátrica. É notório o número crescente de ecocardiografistas no Brasil nos últimos 20 anos, e consequentemente o número de centros formadores [15]. Com a necessidade também crescente de profissionais da área, existe cada vez mais a preocupação por sistemas com banco de dados unificados, e de automatização de processos, por conta da crescente preocupação com a qualidade de vida. Todos os dias, pacientes do Brasil todo realizam exames de ecocardiograma.

Com isso são gerados muitos dados de medidas que, se reunidos, podem se tornar um valioso banco de dados para pesquisas e conhecimento do coração brasileiro, trazendo diagnósticos muito mais precisos, caso sejam usados parâmetros brasileiros. No cenário atual, são usados parâmetros americanos para o processo de laudo do paciente que tem origens étnicas, costumes, hábitos alimentares e outras características exclusivamente brasileiras.

O exame ecocardiográfico ajuda a descobrir o tamanho e a forma do coração, espessura e movimento das paredes, força de bombeamento de sangue, avaliação de válvulas cardíacas e outros fatores. Por exemplo, um coração com um tipo de insuficiência cardíaca (coração fraco e grande), anomalias congênitas e ataque cardíaco são detectados por meio do ecocardiograma. Portanto, a elaboração eletrônica do relatório facilita a adaptação das recomendações acima mencionadas.

O ecocardiograma também é usado rotineiramente para avaliar bebês, crianças e jovens adultos com suspeita de doença cardíaca. Para detectar alterações na estrutura e funcionamento do coração produzido pela doença, é importante determinar com precisão o efeito do crescimento e desenvolvimento normais nas medidas ecocardiográficas do tamanho, espessura e função da câmara. [16] Uma insuficiência cardíaca (coração fraco e grande), anomalias congênitas, sopro ou infarto, são acompanhadas por meio do ecocardiograma. Avaliação de sintomas, como falta de ar, cansaço excessivo, inchaço e palpitações são realizados por meio do ecocardiograma.

As metodologias de realização de laudos feitas pelos médicos são baseadas em es-

tudos americanos, em que as amostras são extraídas de pacientes americanos. Atualmente existem poucos estudos brasileiros amplos e de impacto feitos e aplicados no Brasil com amostras de pacientes brasileiros, justamente pela falta de quantidade necessária de laudos e falta de padronização de informações, pois existem vários bancos de dados de diferentes sistemas médicos.

O problema maior talvez seja a descentralização desses dados, em que muitos são gerados em sistemas incapazes de guardar e gerar relatórios, e outros sequer são inseridos em algum sistema. A falta de padronização dessas informações é outro fator que impossibilita a mineração desses dados. Com base nessas conclusões, viu-se o desafio de coletar esses dados, de maneira anônima, mas certificando-se que os dados sejam de pacientes brasileiros.

O Brasil, por meio da Sociedade Brasileira de Cardiologia, utiliza os valores de referência do coração do americano para elaboração de laudos. São os valores que a Sociedade Americana de Ecocardiografia determina. As medidas disponibilizam resultados satisfatórios nos laudos, no entanto, não são usadas as do coração brasileiro, que possuem origens étnicas diferentes do americano, diversidade de raças e miscigenação muito maior, costumes e hábitos alimentares distintos. [17] Além disso a estatura média do americano é ligeiramente maior que a do brasileiro [18], contribuindo para que o coração, e consequentemente os parâmetros sejam diferentes. [19]

Também existe o fator climático, onde, por meio deste estudo, foi evidenciado que condições climáticas podem influenciar no valor da fração de ejeção, talvez até em outros parâmetros, por mais que o corpo humano tenha a habilidade da homeostase. [20] Valores de referência americanos foram extraídos de pacientes que vivem eu um país de clima temperado, enquanto o brasileiro vive em um país tropical, e além disso, variando de climas frios e muito quentes.

Dada a problemática apresentada acima, nos próximos capítulos serão apresentados os processos para obtenção dos valores de referência do coração do brasileiro, com o objetivo de obter maior precisão no diagnóstico médico, se tratando de laudo de ecocardiograma em seus parâmetros de medidas do coração, por meio de técnicas de mineração de dados, e utilizando a metodologia KDD (*Knowledge Discovery in Databases*) para garantir resultados confiáveis, utilizando amostras apenas de pacientes saudáveis.

2 MINERAÇÃO DE DADOS

A mineração de dados lida basicamente com dados. Uma enorme quantidade de informações é processada para extrair padrões desconhecidos, mas que serão úteis se agrupadas corretamente. A palavra "dados"é o plural latino de *"datum"*, proveniente do verbo "ousar = dar"[21, 22]. O termo "mineração de dados"ou *"data mine"* tem sido usado principalmente por estatísticos, analistas de dados e comunidades de Sistemas de Informações Gerenciais (MIS), é uma forma de descobrir padrões em dados e apresentá-los de forma compreensível e útil [23].

Com isso, ganhou popularidade no campo de banco de dados. Os primeiros usos do termo vêm de estatísticas e, na maioria dos casos, o uso foi associado a conotações negativas da exploração às cegas de dados, sem hipóteses a priori para verificar e comprovar. No entanto, exceções notáveis podem ser encontradas. Em 1978, [24] foi usado o termo em uma demonstração para apresentar como a regressão linear generalizada pode ser usada para resolver problemas que são muito difíceis para os seres humanos e para técnicas estatísticas tradicionais da época resolver. O termo KDD foi cunhado no primeiro workshop do KDD em 1989 [25] para enfatizar que "o conhecimento é o produto final de uma descoberta baseada em dados".

Os dados “brutos”, que são obtidos diretamente por vários processos de aquisição, referem-se a números, figuras, imagens, sons, programas de computador (vistos como coleções de dados interpretados como instruções), etc. Esses dados, uma vez coletados, são então processados, obtendo assim informações que são armazenadas. Após isso, são usadas ou transmitidas posteriormente em um processo do tipo *loop*, ou seja, com a possibilidade de que alguns dos dados processados representem dados ’brutos’ para processos subsequentes [21].

Esses dados "brutos", no contexto da computação, podem ser provenientes de qualquer banco de dados de um sistema que mantém informações com relacionamentos entre eles, de maneira consistente, sem ambiguidades e com um mínimo de período sendo populado. O sistema usado para a apresentação deste trabalho atende a essas características, pois contribui com uma base consistente, com mais de 5 anos de utilização, sendo populado com frequência e a partir de diversas regiões do Brasil. Isso possibilita o trabalho com o banco de dados. Todas as variáveis incluídas na mineração realizada, se mantém desde a publicação do sistema, tornando os dados consistentes.

2.1 KDD - *Knowledge Discovery in Databases*

O KDD se refere ao processo geral de descobrir o conhecimento dos dados que serão relevantes na pesquisa, enquanto a mineração de dados se refere a uma etapa específica desse processo. A mineração de dados é a aplicação de algoritmos específicos para extrair padrões de dados. As etapas adicionais no processo KDD, como preparação de dados, seleção de dados, limpeza de dados, incorporação de conhecimento prévio adequado e interpretação adequada dos resultados da mineração, são essenciais para garantir que o conhecimento seja extraído dos dados.

A aplicação cega de métodos de mineração de dados, isto é, sem o devido processo sugerido pelo KDD, pode ser uma atividade perigosa que facilmente leva à padrões sem sentido. Na Figura 1, é apresentada uma visão geral do que é o processo KDD [26]. Observe que, no processo normalmente se repete várias vezes em relação às etapas anteriores e é bastante confuso, com muita experimentação. Por exemplo, pode-se selecionar, amostrar, limpar e reduzir dados apenas para descobrir, após a mineração, que uma ou várias das etapas anteriores precisam ser refeitas. Foram omitidas setas que ilustrariam essas iterações em potencial, para manter a figura simples.

O processo KDD consiste nas seguintes etapas:

1. Limpeza de dados: etapa de pré-processamento, em que são eliminados *noises* (dados considerados inválidos), *outliers*, informações inconsistentes.

2. Integração de dados: etapa opcional em que diferentes fontes de dados podem ser unificados, tornando uma única base. No caso do sistema utilizado neste estudo, existe apenas um banco centralizado, em que foram trabalhadas as relações entre tabelas e regras de inserção, garantindo integridade de informações importantes.

3. Seleção: etapa de seleção de atributos que irão compor a análise.

4. Transformação de dados: etapa em que os dados são transformados em formato apropriado para aplicação de algoritmos de mineração. Como exemplo, por meio de operações de agregação.

5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse.

6. Avaliação ou Pós-processamento: etapa em que são identificados os padrões interessantes, de acordo com algum critério.

7. Visualização de resultados: etapa onde são utilizadas técnicas de representação de conhecimento, a fim de apresentar ao usuário o conhecimento minerado.

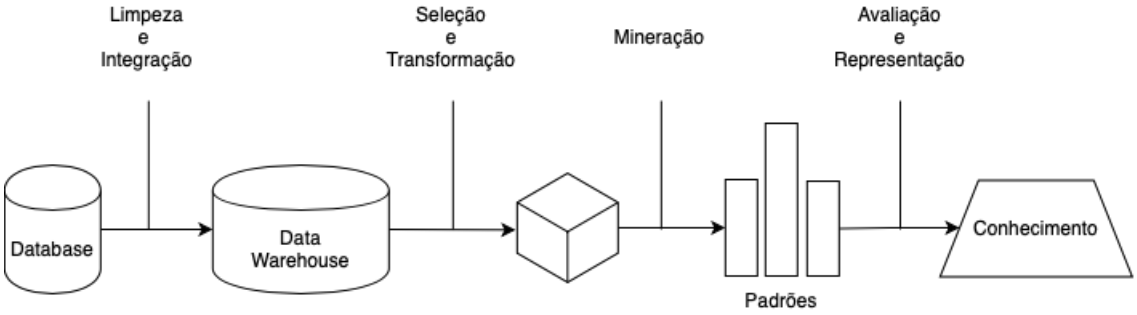

Figura 1 – Visão geral das etapas do processo KDD.

2.2 Técnicas de mineração de dados

Podemos dividir em 5 técnicas de mineração de dados, conforme listadas abaixo. Embora algumas dessas técnicas tenham sido historicamente definidas para trabalhar com dados residentes na memória (normalmente lidos em arquivos simples), algumas dessas técnicas estão começando a ser dimensionadas para operar em bancos de dados [26].

2.2.1 Modelagem Preditiva

O objetivo é prever alguns campos em um banco de dados com base em outros campos. Se o campo previsto é uma variável numérica (contínua) (como uma medição física de, por exemplo, altura), o problema de previsão é um problema de regressão. Se o campo for categórico, será um problema de classificação. Existe uma grande variedade de técnicas para classificação e regressão. O problema em geral é lançado, como a determinação do valor mais provável da variável prevista, considerando os outros campos (entradas), os dados de treinamento (nos quais a variável alvo é fornecida para cada observação) e um conjunto de suposições que representam conhecimento prévio do problema.

2.2.2 *Clustering*

O *Clustering* que também é conhecido como segmentação ou análise de agrupamento de dados, possui o armazenamento em cluster que não especifica os campos a serem previstos, mas tem como objetivo separar os itens de dados em subconjuntos que são semelhantes entre si. Como, diferentemente da classificação, não sabemos o número de *"clusters"* desejados, os algoritmos de cluster geralmente empregam uma pesquisa em dois estágios: um loop externo sobre os números possíveis do cluster e um loop interno para ajustar o melhor cluster possível para um determinado número de *clusters*.

2.2.3 Sumarização de dados

Por vezes, o objetivo é simplesmente extrair padrões compactos que descrevem subconjuntos de dados. Existem dois métodos que representam a obtenção de fatias hori-

zontais (casos) ou verticais (campos) dos dados. No primeiro, são produzidos resumos de subconjuntos: estatísticas suficientes ou condições lógicas válidas para subconjuntos. No segundo caso, são previstas relações entre campos. Essa classe de métodos se distingue das anteriores, pois, em vez de prever um campo especificado (por exemplo, classificação) ou agrupar casos. O objetivo é encontrar relações entre os campos.

2.2.4 Modelagem de Dependências

A compreensão dos dados geralmente é obtida com a derivação de estrutura causal dentro dos dados. Modelos de causalidade podem ser probabilísticas (como derivar alguma afirmação sobre a distribuição de probabilidade que governa os dados) ou eles podem ser determinísticos como na derivação funcional de dependências entre campos nos dados.

2.2.5 Detecção de alterações e desvios

Esses métodos são responsáveis pela informação da sequência, seja em séries temporais ou em alguma outra ordem (por exemplo, sequenciamento de proteínas no mapeamento do genoma). O caráter distintivo dessa classe de métodos é que a ordenação de observações é importante e deve ser considerada.

3 MINERAÇÃO DE DADOS NA ÁREA MÉDICA

A mineração de dados aplicada à dados médicos possui dificuldades particulares. Como mencionado neste documento, a heterogeneidade de bases de informações e a complexidade estrutural dessas bases, tanto pela quantidade de dados e como elas devem ser interpretadas, quanto pela dificuldade em reunir equipes de profissionais médicos qualificados e dedicados para acompanhamento junto a cientistas de dados, se torna um desafio. As regras de negócio da área médica são muito específicas e exigem um cuidado especial por se tratar de vidas humanas.

Os cientistas de dados, bem como equipes de desenvolvimento obviamente não possuem tal conhecimento. Em paralelo, especialistas médicos têm sua vida atípica, ocupada por plantões e longas jornadas de trabalho. Isso impossibilita um acompanhamento adequado junto a desenvolvimento, tampouco sua presença em reuniões de alinhamento que as metologias ágeis de desenvolvimento propõem, em que são necessárias reuniões a cada entrega parcial do projeto. Os projetos acabam sendo inacabados, com erros ou um alto custo da equipe desenvolvedora, por conta de impedimentos e retrabalhos.

Também poucas empresas conseguem recrutar médicos para se dedicarem totalmente ao projeto. Os custos para manter um profissional médico é alto, e eles têm interesse maior em seguir carreira atuando direto em sua área, em que realmente pode adquirir experiência. Os protocolos de exames para diagnósticos são, em geral, complexos e possuem vários atributos diferentes. Exames e testes são pedidos de acordo com a experiência pessoal do médico e disponibilidade de recursos. Muitas vezes, os pacientes não realizam os procedimentos de exames requisitados e deixam lacunas nos prontuários.

As bases de dados de doenças provém de diversas fontes diferentes, como entrevista com o paciente, testes laboratoriais, resultados de equipamentos e exames. Isso tende a produzir dados muito variados e difíceis de serem analisados, demandando o uso de diferentes técnicas e ferramentas necessárias para serem exploradas de maneira eficiente. Existem ainda envolvidas restrições éticas, legais e sociais relativas à privacidade e a validação clínica [27].

Outra preocupação relevante em que a Mineração de dados e outras tecnologias poderiam ajudar a amenizar, são os custos dos cuidados com a saúde. Segundo um estudo de Groves et al. [28] os gastos estão em um escalonamento de 17,6% do PIB dos EUA. É uma tendência observável também nos países em desenvolvimento, conforme relatado pelas Nações Unidas. [29] De acordo com a Organização Mundial da Saúde, os custos são significativamente afetados por gastos ineficazes resultantes de decisões mal informadas, mau gerenciamento e oportunidades perdidas de prevenção de doença [30]. Além do

custo, e apesar da complexidade citada nos parágrafos acima, a demanda por diagnóstico auxiliado por computador intensificou-se notavelmente por três razões [31]:

1. Desafios crescentes na gestão da informação e do conhecimento na prática clínica;
2. Pressão para adotar registros médicos eletrônicos, conforme observado na lei de tecnologia da informação em saúde dos Estados Unidos (HITECH) [32];
3. Fornecer serviços de saúde personalizados.

No Brasil o registro médico eletrônico também está sendo adotado [33], ainda que não seja uma obrigatoriedade, mas os avanços tecnológicos acabam levando todas as áreas para esse caminho, inclusive a médica. Isso ameniza um pouco a dificuldade na obtenção de big datas, que formam a base para uma uma mineração de dados precisa, e aprendizado de máquina para futuras pesquisas relacionadas à prevenções de doenças.

3.1 Desafios para mineração de dados na área

Durante os anos 90 e início dos anos 2000, a mineração de dados foi um assunto de grande interesse para os pesquisadores e cientistas de dados da área de saúde, pois a mineração de dados mostrou alguma promessa no uso de suas técnicas preditivas para ajudar a modelar o sistema de saúde e melhorar a prestação de serviços [27, 34]. No entanto, logo foi descoberto que a mineração de dados de assistência médica apresentava muitos desafios relacionados à veracidade dos dados e limitações da modelagem preditiva. Isso levava a falhas nos projetos de mineração de dados [34], desencorajando pesquisadores e cientistas de dados a defenderem seus trabalhos.

Porém com a entrada do conceito de big data de maneira forte no anos 2010, a mineração de dados voltou a ser um assunto amplamente comentado. Até então, muito foi escrito sobre os impactos positivos da mineração de dados nas práticas de saúde, relacionados a questões de melhores práticas, detecção de fraudes, gerenciamento de doenças crônicas e tomada de decisões em saúde em geral. [34] Os modelos preditivos voltaram ser utilizados, dessa vez com técnicas de análises mais avançadas.

Eles são capazes de fornecer cálculos e algoritmos para avaliações mais complexas de dados, permitindo verificar de maneira assertiva o que está acontecendo, utilizando maior quantidade de dados, análises e raciocínio sistemático graças ao maior acesso às informações por meio da tecnologia. [35]

A dificuldade de minerar dados médicos pode ser dividida em 4 esferas [36]:

1. Heterogeneidade dos dados;

2. Questões éticas, legais e sociais;

3. Filosofia estatística;

4. Status especial da medicina.

3.1.1 Heterogeneidade dos dados

Os dados iniciais para amostras têm como origem várias fontes, como exames clínicos e laboratoriais, de imagens, histórico médico, muitas vezes apenas obtidos por meio de conversa entre médico e paciente, e transcritos em documentos de texto. Todos esses fatores influenciam diretamente no diagnóstico, prognóstico e tratamento do paciente. Isso evidencia a grande heterogeneidade dos dados, além de serem volumosos [36]. Entrevistas, prontuários e laudos médicos são particularmente desafiadores devido às diferentes maneiras de se referir a um mesmo achado clínico, ou mesmo a diferentes interpretações por parte dos médicos que escrevem [27], o que pode tornar os resultados insatisfatórios.

A interpretação de imagens, sinais ou qualquer outro dado clínico pelo médico, é outro ponto de dificuldade, pois é feito por meio de texto livre, impossibilitando gerar padrões e relacionamento entre informações. Eles não apenas usam nomes diferentes (sinônimos) para descrever a mesma doença, mas também tornam a tarefa ainda mais assustadora, usando diferentes construções gramaticais para descrever os relacionamentos entre as entidades médicas [36]. Quando falamos ainda de um país como o Brasil, em que a gramática diferencia muito de acordo com a região, o trabalho se torna ainda mais difícil.

Foi sugerida uma forma de resolver esses problemas, em que a interpretação do computador pode conter parte da solução para processar a interpretação do médico [37, 38, 39]. Os princípios da tradução por computador podem ser resumidos da seguinte forma [40]:

1. A tradução automática é tipicamente composta por três etapas: análise de uma sentença no idioma de origem, transferência de um idioma para outro, e geração de uma sentença no idioma alvo.

2. Qualquer idioma pode ter um conjunto de expressões. O máximo possível de expressões deve ser coletado no dicionário.

3. Os sistemas de tradução atuais podem analisar e traduzir frases compostas de menos de 10 palavras. Mesmo um ser humano não consegue entender o significado de uma frase longa na primeira leitura. Uma outro ponto de atenção é a ambiguidade.

4. As regras gramaticais na tradução automática podem ser consideradas regras para a inteligência artificial também.

3.1.2 Questões éticas, legais e sociais

Como os dados médicos são coletados de seres humanos, existe uma enorme questão ética e legal projetada para impedir o abuso em pacientes e o uso indevido de seus dados, além da rigorosidade variar muito para cada país. Privacidade e segurança da informação sofrem com a burocracia, tendo processos redobrados, para viabilizar uma pesquisa [36]. Existe uma questão bastante discutida sobre a propriedade das informações médicas na mineração de dados. Na área jurídica, a propriedade é determinada por quem tem o direito de vender um item específico de propriedade [41]. Porém, é eticamente impróprio vender dados ou tecidos humanos, tornando a propriedade dos dados algo confuso.

Milhares de terabytes são gerados anualmente na América do Norte e Europa. Todos esses dados são espalhados por bancos de dados heterogêneos, muitos sem nenhuma identificação de origem ou princípios comuns de organização. Também no ambiente da Computação, é muito fácil replicar dados e compartilhar essas informações pela rede. A questão da propriedade das informações do paciente é incerta e é objeto de ações judiciais recorrentes e altamente divulgadas e inquéritos do congresso [36]. Os pacientes individuais têm recebido seus próprios dados? Eles assinam termos de liberação desses dados para pesquisas? Os médicos são os proprietários dos dados, uma vez que foram eles que extraíram?

Outra característica da mineração de dados médicos, é o medo de ações judiciais contra médicos e outros profissionais de saúde. Os médicos e cientistas de dados médicos são relutantes em entregar seus dados aos mineradores de dados. Anomalias aparentes no histórico médico de um paciente podem desencadear uma investigação [36]. Em muitos casos, nem todos os resultados anormais na medicina são necessariamente o resultado de um comportamento negligente do provedor. No entanto, uma investigação inevitavelmente consome o tempo e a energia emocional dos médicos. Por isso eles acabam não fornecendo esses dados para não se exporem a esse risco.

Em relação às diretrizes para privacidade do paciente, existem várias políticas e procedimentos administrativos que normalmente não seriam necessários para a mineração de dados não médicos [42]. Deve haver políticas para avaliar e certificar que medidas de segurança apropriadas estão em vigor na instituição de pesquisa, contratos legais entre a organização e quaisquer partes externas com acesso a informações de saúde identificáveis individualmente, exigindo que as partes externas protejam os dados.

Devem haver planos de contingência para resposta a qualquer emergência, incluindo um plano de backup de dados e um plano de recuperação de desastres, um sistema de controle de acesso a informações que inclua políticas para a autorização, estabelecimento e modificação de privilégios de acesso a dados. Deve haver também uma revisão interna contínua dos registros de acesso a dados, a fim de identificar possíveis violações

de segurança.

A organização deve garantir a supervisão do pessoal que executa atividades de manutenção de sistemas técnicos, a fim de manter registros de autorização de acesso, garantir que o pessoal que cuida da operação e também manutenção, tenha acesso adequado. Garantir o adoção de procedimentos de segurança de pessoal e de maneira alguma haver qualquer descuido em qualquer processo de treinamento de usuários para a segurança do sistema.

Deve haver procedimentos de rescisão que são executados quando um funcionário deixa ou perde o acesso aos dados. Também é necessário treinamento de segurança para todo o pessoal, incluindo treinamento de conscientização para todo o pessoal, lembretes periódicos de segurança, treinamento do usuário sobre proteção contra vírus, treinamento do usuário sobre a importância de monitorar falhas de acesso, gerenciamento de senhas e como relatar discrepâncias [42].

Além dessas e outras regras, os mineradores de dados médicos consideram o trabalho de mineração de dados médicos muito desgastante e oneroso. Eles devem avaliar muito bem a real necessidade e retorno da pesquisa, além da suspeita de o estudo pode ser um indício de epidemia ou problema em massa, sendo necessário antes de tudo, combinar informações, entre eles a origem de cada informação por meio de dados de endereço do paciente.

Essas e muitas outras regras impõem restrições aos mineradores de dados médicos que outros pesquisadores acadêmicos considerariam onerosos e sufocantes à criatividade da pesquisa científica. Os pesquisadores devem avaliar cuidadosamente a necessidade percebida de informações, como códigos postais (que podem ser necessários para estudos epidemiológicos), que também podem tornar os dados re-identificáveis em combinação com outras informações [43].

3.1.3 Filosofia estatística

Bases heterogêneas são limitadas ou até inviabilizadas de serem trabalhadas, quando utilizados em uma metodologia estatística. Até mesmo métodos de mineração de dados mais flexíveis, muitas vezes não conseguem extrair informações de sua base de dados bruta. Como uma mineração de dados médicos exige estudos complexos, é necessário que uma base estatística seja aplicada, tornando esse relacionamento entre os dois algo difícil de mantê-los juntos como requisitos de uma pesquisa [27, 41].

Os testes estatísticos clássicos são projetados a partir da ideia de um experimento repetível, com regras previamente definidas. Não é correto alterar regras no meio de um experimento, porque as fórmulas e distribuições se tornam sem sentido. Assim, os testes estatísticos clássicos empregados na medicina podem estar sujeitos a problemas, sendo

agrado pela heterogeneidade. O paradigma intelectual da estatística clássica dependa não apenas dos números reais coletados, mas também do estado de espírito de uma pessoa no início da investigação estatística. Se alguém muda de ideia durante a investigação, polui a interpretação dos dados, mesmo que nenhum dos valores observados seja alterado [36].

Vários estudos patrocinados pelo governo federal americano se tornaram inválidos por conta dessas circunstâncias, incluindo: quimioterapia para câncer de próstata [44]; quimioterapia para câncer de mama [45]; terapia hipoglicêmica oral de diabetes de início adulto [46]; e terapia com esteróides para fibrose cística [47]. Muitos ensaios clínicos patrocinados pelo governo federal agora exigem a posição de um *ombudsman* [48], que tem o poder de interromper a investigação quando o bem-estar dos pacientes do estudo está potencialmente ameaçado.

Em nosso estudo, que será explicado mais adiante, foram feitos tratamentos e transformações para se extrair apenas informações necessárias, isolando pacientes saudáveis para um resultado assertivo. Foi necessário um tratamento detalhado, pois para alcançar o objetivo de obter valores de referências, devem haver apenas pacientes sem problemas cardíacos. Como se trata de uma base unificada, não foi um problema a ser tratado o fator "Filosofia estatística"em nosso estudo, O banco é unificado e padronizado de um único sistema, evitando a base heterogênea, mas que abrange o Brasil todo.

3.1.4 Status especial da medicina

Por se tratar de uma área que envolve vidas, e muitas vezes os eventos médicos são de vida ou morte, a medicina possui esse status especial na visão científica a ser considerada. A cobrança por resultados perfeitos ou próximos disso, com pouca tolerância para erros é de certa forma justificável, por se tratar de vidas em jogo. A saúde humana depende e sempre dependerá desse pilar para que exista evolução em todas as outras áreas da ciência. [42].

Todas as pessoas desfrutam dos benefícios de pesquisas médicas realizadas em outros pacientes, mas muitas vezes relutamos em contribuir ou divulgar nossas próprias informações para tais fins. Assim, quando os dados de estudos médicos são publicados, espera-se que os pesquisadores mantenham a dignidade de cada paciente e que os resultados sejam utilizados apenas para fins socialmente benéficos. [42]

Com o avanço e popularização dos aparelhos *wearables* (vestíveis), como *smart watches*, mais informações estão sendo obtidas, tornando o Big Data ainda mais rico. Com isso a preocupação com dados sensíveis se torna algo relevante, e grande empresas como Apple [49] e Google [50] passam a serem responsáveis por rígidas regras de segurança e privacidade das informações. Essa grande carga repentina de informações contribui para que os dados tenham um controle melhor supervisionado, rastreável, criptografado e em constante monitoramento de como e onde são utilizadas essas informações.

4 METODOLOGIA

A base de dados utilizada para este projeto, em que contém todas as informações de clínicas, médicos, pacientes e seus valores obtidos no laudo, teve a autorização do acesso total e utilização por parte dos proprietários do Sistema EcoCloud. Desde que seu uso seja feita de maneira anônima, isto é, sem expor nome algum de paciente, médico ou clínica. Apenas utilizando seus valores para serem extraídas informações de maneira macro (médias gerais de cada parâmetro, quantidade de amostras e por localização).

O Sistema EcoCloud é acessado de maneira online, via navegador e tem dados desde 2016, ano do seu lançamento. Utilizam o sistema, médicos de várias regiões do Brasil, concentradas na região Sul e Sudeste e em algumas partes do Norte e Nordeste. Cada médico tem seu acesso restrito, em que é possível realizar o cadastro e fluxo de laudo inserindo os valores dos parâmetros do coração medidos, que por sua vez são gravados no banco de dados. Essas informações são as utilizadas neste projeto. Todo médico, ao realizar o cadastro no sistema, aceita um termo em que os dados serão utilizados de forma anônima para pesquisas.

O banco de dados do Sistema EcoCloud é todo desenvolvido em MySQL. Para a conexão com o banco e extração dos dados foi utilizado o software MySQL Workbench. A mineração de dados foi realizada por meio da linguagem Python, por ser otimizada para o objetivo. A IDE (*Integrated Development Environment* ou Ambiente de Desenvolvimento Integrado) utilizada para Python foi o programa PyCharm, própria e que dá suporte para a linguagem.

Como pesquisa principal, é apresentado todo o processo realizado para a mineração de dados, desde o pré-tratamento, removendo *outliers*, dados de testes, apagados, errados, e principalmente pacientes doentes que podem interferir nos resultados de valores de referência para uma pessoa sadia. Conhecendo a necessidade de obter dados relevantes para que seja possível extrair relatórios, e, a partir dele, realizar métodos para determinar novos parâmetros e intervalos de medidas do coração, foram necessárias ações na base de dados do sistema. Para que a mineração no banco de dados do sistema utilizado entregue resultados satisfatórios, foi respeitado o processo KDD. Abaixo é descrita cada etapa, e como foi realizada.

4.1 Extração de dados

Dentre as 89 tabelas existentes no sistema, foram utilizadas 5 tabelas que contém as informações necessárias para a pesquisa. A tabela "relatorio"se trata da tabela principal, que possui os valores relacionados às medidas do coração inseridas pelo médico, e

seus relacionamento com outras tabelas. Essa é a tabela mais importante para o estudo. Também possui os identificadores *ID* que se relacionam por meio de chaves estrangeiras com as tabelas do médico responsável e consulta.

Na sequência, existe outra tabela que traz informações relacionadas aos dados antropométricos do paciente, como sexo, data de nascimento (para definir a idade), peso e altura. por meio desses dados são feitos cálculos no sistema para definir a fração de ejeção e diagnósticos gerados automaticamente, que são interferidos pelo peso altura e idade do paciente. Existe também a tabela "medico", que serve exclusivamente para obter a região em que foi feita a consulta, consequentemente determinando a região do paciente, sendo possível gerar resultados separados por localidades. Existem também várias colunas relacionadas a negócios do sistema, irrelevantes para a pesquisa.

Por fim, foi utilizada a tabela "estado". Se trata de uma tabela bem simples, mas por ela foi possível separar as amostras por estados do Brasil, permitindo assim, juntamente com a tabela "medico", a separação de laudos por regiões. A *query* utilizada para extração de dados, e na sequência a geração de um arquivo CSV, foi a seguinte:

```
SELECT rel.peso, rel.altura, rel.imc, rel.superficie_corporea, rel.raiz_aortica,
       rel.atrio_esquerdo,  rel.septo_interventricular, rel.parede_posterior,
       rel.diametro_diastolico, rel.diametro_sistolico,
       rel.vol_indexado_atrio_esquerdo, rel.diametro_basal_ventriculo_direito,
       rel.diametro_medio_ventriculo_direito, rel.FE_teicholz, rel.FE_simpson,
       rel.FE_subjetiva, rel.genero, rel.atrio_direito_dilatado,
       rel.ventriculo_direito_dilatado, rel.data_cadatro,
       rel.contratilidade_segmentar_normal, rel.tipo_laudo, con.tipo_exame,
       con.tipo, con.data, con.horario, con.medico_solicitante_nome,
       con.medico_solicitante_uf, con.pdf, con.removed,
       pac.sexo, pac.nascimento, pac.peso, pac.altura, uf.estado, uf.uf
FROM ecocloud.relatorio as rel
left join consulta as con on rel.consulta_id = con.consulta_id
left join paciente as pac on con.paciente_id = pac.paciente_id
left join medico as med on med.medico_id = rel.medico_id
left join estado as uf on med.estado_id = uf.estado_id
```

4.2 Limpeza de dados

A primeira parte da limpeza de dados começou por meio de *queries* no SQL. Foi feita uma busca de dados inválidos como palavras ou sentenças que contém "teste"em cadastros de médicos, pacientes, clínicas e hospitais. Esses dados são cadastrados pelos próprios usuários médicos ou secretárias que utilizam a ferramenta, em caráter de teste ou

treinamento. Foram eliminados também cadastros de pacientes com data de nascimento incoerente, com "00-00-0000", ou então datas de nascimento com mais de 120 anos atrás.

A partir disso, foi realizada a exclusão em cascata de dados inválidos, excluindo todos os relacionamentos, até chegar aos laudos e seus valores inseridos. Com isso foram eliminados 1215 laudos inválidos, dos 22050 iniciais. Na segunda parte, já após a extração do arquivo CSV, que posteriormente será usada para a mineração de dados, foram eliminados frases que apontam algum indício ou problema em parte do coração. Qualquer paciente que esteja com qualquer anormalidade no coração pode interferir nos resultados finais.

1. Pacientes com átrio direito dilatado;
2. Pacientes com ventrículo direito dilatado;
3. Pacientes com qualquer situação moderada ou mais severa na valva ou função;
4. Pacientes com problemas de estenose;
5. Pacientes com problemas de gradiente;
6. Pacientes com com qualquer grau de derrame;
7. Pacientes com com qualquer grau de espessamento;
8. Pacientes com fração de ejeção menor que 0,55.
9. Pacientes com superfície corpórea maior que 3.

Essas frases indicam pelo menos um grau mínimo, leve, moderado ou importante de qualquer anormalidade do coração. São apontadas pelo próprio sistema por meio da inserção de valores de medidas do coração, ou então selecionadas de maneira manual pelo médico, em seu processo de diagnóstico. Desse mesmo CSV, também foram retirados alguns laudos considerados inválidos, com valores de medidas zerados, por conta de alguns possíveis testes que sobraram, ou então por falha do próprio médico no cadastro. Os resíduos de erro estudados foram utilizados para detectar valores extremos a serem excluídos da análise.

Luis Alvarez et al. [51], em sua pesquisa envolvendo crianças e adolescentes de 15 a 18 anos, com 367 corações, relatou que existem diferenças significativas e características diferentes em seus parâmetros em relação a adultos. Por conta disso, foram descartados os resultados com crianças abaixo de 18 anos, pois os valores podem interferir na média geral. O motivo é pelo fato do sistema não ter disponível até a publicação deste documento, a possibilidade de laudar e utilizar valores de referências para crianças, bem como as frases pré-determinadas de diagnóstico para a faixa etária. Por isso, apesar da base de dados

possuir laudos de pacientes com idade inferior a 18 anos, não são adequados para este estudo, pois os valores de referência disponíveis são para adultos.

Pacientes com superfície corpórea maior que 3 foram descartados por serem caracterizados como dados inválidos (erro de cadastro). Posteriormente, na etapa de remoção de outliers, foram removidos os pacientes que tenham superfície corpórea maior que 2,5, conforme mostrado na Figura 11. Após todos os processos descritos acima, foram removidos 1253 laudos de pacientes doentes ou com dados de medidas inválidos, que poderiam influenciar no resultado final da pesquisa.

4.3 Seleção

Para a seleção de atributos, foram analisadas e separadas as tabelas relacionadas e que contém as informações necessárias para a análise:

1. Tabela médico;
2. Tabela consulta;
3. Tabela relatório;
4. Tabela paciente;
5. Tabela estado.

A tabela médico grava todos os cadastros dos usuários médicos, que é a principal entidade do banco. A partir dela são criados e relacionados os registros de consulta, que por sua vez se relacionam com relatório, paciente e estado. A tabela relatório possui todas as medidas e frases que compõem o laudo, sendo a mais importante para a análise. As tabelas paciente e estado, identificam o próprio paciente e sua região.

4.4 Transformação de dados

Nesta etapa não foi necessário grande esforço, pois o banco já reproduz de maneira satisfatória todos os campos, em tabelas relacionais já previstas, com seus tipos de dados apropriados. No entanto, foi preciso realizar a junção de dados de fração de ejeção, que possuem três colunas diferentes. No sistema usado para a mineração, existe a possibilidade de que no laudo final apareça a fração de ejeção gerada automaticamente pelo cálculo de *Teicholz*, conforme mostra a Figura 2, ou então que seja feita a entrada de dado de FE de maneira manual. Neste caso o médico opta pelo cálculo Subjetivo ou *Simpson*, fazendo de maneira manual e inserindo no campo apropriado durante o processo de laudo.

```
// FE
float B7 = medidaDiametroDiastolico.medidaValor.floatValue; // diametroDiastolico
float B10 = medidaDiametroSistolico.medidaValor.floatValue;// diametroSistolico
float FE_teicholz =
    ((powf(B7,3)*7)/(2.4+(B7/10))-(powf(B10,3)*7)/(2.4+(B10/10)))/((powf(B7,3)*7)/(2.4+(B7/10)))*100;
```

Figura 2 – Cálculo de fórmula *Teicholz*, feito em *Objective-C* para iOS, usando as variáveis Diâmetro diastólico e sistólico.

```
# Médias, Medianas e Desvios Padrões
df_std = df.std(axis=0)
df_mean = df.mean(axis=0)
df_med = df.median(axis=0)

viae_std = np.std(np.array(vol_indexado_atrio_esquerdo))
viae_mean = np.mean(np.array(vol_indexado_atrio_esquerdo))
viae_med = np.median(np.array(vol_indexado_atrio_esquerdo))
dbvd_std = np.std(np.array(diametro_basal_ventriculo_direito))
dbvd_mean = np.mean(np.array(diametro_basal_ventriculo_direito))
dbvd_med = np.median(np.array(diametro_basal_ventriculo_direito))
dmvd_std = np.std(np.array(diametro_medio_ventriculo_direito))
dmvd_mean = np.mean(np.array(diametro_medio_ventriculo_direito))
dmvd_med = np.median(np.array(diametro_medio_ventriculo_direito))
```

Figura 3 – Cálculo para extração de médias, medianas e desvios padrões.

Com isso, é exibido no laudo e consequentemente guardado na base de dados, apenas o resultado da FE a partir de apenas um dos três métodos. o cálculo em Teicholz é sempre realizado de maneira automática. Caso o médico opte pela inserção manual do valor de FE calculado com a fórmula Simpson ou Subjetiva, esses são os utilizados, desconsiderando a fórmula Teicholz. Do contrário é levado em considera ção a resultado em Teicholz. O dataset final sempre considera a decisão do médico, no caso, qual fórmula ele escolheu para chegar ao valor final de FE.

No banco de dados eles são armazenados em colunas diferentes, tendo a necessidade de fazer a junção dessas colunas. Independente do método de cálculo usado, o resultado final é tratado de maneira igual. Feita essa junção, os dados de fração de ejeção puderam então serem normalizados. Dessa forma, foi adicionada a coluna única de fração de ejeção junto às demais colunas de valores de medidas do coração.

4.5 Mineração

Nesta importante etapa, foram extraídos valores como as médias, medianas e desvios padrões conforme ilustra a Figura 3, e também outras informações que servem de

apoio para a análise, como a quantidade total de amostras e separadas por sexo, região do Brasil. Na sequência, são gerados os gráficos do tipo *BoxPlot*, apresentados no próximo capítulo, e também *ScatterPlot*, em que é aplicada a regressão linear. Em todos os gráficos *ScatterPlot*, são trabalhadas com a superfície corpórea como parâmetro de comparação para as medidas.

A área de superfície corpórea (ASC) foi usada como variável independente em uma análise de regressão não linear para o valor médio previsto de cada uma das estruturas medidas ecocardiograficamente, em cada laudo, após o pré-processamento dos dados. Devido ao problema de variações heterogêneas com essas medições em toda a faixa de ASC (quanto maior a ASC, maior a variância). A superfície corpórea é frequentemente usada em fins clínicos em relação ao peso corporal, porque é um indicador mais preciso da massa metabólica (a necessidade de energia do corpo), em que a massa metabólica pode ser estimada como massa livre de gordura, uma vez que a gordura corporal não é metabolicamente ativa. Na cardiologia é usada para determinar o índice cardíaco. A Figura 4 apresenta o cálculo para obtenção.

4.6 Avaliação e visualização de resultados

No próximo capítulo, são exibidos todos os gráficos gerados, evidenciando os *BoxPlots* que resultaram em gráficos relativamente simétricos, compatível com dados reais, trazendo outliers, que são desconsiderados dos comparativos finais. Também como dito, são apresentados os gráficos *ScatterPlots*, que evidenciam em quais intervalos de medidas existem mais pacientes, sendo possível extrair valores de referência, tanto para o sexo masculino quanto feminino, que são apresentados no capítulo subsequente.

4.7 Uso das técnicas mencionadas em outras bases de dados

Por meio dos processos descritos acima, é possível utilizar as mesmas técnicas em estudos de outras áreas da medicina que envolvam valores de referência, com algumas adaptações. Os processos a serem seguidos, se resumem nas seguintes etapas:

1. Exclusão de dados como datas e parâmetros inválidos, informações duplicadas e tabelas e colunas desnecessárias, utilizando queries SQL;

2. Extração da planilha (em arquivo CSV), e utilização de expressão regular para remoção de pacientes não sadios, identificados por meio de frases e termos que são inseridos pelos médicos. É essencial que o*Dataset* final não tenham dados de pacientes com qualquer problema que possa interferir nos resultados finais;

3. Exclusão de *outliers* a partir de regras definidas;

$$SC(m2) = 0,007184X(altura)^{\,0,725})X(peso)^{\,0,425}$$

Figura 4 – Cálculo para obtenção da superfície corpórea.

4. Obtenção de médias, medianas simples por meio da linguagem Python, para cada uma dos parâmetros a serem analisados;

5. Utilização da função *boxplot* para geração de gráficos que tem como objetivo a pré-análise dos dados obtidos;

6. Uso de Regressão Linear para obter os desvios padrões, como descrito abaixo. Os vetores superficiecorporea[] e medida[], podem ser substituídos pelos preditos e coeficientes necessários para a pesquisa.

```
X = np.array(df['superficiecorporea'])
y = np.array(df[medida])

X_train, X_test, y_train, y_test = train_test_split(X, y)

lin = LinearRegression()
lin.fit(X_train, y_train)
print(r2_score(y_test, lin.predict(X_test)))

a = lin.predict
b = lin.coef
x0 = X[0][0]
stdAE = df_std[medida]
z_plus_1 = a + stdAE - b * x0
z_minus_1 = a - stdAE - b * x0
z_plus_2 = a + 2 * stdAE - b * x0
z_minus_2 = a - 2 * stdAE - b * x0
```

Um diagrama macro com fluxo genérico das etapas, é apresentada na Figura 5, como auxílio para obtenção de resultados.

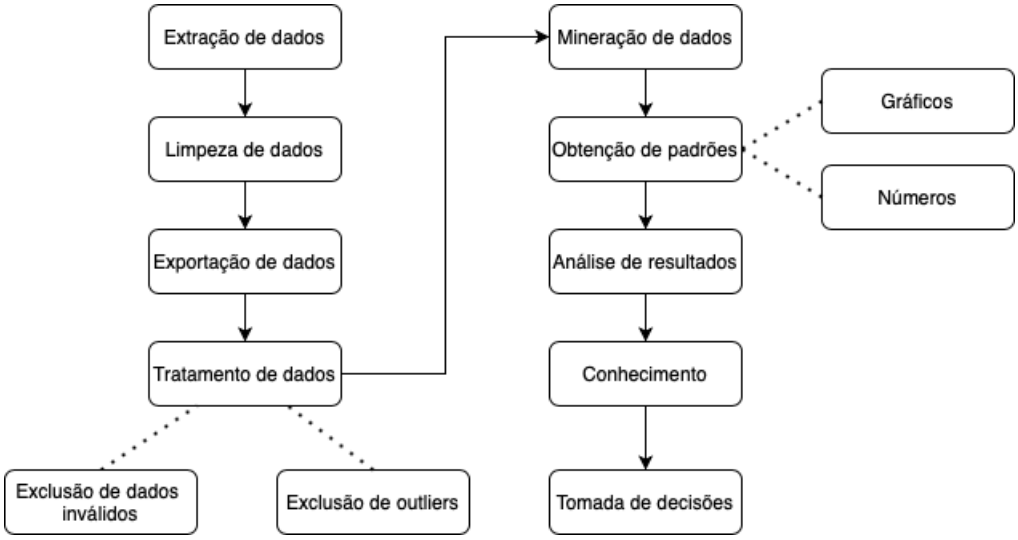

Figura 5 – Diagrama macro com fluxo genérico das etapas para projetos de obtenção de resultados.

5 RESULTADOS

Foram criadas regras para a análise de dados junto aos médicos responsáveis pelo sistema, e, que são isoladas pessoas saudáveis e não saudáveis, pois interferem nos resultados da obtenção de valores de referências. Foram calculados os *Z-scores* de cada parte do coração, para apresentar os gráficos e traçar as referências dentro deles. Também foram feitas regras para analisar em como as regiões onde moram os pacientes interferem nos resultados.

As amostras realizadas e examinadas com a ajuda da automatização para realização de diagnóstico, que é característica dos laudos feitos por meio do sistema EcoCloud, foi perto de uma distribuição igual, entre homens e mulheres, com 47% e 53%, respectivamente. Em média eles tinham 55 anos, um índice de massa corporal (IMC) de 27,56 e uma superfície corporal de 1,86. A Figura 6 demonstra a porcentagem de exames realizados por intervalos de idade.

Segundo o gráfico, boa parte dos pacientes (39,92%) tinha entre 51 e 70 anos. Em contraste, a menor parte (10,50%) possuía um intervalo menor ou igual a 30 anos. Os demais tiveram entre 31 e 50 (27,88%) e foram maiores ou iguais a 71 (21,70%) anos. Para uma melhor análise desses grupos de pacientes, o trabalho apresenta o IMC de acordo com cada faixa etária.

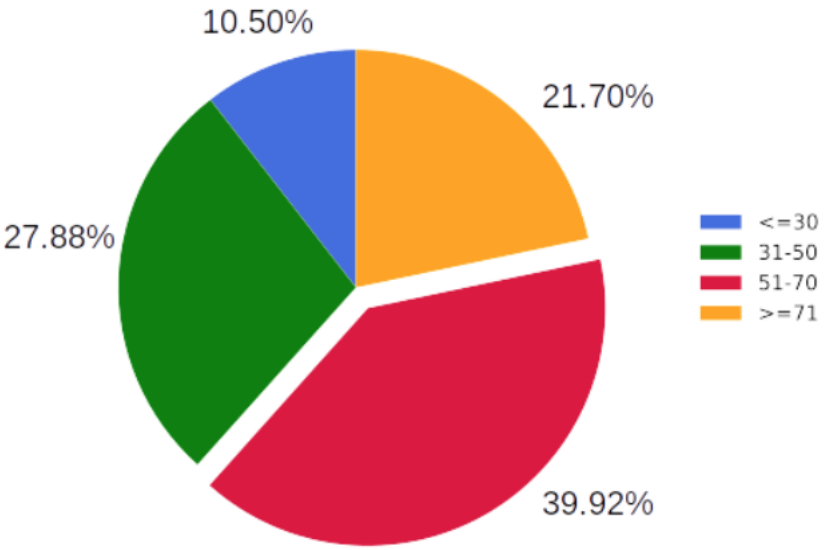

Figura 6 – Percentual de pacientes agrupados por idade.

5.1 Análise preliminar dos dados

Abaixo são apresentados os gráficos do tipo *Boxplot*, em que é possível observar que existe relativa simetria entre os dados. Os gráficos foram gerados utilizando curvas de distribuições normais, que são representações gráficas do teorema da distribuição normal, afirmando que as médias de variáveis aleatórias extraídas independentemente de distribuições independentes convergem na distribuição para a normal, isto é, tornam-se normalmente distribuídas quando o número de variáveis aleatórias é suficientemente grande. [52]

Um *Boxplot* é uma maneira padronizada de exibir a distribuição de dados com base em um resumo de cinco números ("mínimo", primeiro quartil (Q1), mediana, terceiro quartil (Q3) e "máximo"). Ele pode evidenciar sobre seus valores extremos e quais são seus valores. Também pode dizer se seus dados são simétricos, com que grau de rigidez eles são agrupados e como seus dados são distorcidos [53], reforçando junto com os gráficos de testes de normalização, a simetria dos valores.

Para algumas distribuições/conjuntos de dados, é possível encontrar mais informações do que as medidas de tendência central (mediana, média e modo), como são analisadas abaixo. O gráfico estilo *Boxplot* foi escolhido por ele apresentar de maneira clara a localização de cada valor dentro da área, e também a dispersão entre os pontos. Com isso, o gráfico consegue entregar de maneira valiosa a verdadeira distribuição, onde em uma planilha com números é impossível de ter essa visão.

Os *outliers* são facilmente visíveis e identificados pelo gráfico *Boxplot*, mesmo depois da etapa de pré-processamento, em que são evidenciados valores que não correspondem a um paciente sadio. Esses pontos desgarrados podem afetar de forma adversa as decisões a serem tomadas a partir da análise dos dados se não forem devidamente considerados [54]. Foi considerada a média de +/- 3SD (distribuição normal) e mediana +/- 2,5MAD (distribuição não normal). No gráfico de idade da Figura 7, é apresentada de maneira simétrica, em que há apenas 1 *outlier*, que se trata de um dado teste ou erro de cadastro. Assim como outros *outliers* de outras medidas, eles são desconsiderados do cálculo final.

Já no gráfico da Figura 8, para altura do paciente, existem alguns *outliers* de pacientes com menos de 1,4 m, que provêm de uma porcentagem que o filtro eliminou crianças, ou dados de testes. Além disso, existem alguns *outliers* de pessoas com mais de 1,9 m. Eles podem ser reais, porém estão fora do primeiro quartil. O *BoxPlot* para dados de peso do paciente (Figura 9) possui muitos *outliers* no primeiro quartil, que são considerados obesos ou obesos mórbidos. No terceiro quartil, existem dados testes e pesos de crianças. Os laudos desses pacientes também foram removidos.

Dados de laudos de crianças não entregam um dado válido para referências de

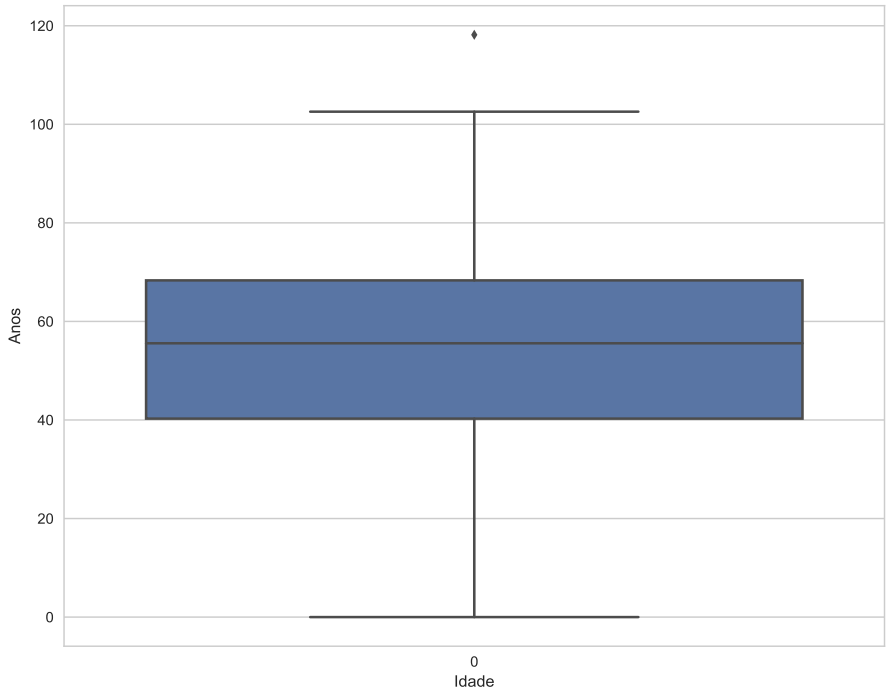

Figura 7 – *Boxplot*: idade do paciente, em anos.

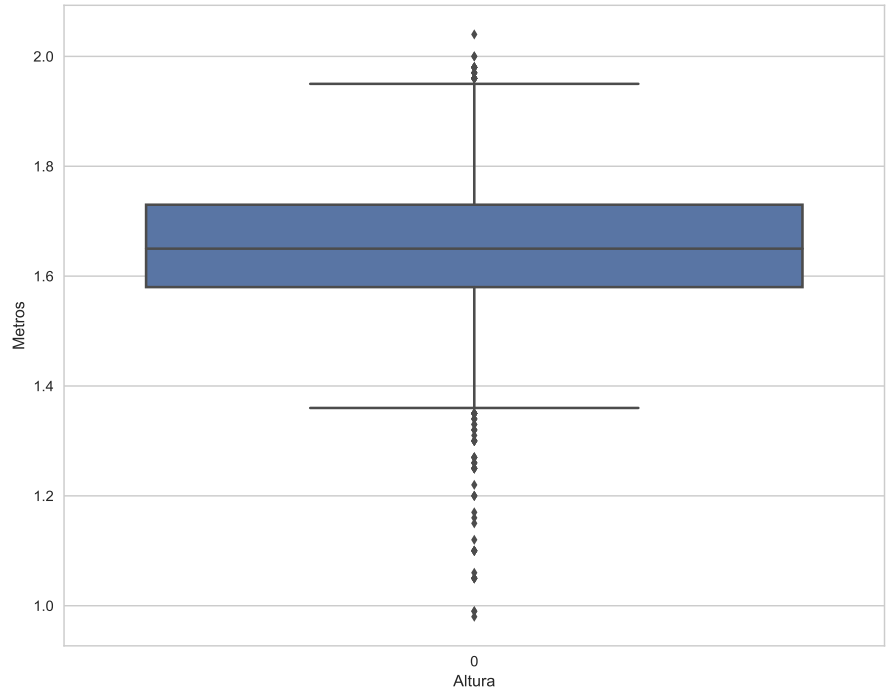

Figura 8 – *Boxplot*: altura do paciente, em metros.

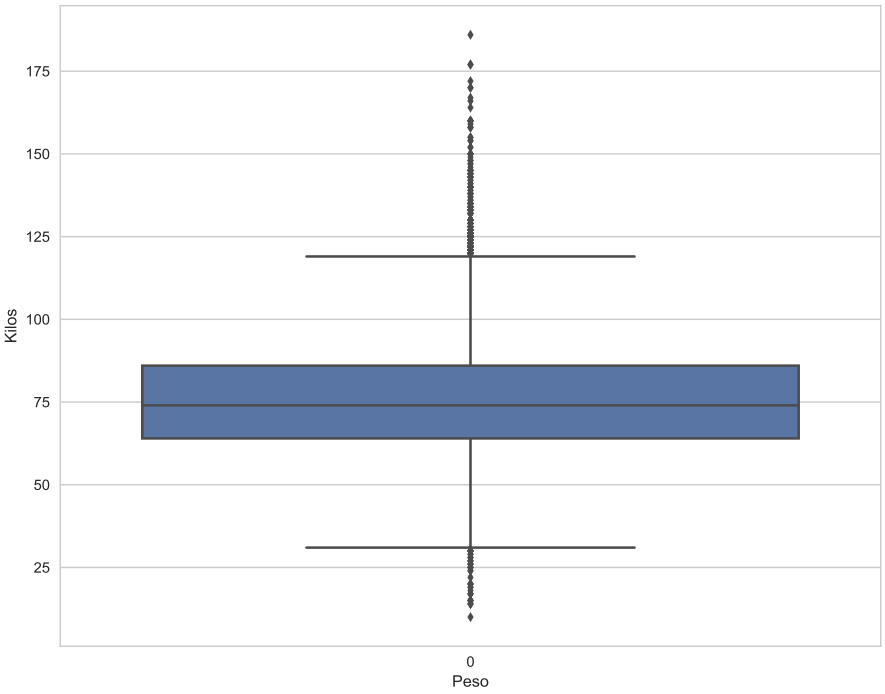

Figura 9 – *Boxplot*: peso do paciente, em quilos.

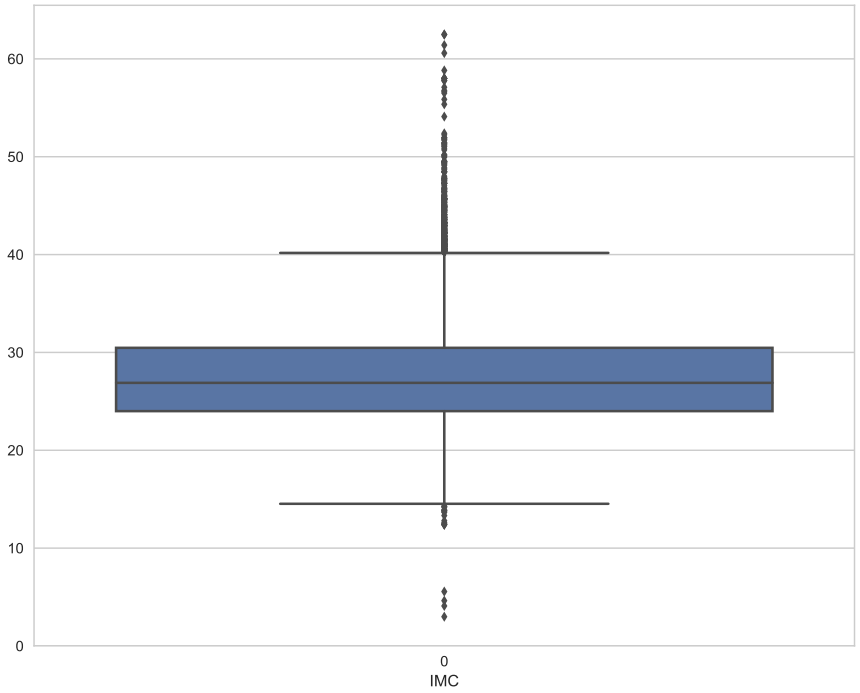

Figura 10 – *Boxplot*: IMC do paciente.

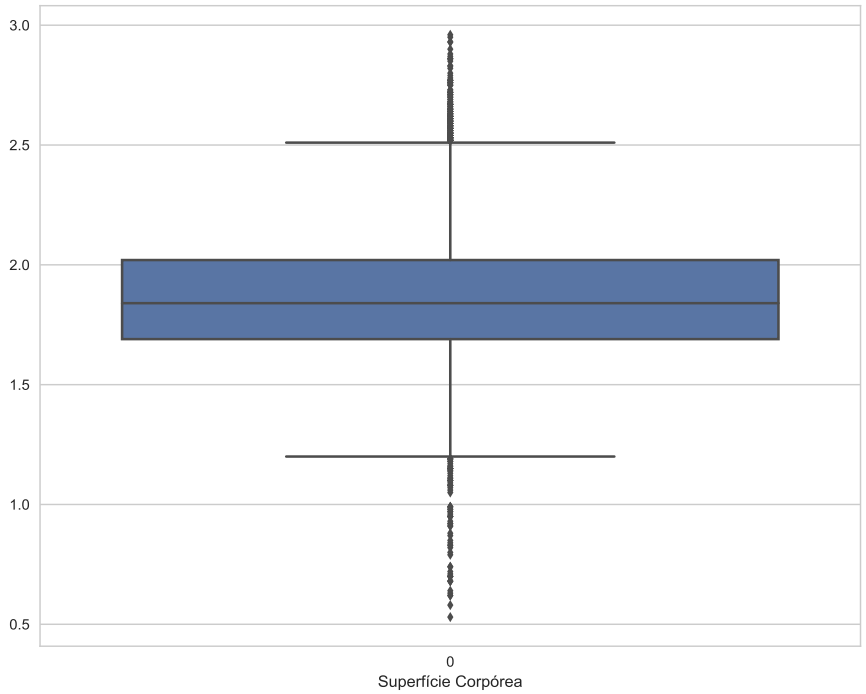

Figura 11 – *Boxplot*: superfície corpórea do paciente.

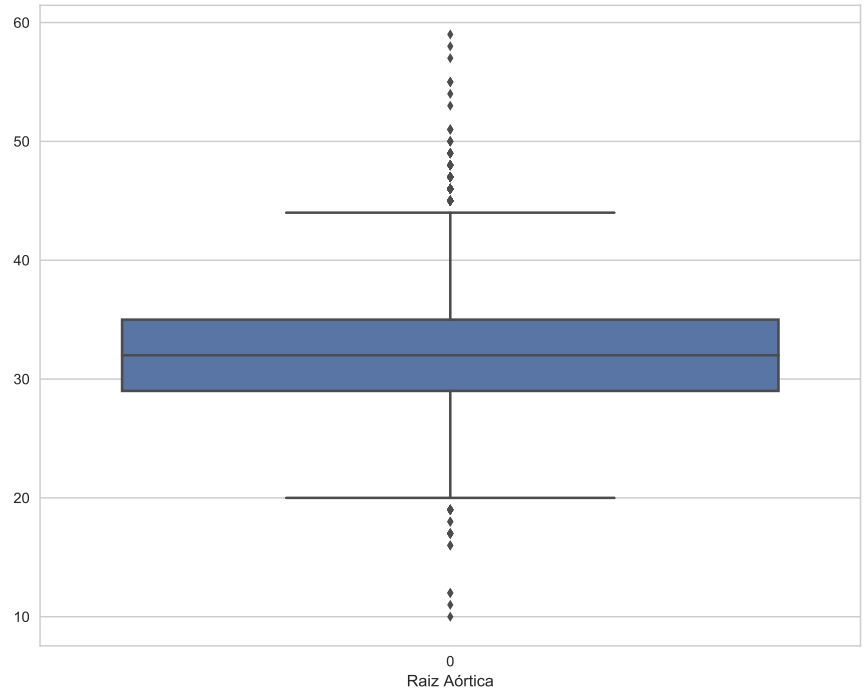

Figura 12 – *Boxplot*: raiz aórtica do paciente, em mm.

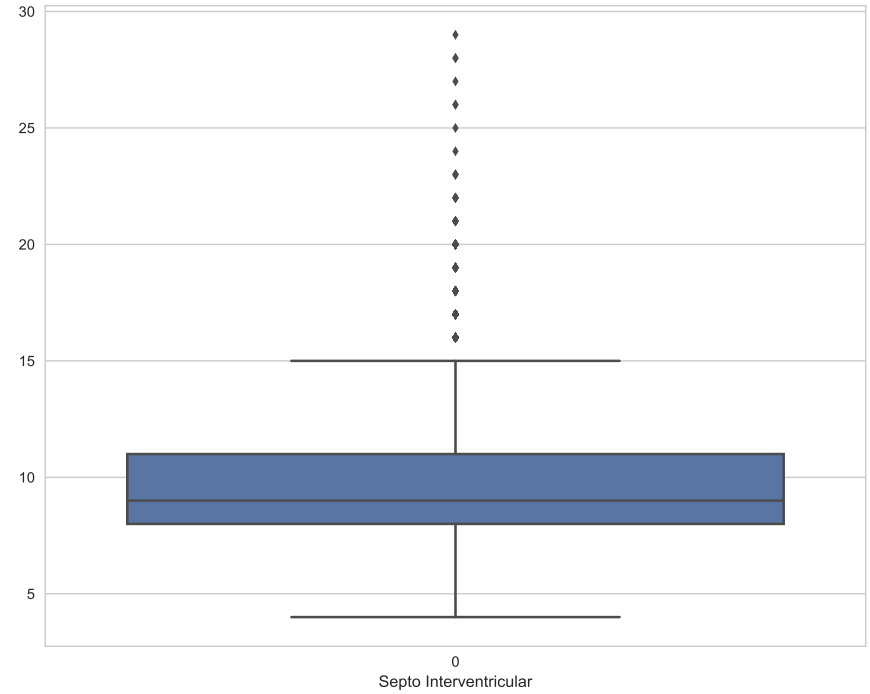

Figura 13 – *Boxplot*: septo do paciente, em mm.

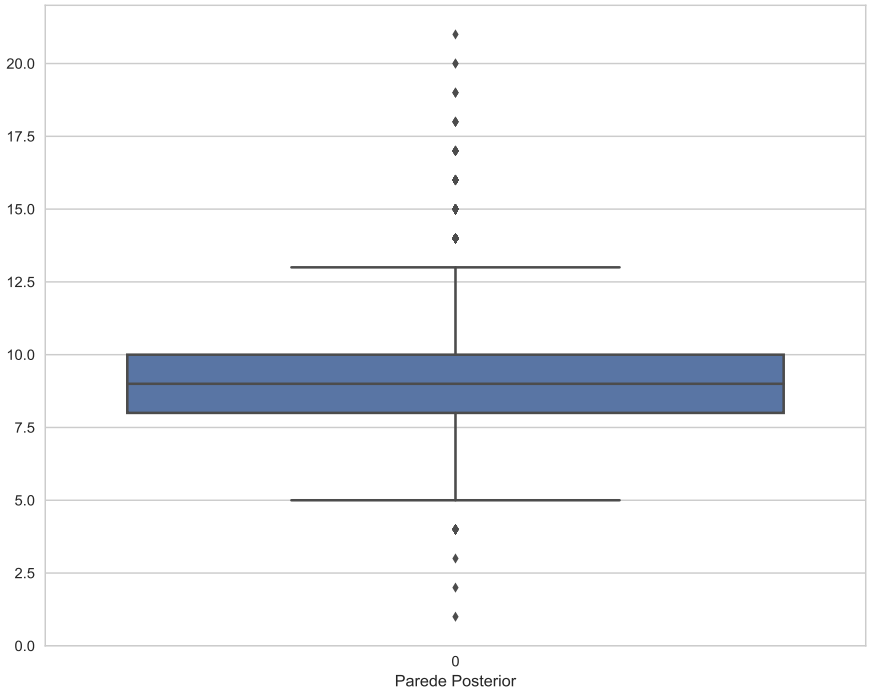

Figura 14 – *Boxplot*: parede do paciente, em mm.

valores, pois estão em desenvolvimento, tendo o coração com tamanho e comportamento diferentes [16], assim como os obesos. O gráfico para IMC mostrado na Figura 10, acompanha os gráficos de peso e altura, em que também existem outliers no primeiro e terceiro quartil. Neste caso, ele entrega muitos *outliers* no seu primeiro quartil, fazendo com que o gráfico fique assimétrico. Na outra extremidade também foram considerados fora dos cálculos os pacientes com IMC abaixo de 14. Os dados do gráfico de IMC está intimamente relacionado a obesos e obesos mórbidos apresentados no *BoxPlot* para dados de peso, pois influencia diretamente no cálculo de IMC.

O IMC é calculado dividindo o peso pela altura elevada ao quadrado. Na Tabela 1, é apresentado o quadro para grau de obesidade, em que aponta o maior grau para pessoas com IMC maior que 40. No *BoxPlot* de IMC mostrado, existe um indício de muitas pessoas com o grau maior, que é um ponto de atenção para a saúde do brasileiro, pois o grau maior que 40 é considerado grave [55]. Continuando o fator de gráficos influenciados pelo peso dos pacientes, a justificativa para a quantidade grande de outliers no *BoxPlot* para superfície corpórea na Figura 11, é por conta de estar também diretamente relacionado com o peso e altura do paciente para o cálculo, onde esses valores são variáveis da fórmula para obter o resultado de superfície corpórea.

Entrando nos gráficos das medidas do coração extraídas, a Figura 12 mostra alguns poucos *outliers* para a raiz aórtica, que são dados de pessoas que realmente podem ter algum tipo de problema em sua raiz aórtica. Neste caso, esse registro passou pelo pré-processamento, mesmo tendo um valor que caracteriza um paciente doente. A justificativa é que no processo de laudo, o médico pode não ter apontado qualquer doença para ele, que depende também de histórico clínico dele.

Essa regra também se aplica para os resultados para as medidas de septo, parede, átrio, sístole e diástole, FE, volume indexado do átrio esquerdo, diâmetro médio e basal do ventrículo direito. Alguns deles apresentam alguns poucos *outliers*, que também foram removidos. A fração de ejeção é apresentada na Figura 15, em que o gráfico reúne valores calculados por meio das três diferentes diferentes fórmulas (Teicholz, Simpson e Subjetiva), porém seus resultados têm o mesmo intervalo de referências, possibilitando que seja mostrado em apenas um. *BoxPlot*

Classificação	**IMC**
Magreza grau III	< 16
Magreza grau II	16.0 - 16.9
Magreza grau I	17.0 - 18.4
Adequado	18.5 - 24.9
Sobrepeso	25.0 - 29.9
Obesidade grau I	30.0 - 34.9
Obesidade grau II	35.0 - 39.9
Obesidade grau III	>= 40 height

Tabela 1 – Quadro de grau de obesidade. Fonte: Organização Mundial da Saúde (1995, 1997)

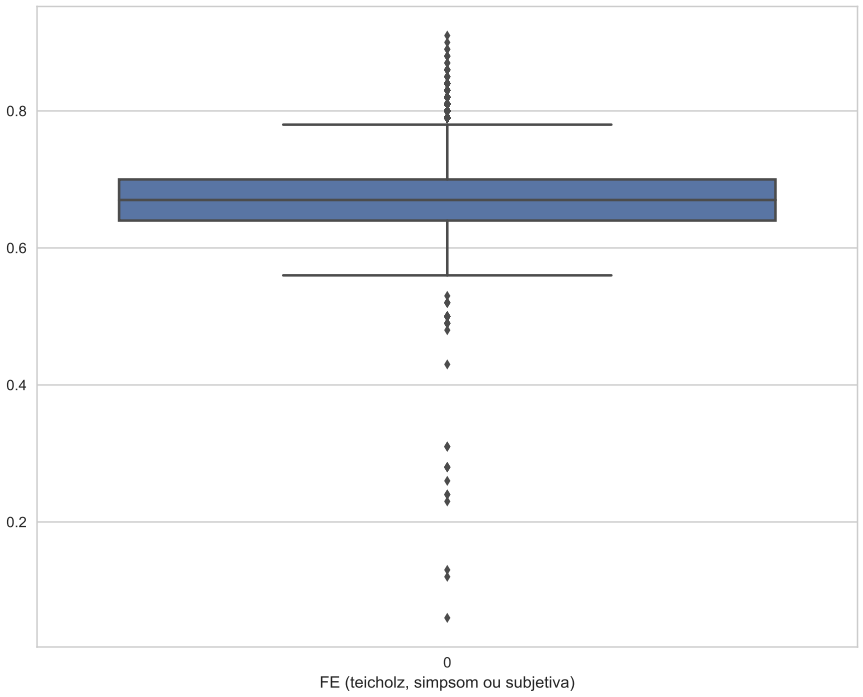

Figura 15 – *Boxplot*: FE do paciente.

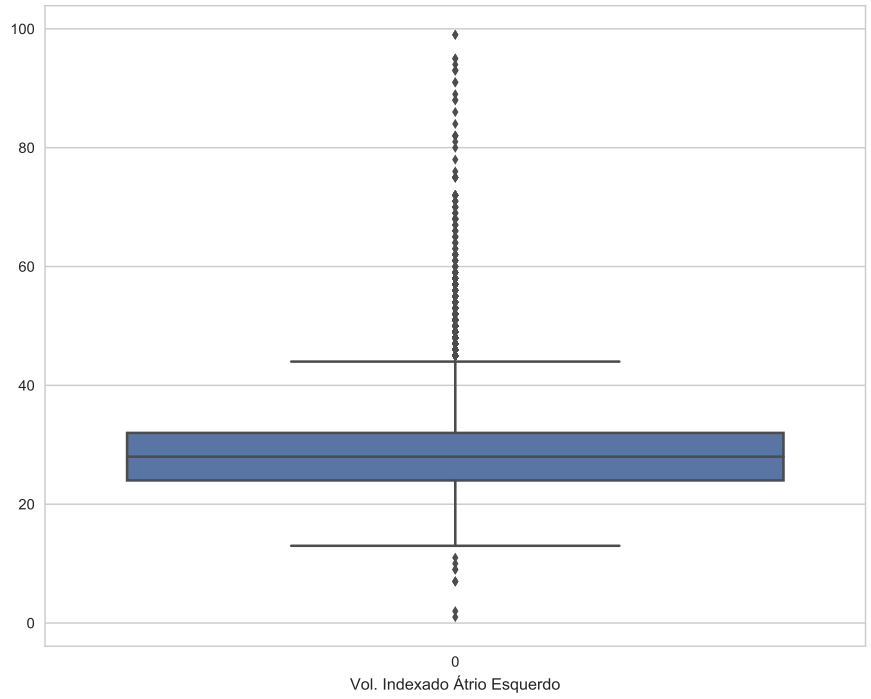

Figura 16 – *Boxplot*: volume indexado do átrio esquerdo do paciente, em mm.

O volume indexado do átrio esquerdo possui uma grande concentração de *outliers* acima de 40mm, por conta de ser um valor opcional a ser cadastrado, em que os médicos dão maior importância no cadastro do valor, justamente quando o valor se mostra fora do intervalo de referências. Com isso, se reflete no gráfico apresentado na Figura 16, em que é evidente a quantidade de pontos no primeiro quartil.

5.2 Pré-análise de desvio padrão

Abaixo são apresentados os gráficos do tipo *Scatter Plot*, que foram gerados por meio da linguagem Python, com o objetivo de analisar e obter o desvio padrão de cada medida do coração. Os pontos em verde escuro representam pacientes do sexo masculino, enquanto os pontos em verde claro, os pacientes do sexo feminino. Todos os gráficos apresentados nas figuras 16 a 25 apresentam cinco linhas. A primeira linha superior se refere ao *Z-score* 2, a segunda linha representa o *Z-score* 1. A terceira linha e também a central, é o *Z-score* 0, que quanto mais pontos próximos a ele, mais normal e comum é o valor de medida do paciente. A quarta e quinta linha se refere ao *Z-score* -1 e *Z-score* -2 respectivamente.

O fator idade foi desconsiderado para a análise, por conta de haver interferência mínima com os valores de referência. Pode haver alguma mudança a cada década somente

[56, 57, 58, 59, 60]. dando espaço para os gráficos a utilização de superfície corpórea para uma melhor visualização, e seguindo o mesmo tipo de gráfico utilizado no trabalho da Sociedade Americana de Cardiologia. [61] Os valores de *Z-score* -1 ao *Z-score* 1 são as referências que buscamos ao longo desse trabalho, e são fielmente representadas nos gráficos de dispersão apresentados nas figuras 20 a 29. A área de superfície corpórea foi usada como variável independente em uma análise de regressão linear para o valor médio previsto de cada uma das 10 estruturas medidas ecocardiograficamente.

O primeiro passo para geração de gráficos *ScatterPlot* ou de dispersão, foi criar uma função dinâmica para recebimento de um valor de medida variável que foi usado no eixo Y, mantendo a superfície corpórea no eixo X, de maneira fixa sendo usado por todos os gráficos. Após isso, foi ajustado o modelo em um conjunto de treinamento para fazer previsões de dados que não foram treinados. Dividimos nossos dados em dois subconjuntos: dados de treinamento e dados de teste, e ajustamos nosso modelo nos dados de treinamento, para fazer previsões sobre os dados de teste.

Por fim, cada medida do coração dentre as dez estudadas, passou pelo método para geração dos gráficos, e em cada um deles foram gerados os valores de desvios padrões, calculados e apresentados como resultado final no próximo capítulo, evidenciando as diferenças entre padrões de referências de medidas do coração do brasileiro e do americano.

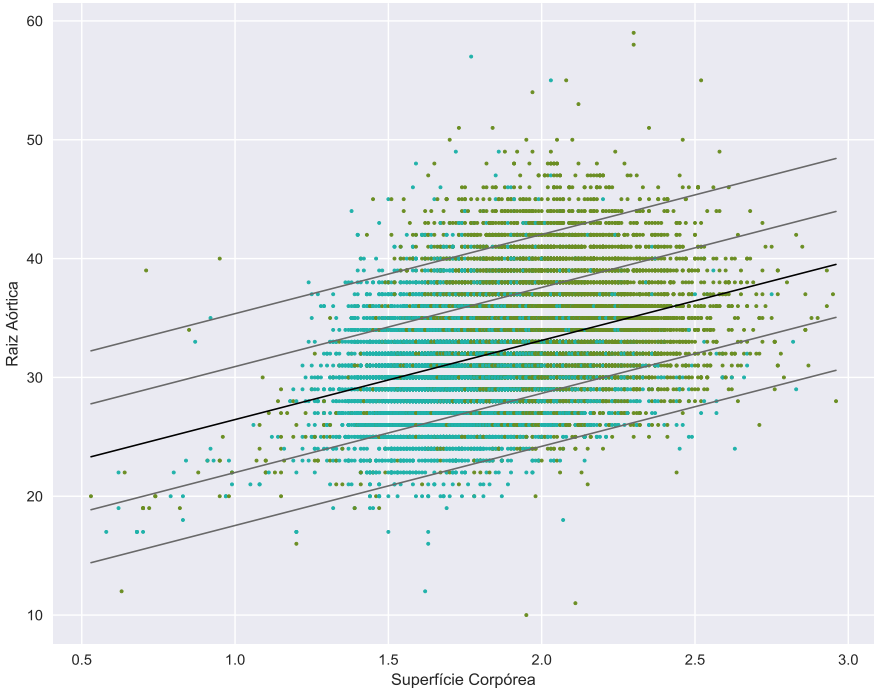

Figura 17 – *Scatter Plot*: raiz aórtica, em mm.

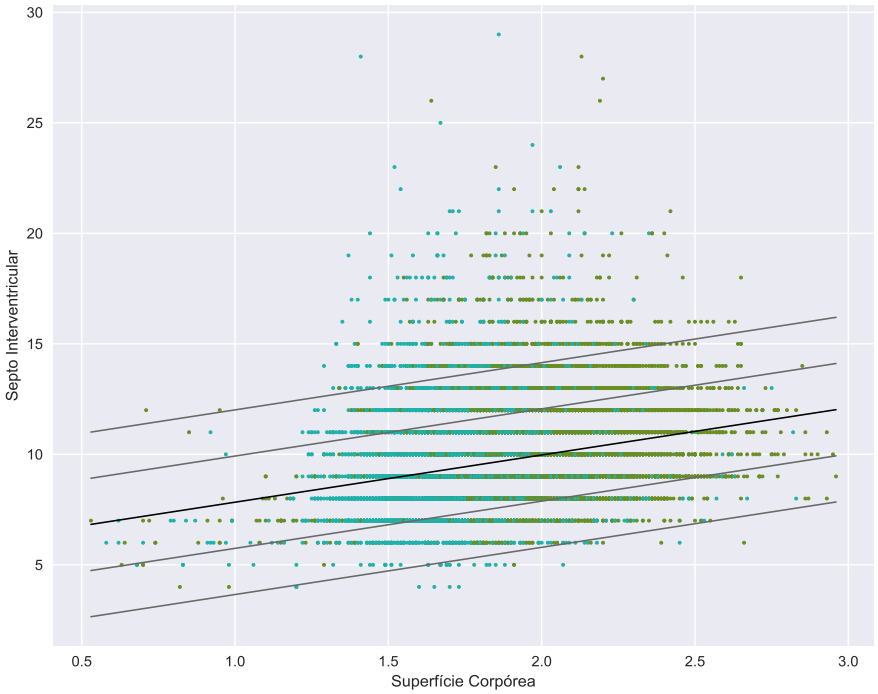

Figura 18 – *Scatter Plot*: septo, em mm.

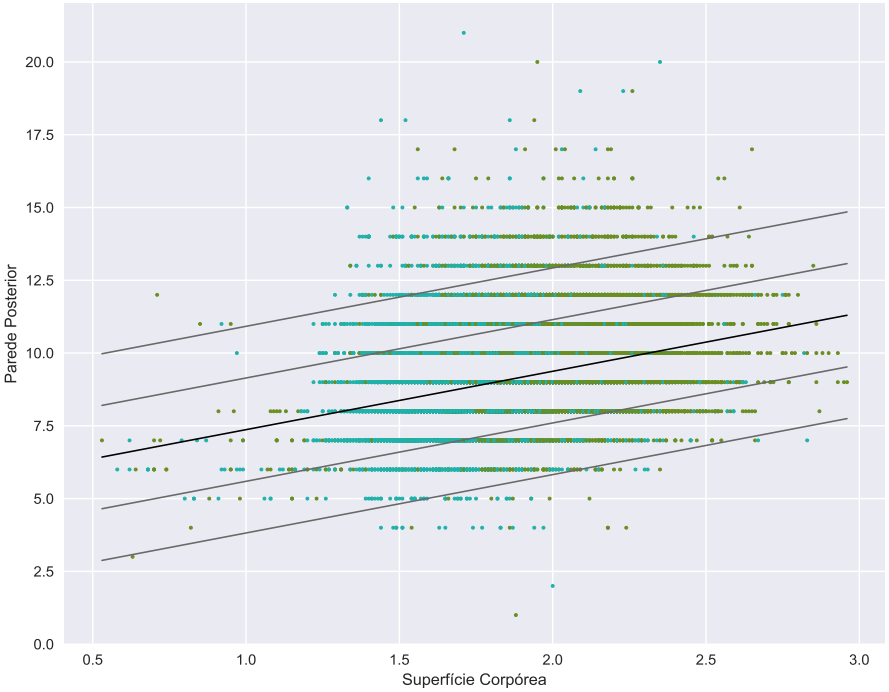

Figura 19 – *Scatter Plot*: parede posterior, em mm.

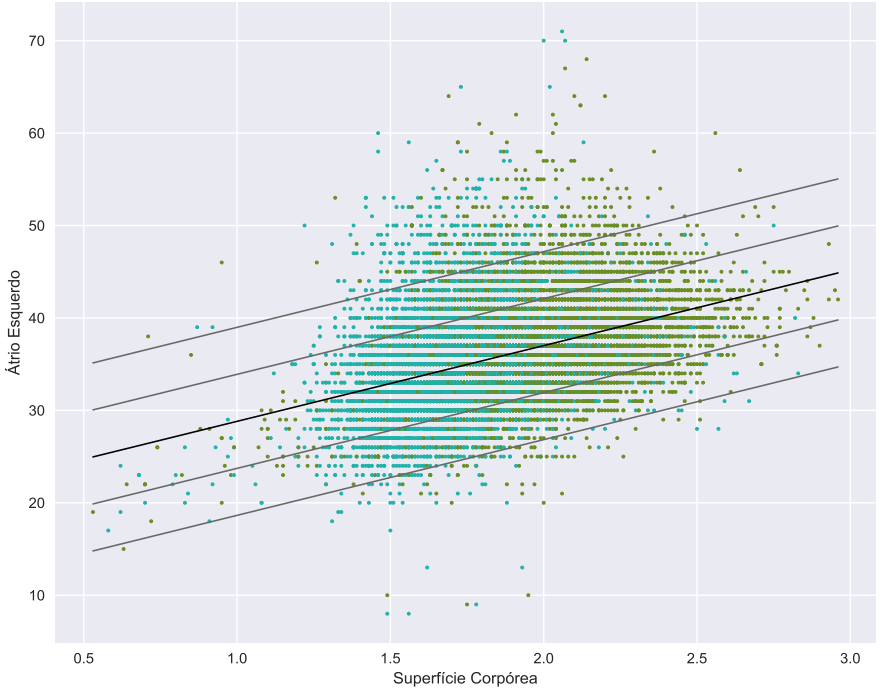

Figura 20 – *Scatter Plot*: átrio esquerdo, em mm.

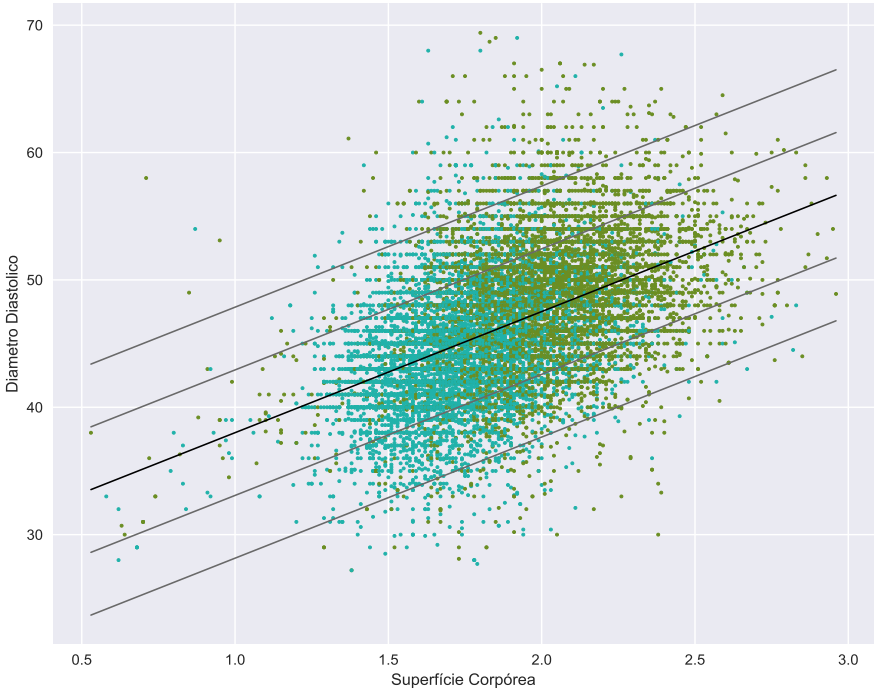

Figura 21 – *Scatter Plot*: diâmetro diastólico, em mm.

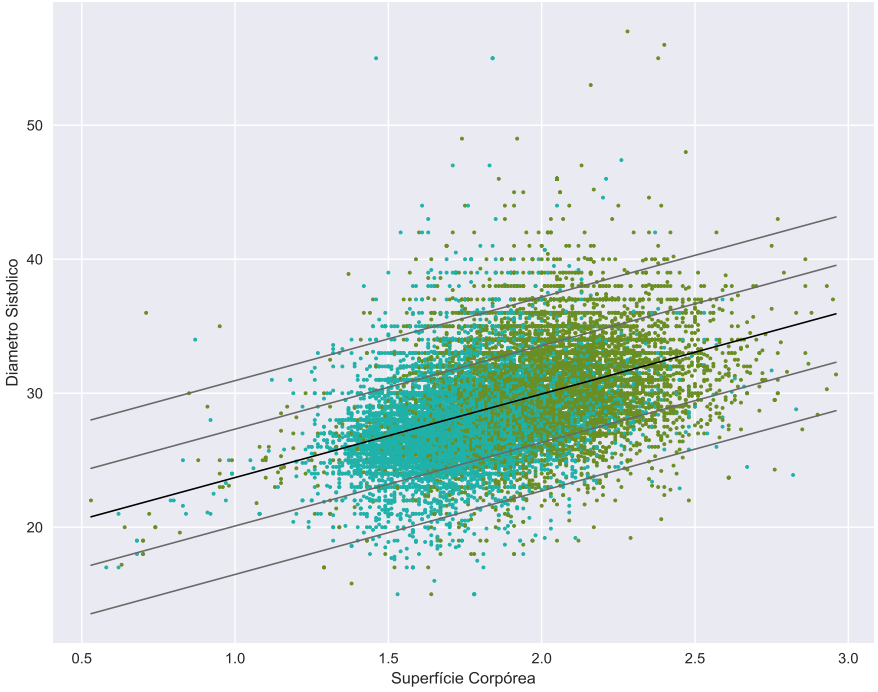

Figura 22 – *Scatter Plot*: diâmetro sistólico, em mm.

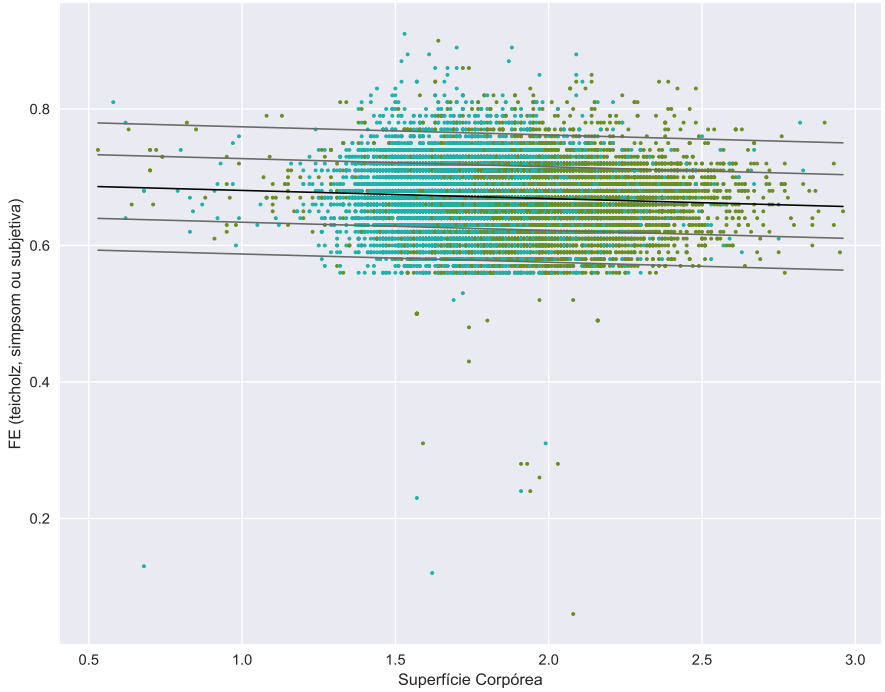

Figura 23 – *Scatter Plot*: FE, em mm.

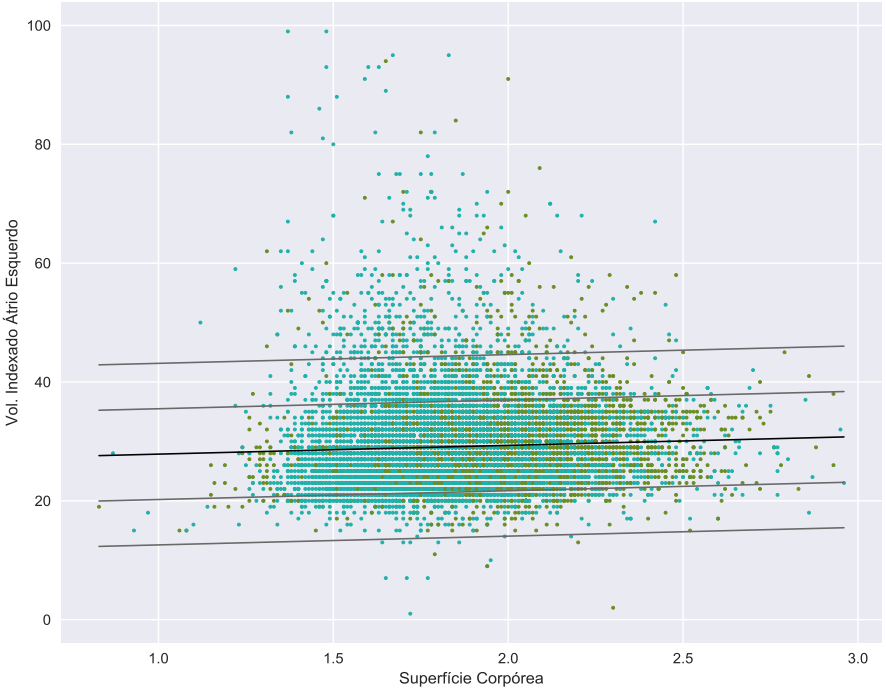

Figura 24 – *Scatter Plot*: volume indexado do átrio esquerdo, em mm.

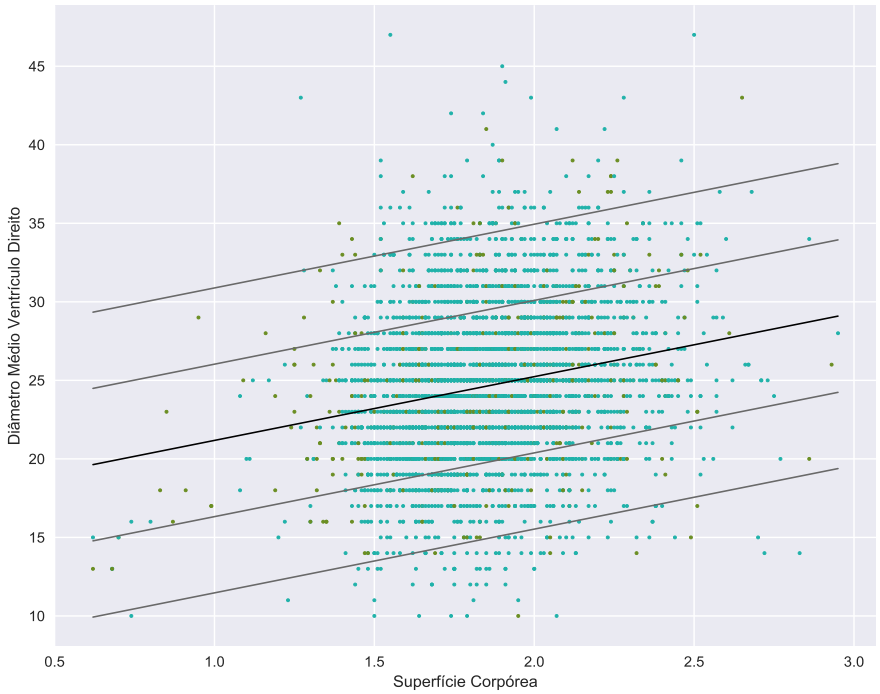

Figura 25 – *Scatter Plot*: diâmetro médio do ventrículo direito, em mm.

Figura 26 – *Scatter Plot*: diâmetro basal do ventrículo direito, em mm.

6 COMPARATIVO DE VALORES DE REFERÊNCIAS

A *American Society of Echocardiography* com apoio da *European Association of Cardiovascular Imaging* realizou um grande estudo para obtenção de valores normais para todas as quatro câmaras cardíacas, que serve como os valores de referências oficiais nos Estados Unidos e também no Brasil. Utilizaram como base vários bancos de dados, compilando e considerando o maior número possível de pacientes normais, garantindo um relatório bastante preciso. [61]

Após a pesquisa feita pelas duas instituições, o estudo de Roberto M. Lang et al. [61] buscou eliminar várias pequenas discrepâncias em relação às orientações publicadas anteriormente. O fato de usarem um maior número de banco de dados graças à tecnologia, possibilitou melhorar a confiabilidade dos valores de referência. Semelhante ao objetivo desta pesquisa, que também foi mostrado que existem muitas diferenças em relação aos valores americanos.

Na Tabela 2 e Tabela 3 são apresentados os comparativos entre os valores de referência brasileira, extraídas do banco de dados usado na pesquisa, e os valores americanos da Sociedade Americana de Ecocardiografia. Foram 20216 pacientes das regiões Sudeste e Sul, e 699 da região Nordeste e Norte. Na base de dados utilizada, não haviam clínicas e hospitais da região Centro-Oeste. Dentre os dados, é possível observar que existem diferenças entre todas as áreas do coração.

Originalmente os dados brutos somavam 24450 laudos, tendo em seu pré-processamento a remoção de 2215 laudos inválidos, 1110 pacientes apontados como doentes nos laudos gerados, por meio de expressão regular e também com a remoção de *outliers*, além de 105 laudos de pacientes com menos de 17 anos. Na Figura 27 é apresentado um fluxograma detalhado.

Os dados foram separados entre pacientes do sexo masculino e feminino, seguindo a forma de referenciar dos americanos, que também diferenciam e dão valores diferentes para os sexos. Assim como o Sistema EcoCloud segue, no cadastro do paciente é exigido que se preencha o sexo, e com isso, ao gerar o laudo final, de maneira dinâmica são entregues os valores de referência de acordo com o sexo. Os valores de ventrículo direito basal, ventrículo direito médio e átrio esquerdo (volume indexado), tem seus valores iniciais em zero nas referências americanas. Isso porque não existem um valor mínimo fora do normal.

No caso da referência brasileira, o sistema conseguiu entregar um valor mínimo, porém não necessariamente caracteriza anormal um valor abaixo do mínimo nesses três casos. Valores de referência normais para FE e também para ventrículo esquerdo (a partir da ecocardiografia bidimensional), têm sido constantemente atualizadas usando estudos

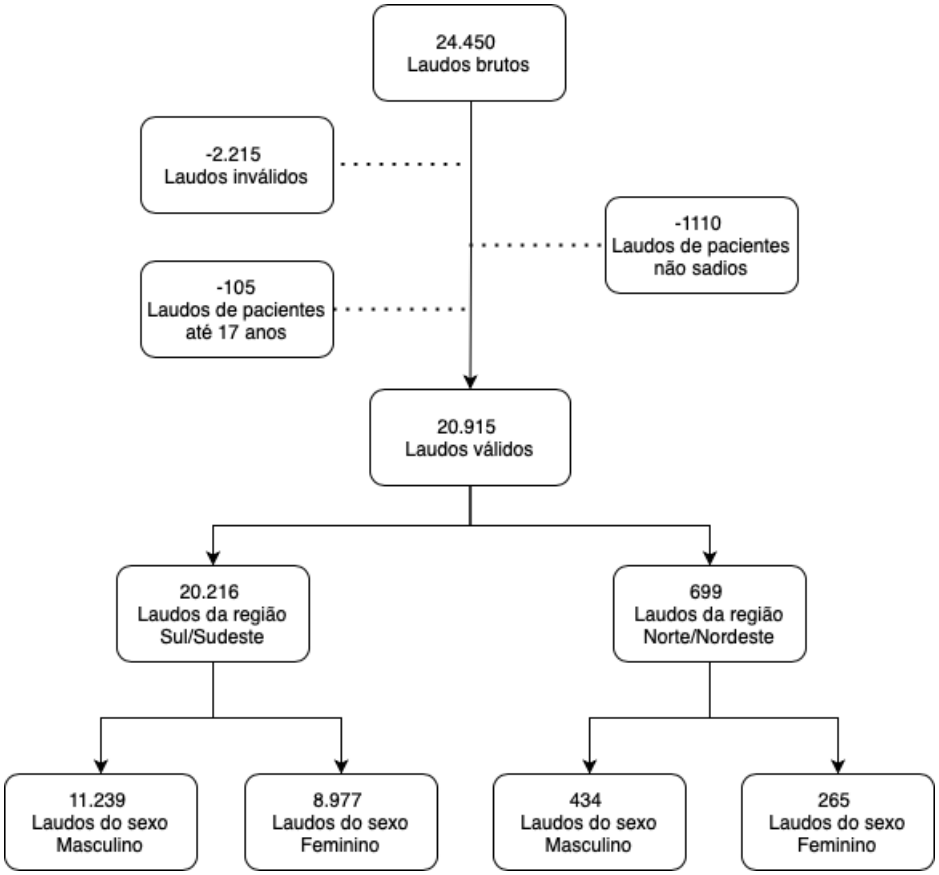

Figura 27 – Fluxograma detalhado do pré-processamento e limpeza de dados.

	Referência americana	**Referência brasileira**
Raíz aórtica	31-37 mm	29-38 mm
Átrio esquerdo	30-40 mm	32-42 mm
Díastole	42-58 mm	43-53 mm
Sístole	25-40 mm	26-34 mm
Septo intraventricular	6-10 mm	8-12 mm
Parede posterior	6-10 mm	7-11 mm
Fração de ejeção	0.52-0.72	0.62-0.71
Átrio esquerdo (Volume indexado)	<=34 mm	22-36 mm
Ventrículo direito basal	<=41 mm	28-41 mm
Ventrículo direito médio	<=35 mm	20-30 mm

Tabela 2 – Quadro comparativo entre valores de referências brasileiras extraídas do Eco-Cloud e referências americanas para pacientes de sexo masculino.

baseados em populações. A FE não necessariamente é somente relacionada ao sexo, idade e tamanho do indivíduo. Em paralelo, os valores para ecocardiografia tridimensional também têm seus valores publicados para diferentes populações étnicas [61], conforme mostra a Figura 28.

	Referência americana	**Referência brasileira**
Raíz aórtica	27-33 mm	26-34 mm
Átrio esquerdo	27-38 mm	29-39 mm
Díastole	38-52 mm	40-48 mm
Sístole	22-35 mm	24-31 mm
Septo intraventricular	6-9 mm	8-12 mm
Parede posterior	6-9 mm	7-11 mm
Fração de ejeção	0.54-0.74	0.62-0.72
Átrio esquerdo (Volume indexado)	<=34 mm	20-36 mm
Ventrículo direito basal	<=41 mm	27-40 mm
Ventrículo direito médio	<=35 mm	19-28 mm

Tabela 3 – Quadro comparativo entre valores de referências brasileiras extraídas do Eco-Cloud e referências americanas para pacientes de sexo feminino.

6.1 Utilização de *Z-Score* para análise

Simplificando, um *Z-Score* (também chamado de escore padrão) dá uma ideia de quão longe está o ponto médio de dados. Mais tecnicamente, é uma medida de quantos desvios padrões abaixo ou acima da população significam uma pontuação bruta. É uma maneira de comparar os resultados com uma população "normal". Dada essa premissa, que é utilizada também pela Sociedade Americana de Ecocardiografia por meio do documento de diretrizes e normas [61], foi utilizada nesse estudo também para análise e comparação de valores.

Assim como Michael D. Pettersen apresentou em seu trabalho [62], diagramas de dispersão foram usados para analisar dados, comparando valores de referências de determinada área do coração com a superfície corpórea dos pacientes. Ele coloca em discussão a importância da medição detalhada das estruturas cardíacas, sendo um aspecto crucial na gestão de crianças com vários problemas congênitos e doenças cardíacas adquiridas em crianças. As decisões sobre o tipo e o momento das intervenções muitas vezes dependem em grande parte das medições do coração [62].

O trabalho cita também a importância dessas medidas na prática clínica, e que existe uma preocupação, pois está faltando um conjunto grande de valores de referência derivados de uma grande parte de pacientes infantis em geral [62]. Muitos grandes laboratórios de ecocardiografia pediátrica têm desenvolvido seus próprios valores de referência para usar em seus relatórios ecocardiografia e se responsabilizam de sua tomada de decisão clínica.

	Aune et al., 2010	Fukuda et al., 2012	Chahal et al., 2012	Muraru et al., 2013
Number of subjects	166	410	978	226
Ethnic makeup of population	Scandinavian	Japanese	51% European White 49% Asian Indian	White European
EDVi, mL/m²				
Men, mean (LLN, ULN)	66 (46, 86)	50 (26, 74)	White: 49 (31, 67) Indian: 41 (23, 59)	63 (41, 85)
Women, mean (LLN, ULN)	58 (42, 74)	46 (28, 64)	White: 42 (26, 58) Indian: 39 (23, 55)	56 (40, 78)
ESVi, mL/m²				
Men, mean (LLN, ULN)	29 (17, 41)	19 (9, 29)	White: 19 (9, 29) Indian: 16 (6, 26)	24 (14, 34)
Women, mean (LLN, ULN)	23 (13, 33)	17 (9, 25)	White: 16 (8, 24) Indian: 15 (7, 23)	20 (12, 28)
EF, %				
Men, mean (LLN, ULN)	57 (49, 65)	61 (53, 69)	White: 61 (49, 73) Indian: 62 (52, 72)	62 (54, 70)
Women, mean (LLN, ULN)	61 (49, 73)	63 (55, 71)	White: 62 (52, 72) Indian: 62 (52, 72)	65 (57, 73)

BSA, body surface area; *EDVi*, left ventricular end-diastolic volume index; *EF*, left ventricular ejection fraction; *ESVi*, left ventricular end-systolic volume index; *LLN*, lower limit of normal; *NR*, not reported; *RT3DTTE*, real-time three-dimensional transthoracic echocardiography; *SVi*, left ventricular stroke volume index; *ULN*, upper limit of normal. LLN and ULN are defined as mean ± 2 standard deviations.

Figura 28 – Valores normais para parâmetros do ventrículo esquerdo obtidos com o ecocardiograma 3D. Fonte: *Chamber Quantification*, 2015.

6.2 Correlações entre FE e fatores climáticos

Existe um estudo que defende a ideia de que fatores geográficos e climáticos podem influenciar nos valores de referência do coração. No artigo de Jing Jang [20], ele mostra uma complexa relação entre a homeostase do corpo e ambiente geográfico. A homeostase é a habilidade de manter o meio interno em um equilíbrio quase constante, independentemente das alterações que ocorram no ambiente externo [63]. Ele defende fortemente que as correlações entre FE e ritmo cardíaco de sujeitos chineses estão diretamente relacionados a fatores ambientais geográficos, incluindo humidade relativa do ar, e quantidade de precipitação.

O estudo evidencia que ao sul da China, onde existem mais chuvas e menos calor, os valores de FE são maiores. E quanto mais ao norte da China, onde existem menos chuvas e é mais quente, os valores são menores. A Figura 29 ilustra de maneira clara o resultado do estudo. Assim como Jing Jang et al. [20] afirma, nesta pesquisa foi possível comprovar que os fatores geográficos influenciam nos valores de referência do coração. Usando a base de dados apresentada neste trabalho, e por meio da mineração de dados, foi possível agrupar laudos da região Sul/Sudeste e Norte/Nordeste, e evidenciar as diferenças, conforme mostra o gráfico da Figura 30.

Da mesma forma, no Norte e Nordeste do Brasil em que o calor é maior, os valores de FE tendem a ser menores que nas regiões Sul e Sudeste. A temperatura no Norte do Brasil varia entre 24º e 26 ºC, e no Nordeste entre 20º e 28 ºC no Verão. Enquanto no Sudeste do Brasil, a média anual é de 20ºC e no Sul a temperatura média anual está entre

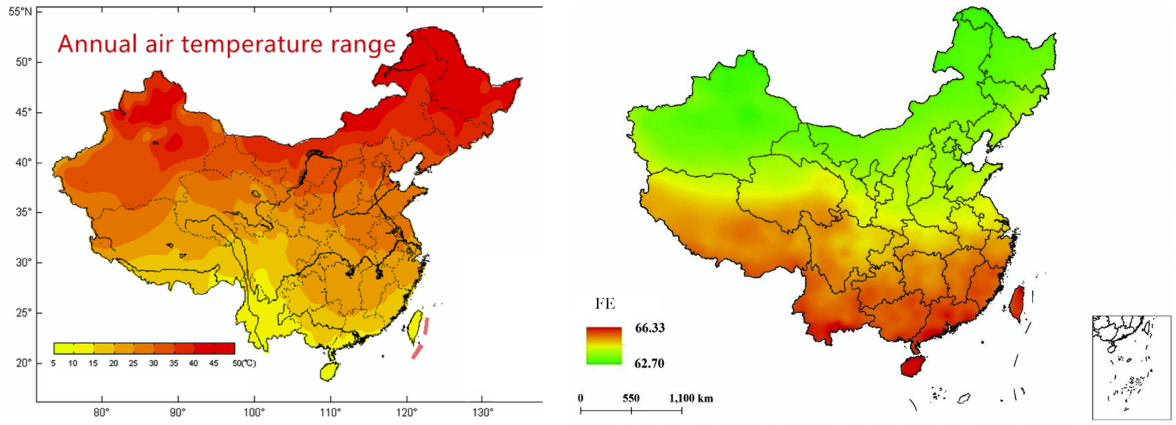

Figura 29 – Evidência de maior FE em pacientes nas regiões mais frias da China. Fonte: *A new method to get the lvef referencevalues of the healthy adult male by heart rate and geographical environment factors*, 2018.

	Sul/Sudeste	**Norte/Nordeste**
Raíz ártica	27-36 mm	26-34 mm
Átrio esquerdo	30-40 mm	28-39 mm
Díastole	41-50 mm	38-55 mm
Sístole	25-32 mm	22-37 mm
Septo intraventricular	7-11 mm	7-11 mm
Parede posterior	7-10 mm	7-11 mm
Fração de ejeção	0.62-0.71	0.60-0.73

Tabela 4 – Quadro comparativo entre valores de referências de pacientes das regiões Sul/Sudeste, com pacientes das regiões Norte/Nordeste.

14 e 22 °C. [64] Com resultados semelhantes ao trabalho de Jing Jang, os valores de FE em regiões mais frias da China (Sul), tendem a ser menores (média de 0,62), enquanto ao norte mais quente, a FE tem média de 0,66. Na mineração de dados desta pesquisa com o banco de dados utilizado, temos as regiões Sul/Sudeste que são mais frias com média de 0,67, enquanto nas regiões Norte/Nordeste que são mais quentes, com média de 0,66.

Analisando os resultados dos dois países, a pesquisa deste trabalho reforça o estudo feito na China de que o fator temperatura influencia na fração de ejeção do ventrículo esquerdo. Na Figura 30 é apresentada de maneira clara a confirmação de que no Brasil ocorre o mesmo fenômeno ao comparar com pesquisa feita na China. A Figura 29 apresenta em seu primeiro mapa da China que o Norte é mais quente que o Sul, entre os períodos de 1970 a 2000, data em que foram extraídos os dados. Hoje por conta da industrialização, a situação se inverteu, com o Sul mais quente. No segundo mapa mostra que a FE ao Sul é maior que a FE ao Norte da China.

Outros dados além da FE também tem valores diferentes quando comparamos pacientes do Norte com o Sul do Brasil, como mostra a Tabela 4. Existe também diferenças

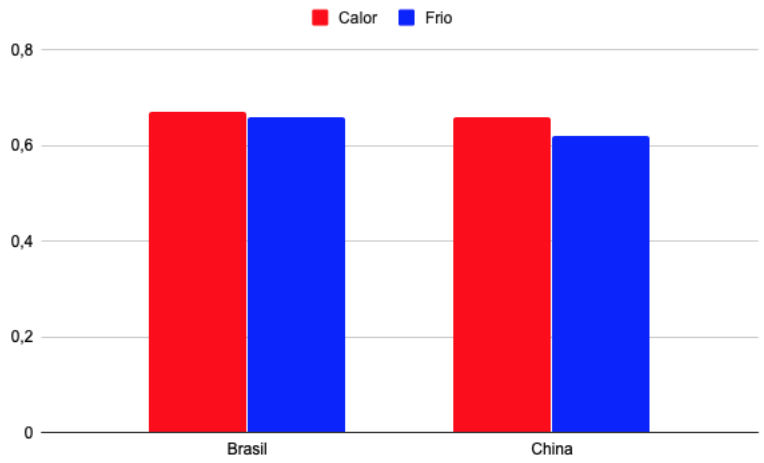

Figura 30 – Gráfico comparativo de valor médio da fração de ejeção nas regiões frias e quentes do Brasil e China.

	Sul/Sudeste	**Norte/Nordeste**
Peso	58-93	52-84
Altura	155-176	149-170
IMC	22,2-32,8	21,7-31,4
Superfície corpórea	1,61-2,11	1,47-1,99 mm

Tabela 5 – Quadro comparativo entre valores de dados antropométricos de pacientes das regiões Sul/Sudeste, com pacientes das regiões Norte/Nordeste.

nos valores referência para peso, altura, IMC e superfície corpórea, de acordo a Tabela 5. Fatores que interferem diretamente nos resultados do laudo médico, em que é exibido o índice de massa do paciente e servem para avaliação clinica.

O propósito principal dessa pesquisa não é exatamente mostrar que fatores externos influenciam nos parâmetros. Apesar disso, as evidências apresentadas servem para que em futuros estudos, possa ser considerada a região em que vive o paciente para determinar os valores de referência do coração, principalmente FE.

7 CONCLUSÃO

O presente documento mostra que a centralização de informações, aliada a análise de dados, tem um poder muito grande em entregar informações que antes era uma tarefa quase impossível na quantidade de laudos aqui apresentado. Essas informações são importantes para tomadas de decisões. As tecnologias utilizadas, dispositivos móveis, infraestrutura robusta como a *Amazon Web Services*, onde foi hospedada a aplicação, proporcionam ferramentas que podem, além de otimizar processos, entregar dados ricos em informações e abrangência, possibilitando uma análise em que se vê notoriamente que existem diferenças entre povos, quando se trata de medidas do coração.

Mesmo com a falta de centralização de informações de laudos no Brasil, e também da digitalização dos mesmos, o Sistema EcoCloud conseguiu reunir mais de 20 mil laudos de pacientes sadios, isolando pacientes que tenham algum indício de problema cardíaco que possa interferir nos resultados. Foram apresentados 10 resultados de medidas do coração brasileiro, além de dados antropométricos, com pacientes das regiões Sul, Sudeste, Nordeste e Norte e divididos por sexo.

Neste estudo, não pôde-se realizar com pacientes abaixo de 18 anos, por motivos mencionados. Também não houveram laudos de pacientes da região Centro-Oeste, por conta da ausência de médicos usuários do Sistema EcoCloud. Para projetos futuros, outras técnicas de mineração de dados podem ser aplicadas para que sejam gerados resultados alternativos, e, que seria possível realizar comparações de análises, obtendo valores ainda mais precisos ou outras conclusões, além de estudos mais aprofundados com dados de pacientes infantis. Estudos para detecção de doenças por meio dos novos valores de referência, poderiam ser feitos com o objetivo de obter maior assertividade na prevenção de doenças cardíacas. por meio de *Machine Learning*.

Como pontos fortes e vantagens do estudo, foi reforçada a evidência, desta vez brasileira, de que os fatores climáticos influenciam nos resultados, principalmente a fração de ejeção. Com isso, um projeto futuro em que os valores de referência possam ser dinâmicos conforme a região do paciente, seria um grande passo.

Também resultados de valores de referência de pacientes brasileiros foram comparados com os pacientes americanos, e mostraram que existem variações significativas. Os milímetros de diferença podem ser decisivos no diagnóstico mais preciso de doenças, e até salvar vidas. Levar este estudo e outros que reforcem a necessidade de utilização de parâmetros brasileiros até a Sociedade Brasileira de Cardiologia, é um passo importante para mudarmos a forma de laudar um exame de ecocardiograma no Brasil.

REFERÊNCIAS

[1] ALMEIDA, M. F. de. Descentralização de sistemas de informação e o uso das informações a nível municipal. *Inf. Epidemiol. Sus v.7 n.3 Brasília set. 1998*, 1998.

[2] SHAROVSKY, R. *Digital Echocardiographic laboratory: when and how to have it. Rev Bras Ecocardiogr.* 2008. 36-40 p.

[3] GOTTDIENER J. S., B. J. D. R. G. J. K. A. M. W. J. . . S. N. B. *American Society of Echocardiography recommendations for use of echocardiography in clinical trials: A report from the american society of echocardiography's guidelines and standards committee and the task force on echocardiography in clinical trials.* [S.l.]: Journal of the American Society of Echocardiography, 2004. 1086-1119 p.

[4] PELLIKKA P. A., N. S. F. E. A. A. K. C. A. . S. S. G. *American Society of Echocardiography recommendations for performance, interpretation, and application of stress echocardiography.* [S.l.]: Journal of the American Society of Echocardiography, 2007. 1021-1041 p.

[5] PICARD M. H., A. D. B. S. M. D. J. M. D. P. S. G. L. D. . . P. P. A. *American Society of Echocardiography recommendations for quality echocardiography laboratory operations.* [S.l.]: Journal of the American Society of Echocardiography, 2011. 1-10 p.

[6] CURCIN V., F. E. D. R. . C. D. *Templates as a method for implementing data provenance in decision support systems.* [S.l.]: Journal of biomedical informatics, 2017. 1-21 p.

[7] CHAUDHRY B., W. J. W. S. M. M. M. W. R. E. . . S. P. G. *Systematic review: impact of health information technology on quality, efficiency, and costs of medical care.* [S.l.]: Annals of internal medicine, 2006. 742-752 p.

[8] SUJANSKY, W. V. *The benefits and challenges of an electronic medical record: much more than a" word-processed" patient chart.* 1998.

[9] OTTO, C. M. *The practice of clinical echocardiography. Elsevier Health Sciences.* 2012.

[10] M WANG H, L. R. e. a. N. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death.

[11] LENTSCK, T. A. d. F. M. M. H. Internações por doenças cardiovasculares e a cobertura da estratégia saúde da família. 2015.

[12] SOUZA DEBORAH CARVALHO MALTA, E. B. F. M. L. B. Maria de Fátima Marinho de. Transição da saúde e da doença no brasil e nas unidades federadas durante os 30 anos do sistema Único de saúde. 2018.

[13] APPLICATION of appropriateness criteria in outpatient transthoracic echocardiography.

[14] QUALIDADE em cardioimagem: critérios de appropriateness aplicados à ecocardiografia.

[15] LIRA-FILHO SAMIRA MORHY, A. C. C. D. L. B. J. L. A. J. A. E. *Serviços de Ecocardiografia no Brasil: Uma Visão Geral*. 2014.

[16] HENRY M.D., J. W. P. J. M. G. M. S. I. H. M. J. M. R. W. L. Echocardiographic measurements in normal subject. 2015.

[17] ONLINE, R. *As principais diferenças culturais entre Brasil e Estados Unidos*. Tese (Doutorado), 2018. Disponível em: <https://www.remessaonline.com.br/blog/diferencas-culturais-entre-brasil-e-estados-unidos/>. Acesso em: 27.6.2020.

[18] BBC. *Brasileiro cresce em altura nos últimos cem anos*. Tese (Doutorado).

[19] PFAFFENBERGER PHILIPP BARTKO, A. G. E. P. J. B. E. L. D. B. H. B. G. M. S.; MASCHERBAUER, J. Size matters! impact of age, sex, height, and weight on the normal heart size. p. 6:1073–1079, 2013.

[20] A B, M. G. a. Z. Y. b. P. L. a. D. W. a. J. J. A new method to get the lvef reference values of the healthy adult male by heart rate and geographical environment factors. *International Journal of Gerontology*, 2018.

[21] GORUNESCU, F. Data mining: Concepts, models and techniques. 2011.

[22] EU- clid. Encyclopdia Britannica. 2010. <http://www.britannica.com/EBchecked/topic/194880/Euclid>. Accesado: 30/01/2020.

[23] HASTIE T.; TIBSHIRANI, R. . F. J. The elements of statistical learning: Data mining, inference, and prediction. springer, second edition. 2009.

[24] LEAMER EDWARD, E. Specification searches: cd hoc inference with nonexperimental data. *Wiley*, 1978.

[25] PIATETSKY-SHAPIRO, G.; FRAWLEY, W. Knowledge discovery in databases. *MIT Press*, 1991.

[26] FAYYAD, U. Data mining and knowledge discovery in databases: Implications for scientific databases. 1997.

[27] FERREIRA, E. A. L. Mineração de dados aplicada à dados médicos. UNIVERSIDADE FEDERAL DE MINAS GERAIS, 2015.

[28] P KAYYALI B, K. D. V. K. S. G. The big- data revolution in us health care: accelerating value and innovation. p. 1–19, 2013.

[29] ASSEMBLY, U. N. G. Prevention and control of non-communicable diseases - report of the secretary-general a/66/83. 2011.

[30] K., P. Better health care through data how health analytics could contain costs and improve care.

[31] MA, S. Y. M. Shortliffe eh. clinical decision- support systems. in: Biomedical informatics.shortliffe eh. p. 643–674, 2014.

[32] RECOVERY, A.; 2009, R. A. of. Title xiii of division a and title iv of division b. p. 112–382.

[33] CONSELHO FEDERAL DE MEDICINA - Resolução CFM No 1.821/2007.

[34] M, A. B. H. The hazards of data mining in healthcare. 2017.

[35] BUSINESS, P. S. *Modelos de Avaliação Preditiva na Saúde*. Tese (Doutorado), 2018. Disponível em: <https://saudebusiness.com/gestao/modelos-de-avaliacao-preditiva-na-saude/>. Acesso em: 25.6.2020.

[36] CIOS K.J., M. G. Uniqueness of medical data mining," artif intell med. p. 26(1–2), 1–24, 2002.

[37] CD, S. H. M. Foundations of statistical natural language processing. 2000.

[38] W., C. Medical natural language understanding as a supporting technology for data mining in healthcare. p. 32–60, 2000.

[39] C, H. G. F. Evaluating natural language processors in the clinical domain. p. 37:334–44, 1998.

[40] M., N. Machinetranslation. p. 898–902.

[41] GW, B. J. M. Anatomic pathology data mining. p. 61–108 [chapter 4], 2000.

[42] JM, S. Legal policy and security issues in the handling of medical data. p. 17–31, 2000.

[43] L, S. Computational disclosure control: a primer on data privacy protection. p. 17–31, 2001.

[44] US Veterans Administration Co-operative Urological Research Group. Treatment and survival of patients with cancer of the prostate. The Veterans Administration Co-operative Urological Research Group. Surg Gynecol Obstet. p. 124(5):1011–17.

[45] EG GRAY R, S. A. O. C. T. D. G. K. C. M. F. G. M. Efficacy of adjuvant chemotherapy in high-risk node-negative breast cancer: an intergroup study. p. 320(8):485–9, 1989.

[46] MG KNATTERUD GL, P. T. G. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. 3. clinical implications of ugdp results. p. 218(9):1400–10.

[47] HC FITZSIMMONS SC, A. D. K. M. R. B. C. P. F. P. L. Risk of persistent growth impairment after alternate-day prednisone treatment in children with cystic fibrosis. p. 342(12):851–9, 2000.

[48] US Food and Drug Administration. Delegations of authority and organization; Office of the Commissioner—FDA. p. 56(225):58758, 1991.

[49] APPLE, S. *About the privacy and security of your health records*. Tese (Doutorado), 2020. Disponível em: <https://support.apple.com/en-us/HT209519>. Acesso em: 25.6.2020.

[50] HEALTHCARE, D. *Google in healthcare: Data privacy and cybersecurity concerns*. Tese (Doutorado), 2020. Disponível em: <https://blog.definitivehc.com/google-in-healthcare-data-privacy-and-cybersecurity-concerns>. Acesso em: 25.6.2020.

[51] ALVAREZ ANTONIA ARHNEGA, R. S. J. A. C. L. The quantitative anatomy of the normal human heart in fetal and perinatal life. *Internattonal Journal of Cardiology*, 1987.

[52] TONG, Y. The multivariate normal distribution.

[53] GALARNYK, M. Understanding boxplots. In: _____. [s.n.], 2018. Disponível em: <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>. Acesso em: 4.2.2020.

[54] PETENATE, M. *BoxPlot: Saiba tudo sobre o Diagrama de caixa e como interpretar esse gráfico*. Tese (Doutorado), 2019. Disponível em: <https://www.escolaedti.com.br/o-que-e-um-box-plot>. Acesso em: 9.3.2020.

[55] SAúDE, O. M. da. 1995, 1997.

[56] NP WITTE KK, T. S. d. S. R. C. A. C. J. N. Longitudinal ventricular function: normal values of atrioventricular annular and myocardial velocities measured with quantitative two dimensional color doppler tissue imaging. p. 16:906–21, 2003.

[57] GO REICHEK N, B. D. D. P. B. Effects of sample volume location, imaging view, heart rate and age on tricuspid velocimetry in normal subjects. p. 65:1026–30, 1990.

[58] P ESPOSITO R, O. M. N. S. G. M. I. The impact of ageing on right ventricular longitudinal function in healthy subjects: a pulsed tissue doppler study. p. 10:491–8, 2009.

[59] ZOGHBIWA HABIB GB, Q. M. Doppler assessment of right ventricular filling in a normal population. comparison with left ventricular filling dynamics. p. 82:1316–24, 1990.

[60] AL., L. G. R. et. Diretrizes para avaliação ecocardiográfica do coração direito em adultos: um informe da sociedade americana de ecocardiografia. p. 23:685–713, 2010.

[61] LANG LUIGI P. BADANO, J. A. e. a. R. M. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Journal of the American Society of Echocardiography*, 2015.

[62] PETTERSEN MD, W. D. P. M. E. S. M. M. D.; HUMES, M. R. A. Regression equations for calculation of z scores of cardiac structures in a large cohort of healthy infants, children, and adolescents: An echocardiographic study. 2008.

[63] HOMEOSTASE. 2020. <https://brasilescola.uol.com.br/biologia/homeostase.htm>. Accesado: 13/01/2020.

[64] CIENTíFICO, C. *Climas do Brasil: conheça o clima típica de cada região brasileira*. Tese (Doutorado), 2018. Disponível em: <https://conhecimentocientifico.r7.com/climas-do-brasil-conheca-o-clima-tipica-de-cada-regiao-brasileira/>. Acesso em: 26.6.2020.

Anexos

.1 EcoCloud

EcoCloud é um sistema que tem como objetivo principal otimizar e automatizar o fluxo de geração de laudo ecocardiograma. Além disso, possui ferramentas de auxílio, como agenda online compartilhada com a secretária, aplicativos capazes de diagnosticar, gerar laudos e enviar por e-mail, aplicativo para médico solicitante que recebe em tempo real um laudo solicitado e finalizado, além de outras funcionalidades. O sistema é usado por vários médicos em hospitais, clínicas e unidades de pronto atendimento do Paraná e Brasil.

Ao longo desse período, gerou um banco de dados de mais de 20 mil laudos de diferentes pacientes contendo várias características. Os resultados demonstraram que o sistema é uma alternativa viável para os cardiologistas terem agilidade na assistência ao paciente e assistência no processo de tomada de decisão. As metodologias que são usadas para preparar os relatórios no sistema são baseadas em alguns trabalhos propostos na literatura científica [3, 4].

De acordo com a American Society of Ecocardiography (ASE) [4, 5], é recomendado que todos os relatórios ecocardiográficos sigam um padrão de apresentação uniforme e uma linguagem comum que inclui dados dos principais elementos das estruturas e medidas cardíacas, apresentando relatórios laboratoriais de forma semelhante em estrutura e escrita.

.1.1 Estudo de viabilidade

Médicos envolvidos participaram de todo o processo de planejamento do projeto, que, por meio dos seus conhecimentos das rotinas de atendimento aos pacientes, foi feito um estudo de viabilidade, em que o objetivo foi elaborar um fluxo condizente com o processo de medição de áreas do coração e exibir uma visualização de tela agradável, limpa e intuitiva. Inicialmente foram listados os objetivos principais do projeto, analisando quais atividades poderiam ser realizadas de maneira independente e por diferentes usuários além do médico, e depois disso foram isoladas.

Após essa listagem, foram denominados quais os módulos mais são os mais importantes, a fim de que tenham locais de acesso privilegiados na resultado final das telas. Foi criada a identidade visual do sistema, seu protótipo e apresentação. Nesse momento os médicos envolvidos tiveram um papel crucial na validação. Após várias simulações e mudanças, por fim o protótipo é encaminhado para o desenvolvimento front-end. A Figura 31 mostra o fluxo de como foi feito o processo de maneira macro.

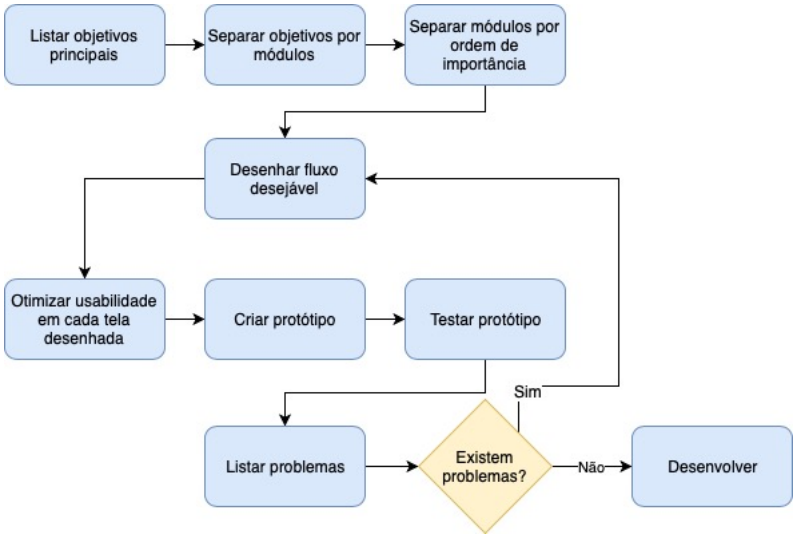

Figura 31 – Diagrama de planejamento e execução de estudo de usabilidade.

.1.2 Resultado de otimização de processo

Como resultado, o sistema se mostrou eficaz, por conta das várias automatizações e frases pré-determinadas, além da geração automática de laudo. A Figura 32 mostra um comparativo entre o método tradicional, feito de maneira manual, envolvendo mais pessoas e tempo, e o método utilizando o sistema.

.1.3 Cálculos

Existem no projeto, cálculos específicos para que seja possível alcançar valores necessários para apresentação do laudo final. A Figura 33 mostra a planilha utilizada para chegar ao valor de fração de ejeção (percentual de sangue que o ventrículo ejeta para a aorta na sístole), em que, por meio de correções da fórmula de cubo, o volume diastólico final do ventrículo esquerdo é calculado por meio do diâmetro diastólico do ventrículo esquerdo. Além do cálculo da fração de ejeção com o método Teicholz, é possível como alternativa a inserção da FE (fração de ejeção) utilizando outros dois métodos: Subjetiva e Simpson. Esses dados também são utilizados na análise.

Existe também no projeto outros cálculos automáticos, como a porcentagem de encurtamento, índice de massas do ventrículo esquerdo, espessura relativa, e superfície corpórea. Todos eles relacionados aos dados antropométricos do paciente e seu sexo. Seus intervalos de referência também variam de acordo com a idade e sexo do paciente.

.1.4 Preenchimento automático e manual

Outros campos de cálculos também podem ser inseridos, porém não são automatizados, pois exigem do médico outras técnicas para chegar aos valores e também histórico clínico do paciente, como gradiente diastólico máximo e médio pela válvula mitral, área

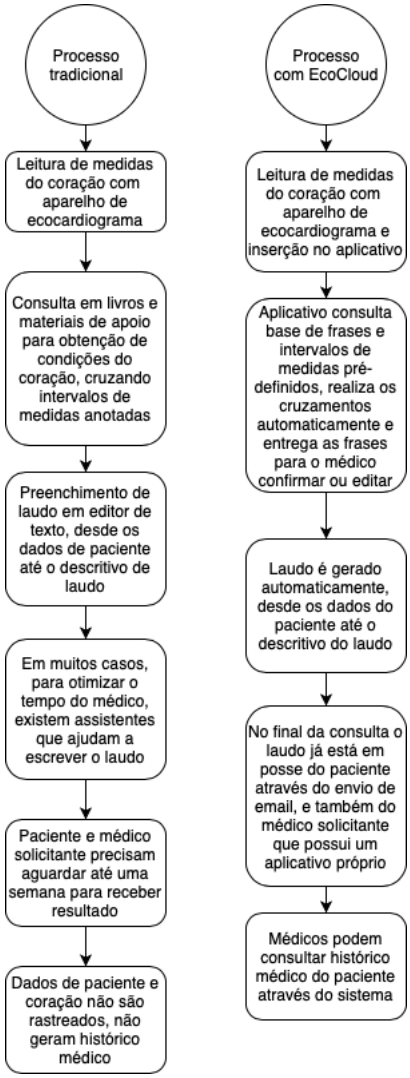

Figura 32 – Comparação de método tradicional X método com EcoCloud.

fx =((POWER(B8;3)*7)/(2,4+(B8/10))-(POWER(B9;3)*7)/(2,4+(B9/10)))/((POWER(B8;3)*7)/(2,4+(B8/10)))

	A	B	C	D	E	F	G	H
1	Parâmetros			V.N.	Parâmetros			V.N.
2	Sexo	F						
3	Peso	46	kg		FE (Teichholz)	67%		(>55%)
4	Altura	160	cm		% de encurtamento	37%		(30-40%)
5	Raiz Aórtica	32	mm	(21-37 mm)				
6	AE - Diâmetro AP	10	mm	(28-40 mm)	Massa do VE	92,77	g	H <224g
7	AE - Volume indexado		ml/m2	(<34 ml/m2)				M <162g
8	DDVE	38	mm	(38-54 mm)	Índice de Massa do VE	64,12	g/m2	H <115
9	DSVE	24	mm	(26-34 mm)				M <95
10	Septo IV	9	mm	(7-11 mm)	Espessura relativa	0,42		<0,42
11	Parede posterior VE	8	mm	(7-11 mm)				
12	VD	-	mm	(10-26 mm)	Superfície Corpórea	1,45	m2	
13	DDVE = Diametro Diastolico							
14	DSVE = Diametro Sistolico							
15	AE = Átrio Esquerdo							
16	VD = Ventrículo Direito							

Figura 33 – Planilha para cálculo de valores.

valvar mitral pelo PHT e planimetria, vena contracta do refluxo mitral, EROA, volume e fração regurgitante.

Além de valores, o médico pode também verificar situações e simplesmente assinalar o sintoma ou situação. Situação das funções do coração, como contratilidade segmentar anormal, contratilidade ventricular direita diminuída, situação da função diastólica, hipocinesia, acinesia e discinesia alteradas.

O modelos de laudo se dividem em quatro, atualmente. Os dois primeiros modelos tem caráter automatizado, em que, ao inserir os valores de medidas, o médico, teoricamente, tem a possibilidade de gerar o laudo com o sistema realizando todos os cálculos necessários. Normalmente essa situação se dá quando o paciente é saudável. Já em casos de problemas cardíacos, principalmente avançados, existe a necessidade do médico detalhar melhor a situação do coração. Esses são os laudos que serão utilizados para esta pesquisa, pois os valores são rastreáveis, a nível de banco de dados, e existem relacionamentos para o cruzamento de informações.

Já os outros dois laudos possuem frases de diagnóstico pré-determinadas, porém elas devem ser selecionadas após as impressões diagnósticas do médico. Esses modelos de laudos não serão utilizados nesta pesquisa, por conta da alta personalização dos textos por parte dos médicos, impossibilitando rastreabilidade e tratamento das informações do banco de dados.

Após mais de um ano de desenvolvimento, o EcoCloud foi implantado para um período de teste em hospitais, clínicas e unidades de pronto atendimento do Paraná. Testes foram feitos durante o segundo ano. Desde então, o sistema tem sido usado por vários médicos para examinar pacientes reais e dar sugestões sobre o desempenho do software para seus desenvolvedores. No entanto, durante a maior parte deste primeiro ano, o EcoCloud foi usado com uma dupla verificação.

Isso significa que, além do software, os médicos também usaram seu modo preferido de realizar o diagnóstico, com o objetivo de detectar possíveis imprecisões nos relatórios do sistema, assim como mostra a Figura 34.

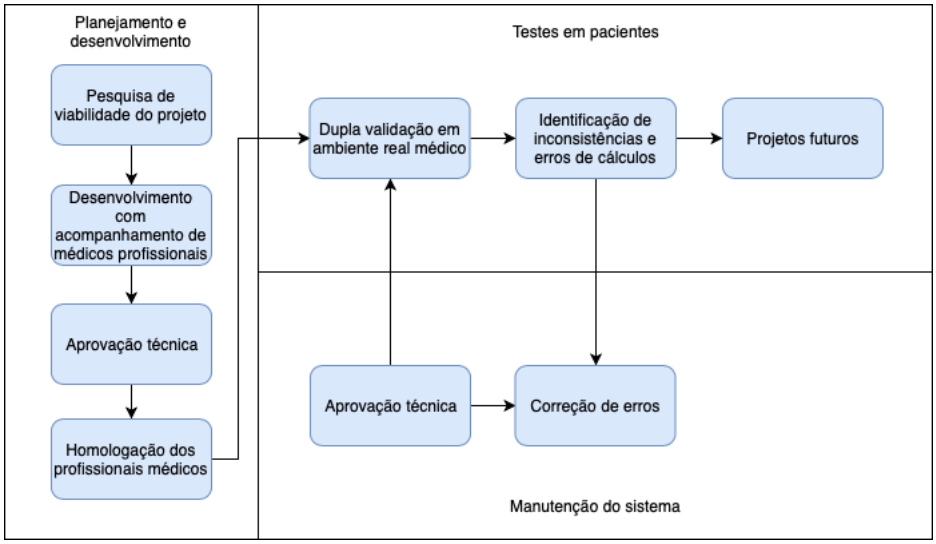

Figura 34 – Metodologia aplicada para o desenvolvimento do projeto.

.2 Principais tabelas utilizadas no sistema

Abaixo são apresentadas as quatro tabelas, que a partir delas foram extraídos todos os resultados desta pesquisa. Na tabela da Figura 35 são extraídas todas as medidas do coração, além de associações para as outras tabelas de médico (Figura 37) e estado (Figura 38), que por sua vez guarda a região do país em que a consulta foi realizada, e a tabela do paciente (Figura 36), que tem a função de de gravar os dados antropométricos.

Column	Type	Default Value	Nullable
relatorio_id	int(11)		NO
medico_id	int(11)		NO
consulta_id	int(11)		NO
peso	decimal(10,2)		YES
altura	decimal(10,2)		YES
imc	decimal(10,2)		YES
superficie_corporea	decimal(10,2)		YES
raiz_aortica	decimal(10,2)		YES
atrio_esquerdo	decimal(10,2)		YES
septo_interventricular	decimal(10,2)		YES
parede_posterior	decimal(10,2)		YES
diametro_diastolico	decimal(10,2)		YES
diametro_sistolico	decimal(10,2)		YES
vol_indexado_atrio_esquerdo	decimal(10,2)		YES
diametro_basal_ventriculo_direito	decimal(10,2)		YES
diametro_medio_ventriculo_direito	decimal(10,2)		YES
FE_teicholz	decimal(10,2)		YES
FE_simpson	decimal(10,2)		YES
FE_subjetiva	decimal(10,2)		YES
genero	int(11)		YES
atrio_direito_dilatado	tinyint(4)		YES
ventriculo_direito_dilatado	tinyint(4)		YES
data_cadatro	timestamp	CURRENT_TIMESTAMP	YES
contratilidade_segmentar_normal	tinyint(4)		YES
tipo_laudo	int(11)	1	YES

Figura 35 – Tabela principal em que são guardadas as medidas do coração do paciente.

Column	Type	Default Value	Nullable	Extra
paciente_id	int(11)		NO	auto_increment
nome	varchar(45)		YES	
email	varchar(45)		YES	
sexo	varchar(45)		YES	
nascimento	date		YES	
peso	int(11)		YES	
altura	decimal(3,2)		YES	
cns	varchar(18)		YES	
cpf	varchar(45)		YES	
telefone	varchar(45)		YES	
senha	varchar(70)		YES	
token	varchar(100)		YES	

Figura 36 – Tabela "paciente" em que são guardados os dados antropométricos.

Column	Type	Default Value	Nullable	Extra
medico_id	int(11)		NO	auto_increment
nome	varchar(70)		NO	
sobrenome	varchar(70)		NO	
crm	varchar(45)		NO	
email	varchar(150)		NO	
senha	varchar(70)		NO	
estado_id	int(11)		NO	
especialidade_id	int(11)		NO	
endereco	varchar(250)		YES	
data_cadastro	timestamp	CURRENT_TIMESTAMP	NO	
sign	text		YES	
token	varchar(100)		YES	
pagamento_id	int(11)		YES	
pagamento_status	tinyint(1)	0	YES	
require_cns	tinyint(4)	0	YES	
require_email	tinyint(4)	0	YES	
require_medico_solicitante	tinyint(4)	0	YES	
require_medico_email_solicitante	tinyint(4)	0	YES	
genero	tinyint(4)		YES	
logado	tinyint(4)	0	YES	
laudo_selecionado	tinyint(4)	0	YES	
require_cpf	tinyint(4)	0	YES	
require_convenio	tinyint(4)	0	YES	
is_trial	tinyint(4)		YES	
env	varchar(45)		YES	
data_inscricao	date		YES	
metodo_inscricao	varchar(45)		YES	
carimbo	text		YES	
telefone	varchar(45)		YES	
assinatura_status_label	varchar(150)		YES	
ultima_cobranca	date		YES	
first_laud	tinyint(4)	1	YES	
subscribe_mail	tinyint(4)	1	YES	
is_residente	tinyint(1)		YES	
cabecalho_pdf	tinyint(4)	1	YES	
show_relatorio	tinyint(4)	0	YES	
pagarme_status	tinyint(4)	0	YES	

Figura 37 – Tabela "medico", em que é extraída localidade da consulta realizada.

Column	Type	Default Value	Nullable	Extra
estado_id	int(11)		NO	auto_increment
estado	varchar(75)		YES	
uf	varchar(5)		YES	

Figura 38 – Tabela "estado", em que é extraída a localidade da consulta realizada.

TRABALHOS PUBLICADOS PELO AUTOR

Trabalhos publicados pelo autor durante o programa.

Publicações principais do trabalho.

1. Renato Kuroe, Luiz Rodrigues, Robson Bonidia, Danilo Sanches, Jacques Brancher, Fabrício Furtado, Willyan Nazima. **EcoCloud: A Specialized Computer System for Elaboration Echocardiography Reports**, Twenty-fourth Americas Conference on Information Systems, New Orleans, 2018.

Publicações complementares.

1. Renato Kuroe, Marcelo Pereira da Silva, Jacques Duílio Brancher. **Rastreabilidade de Requisitos Usando Ferramenta de BI**, SETII - Seminário em Tecnologia da Informação Inteligente, UNINOVE, São Paulo-SP, 2017.