**Universidade do Estado do Rio de Janeiro**
Centro Biomédico
Instituto de Biologia Roberto Alcântara Gomes

Thiago da Silva Pereira de Souza

# Desenvolvimento de ferramenta para auxiliar a determinação de intervalo de referência em parâmetros laboratoriais a partir de um banco de dados de grande porte

Rio de Janeiro
2020

Thiago da Silva Pereira de Souza

# Desenvolvimento de ferramenta para auxiliar a determinação de intervalo de referência em parâmetros laboratoriais a partir de um banco de dados de grande porte

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Saúde, Medicina Laboratorial e Tecnologia Forense, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Luís Cristóvão de Moraes Sobrino Pôrto

Rio de Janeiro

2020

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CB-A

S729 Souza, Thiago da Silva Pereira de
Desenvolvimento de ferramenta para a determinação de intervalo de referência de parâmetros laboratoriais a partir de um banco de dados de grande porte / Thiago da Silva Pereira de Souza - 2020.
65f.

Orientador: Prof. Dr. Luis Cristóvão de Moraes Sobrino Pôrto

Mestrado (Dissertação) - Universidade do Estado do Rio de Janeiro, Instituto de Biologia Roberto Alcântara Gomes. Pós-graduação em Saúde, Medicina Laboratorial e Tecnologia Forense.

1. Laboratórios Clínicos – Teses. 2. Valores de Referência. 3. Técnicas de Laboratório Clínico. 4. Técnicas e procedimentos diagnósticos. I. Pôrto, Luis Cristóvão de Moraes Sobrino. II. Universidade do Estado do Rio de Janeiro. Instituto de Biologia Roberto Alcântara Gomes. III. Título.

CDU 616-074

Bibliotecária: Ana Rachel Fonseca de Oliveira
CRB7/6382



_____ _____

Assinatura Data

Thiago da Silva Pereira de Souza

**Desenvolvimento de ferramenta para auxiliar a determinação de intervalo de referência em parâmetros laboratoriais a partir de um banco de dados de grande porte**

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Saúde, Medicina Laboratorial e Tecnologia Forense, da Universidade do Estado do Rio de Janeiro.

Aprovada em 17 de novembro de 2020.

Banca Examinadora: _____

Prof. Dr. Luís Cristóvão de Moraes Sobrino Pôrto (Orientador)

Instituto de Biologia Roberto Alcântara Gomes – UERJ

_____

Prof. Dr. Alexandre da Costa Sena

Instituto de Matemática e Estatística – UERJ

_____

Prof.ª Dra. Maria Fernanda Miguens Castelar Pinheiro

Instituto Estadual de Diabetes e Endocrinologia

Rio de Janeiro

2020

## DEDICATÓRIA

Dedico esse presente trabalho à minha filha Lara, que esta ferramenta seja um objeto de uso para melhor servir a população. Ela que fez eu entender o momento, o valor da oportunidade, o valor do segundo.

A você, Lara, todo o meu amor.

## AGRADECIMENTOS

A Deus pela força, ânimo, desmedido cuidado e atenção a minha vida, sempre foi e será a lâmpada para os meus pés e é luz para o meu caminho.

Aos meus pais por sempre estarem disponíveis para me ajudar, são anjos enviados na minha vida onde, não importa o onde e o quando, sempre se abrem a ajudar.

A minha esposa e meus filhos pela compreensão, pela ausência nas noites e dias, de semana a final de semana, com uma mão auxiliadora em ânimo e força, seu café a minha mesa sempre será lembrado.

Ao Dr. Izidro Bendet e a Dra Fernanda Pinheiros, meus mentores no conhecimento científico, meu primeiro contato, o primeiro passo na ciência da saúde.

A Dra. Monica Di Calafiori Freire por me dar esse desafio e proposta científica.

Ao professor Dr. Cristóvão Pôrto no qual me proporcionou aquecidos debates técnicos em medicina, saúde, estatística, parafraseando Newton "*Se cheguei até aqui foi porque me apoiei no ombro dos gigantes"*

Ao professor Dr. Alexandre Senna, que me esclareceu diversos aspectos da estatística aplicada à medicina, talvez não conseguisse dar o passo seguinte se não fosse pela nossa conversa.

Ao professor Dr. Paulo Telles pelos primeiros passos a esclarecer a aplicação metodológica da estatística à saúde.

Ao Guilherme Van der Velde por ver meu potencial e permitir abrir mão de horários de trabalho para que conseguisse realizar as aulas presenciais, foi muito além de chefe, um líder.

Ao professor Armênio Cardoso, onde este me proporcionou a maior gama de conhecimento e especialidades na área de tecnologia da informação.

A todos os docentes do curso de mestrado, foram maravilhosos todos os conhecimentos passados.

À equipe da secretaria Programa de Saúde, Medicina Laboratorial e Tecnologia Foresnse, por nunca medir esforços para me ajudar ao me passar informações importantes.

Vocês fizeram tudo acontecer!

Saruman acredita que apenas um grande poder pode manter o mal sob controle. Mas não foi isso que descobri. Eu descobri que são as coisas pequenas, os feitos diários de pessoas comuns é que mantêm o mal afastado. Simples atos de bondade e amor.

*Gandalf, the White. The Hobbit – Lord of the Rings*

## RESUMO

SOUZA, Thiago da Silva Pereira de. **Desenvolvimento de ferramenta para auxiliar para a determinação de intervalo de referência de parâmetros laboratoriais a partir de um banco de dados de grande porte**. 2020. 65 f. Dissertação (Mestrado em Saúde, Medicina Laboratorial e Tecnologia Forense) – Instituto de Biologia Roberto Alcântara Gomes, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2020.

A determinação dos intervalos de referências (IR), para cada análito, provém do estudo na população para cada situação fisológica ou patológica, em seus vários estágios, e é de grande importância para a tomada de decisão médica. Este estudo tem por objetivo avaliar a possibilidade da determinação de intervalos de referência (IR) para exame clínico laboratorial, utilizando metodologia indireta a partir de um banco de dados de grande porte com os recursos da tecnologia da informação. A metodologia direta, apesar de apresentar resultados mais confiáveis, é um método mais oneroso e complexo do que a metodologia indireta, exigindo o recrutamento, avaliação clínica e custo de materiais com a realização dos exames nas situações controladas. A possibilidade de avaliar o IR utilizando a metodologia indireta, permite que exames de rotina onde contemplam (que contemplem) tão somente o laudo de paciente, pertença (contenha) a um *pool* de informações onde o dado se torna informação, de forma muito mais rápida e econômica, possibilitando o aperfeiçoamento nos critérios clínicos utilizado para diagnóstico, métodos de estimativa indireta (retrospectiva), que calculam o IR a partir de informações existentes em banco de dados são uma alternativa a metodologia direta. Utilizando o grande volume de dados armazenados no banco de dados do laboratório clinico, *Big Data*, podemos realizar atividades de extração, mineração, onde conseguimos, por parâmetros ajustados, filtrar a população clinicamente saudável, curadoria, permitindo trabalhar os dados subsequentes somente de uma população onde (que) obedeçam ao critério de restrição. As amostras selecionadas foram randomizadas e todas realizadas na mesma plataforma, Cobas 6000 Roche, no período completo de 2018, ambos sexos e abrangendo de zero a 99 anos de idade como amostragem de validação do modelo estatístico e critério de exclusão. Em relação a metodologia estatística, é realizado o teste homogeneidade das variâncias a partir do teste P de Shapiro, posteriormente avaliado a relação do analito em seus diversos sistemas de origem para verificar se há diferença partir do teste P de Kolmogorov-Smirnov. Em seguida procedeu a exclusão dos extremos (*outliers*) tomando como valores acima e abaixo de 3 desvios padrões, por fim aplicou-se o modelo estatístico de Bhattacharya para obter o valor de referência. Em todos os testes, o nível de significância adotado foi de 5%. As análises estatísticas foram realizadas utilizando o sistema operacional Linux, distribuição Mint 18.1, linguagem de programação R e banco de dados PostgreSQL. Os valores encontrados para o Glicose total, nos percentis 75 e 95 foram: 73 a 108 mg/dL, respectivamente. Para insulina, nestes percentis foram de 1.2 a 21 mcU/mL, respectivamente. Os valores dos parâmetros aqui avaliados, definidos em diferentes faixas etárias de brasileiros da cidade do Rio de Janeiro, podem representar limites de decisão para a população brasileira contribuindo para aprimorar o diagnóstico em nosso país.



Palavras-chave: Intervalo de referência. Valores normais. Laboratório clínico. Normalidade. Intervalos de significância.

## ABSTRACT

SOUZA, Thiago da Silva Pereira de. **Development of a tool to evaluate the determination of reference range in laboratory parameters from a large database**. 2020. 65 f. Dissertação (Mestrado em Saúde, Medicina Laboratorial e Tecnologia Forense) – Instituto de Biologia Roberto Alcântara Gomes, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2020.

The determination of the reference interval (IR) comes from an extensive population study, however it is of great importance for medical decision-making. This study aims to evaluate the possibility of determining reference intervals (RI) for clinical laboratory examination, using indirect methodology from a large database with information technology resources. The direct methodology, despite presenting more reliable results, is a more costly and complex method than the indirect methodology, requiring recruitment, clinical evaluation and cost of materials with the patient. The possibility of evaluating the RI using the indirect methodology, allows routine exams where only the patient's report is considered, to belong to a pool of information where the data becomes information, much more quickly and economically, enabling improvement in clinical criteria used for diagnosis, indirect (retrospective) estimation methods, which calculate the RI from information in the database, are an alternative to the direct methodology. Using the large volume of data stored in the database of the clinical laboratory, Big Data, we can perform extraction activities, Data Mining, where we managed, by adjusted parameters, to filter the clinically healthy population, Data Cleanring, allowing to work the subsequent data only from a population where they meet the restriction criterion. This study is framed in Law No. 12,527 / 2011, which seeks to conduct the research in a way that does not reveal the patient's identification and make it anonymous, no data beyond what was necessary was and will be used in order to individualize the patient, allowing the use of given under the cover of the law. The selected samples were randomized and all performed on the same platform, Cobas 6000 Roche, in the full period of 2018, both sexes and covering from zero to 99 years of age as a validation sample of the statistical model and exclusion criteria. Regarding the statistical methodology, the homogeneity test of variances is performed using the Shapiro P test, subsequently evaluating the relationship of the analyte in its various systems of origin a to verify if there is a difference from the Kolmogorov-Smirnov test, then proceeded to the exclusion of extremes (outlies) taking as values above and below 3 standard deviations, finally the statistical model of Bhattacharya was applied to obtain the reference value. In all tests, the level of significance adopted was 5%. Statistical analyzes were performed using the Linux operating system, Mint 18.1 distribution, R programming language and PostgreSQL database. The values found for total glucose, in the 75th and 95th percentiles were: 73 to 108 mg / dL, respectively. For insulin, in these percentiles were 1.2 to 21 mcU / mL, respectively. The values of the parameters evaluated here, defined in different age groups of Brazilians in the city of Rio de Janeiro, may represent decision limits for the Brazilian pediatric population, contributing to improve the diagnosis in our country.



Keywords: Reference range. Normal values. Clinical laboratory. Normality. Significance intervals.

## LISTA DE FIGURAS

## LISTA DE GRÁFICOS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| bhat | *Bhattacharya* |
| BPLC | Boas Práticas em Laboratório Clínico |
| CLSI | *Clinical Laboratory Standard Institute* |
| DG | Distribuição Gama |
| IBGE | Instituto Brasileiro de Geografia e Estatística |
| IFCC | Federação Internacional de Química Clínica - *International Federation of Clinical Chemistry* |
| IR | Intervalo de Referência |
| NCCLS | Comitê Nacional para Padrões de Laboratórios Clínicos |
| OMS | Organização Mundial da Saúde |
| DNA | Ácido Desoxirribonucleico |
| SBPC/ML | Sociedade Brasileira de Patologia Clínica/Medicina Laboratorial |
| SD | Desvio Padrão |
| SQL | Linguagem de Consulta Estruturada |
| Unidades SI | Sistema Internacional de Unidades |
| h | Habitantes |
| Hbg | Hemoglobina glicada |

**SUMÁRIO**

## INTRODUÇÃO

Os exames dos fluidos biológicos dos pacientes variam (podem ser diferentes) quando observada em uma determinada população (hábitos alimentares e culturais, clima, altitude, aspectos geográficos) assim sendo os profissionais de saúde precisam de uma medida de referência no qual eles podem comparar e tomar conclusões destinando uma correta decisão clínica[1] .

Forsman[2], em seu estudo, mostrou que mais da metade das decisões médicas feitas dentro do ambiente hospitalar eram dependentes dos resultados dos exames laboratoriais. Com a evolução das práticas de qualidade e com a evolução da tecnologia, possivelmente, teremos um percentual maior dessa influência.

Para que o resultado possa ser corretamente interpretado e, portanto, útil na prática médica, ele precisa preencher alguns requisitos, como ter sido obtido por uma metodologia confiável e robusta, ter valores preditivos positivos e negativos relevantes e altos níveis de sensibilidade e especificidade, além disso, é essencial que seus limites de significância e limitações sejam bem conhecidos, o que significa que é crucial ter seus intervalos de referência (IR) bem determinados[3].

A Organização Mundial da Saúde (OMS), a Federação Internacional de Química Clínica (IFCC - *International Federation of Clinical Chemistry*) e o *Clinical Laboratory Standard Institute* (CLSI) definem intervalo de referência como o conjunto de resultados obtidos por observação ou medição quantitativa de um analito em um grupo selecionado de indivíduos com base em critérios bem definidos[1][4][5].

Recomendam, ainda, ser oportuno ao laboratório que estabeleça seus próprios intervalos, valide os valores fornecidos pelo fabricante dos insumos, ou ainda adote os valores disponíveis na literatura[5]. O fato é que a utilização de IRs predefinidos, dependendo do analito, poderá levar a um erro de interpretação, pois sua geração não levou em consideração particularidades e variáveis como demografia, cultura alimentar, diferenças climáticas e doenças pertinentes a uma específica etnia[6].

Determinar seus próprios IR, embora muito desejável, é mais trabalhoso e custoso quando utilizamos a metodologia direta do que as outras opções, porque implica em revisão de literatura, seleção de indivíduos de referência, aplicação de questionários detalhados e análise de variáveis biológicas como sexo, idade e genética, variabilidade e estado fisiológico, entre outras tarefas[6].

O processo recomendado para realização do estudo no qual o laboratório poderá definir seu próprio intervalo de referência é chamado de abordagem direta, esta abordagem realiza um catálogo populacional onde os indivíduos representam a população de referência, estes são selecionados e definidos a partir de coleta de amostra biológica que serão analisadas para este fim[7].

Porém para laboratórios que possuem um sistema de informação robusto, com grande volume de dados de um longo período, para avaliação de sazonalidade, poderá trazer à luz um intervalo de referência utilizando conceitos como a derivação de IR utilizando a metodologia indireta, *big data* e uma modelagem estatística norteada pela literatura chamada de abordagem indireta[8].

Na abordagem indireta no qual os resultados das amostras são coletados com objetivo de exames de rotina, diagnóstico, triagem ou monitoramento poderão ser usados para definição de um IR[8].

Mineração de dados (*data mining)* e *big data* são conceitos que viabilizam o processo de usar dados previamente gerados para identificar novas informações. O banco de dados de grandes sistemas laboratoriais contém muitos milhares de resultados de pacientes colhidos em condições pré-analíticas corretas e validadas e usar esses dados para validação e consolidação do intervalo de referência é um exemplo de uso das ferramentas atuais dessas metodologias[9].

No entanto, a promessa do *big data* ainda não foi realizada em seu potencial, já que a mera disponibilidade dos dados não se traduz em conhecimento ou prática clínica. Além disso, devido à variação na complexidade e estrutura dos dados, à indisponibilidade de tecnologias computacionais e à preocupação em compartilhar dados particulares de pacientes, protegidos por confidencialidade, poucos projetos de grandes conjuntos de dados clínicos são disponibilizados aos pesquisadores em geral[10].

O uso das atuais ferramentas computacionais viabiliza não somente a determinação do IR pela abordagem indireta, como também viabiliza o melhor controle interno de qualidade,(?) determinação de variação biológica, indo muito além somente do uso clínico como também abrindo oportunidade para aprendermos mais sobre mudanças fisiológicas, efeitos de interferentes e estudos epidemiológicos[11].

O presente estudo de dados laboratoriais clínicos existente na base de dados da empresa Dasa da população do estado do Rio de Janeiro no períodos de janeiro a dezembro de 2018 para os exames glicose, insulina e hemoglobina glicada (Hbg/HBA1c), e sua adequação para a determinação do IR pode ser utilizada para desenvolver e definir faixas de referência próprias para laboratórios clínicos.

**Justificativa**

Os resultados do estudo poderão ajudar na elaboração de valores de referência mais precisos (adequados) para exames laboratoriais, com benefícios para um diagnóstico mais acurado e tratamento (conduta) mais adequado (acertada). Uma vez que os resultados dos testes clínicos são interpretados de acordo com valores de referência, as determinações de IR específicos para uma determinada população, se propõem a serem mais adequados à nossa população e para faixas etárias específicas. Esta metodologia poderá, ainda, ser utilizada para desenvolver e definir faixas de referência próprias para outros laboratórios clínicos para uso no Brasil pois reflete melhor que os valores de referência estabelecidos em outros países.

# 1 **REVISÃO DE LITERATURA**

Para melhor compreensão do estudo proposto e qual sua afinidade com a bioinformática, será descrita uma revisão da literatura para compreensão do modelo estatístico e da bioquímica, do hormônio e fisiopatologia envolvida na abordagem e desenvolvimento do intervalo de referência no território estudado.

## 1.1 **Geografia e Demografia**

O estado do Rio de Janeiro está situado na parte sudeste do território brasileiro, é uma das 27 unidades da República Federativa do Brasil, tendo como limite os estados de Minas Gerais ao norte e noroeste, Espirito Santo ao norte e São Paulo ao sudoeste, sua costa, é banhada pelo oceano atlântico, sendo o terceiro maior estado em extensão costeira. Apesar de ter relativamente pequena superfície, é o terceiro estado com menor território, concentra 8,4% da população brasileira com 16.718.956 habitantes, segundo censo do Instituto Brasileiro de Geografia e Estatística (IBGE) em 2017, sendo o estado com maior densidade demográfica[12].

Figura 1 – Mapa do Estado do Rio de Janeiro


Fonte: IBGE, 2008 (12).

No Brasil, no censo de 2007 apresentou 183.987.291 vidas e este pode chegar, em estimativa a 220 milhões de brasileiros. A distribuição da faixa etária será o principal fator a ser observado de mudança nesse cenário. Serve-se de exemplo a mudança da população acima de 70 anos, que em 2008 representava 7,95% e em 2050 estima-se estar em 32,9% da população[12].

O estado do Rio de Janeiro é formado por uma população etnicamente e culturalmente diversificada[13][14]. Índios nativos de quatro etnias já haviam fixado residência antes da colonização, o aumento demográfico posterior ocorreu principalmente devido ao período onde o estado foi residência da monarquia portuguesa e depois abrigou o distrito federal, o tornando um atrativo para migração durante o período expansionista com a chegada de africanos, europeus e asiáticos e nas duas grandes guerras onde chegaram refugiados políticos de toda parte do mundo[15].

Os dados utilizados neste estudo são oriundos da população do Rio de Janeiro, coletados do período de 01 de janeiro de 2018 até 31 de dezembro de 2018 somente para validar a possibilidade do desenvolvimento de uma ferramenta analítica, o período poderá se diversificar, dependendo da necessidade do estatístico que utilizar a ferramenta.

O que justifica a utilização dessa população é sua variada miscigenação debatida previamente, e o conteúdo do banco de dados que contém as informações pertinentes de quatro laboratórios (Públicos? Privados? Hospitais? Quais laboratórios) de análises clínicas do estado do Rio de Janeiro.

## 1.2 **Diabetes e fisiopatologia**

O diabetes melito é conhecida como a doença mais comum por comprometimento da liberação de um determinado hormônio pancreático. As formas conhecidas, tipo 1 e tipo 2, são determinadas pela acometimento da liberação da insulina, o tipo 1 no qual também é conhecido como insulino-dependente é resultado das destruições das células betas, o tipo 2 é resultado da perda da regulação normal da secreção de insulina, resultado de um complexo processo fisiopatológico que culmina com a resistência insulínica, gestacional e secundaria.[16,17]

A fisiopatologia da doença envolve um problema associado a comprometimento da entrada da glicose nas células e resultando no acumulo de glicose no sangue, esse processo causa o aumento da osmolaridade plasmática e perda de função urinária na retenção da glicose, acompanhada da excessiva perda de agua e solutos, patologicamente conhecido como poliúria, a desidratação causada pela poliúria resulta no desencadeamento de processos compensatórios, o mais comum é o aumento da sede, conhecido como polidispsia.[18]

A incapacidade da célula de consumir a glicose plasmática, assemelha ao processo de inanição, conhecido como inanição celular, resultado no aumento do apetite, chamado de polifagia.[18]

A ausência da insulina resulta em aumento dos níveis energéticos na ativação da lipólise e da proteólise, aumenta os níveis circulantes de ácidos graxos livres e aminonácidos gliconeogénicos que extrapolam a capacidade do fígado processamento metabólico, levando ao aumento de corpos cetônicos no sangue.[18]

O diagnóstico clinico da diabetes tem como base em consensos das associações médicas nacionais, como a Sociedade Brasileira de Diabetes[19] e internacionais como a OMS e European Society of Cardiology[20], estas preconizam a medição da glicose em jejum com limite referente a 126ml/dL, ...com limite referente a 99 mg/dL para considerar o indivíduo como não portador de anormalidades no metabolismo da glicose, considerando-se ) e níveis de glicose em horários livres (aleatórios) referentes limitados as 200 ml/dL, as associações clinicas dos sinais e sintomas, tais como, poliúria, polidipsia ou polifagia, ou uma elevação persistente de níveis plasmáticos de glicose após uma carga de glicose oral acima de 200ml/dL após duas horas da coleta.[19,20]

## 1.2 **Desenvolvimento do intervalo de referência**

O IR serve para nortear os profissionais de saúde a tomar decisões de acordo com a mensuração fisiopatológica do paciente no momento da coleta. Durante a fase analítica de um laboratório de análises clínicas os resultados da avaliação do material biológico são comparados aos valores de referência, valores esses que podem ser uma faixa, um resultado ou uma indicação dicotômica para a avaliação de um valor “normal”[21].

Os valores normais são extraídos de uma população saudável, porém alguns autores mantêm um debate aberto do que é uma população saudável, a indicação clínica do que é

normal ou não é normal. Schneider em 1960[22] foi um dos primeiros autores a tocar no assunto da determinação de uma população saudável e suas características, em seu artigo resume que “Pessoas saudáveis são definidas como aquelas que têm valores de atributos específicos e selecionados, não característicos dos estados definidos, que parecem importantes para os objetivos imediatos do médico que faz a classificação”, explicando que é mais importante definir quem pertence ao grupo dos pacientes “não sadios” do que determinar o que é um paciente sadio[22]. Gräsbeck em 2004[21] Gräsbeck R, Saris NE. Establishment and use of normal values. Scand J Clin Lab Invest 1969;26(Suppl. 110): p.62–3. discutiu extensivamente o conceito de saúde e doença, onde afirmou que “a saúde é relativa, e o mesmo indivíduo pode ser considerado saudável e doente, dependendo da situação”. Barth em 2009[23] dissertou sobre os IR que devem fornecer informações críticas para nortear os médicos em suas tomadas de decisões, também discutiu sobre o desenvolvimento e uso de faixas de referência por considerar como os médicos utilizam esses dados na tomada de decisão.

A importância de determinar corretamente o IR foi amplamente enfatizada por Friedberg, quando mencionou que os resultados dos testes são "enquadrados" por IR e o uso de faixas aberrantes pode influenciar a tomada de decisões [24].

Existem, basicamente, três principais formas de prover um intervalo de referência, o modelo convencional ou *a priori* que segue um estudo transversal, com a seleção dos indivíduos de referência antes da fase analítica usando o protocolo recomendado pela IFCC (referência), o modelo *a posteriori* onde são analisados pacientes pre (pré)-testados porém com capacidade de aplicar os critérios de seleção com os dados disponíveis e o modelo indireto onde são analisados um volume grande de dados laboratoriais recuperados dos bancos de dados e aplicados modelos estatísticos que garantam a confiabilidade do resultado obtido[25].

Os laboratórios geralmente usam várias técnicas para determinar os IR associados à analitos para indivíduos gozando de *boa saúde* (isso não é uma verdade). Isso inclui transferência e validação de valores de referência de dados publicados ou de outros laboratórios e adotando padrões recomendados pelas agências reguladoras ou baseando-se valores de referência dos laboratórios na análise estatística das próprias amostras testadas pelos laboratórios resultados[25].

Embora a opção *a priori* seja indicada como a melhor opção como o método indicado para produção de intervalo de referência, os estudos convencionais são normalmente realizados com a disponibilização de recursos financeiros, interação humana e tempo do

colaborador, esses quando disponíveis ainda devem ser viabilizados pela definição e recrutamento de pessoas saudáveis e na aprovação por um comitê de ética que geralmente se mostra inviável para os pequenos e médios laboratórios[8].

É evidente, de muitas fontes, que a importância dos valores de referência não é tão importante definir “normal” e “anormal” ou saúde ou doença, mas ajudar os médicos a tomar uma decisão no acompanhamento para o tratamento dos pacientes. Desse modo, pode se inferir que os médicos exigem valores para comparação com os resultados do paciente para tomadas de decisão, ao invés que "valores normais"[26].

Bock et al.[27] mostraram que 60% a 70% de todas as decisões críticas em medicina são tomadas com base nas informações fornecidas através de resultados de laboratório. Afirma-se também que, embora a tomada de decisões médicas seja cada vez mais com base nos resultados laboratoriais, enfatiza a importância dos valores de referência laboratoriais, porém a abordagem para a geração desses valores permaneceu inalterada nos últimos 20 anos[27].

No nosso presente estudo, utilizaremos o modelo indireto, pois aplicaremos recursos atuais de tecnologia da informação para recuperar resultados de banco de dados, aplicar regras de filtro para determinar uma população saudável e estratificar o intervalo de referência.

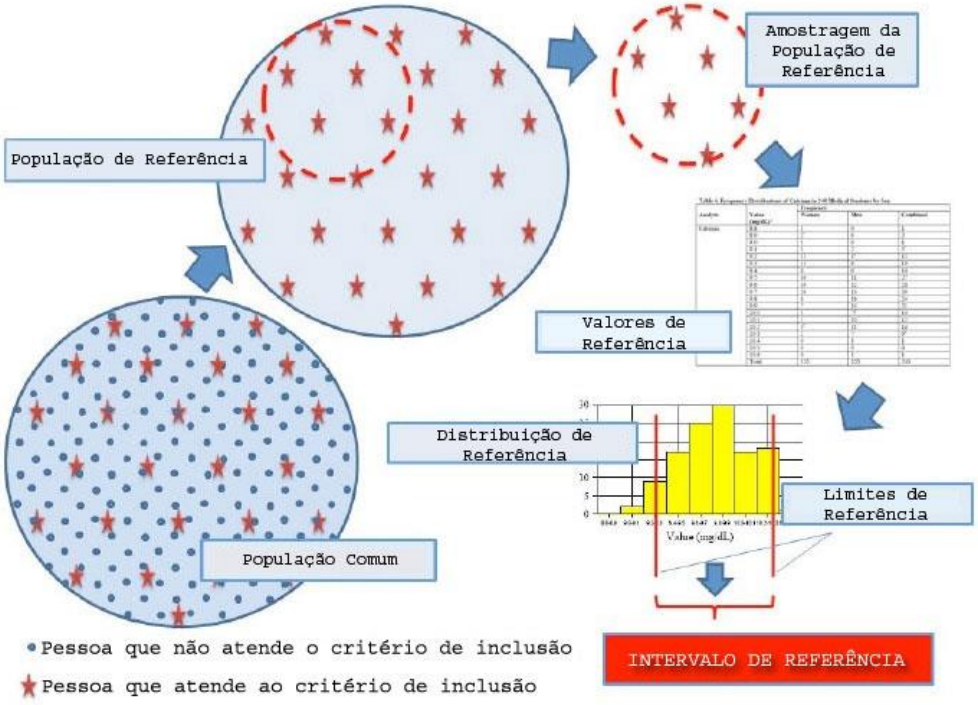
## 1.3 **Indivíduos de referência**

De acordo com CLSI/IFCC e suas recomendações, o indivíduo de referência é uma seleção de pessoas dentro de um critério determinado, a partir de um analito este grupo é estratificado para determinar qual será seu alvo dentro a população comum, este é o primeiro passo para determinação do que está sendo pesquisado, o modelo de pesquisas populacional utilizado neste estudo é a determinação de intervalo de referência a partir do método indireto[25].

População de referência refere-se a todos os indivíduos de referência de uma população em geral, esta procura determinar dentro da população comum a população sadia. Grupo de amostras de referência são indivíduos dentro de um conjunto da população de referência, uma amostragem que pode ser realizada por idade, sexo, etnia, e outros valores que podem determinar um filtro dentro da população de referência. Os valores de referências são obtidos a partir do resultado de testes do grupo de amostras de referência. A distribuição

de referência é avaliada a partir da amostra de referência, neste ponto podem existir diversos modelos estatísticos que atendem a esta tarefa. Finalmente os intervalos de referência são obtidos a partir dos valores que se encontram entre os limites inferiores e superiores dos limites de referência como segue o entendimento na Figura 2[5].

Figura 2 – Processo de seleção de indivíduos de referência e determinação do intervalo de referência




Fonte: Adaptado de Aytekin e Ermerk, 2008[28].

## 1.4 **População de referência**

O debate que envolve como podemos determinar um indivíduo saudável é um debate muiti-profissional, a OMS define que "saúde é um estado de completo bem-estar físico, mental e social e não apenas a ausência de doença ou enfermidade"[29].

A definição do status "saudável" é particularmente difícil de estabelecer pois podemos encontrar diversas dificuldades para estabelecer e assumir que pode haver uma multiplicidade de condições a serem encontradas[5]. Petitclerc em 2004[30] mencionou que é um trabalho árduo e quase impossível de determinar fielmente um grupo saudável onde apresenta toda a

diversidade biológica de uma região em tamanho amostral relevante. Ele também criticou a definição de saúde da OMS por ser rígida, reiterada por diversos autores. Adeli em 2008[31] afirmou que amostra de populações saudáveis em estudos de referência devem abranger uma série de parâmetros antropométricos, incluindo etnia, gênero e idade, uma vez que muitos parâmetros exibem interdependências que podem afetar o resultado.

"Saúde" é um conceito difícil de definir. Uma definição bem conhecida é dada pela OMS: "Saúde é um estado de completo bem-estar físico, mental e social, e não apenas a ausência de doença ou enfermidade.". Alguns autores, por exemplo, Gräsbeck[21], criticaram essa definição como irrealista e propuseram outras que enfatizam a ausência de valores conceituais indesejáveis, as chamadas definições privativas.

Todos os autores concordam que a saúde é um estado muito individual, por isso parece inadequado defini-lo por "valores normais" na química clínica analítica. No entanto, como geralmente não é possível especificar o valor normal para indivíduos, seria útil alguma forma de limite para detectar desvios brutos da 'saúde'. Obviamente, nenhum limite deve ser tomado como absoluto e não deve ser chamado de valor "normal", mas, em vez disso, um limite de "referência". Ao determinar os limites de referência, é importante garantir que a população de referência corresponda à população de pacientes o mais próximo possível em todos os aspectos, exceto na doença. Atenção especial deve ser dada a distribuições semelhantes de idade e sexo, contextos sociais, dietas etc. É claro que os mesmos métodos de amostragem e análise devem ser usados para ambas as populações. Como os doadores de sangue ou a equipe do laboratório geralmente não têm a mesma distribuição etária dos pacientes hospitalizados, podem ser encontradas grandes discrepâncias se os grupos anteriores forem usados como amostras de referência. A Federação Internacional de Química Clínica (IFCC) deu uma série de recomendações sobre a teoria dos valores de referência[5].

Utilizamos o critério de exclusão para definir um filtro para definição da população saudável a partir dados coletado do banco de dados do período de 2018.

## 1.5 **Abordagens diretas vs. Indiretas**

A abordagem tradicional para o estabelecimento de intervalos é denominada “abordagem direta”[25]. Nesse processo, indivíduos de uma população são selecionados para amostragem com base em critérios pré-definidos. As amostras desses são então coletadas,

analisadas e mensuradas. Esta abordagem foi subdividida em seleção a *priori* e a *posteriori*. A abordagem a *priori* é selecionar indivíduos para coleta e análise de amostras, caso atendam aos critérios de inclusão. Na abordagem a *posteriori*, as amostras coletadas de uma população serão incluídas na análise baseada em outros fatores, como detalhes clínicos ou outros resultados da medição, que não foram utilizados para definir a coleção. Assim, na abordagem a posteriori, nem todas as amostras coletadas seriam incluídas na população de referência para análises adicionais. Idealmente, a abordagem direta usaria membros selecionados aleatoriamente da população de referência; no entanto, isso raramente é alcançado com a população testada geralmente fortemente influenciada por conveniência e fatores de custo. Aleatória verdadeira para procurar um grupo totalmente representativo requer uma extenso planejamento e implementação, como foi usado nas pesquisas canadenses sobre medidas de saúde[6]. Outro fator com abordagens diretas é que em todos os testes selecionados, o resultado é incluído na análise estatística. Isso pode gerar *outlier* que uma parte vital do processo, embora a próprio processo de exclusão pode afetar significativamente os valores determinados[31]. Limitações conhecidas para estudos diretos incluem dificuldade na definição de saúde e na prevalência de doença subclínica, bem como viés de seleção associado a amostras relativamente pequenas[8].

Abordagens indiretas são aquelas realizadas usando resultados de laboratório coletados para outros fins, geralmente para atendimento clínico de rotina, embora também para triagem, em que os IR geralmente são determinados por métodos estatísticos baseados na identificação de uma distribuição no meio dos dados, em vez de exigir avaliação de todos os resultados individuais no banco de dados como pertencentes ou não à população de referência. Embora processos paramétricos ou não paramétricos padrões tenham sido utilizados para estudos com intervalo de referência indireta, essas técnicas sofrem influência dos resultados mais extremos em um conjunto de dados, que também são os que mais provavelmente serão afetados pela doença[8].

Uma questão vital em qualquer projeto de intervalo de referência, usando técnicas diretas ou indiretas, é a compreensão dos fatores que influenciam as variações nas concentrações de analitos. Os efeitos da variabilidade intra e inter-individual, variabilidade analítica e pré-analítica, fisiológica e patológica, bem como a tomada de decisão clínica, precisam ser considerados ao projetar estudos, interpretar os resultados e decidir sobre cada IR. O estabelecimento de IR não deve ser considerado apenas uma estatística descritiva, mas também requer a supervisão de especialistas em medicina laboratorial e fisiologia[7].

É importante notar que os IR não devem ser confundidos com os limites de decisão clínica. Os IR são geralmente considerados como uma distribuição dos valores dos testes na população predefinida, enquanto os limites de decisão clínica são determinados principalmente pela avaliação dos resultados dos pacientes ou da resposta às mudanças no manejo terapêutico. Para alguns parâmetros, no entanto, o termo IR tem sido estabelecido por consensos nacionais e internacionais, como ocorre para os lípides, glicemia, hemoglobina glicada que se definem com base em avaliação de risco, os limites de decisão.

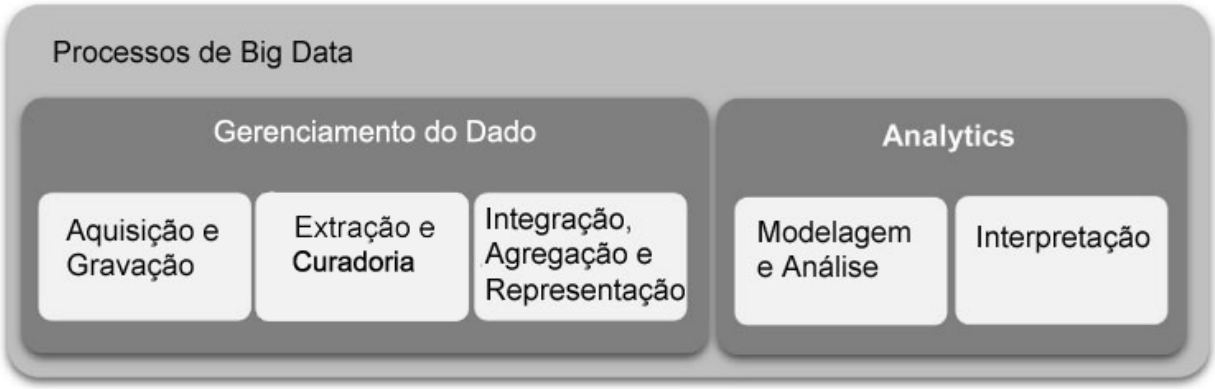
## 1.6 **Estudos sobre intervalos de referência com grande banco de dados**

O processo de informatização ocorreu fortemente no final do século XX, a estrutura dos dados foi modelada para atender ao negócio e não a pesquisa, com isso, temos um grande volume de informação no qual não conseguimos extrair, de forma confiável, rápida e dinâmica os dados necessários, revisto por Gorender[32].

A próxima fronteira para ser transpassada em análise epidemiológica são os grandes bancos de dados ou *big data*. O aumento de estudos multicêntricos, populacionais, medicina de precisão e transparência tem gerado um volume grande de informações complexas e dados desestruturados, demandando novas técnicas como mineração de dados ou *data mining* [33].

Os "4Vs" são um ponto de partida adequado para uma discussão sobre análise de *big data* na área da saúde. Mas há outras questões a serem consideradas, como o número de arquiteturas e plataformas e o domínio do paradigma de código aberto na disponibilidade de ferramentas, considere também o desafio de desenvolver metodologias e a necessidade de interfaces amigáveis ao usuário[9].

Figura 3 – Processos para extrair *insights* de Big Data




Fonte: Adaptado de Gandomi et al. [11].

Os 4 Vs do *big data* são:

*a)* *Volume* – refere-se ao tamanho dos dados;

*b)* *Variedade (Variety)– refere-se ao formato dos dados;*

*c)* *Velocidade (Velocity)* – refere-se à velocidade de streaming dos dados; e

*d)* *Veracidade (Veracity)* – refere-se sobre os dados serem confiáveis ou não.

Porém *big data* tem seu valor potencial apenas quando é utilizado para impulsionar a tomada de decisão. Para permitir essa tomada de decisão com base em evidências, as organizações precisam de processos eficientes para transformar grandes volumes de dados dinâmicos e em movimento rápido em *insights* significativos. O processo geral de extrair insights do *big data* pode ser dividido em cinco estágios, mostrados na Figura 3. Esses cinco estágios formam os dois subprocessos principais: o gerenciamento de dados e o analítico. O gerenciamento de dados envolve processos e tecnologias de suporte para adquirir e armazenar dados e prepará-los e recuperá-los para análise. *Analítico*, por outro lado, refere-se a técnicas usadas para analisar e adquirir inteligência de *big data*. Assim, a análise de *big data* pode ser vista como um subprocesso no processo geral de "extração de insight" a partir de *big data*[11].

Estudos sugerem que o Sistema de Informática Laboratorial (LIS) é uma rica fote de dados onde possui uma combinação de pacientes saudáveis que vão realizar exames de rotina com pacientes crônicos onde mantem o status de sua fisiopatologia atualizada[34].

## 1.7 **Número de participantes**

Não há um modelo estatístico preciso para determinar o número de amostras necessárias. No entanto, para produzir resultados robustos se faz necessário um grande volume de dados. Se um conjunto de dados é composto por quase todos os resultados não afetados e se aproxima da distribuição Gaussiana, por exemplo, sódio ou cálcio sérico em uma população de clínica geral, números menores podem fornecer estimativas confiáveis. Se a distribuição subjacente estiver enviesada ou fortemente contaminada, serão necessários números maiores[35][36]. Se a ferramenta estatística utilizada produzir um intervalo de confiança em torno dos limites de referência derivados, será possível avaliar se os limites gerados são "próximos o suficiente".

Na ausência de tais estimativas, a avaliação de vários subconjuntos de dados para demonstrar a reprodutibilidade pode fornecer evidências de suporte. Os comentários acima são apenas qualitativos. No entanto, para fornecer um ponto de partida para trabalhos adicionais nessa área, 1000 indivíduos podem ser considerados um número pequeno e acima de 10.000 como um número grande, e em populações pouco representadas em um banco de dados (por exemplo, extremos de idade), números menores podem ainda fornece informações úteis. Também foi recomendado o uso de no mínimo 400[37] (trabalho realizado com pacientes internados) resgates de dados de referência para cada partição para um cálculo do intervalo de referência estatisticamente confiável.

## 1.8 **Determinando o modelo estatístico para avaliação do Intervalo de Referência**

Arzideh[38] afirmou que é possível estabelecer faixas de referência usando dados do paciente, se o testes estatísticos apropriados forem realizados, por exemplo, truncamento da distribuição Gaussiana, embora Reed[39] e colaboradores afirmaram que a maioria dos dados biológicos não pode ser adequadamente descrita por Curvas Gaussianas ou log-gaussianas[35].

Muitas distribuições de dados biológicos tendem a unimodal e inclinado positivamente, o que influenciará a determinação da referência intervalo. O método preferido para a análise estatística de um grande número de pontos de dados é o paramétrico, mas nos casos de distribuições que não apresentam características gaussianas, a (o) método não-

paramétrico seria preferível. A faixa "normal" estimada ou faixa de referência depende da distribuição observada e, portanto, do método estatístico apropriado[35].

Como apresentado, diversos estudos têm métodos detalhados de uso de dados gerados em laboratório para desenvolver faixas de referência específicas da população. No entanto, o grande desafio é identificar a população saudável e doente e remover a porção não desejável do total de dados. Embora vários métodos tenham sido relatados, o método mais amplamente referenciado foi proposto por Bhattacharya[40] (1967), pesquisas foram conduzidas usando o método “Bhattacharya” para analisar os dados do paciente e este método permitiu o refinamento dos dados laboratoriais “brutos” e definiu IR que correspondiam à população do estudo.

Nem todos os dados dos pacientes podem ser incluídos em populações gaussianas e onde os dados do paciente parecem estar distorcidos e um modelo gaussiano claramente não é apropriado, Hemel et al.[41] propuseram o uso da distribuição gama. Esses autores também recomendaram a inspeção dos resíduos para avaliar quão bem os dados se encaixam no modelo estatístico assumido.

## 1.9 **Precisão e exatidão**

Precisão é uma medida de como os valores reproduzíveis estão em uma série de medições, enquanto a precisão indica quão próximo um determinado valor está dos valores-alvo. A precisão pode ser determinada para um teste específico por análise de um controle testado em que o valor alvo é conhecido. Isso é normalmente fornecido pelo fabricante ou fabricado internamente, medindo com precisão uma quantidade predeterminada do analito e depois dissolvendo-o em uma quantidade predeterminada de uma matriz de solvente, onde a matriz é semelhante ao plasma. Um ensaio ideal possui excelente precisão e exatidão, mas a boa precisão de um ensaio nem sempre garante uma boa exatidão e vice versa[42].

## 1.10 **Distribuição Gama**

A distribuição Gama DG é uma distribuição extremamente flexível nas variedades de formas, confiabilidade e funções de risco para modelagem de dados. A distribuição gama é

usada para modelar valores de dados positivos que são assimétricos à direita e maiores que 0, ela é comumente usada em estudos de sobrevivência de confiabilidade. Hager e Bain[43] estenderam o estudo das propriedades dos estimadores de máxima verossimilhança do modelo DG e forneceram um processo discriminatório entre o modelo Weibull e o modelo Gama. Tadikamalla e Penn[44] apresentaram um método mais fácil e rápido para amostragem aleatória da DG. Huang e Hwang[45], Dadpay et al.[46] e Nadarajah e Gupta[47] obtiveram algumas medidas de informação para a DG.

Na teoria das estimativas, uma das coisas fundamentais sobre a precisão de um estimador é encontrar um bom limite inferior para a variação do estimador. Em muitos casos, a variação tem uma forma complicada e não podemos computá-la, portanto, por limites inferiores, podemos aproximar-se dela. Muitos estudos foram realizados para os limites inferiores da variância de um estimador imparcial do parâmetro. Os limites inferiores conhecidos e aplicáveis são Cramer-Rao, Bhattacharyya, Hammersley-Chapman-Robbins, Hammersley, Kshirsagar e Koike. De acordo com a utilidade e amplas aplicações da DG e também a importância de encontrar e aproximar um limite inferior para a variância dos estimadores, obtemos a forma geral da matriz Bhattacharyya para a DG [48].

Dentro da distribuição gama uma importante propriedade é o fato de que à medida que θ theta aumenta, essa distribuição se aproxima de uma Normal com média μ e variância θ $\mu^2 = \mu^2/v$ [49]

Este tipo de distribuição geralmente é aplicada quando se quer fazer algum tipo de analise ligada ao tempo de vida de algum tipo de produto[49], em nosso estudo será utilizado para proporcionar a normalização da distribuição quando esta estiver deslocada a direita e for maior que zero.

## 1.11 **Método Bhattacharya**

O método Bhattacharya também é um método gráfico para identificar uma distribuição gaussiana no meio de outros dados. Como o método de Hoffmann[50], ele foi originalmente desenvolvido na era dos pré-computadores usando sistemas manuais baseados em papel. O procedimento é capaz de separar distribuições sobrepostas, dando uma vantagem sobre Hoffmann nessa configuração. O método Bhattacharya demonstrou ser menos influenciado por dados não incluídos na distribuição gaussiana em comparação com o método Hoffmann.

Este método foi sujeito a revisão e também utilizado em vários artigos publicados. O método depende do usuário, exigindo a seleção do tamanho da lixeira para os dados, a localização da lixeira e o número de lixeiras incluídas na análise. Normalmente, dados de quatro a seis compartimentos são usados para determinar a linha de melhor ajuste e um alto grau de linearidade é preferido por exemplo, r2 > 0,99[40].

Usando o modelo indireto com o método Bhattacharya:

a) para determinar um intervalo adequado, dentro do qual todos os valores são assumidos como realizações do subgrupo não patológico; e
b) para estimar os parâmetros da distribuição assumida. Assim, esses métodos são imprecisos nos dois processos.

Além disso, nenhum teste estatístico é usado para provar ou refutar que a distribuição estimada se encaixa "bem" nos dados (no intervalo predefinido), então "nos casos em que todas as observações pertencem a uma amostra populacional específica, existem testes estatísticos relativamente simples (χ2 teste de Kolmogorov-Smirnov, teste de Shapiro-Wilk) para provar ou refutar que uma distribuição é gaussiana[40].

Usando métodos indiretos, com observações que não pertencem a uma distribuição específica, esses testes não são viáveis. No entanto, pode-se sugerir a presença de com uma subpopulação não gaussiana subjacente a partir da não linearidade no gráfico de Bhattacharya

## 1.12 **Teste de R de Shapiro-Wilk**

Os testes de regressão e correlação são baseados no fato de que uma variável Y ∼ N (μ, σ2) pode ser expressa como Y = μ + σX, onde X ∼ N (0, 1). O teste de Shapiro-Wilk[51] é o teste de regressão mais conhecido e foi originalmente restrito para o tamanho da amostra de n ≤ 50. Se X (1) ≤ X (2) ≤ ⋯ ≤ X (n) denota uma amostra aleatória ordenada de tamanho n de uma distribuição normal padrão (μ = 0, σ = 1), seja m = (m1, m2, mn) o vetor dos valores esperados das estatísticas da ordem normal padrão e seja V = (vij) o n × n matriz de covariância dessas estatísticas de ordem. Seja Y = (Y (1), Y (2), ..., Y (n)) denotar um vetor de observações aleatórias ordenadas de uma população arbitrária. Se Y (i) são observações ordenadas de uma distribuição normal com média desconhecida μ e variância desconhecida σ2, então Y (i) pode ser expresso como Y (i) = μ + σX (i) (i = 1, 2, ..., n).

A estatística do teste Shapiro-Wilk para normalidade é definida como:

$$\text{SW} = \frac{\left[\sum_{i=1}^{n} a_i Y_{(i)}\right]^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Onde:

$$\boldsymbol{a} = \boldsymbol{m}\mathbf{V}^{-1}(\boldsymbol{m}'\mathbf{V}^{-1}\boldsymbol{m})^{-1/2}$$

Os $a_i$'s são pesos que podem ser obtidos de Shapiro e Wilk(51) para o tamanho da amostra $n \leq 50$. O valor de SW está entre zero e um. Pequenos valores de SW levam à rejeição da normalidade, enquanto um valor de um (1) indica normalidade dos dados.

O teste de Shapiro-Wilk foi então modificado por Royston[52] para ampliar a restrição do tamanho da amostra.

Ele deu uma transformação normalizada para Shapiro-Wilk como $Y = (1 - SW)\ \lambda$ para algumas opções de λ. O parâmetro λ foi estimado para 50 tamanhos de amostra selecionados e depois suavizado com polinômios em loge (n) - d onde d = 3 para $7 \leq n \leq 20$ ed = 5 para $21 \leq n \leq 2000$. Royston forneceu o algoritmo AS 181 em FORTRAN 66 para calcular a estatística do teste Shapiro-Wilk e o valor p para tamanhos de amostra 3–2000. Mais tarde, Royston observou que a aproximação de Shapiro-Wilk para os pesos utilizados nos algoritmos era inadequada para n> 50. Em seguida, ele forneceu uma aproximação aprimorada dos pesos e forneceu o algoritmo AS R94, que pode ser usado para qualquer n no intervalo $3 \leq n \leq 5000$.

Este estudo utilizou o algoritmo fornecido por Royston[52].

## 1.13 **Teste Kolmogorov-Smirnov**

Dada uma amostra aleatória $x_1,x_2,\cdots,x_n$, uma função de distribuição empírica S(x) é a fração de observações amostrais menores ou iguais ao valor de x. Se $y_1,y_2,\cdots,y_n$ são as estatísticas de ordem da amostra aleatória observada, sem observações repetidas, então a função de distribuição empírica é definida como:

$$S(x) = \begin{cases} 0, & \text{para } x < y_1; \\ \dfrac{k}{n}, & \text{para } y_k \leq x < y_{k+1},\ k = 1, 2, \cdots, n-1 \\ 1, & \text{para } x \geq y_n \end{cases}$$

A estatística de teste Kolmogorov-Smirnov (KS): $D_n = sup_x\ [|F(x) - S(x)|]$ é usada para testar a hipótese nula que a função de distribuição acumulada $F_x$ é igual a alguma função de distribuição, sob hipótese, S(x), ou seja,

$$\begin{cases} H_0 : F(x) = S(x) \\ H_1 : F(x) \neq S(x). \end{cases}$$

em que, $D_n$ é o menor limite superior de todas as diferenças pontuais $/F_n(x) - S(x)/$.

Para os casos nos quais duas (ou mais) observações sejam iguais, ou seja, quando há *nk* observações para $x_k$, a função de distribuição empírica é uma "step" que salta:

$$\left( \frac{nk}{n} \right)$$

na altura de cada observação $x_k$.

Qualquer que seja o caso, a função de distribuição empírica S(x) é a fração dos valores amostrais que são menores ou iguais a x[53].

# 2 **OBJETIVOS**

## 2.1 **Geral**

Avaliar a viabilidade da determinação de IR para insulina, glicose ou insulina utilizando metodologia indireta.

**Desenvolver ferramenta para determinação de intervalo de referência em parâmetros laboratoriais a partir de um banco de dados de grande porte**

## 2.2 **Específicos**

Os objetivos específicos são:

a) Verificar a viabilidade da utilização de um banco de dados laboratorial;
b) Extrair dados laboratoriais clínicos existentes e pré-testados do LIS em um banco de dados em um formato adequado para análise estatística;
c) Produção de algoritmos para extração de dados específicos para pesquisa cientifica, *Data Mining*, em um formato adequado para uso estatístico;
d) Produção de algoritmos para filtro e limpeza do banco de dados, *Data Cleaning;*
e) Produção de algoritmo para determinação do IR por metodologia indireta.

# 3 **MATERIAL E MÉTODOS**

## 3.1 **Questões Éticas e Aprovações**

Comitê de Ética em Pesquisa – Projeto submetido ao Comitê de Ética do CONEP/Plataforma Brasil em 31/07/2018, sob parecer 2.970.023 com a situação final aprovada.

## 3.2 **Critério de Inclusão**

Este estudo extraiu os dados laboratoriais disponíveis do banco de dados existente do sistema de informações laboratoriais do laboratório (qual) e analisou apenas os dados obtidos para os testes listados.

A suposição é de que os dados no sistema de computadores do laboratório são abrangentes o suficiente para fornecer todas as informações necessárias sobre gênero e local de origem das amostras.

Definir melhor os critérios de inclusão

## 3.3 **Critério de exclusão**

Os conjuntos de dados podem ser bioquimicamente filtrados (?) para reduzir a frequência dos resultados de indivíduos nos quais há uma maior probabilidade de doença afetar o resultado. Isso pode se basear em outros resultados (por exemplo, excluir resultados de tiroxina onde o TSH está fora do intervalo de referência), o local da coleta de amostras (ou seja, residentes em altas altitudes, pacientes), pacientes hospitalizados ou informações clínicas fornecidas, como uso de drogas. Dependendo da frequência relativa das amostras e da natureza da técnica estatística, pode não ser necessário excluir subgrupos específicos (por exemplo, clínica lipídica ou clínica renal). Como exemplo, o uso de métodos estatísticos usando a maior parte dos dados próximos ao centro da distribuição, como Hoffman ou

Bhattacharya, será resistente à inclusão de tais grupos, mas métodos paramétricos ou não paramétricos padrão podem ser fortemente influenciados[8].

Uma abordagem adicional recomendada é limitar os resultados a um único resultado por paciente. Como um paciente doente tem mais probabilidade de ser testado novamente do que um paciente não doente, a falha em fazer isso provavelmente levará à super-repressão dos resultados de indivíduos indispostos. Ao selecionar o resultado único, o último resultado de um paciente durante um "episódio de assistência médica" (por exemplo, uma internação hospitalar) é preferido, pois é mais provável que represente um retorno à saúde. Uma extensão dessa abordagem é usar apenas resultados onde uma única coleta foi feita de um paciente durante o período de coleta de dados (amostras “solo”). Isso se baseia na suposição de que um resultado considerado anormal pelo médico assistente tem maior probabilidade de ser repetido. Outras abordagens para “melhorar” os dados são usar os resultados de outros testes essenciais, como o projeto REALAB[34], ou vincular-se a bancos de dados clínicos, que contêm informações de saúde específicas do paciente[31].

## 3.4 **Técnicas laboratoriais empregadas**

Todas as amostras foram provenientes de 23 pontos de coletas quais foram processadas do dia 01 de janeiro de 2018 a 31 de janeiro de 2018.

A metodologia laboratorial empregada nas análises foi a técnica disponível no laboratório clínico da rede DASA, no município de Duque de Caxias no Rio de Janeiro, sendo uma metodologia totalmente automatizada e amplamente utilizada em laboratórios clínicos.

DASA é a maior empresa de Medicina Diagnóstica do Brasil e a terceira maior do mundo. Está presente em todas as regiões do Brasil e realiza mais de 4 milhões de exames/mês. O núcleo técnico onde os exames foram realizados, possuem certificados de qualidade ISO 900, certificado de Boas Práticas em Laboratório Clínico (BPLC) expedido pela Sociedade Brasileira de Patologia Clínica/ Medicina Laboratorial (SBPC/ML) além da nota PALC. Periodicamente este NTO recebe a visita de auditores externos para verificação dos sistemas de qualidade e de meio ambiente. Todas estas certificações e acreditações de qualidade garantem a exatidão e a precisão dos resultados dos exames laboratoriais realizados.

A plataforma, metodologia, material e volume de amostra estão apresentados na tabela 1.

Tabela 1 – Parâmetros utilizados para análise

| Parâmetro | Plataforma | Metodologia | Material | Kit |
|---|---|---|---|---|
| **Glicose** | Cobas® 6000 c 501 | Colorimétrico enzimático | Soro | Roche GLUC3 |
| **Insulina** | Cobas® 6000 cobas e 601 | Colorimétrico enzimático Homogêneo | Soro | Roche Insulin |
| | | | | |

Fonte: O autor, 2020.

Método colorimétrico enzimático: Os métodos colorimétricos utilizam a intensidade da cor para inferir a quantidade de determinado analito em uma amostra. A intensidade da cor destas reações é medida na faixa visível do espectro (380nm - 680nm).

A glicose e a insulina foram dosadas por estas técnicas no equipamento cobas® 6000 (analyser series- Roche Diagnostics). Em linha modular na esteira de Total Lab Automation.

## 3.5 **Analito Glicose**

A glicose é o principal carboidrato presente no sangue periférico. A oxidação da glicose é a principal fonte de energia celular no corpo. A glicose derivada de fontes alimentares é convertida em glicogênio para armazenamento no fígado ou em ácidos graxos para armazenamento no tecido adiposo. A concentração de glicose no sangue é controlada dentro de limites estreitos por muitos hormônios, os mais importantes dos quais são produzidos pelo pâncreas. A causa mais frequente de hiperglicemia é o diabetes mellitus resultante de uma deficiência na secreção ou ação da insulina. Vários fatores secundários também contribuem para níveis elevados de glicose no sangue. Estes incluem pancreatite, disfunção da tireoide, insuficiência renal e doença hepática. Hipoglicemia é menos frequentemente observada. Uma variedade de condições pode causar baixos níveis de glicose no sangue, como insulinoma, hipopituitarismo ou hipoglicemia induzida por insulina. A medição da glicose na urina é usada como procedimento de triagem do diabetes e para auxiliar na avaliação da glicosúria, na detecção de defeitos tubulares renais e no tratamento do

diabetes mellitus. A medição da glicose no líquido cefalorraquidiano é usada para avaliação da meningite, envolvimento neoplásico das meninges e outros distúrbios neurológicos[54].

Teste de UV.

Método de referência enzimático com hexoquinase.

A hexoquinase catalisa a fosforilação da glicose em glicose-6-fosfato pelo ATP.

$$\text{Glucose} + \text{ATP} \xrightarrow{\text{HK}} \text{G-6-P} + \text{ADP}$$

A glicose-6-fosfato desidrogenase oxida a glicose-6-fosfato na presença de NADP em gluconato-6-fosfato. Nenhum outro carboidrato é oxidado. A taxa de formação de NADPH durante a reação é diretamente proporcional à concentração de glicose e é medida fotometricamente.

$$\text{G-6-P} + \text{NADP}^{+} \xrightarrow{\text{G-6-PDH}} \text{gluconate-6-P} + \text{NADPH} + \text{H}^{+}$$

## 3.6 **Analito Insulina**

A insulina é um hormônio peptídico de 51 resíduos com um peso molecular de 5808 Da. É secretado pelas células β das ilhotas de Langerhans no pâncreas e passa a circular pela veia porta e pelo fígado. A insulina é geralmente liberada em pulsos[55].

A molécula de insulina biologicamente ativa é monomérica e consiste em duas cadeias polipeptídicas, a cadeia α de 21 aminoácidos e a cadeia β de 30 aminoácidos unidas por pontes de dissulfeto. A insulina é o produto biossintético da pré-pró-insulina precursora de cadeia única, que é subsequentemente clivada para produzir pró-insulina[55].

Proteases específicas clivam ainda mais a pró-insulina para produzir insulina e o peptídeo C de conexão que passam para a corrente sanguínea simultaneamente em concentrações equimolares. A insulina circulante tem meia-vida de 3 a 5 minutos e é preferencialmente retida e degradada no fígado. Portanto, apenas cerca de metade da insulina atinge a circulação sistêmica. A inativação ou excreção de pro-insulina e peptídeo C ocorre principalmente no rim e praticamente nenhum peptídeo C é retido no fígado. Como resultado, o peptídeo C possui uma concentração plasmática mais alta que a insulina[56].

A ação da insulina é mediada por receptores específicos e consiste principalmente na facilitação da captação de glicose pelas células do fígado, tecido adiposo e musculatura; esta é a base de sua ação hipoglicemia

O ensaio de insulina Elecsys emprega dois anticorpos monoclonais que são específicos para a insulina humana.

Princípio do teste: Princípio Sanduíche.

Duração total do ensaio: 18 minutos.

1ª incubação: Uma amostra de insulina de 20 µL, um anticorpo monoclonal biotinilado específico à insulina e um anticorpo monoclonal específico à insulina marcado com um complexo de rutênio, complexo Tris (2,2'-bipiridil) rutênio (II) (Ru (bpy ) 2/3 +), formam um complexo sanduíche.

2ª incubação: Após a adição de micropartículas revestidas com estreptavidina, o complexo fica ligado à fase sólida por meio da interação de biotina e estreptavidina.

A mistura de reação é aspirada para a célula de medição, onde as micropartículas são magneticamente capturadas na superfície do eletrodo. Substâncias não ligadas são então removidas com solução ProCell. A aplicação de uma tensão no eletrodo induz então a emissão quimioluminescente que é medida por um fotomultiplicador.

Os resultados são determinados por meio de uma curva de calibração gerada especificamente por instrumentos a partir da calibração de 2 pontos e uma curva principal fornecida por meio do código de barras do reagente ou do código de barras.

## 3.7 **Tecnologia Usada**

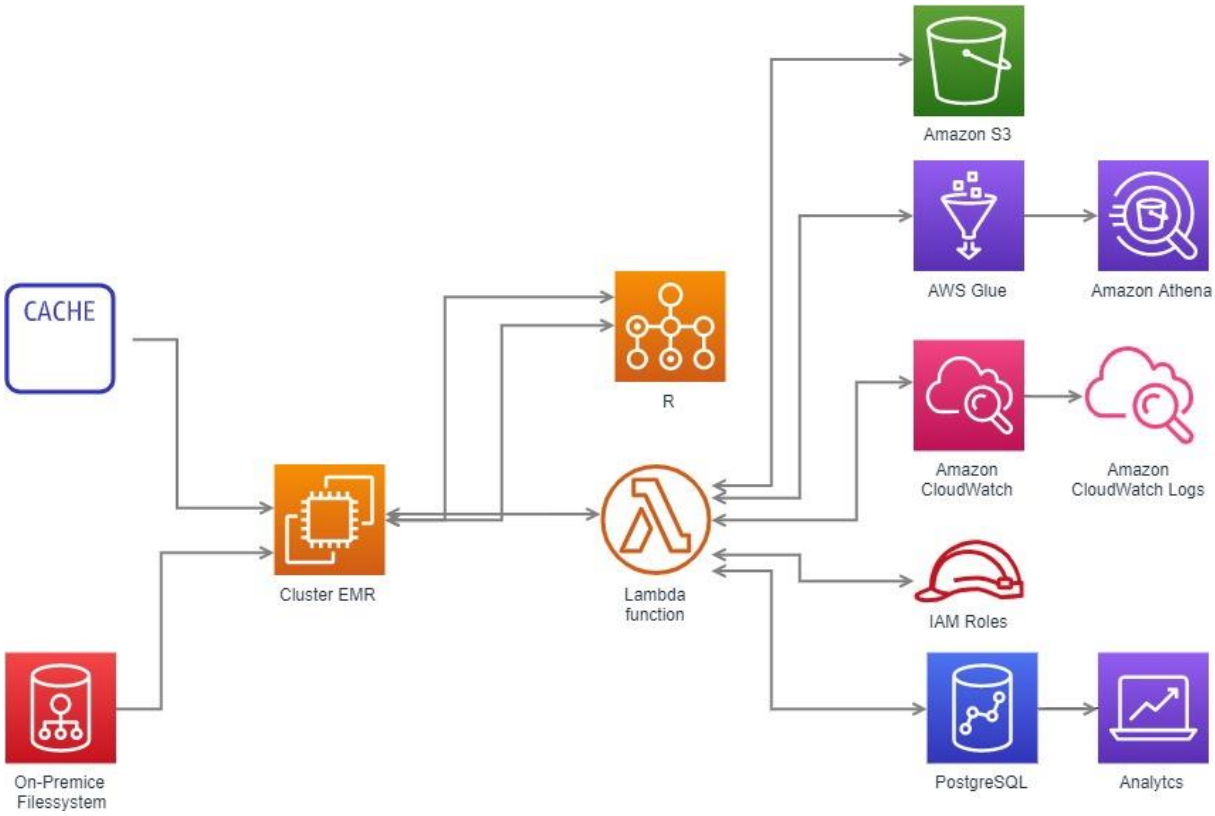
O sistema de informação laboratorial do Laboratório Diagnóstico das Américas construído na linguagem Java utilizando banco de dados Oracle da Sun Microsystems e Caché da Intersystems e destes bancos de dados foi gerado, para obtenção dos dados base desse estudo, foi desenvolvido uma aplicação em COS, Caché Objetct Script no qual foi utilizado para extrair um registro individual de cada paciente contendo, data da realização do exame, sexo, idade, peso, altura como dados antropométricos e resultados de exames foram, Insulina e Glicose e medicamentos.

Um mapeamento dos sistemas de origem dos dados foi realizado e desenvolvido na linguagem R o modelo de conexão e mapeamento das tabelas necessárias, um processo foi

usado para extrair os seguintes dados de cada paciente individual e em dado bruto salvo em uma tabela única com as mesmas características, conforme demonstrado na figura 4.

Os dados foram armazenados na base de dados PostgreSQL.

Figura 4 – Fluxograma da execução programática do sistema desenvolvido




Fonte: O autor, 2020.

Os dados de um paciente (Indexador da amostra, idade, data de nascimento, sexo, região e local de origem da amostra, data e hora do teste, nome e resultado do teste) foi extraído de cada paciente em um teste ou em uma combinação de testes listados na Tabela 2.

Embora a idade dos pacientes estivesse disponível nos dados conjunto, isso não foi considerado e os dados não foram estriados de acordo com as faixas etárias.

Essas informações foram capturadas quando as informações do paciente das amostras de laboratório foram registradas para o LIS.

Tabela 2 – Dados levantados para pesquisa

| TESTE | Ocorrências Antes da remoção do Outlier | Ocorrências após remoção do Outlier | Outliers removidos | % outliers removidos | N = Masc. | % Masc. | N = Fem. | % Fem. |
|---|---|---|---|---|---|---|---|---|
| **Glicose** | 66.079 | 55.620 | 10.459 | 15.8 | 33.693 | 60 | 21.927 | 40 |
| **Insulina** | 66.079 | 43.222 | 22.857 | 34.6 | 20.747 | 47 | 22.475 | 53 |
| | | | | | | | | |
| **Total** | 132.158 | 98.842 | 29446 | 25,2 | 54.440 | | 44.402 | |

Fonte: O autor, 2020.

Os dados da tabela 2 refletem o início do tratamento e curadoria dos valores encontrados, considerando somente os exames Glicose e Insulina, foi possível encontrar 132.158 exames elencáveis para análise estatística, após a remoção dos *outliers* com as regras de exclusões aplicadas, foi selecionado 98.842 exames, estes refletem 25,2% do total. Qualificando por sexo foi possível considerar 54.440 pacientes do sexo masculino e 44.402 do sexo feminino.

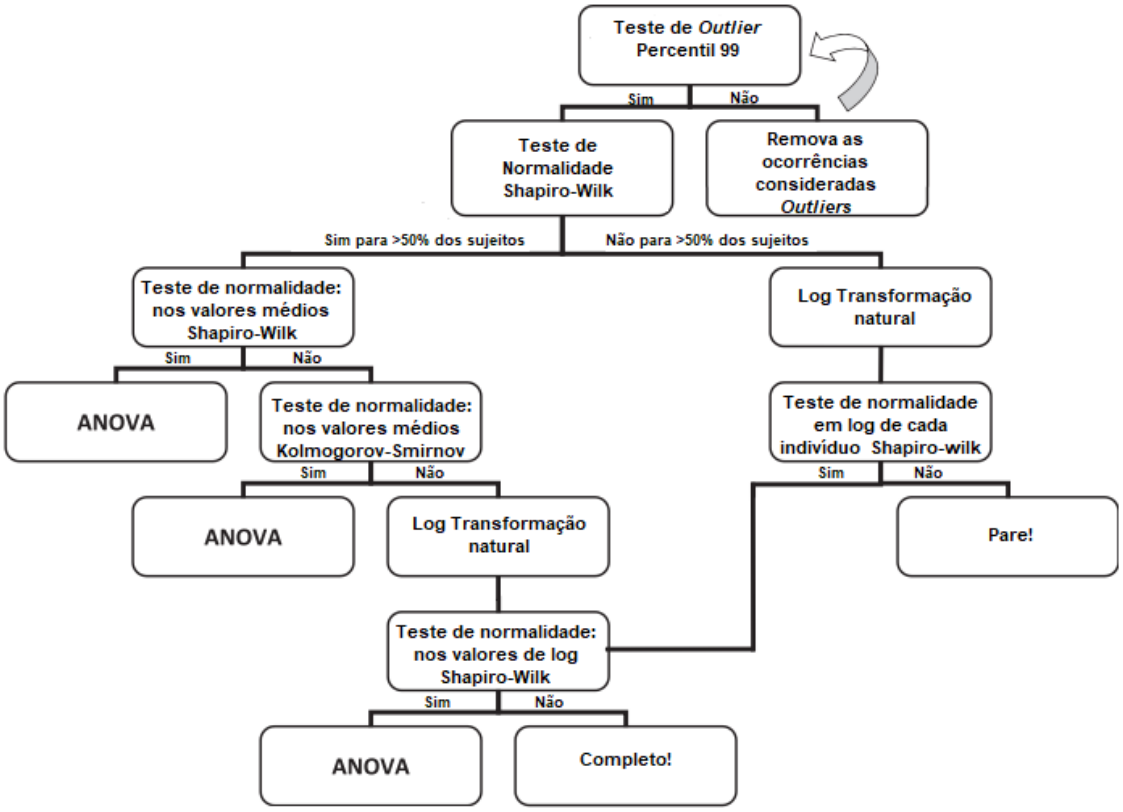
## 3.8 **Transformação Log e Remoção de *Outliers***

Shine em 2008[57] propôs a normalização da transformação usando uma abordagem de três etapas para análise de dados, a primeira etapa é identificar e remover outliers, a segunda etapa é definir a distribuição subjacente dos demais pontos de dados e a terceira etapa é estratificar dados e modelar os efeitos de gênero e idade. Wright[58] também afirmou que essa definição para faixa normal estava entre os percentis de 2,5 e 97,5, que também são os padrões usados no laboratório médico, definindo o intervalo de referência de 95%. A assimetria positiva é comum na distribuição de indivíduos de referência e uma transformação logarítmica é eficaz em reduzir ou removê-la consideravelmente[58].

A maioria dos autores relatou um processo de quatro etapas para analisar os dados usando o método paramétrico:

a) analise os dados brutos e determine se atende aos critérios de distribuição “normais”;
b) tratar por transformação de *log* dos valores originais, se necessário;
c) identifique e remova os *outliers*; e
d) definir IR.

Figura 5 – Fluxograma para remoção de *outliers*




Fonte: O autor, 2020.

A função em R para execução do expurgo dos *outliers*, segundo Wrigth e a construção dos gráficos para glicose é:

```
Glg[is.na(Glg)]=0  ## substitui os NA por zero
upper <- (quantile(Glg, .75) + 1.50 * IQR(Glg))
GlgCorrigido <- Glg[Glg < upper]
lower <- (quantile(Glg, .25) - 1.50 * IQR(Glg))
GlgCorrigido <- GlgCorrigido[GlgCorrigido > lower]
GlgCorrigido <- GlgCorrigido[GlgCorrigido > 0]
```

A função em R para execução do expurgo dos outliers, segundo Wrigth e a construção dos gráficos para insulina é:

```
Ins[is.na(Ins)]=0  ## substitui os NA por zero
upper <- quantile(Ins, .75)  + 1.50 * IQR(Ins)  ## 15.5 = terceiro quartil
```

```
InsCorrigido <- Ins[Ins < upper]
lower <- quantile(Ins, .25)  - 1.50 * IQR(Ins)  ## 6.4 = primeiro quartil
InsCorrigido <- InsCorrigido[InsCorrigido > lower]
InsCorrigido <- InsCorrigido[InsCorrigido > 0]
```

## 3.9 **Testando Normalidade Shapiro**

Na prática, o teste é simples de se aplicar em um computador usando R.

Seja X = (X 1 , ..., X n ) o vetor de dados, representado em R se inserido individualmente como c (X 1 , ..., X n ).

Teste de Shapiro (X) e você verá como resultado uma estatística de teste chamada W (para Wilk) e um valor p.

Se o valor p for menor que o nível convencional 0,05, então rejeita-se a hipótese da normalidade, caso contrário não rejeita-se.

Para aplicar o teste, não é necessário primeiro entender W e satisfaz sempre $0 < W \leq 1$.

Para valores de W perto o suficiente para 1 (dependendo de n), a hipótese de normalidade não será ser rejeitado.

Para W menor, será rejeitado.

Para n = 2, a normalidade nunca pode ser rejeitada, portanto o teste é útil apenas para $n \geq 3$.

A implementação R permite n até 5.000.

Por essa limitação de N=5000, utilizaremos o método de Shapiro em conjunto com a análise randômica, por encaixarem perfeitamente no modelo de dado proposto detalhado no item 4.2.

A função em R usada para o teste de Shapiro-Wilk nas amostras de Insulina foi:

```
Shapiro.test(myfkInsCorrigido)
```

A função em R usada para  no teste de Shapiro-Wilk nas amostras de Insulina foi:

```
Shapiro.test(myfkGlCorrigido)
```

## 3.10 **Testando Normalidade Kolmogorov-Smirnov**

O teste de Kolmogorov-Smirnov(59) é usado para decidir se uma amostra vem de uma população com uma distribuição específica. O teste de Kolmogorov-Smirnov (K-S) é baseado na função de distribuição empírica (ECDF). Dado N pontos de dados ordenados Y1, Y2, ..., YN, o ECDF é definido como $EN=n(i)/N$ onde n (i) é o número de pontos menor que Yi e o Yi é ordenado do menor para o maior valor. Esta é uma função de etapa que aumenta em 1 / N no valor de cada ponto de dados solicitado.

Uma característica atraente desse teste é que a distribuição da estatística do teste K-S em si não depende da função de distribuição cumulativa subjacente que está sendo testada. Outra vantagem é ser um teste exato (o teste de ajuste do qui-quadrado depende de um tamanho de amostra adequado para que as aproximações sejam válidas). Apesar dessas vantagens, o teste K-S possui várias limitações importantes:

a) só se aplica a distribuições contínuas;
b) tende a ser mais sensível perto do centro da distribuição do que nas caudas;
c) talvez a limitação mais séria seja que a distribuição deve ser totalmente especificada. Ou seja, se os parâmetros de localização, escala e forma forem estimados a partir dos dados, a região crítica do teste K-S não será mais válida. Normalmente, deve ser determinado por simulação.

O teste de Kolmogorov-Smirnov é definido por:

a) H0: Os dados seguem uma distribuição especificada;
b) Ha: Os dados não seguem a distribuição especificada.

Estatística do teste: a estatística do teste Kolmogorov-Smirnov é definida como:

$$D = \max_{1 \le i \le N} \left( F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$$

Onde:

F é a distribuição cumulativa teórica da distribuição que está sendo testada, que deve ser uma distribuição contínua (ou seja, nenhuma distribuição discreta como o binomial ou

Poisson) e deve ser totalmente especificada (ou seja, os parâmetros de localização, escala e forma não podem estimado a partir dos dados).

A função em R para este cálculo estatístico é:
ks.test(InsCorrigido,"pnorm",mean(InsCorrigido),sd(InsCorrigido))

A função em R para este cálculo estatístico é:
ks.test(GlgCorrigido,"pnorm",mean(GlgCorrigido),sd(GlgCorrigido))

## 3.11 **Testando Sperman**

A correlação de classificação de Spearman, também conhecida como ρ, é usada para medir a força do relacionamento entre duas variáveis. Você já deve estar se perguntando qual é a diferença entre a correlação de classificação de Spearman e a correlação de momento do produto Person. A diferença é que a correlação de classificação de Spearman é um teste não paramétrico, enquanto a correlação do momento do produto Person é um teste paramétrico.

Um teste não paramétrico não precisa estar em conformidade com as premissas do teste paramétrico, como, por exemplo, os dados normalmente distribuídos. Isso permite que um pesquisador ainda faça inferências a partir de dados que podem não ter normalidade. Além disso, testes não paramétricos são usados para dados que estão no nível ordinal ou nominal. De várias maneiras, a correlação de Spearman e a correlação do momento do produto Pearson se complementam. Um é usado em estatísticas não paramétricas e o outro para estatísticas paramétricas e cada um analisa o relacionamento entre variáveis.

Se você obtiver resultados suspeitos com a análise de correlação do momento do produto Pearson ou se seus dados não tiverem normalidade, a correlação de classificação de Spearman poderá ser útil se você ainda desejar determinar se existe um relacionamento entre as variáveis. A correlação de Spearmen funciona classificando os dados em cada variável. Em seguida, a correlação do momento do produto Pearson é calculada entre os dois conjuntos de variáveis de classificação. Abaixo estão as suposições do teste e as etapas de correlação de Spearman.

As suposições do teste de correlação de Spearman são:

a) Os assuntos são selecionados aleatoriamente.
b) As observações são no nível ordinal, pelo menos.

As etapas da correlação de Spearman são:

**a)** Configure as hipóteses;
**b)** H0: Não há correlação entre as variáveis;
**c)** H1: Existe uma correlação entre as variáveis;
**d)** Defina o nível de significância;
**e)** Calcule os graus de liberdade e encontre o valor crítico t;
**f)** Calcular o valor da correlação de Spearman ou ρ;
**g)** Calcule o valor t e tome uma decisão estatística;
**h)** Conclusão do Estado;

## 3.12 **Testando Método Bhattachaira**

Bhattacharya em 1967[40], desenvolveu um método estatístico para dividir a distribuição na fração considerada de indivíduos "saudáveis" daquela considerada de indivíduos "não saudáveis". Ele afirmou que a distribuição é uma mistura de componentes correspondentes a diferentes populações. Ele relatou que a frequência relativa e a distribuição de frequências devem ser encontradas, o que geralmente é considerado normal. Em seguida, para a resolução da distribuição gaussiana, é utilizada uma abordagem estatística para resolver sobreposições das distribuições gaussianas.

Segundo Baadenhuijsen e Smit[60], um pré-requisito para a aplicação efetiva da técnica de Bhattacharya é a disponibilidade de uma grande quantidade de dados que permitiria a ausência de grandes flutuações estatísticas e permitir o reconhecimento da parte linear da distribuição . Os autores afirmaram que é necessário coletar mais de 1500 valores para cada analito. O algoritmo de Bhattacharya assume que a maior parte de uma população total de amostras não selecionadas pode ser considerada "normal" e que a sobreposição entre a parte "saudável" e a parte anormal (alta ou baixa) é apenas parcial. Hoffmann[50] propôs uma técnica semelhante a ser usada para separar populações diferentes.

Um aspecto abordado diz respeito à transformação dos dados. Baadenhuijsen e Smit[61] avaliaram a eficiência do algoritmo de transformação. Eles exploraram a possibilidade de transformar dados, mesmo nos casos em que a subpopulação subjacente pareça ser gaussiana,

conforme detectado a partir da primeira função derivada. Eles levantaram a hipótese de que uma aplicação mais consistente dessa técnica a todos os dados produziria uma avaliação mais comparável dos dados.

Amador[26] afirmou que várias suposições são feitas geralmente ao examinar os dados obtidos por métodos indiretos. A primeira e principal suposição é que os resultados se ajustam a uma distribuição gaussiana e são iguais ao intervalo de referência da população saudável. No entanto, mesmo quando a média e o desvio padrão (DP) da população de pacientes são comparados com o intervalo de referência em uso, eles geralmente diferem entre si. A segunda suposição é que a proporção de indivíduos afetados pela doença permanecerá estável de um grupo de resultados de pacientes para o próximo. Este também não é o caso, uma vez que diferentes grupos podem apresentar variações na composição e no padrão da doença. Onde foi utilizada a função em R seguinte:

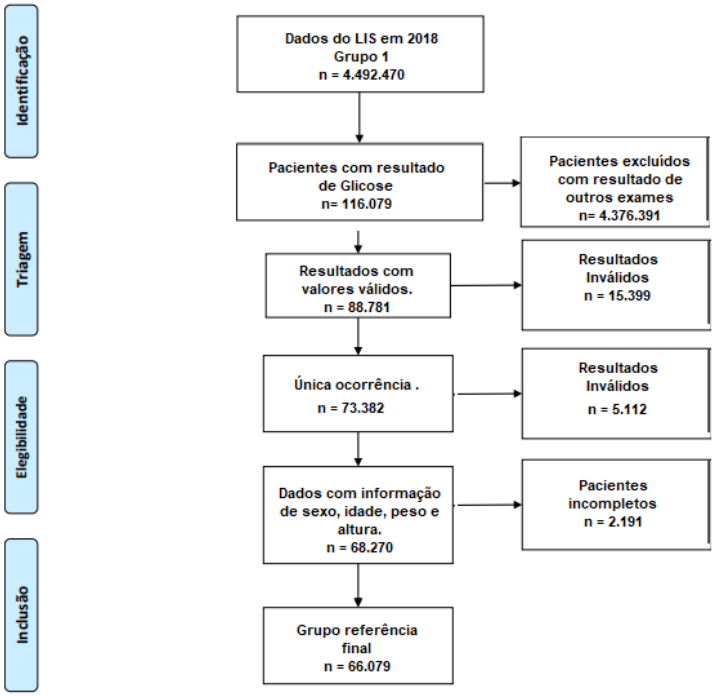
```
dados <- shapiroVector
dhist <- hist(dados,breaks = 30,plot = FALSE)
ly <- log(dhist$counts)
dly <- diff(ly)
df <- data.frame(xm = dhist$mids[-length(dhist$mids)], ly = dly,counts = dhist$counts[-length(dhist$mids)])
h <- diff(df$xm)[1]
linear.bit1 <- subset(df[5:22,])
lm1 <- lm(ly ~ xm, data = linear.bit1, weights = linear.bit1$counts)
lambda1 <- -coef(lm1)[1]/coef(lm1)[2]
mu1 <- lambda1 + h/2
sigma1 <- sqrt(-h/coef(lm1)[2] - h^2/12)
lln.bhat <- qnorm(0.025,mu1, sigma1)
uln.bhat <- qnorm(0.975,mu1, sigma1)
```

# 4 **RESULTADOS**

O sistema de informação laboratorial do Laboratório Diagnóstico das Américas construído na linguagem Java utilizando banco de dados Oracle da Sun Microsystems e Caché da Intersystems foi usado para obtenção dos dados base desse estudo. Foi desenvolvido uma aplicação em COS, Caché Objetct Script, que foi utilizada para extrair um registro individual de cada paciente contendo, data da realização do exame, sexo, idade, peso, altura como dados antropométricos e resultados de exames foram, Insulina, Glicose e medicamentos.

Foram recuperados dados de 36.694 registros de atendimento relativos a 4.492.470 resultados de testes individuais armazenados no LIS do laboratório Diagnóstico das Américas (DASA) durante um período de 12 meses de 2018. Isto constitui a base de dados original para este estudo. Este grupo de amostra incluiu 21.927 mulheres e 9.516 homens, bem como 2.919 resultados de sexo desconhecido. Sexo desconhecido é registrado quando nenhum gênero é indicado no formulário de solicitação do paciente, o n final foi de 66.079 resultados, conforme arvore de decisão demonstrado na Figura 6.

Figura 6 – Resultado após filtros e regras aplicadas



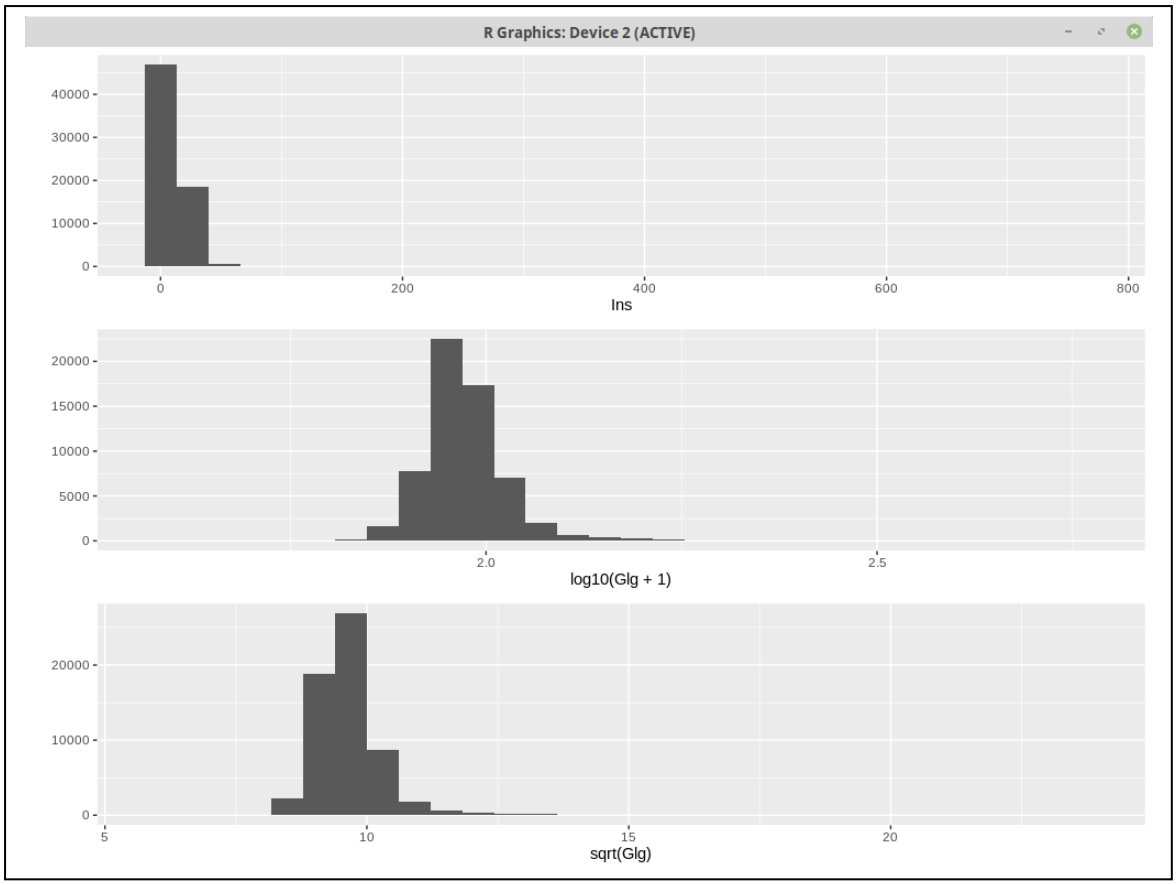

Fonte: O autor, 2020.

Foram analisados 43.222 resultados de glicose no plasma em jejum e 55.620 de insulina. Embora quase 25% dos outliers tenham sido removidos dos resultados da glicose em

jejum, a distribuição ainda não era normal, como pode ser visto nas estatísticas (assimetria = 0,618), a gráfico de histograma e probabilidade e teste de Kolmogorov-Smirnov para normalidade p.

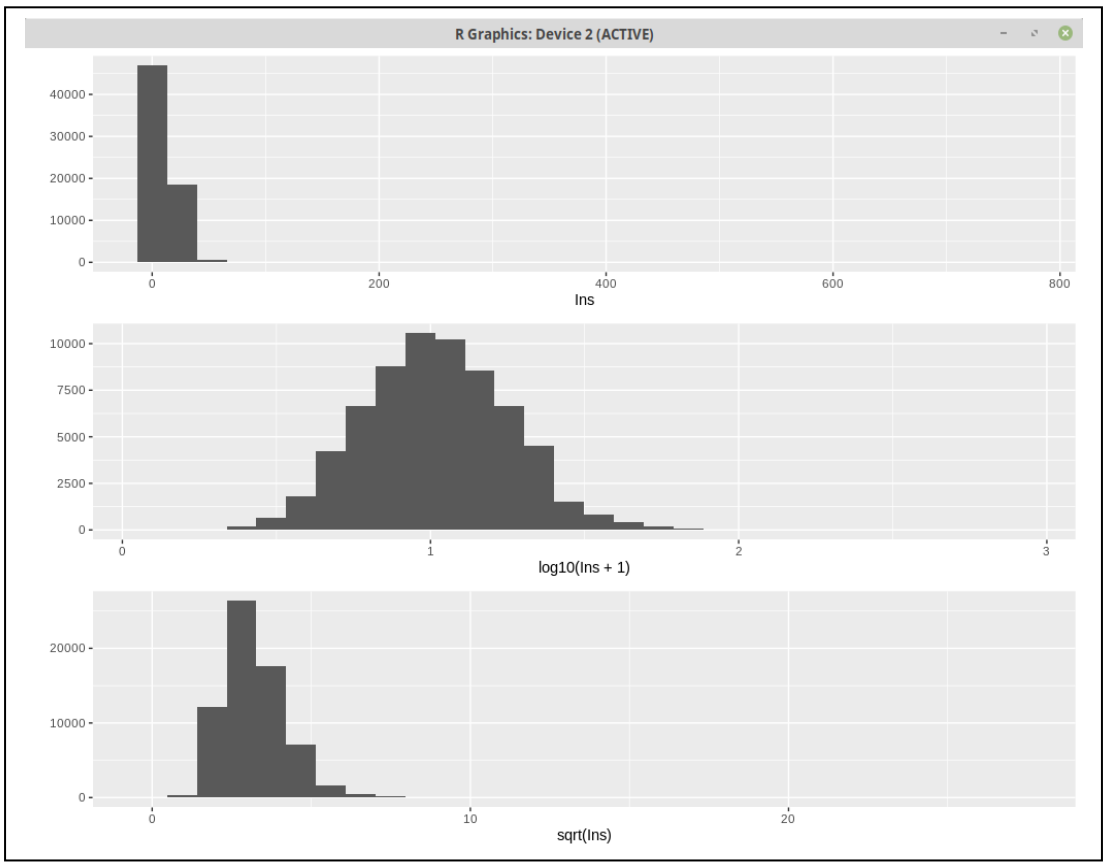
Na aplicação, as primeiras análises são realizadas pelos gráficos disponíveis, nessa versão não está automatizado o melhor modelo a ser usado, então é uma intervenção do usuário.

Figura 7 – Resultado após filtros e regras aplicadas para Glicose




Fonte: O autor, 2020.

Figura 8 – Resultado após filtros e regras aplicadas para Insulina



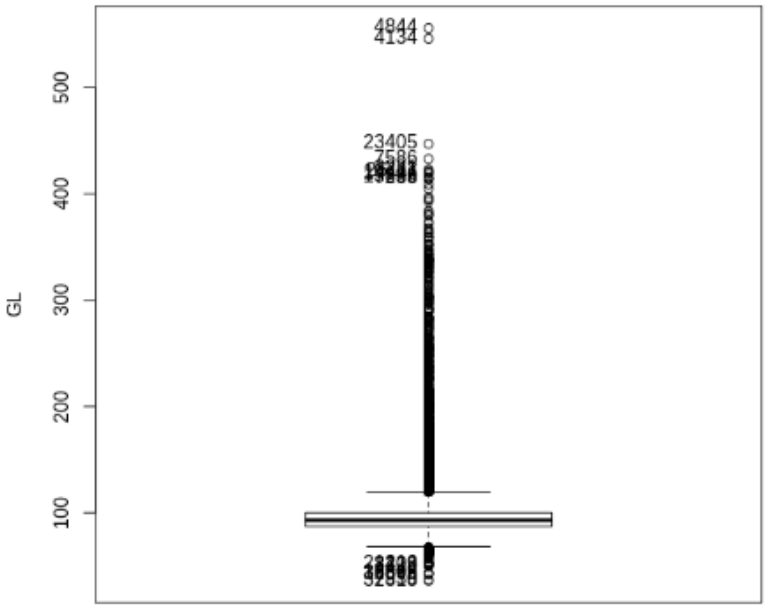

Fonte: O autor, 2020.

A aplicação gráfica auxilia na demonstração do conteúdo no qual leva ao entendimento de qual distribuição será melhor ajustada aos modelos seguintes.

Os gráficos demonstram que utilizando a distribuição em log, padroniza uma distribuição normal melhor ajustada do que o gráfico em distribuição gaussiana e em distribuição sob raiz quadrada.

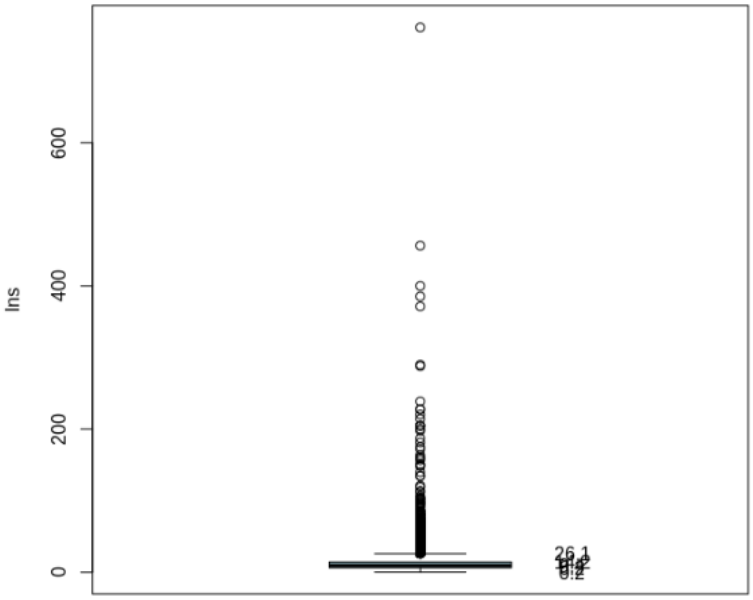
Quando plotados em GGplot, os dados normais não se demonstram uniformemente distribuídos, conforme demonstrado nos Gráficos 1 e 2.

Gráfico 1– BoxPlot dos dados brutos de Glicose no dataset




Fonte: O autor, 2020.

Gráfico 2 – BoxPlot dos dados brutos de Insulina no dataset

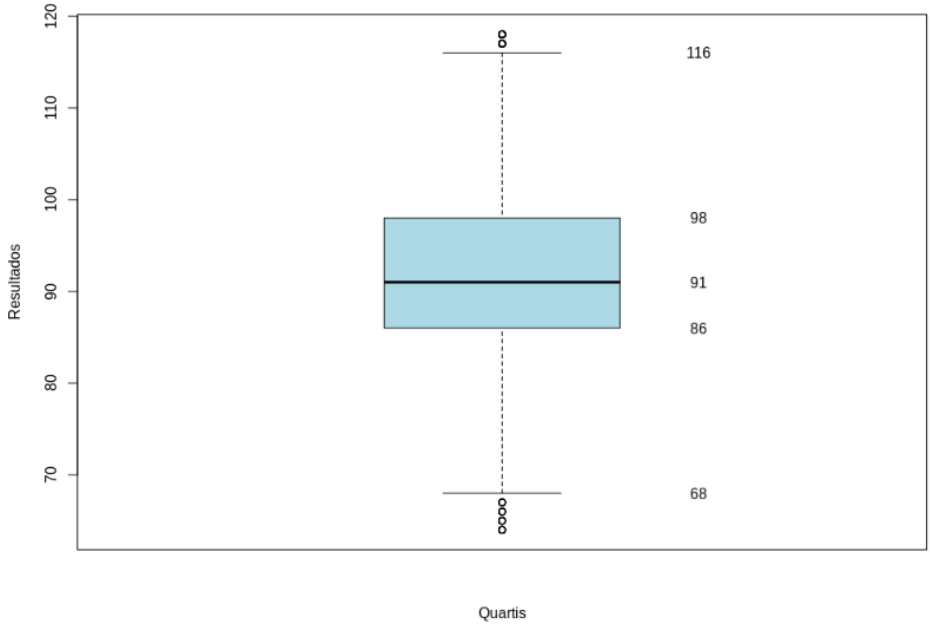



Fonte: O autor, 2020.

Os dados totais sob a distribuição normal passaram sob ajuste para normalização, utilizando a função proposta por Wright(58) também afirmou que essa definição para faixa normal estava entre os percentis de 2,5 e 97,5.

O Resumo estatístico, para o analito Glicose, após a remoção do outlies pode ser interpretado pelo gráfico em BoxBlot onde o valor mínimo é de 68, primeiro quartil de 86, mediana 91, média 92,09, terceiro quartil de 98 e valor máximo de 116 conforme representado no Gráfico 3.

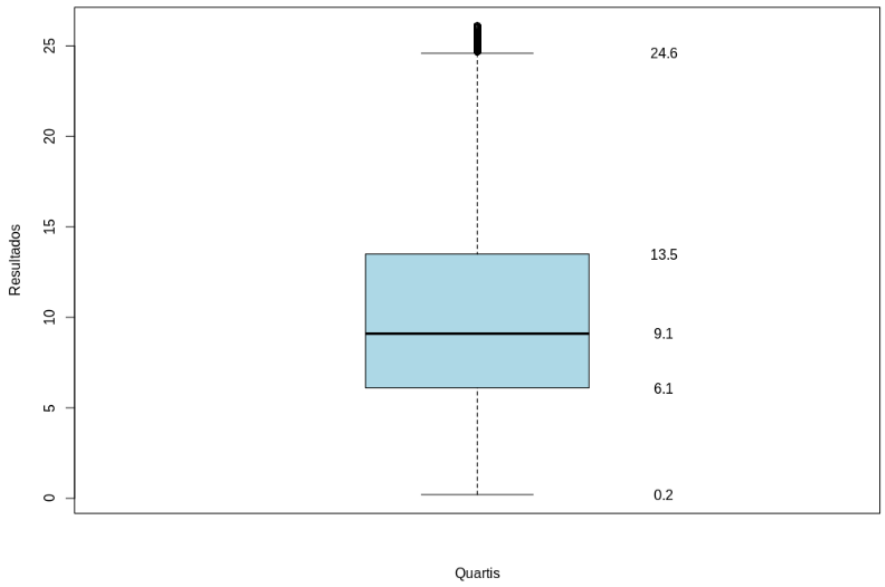
Gráfico 3 – BoxPlot dos tratados com remoção de *outliers* de Glicose no dataset




Fonte: O autor, 2020.

O Resumo estatístico, para o analito Insulina, após a remoção do outlies pode ser interpretado pelo gráfico em BoxBlot onde o valor mínimo é de 0.02, primeiro quartil de 6.1, mediana 9,40, média 11.35, terceiro quartil de 13.5 e valor máximo de 24.6 conforme representado no Gráfico 4.

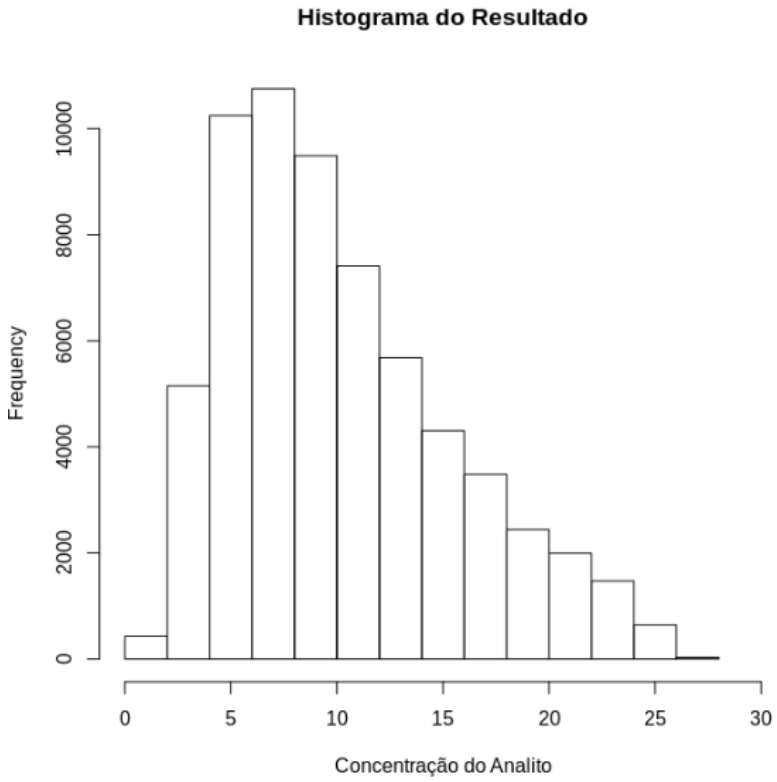
Gráfico 4 - BoxPlot dos tratados com remoção de outliers de Insulina no dataset




Fonte: O autor, 2020.

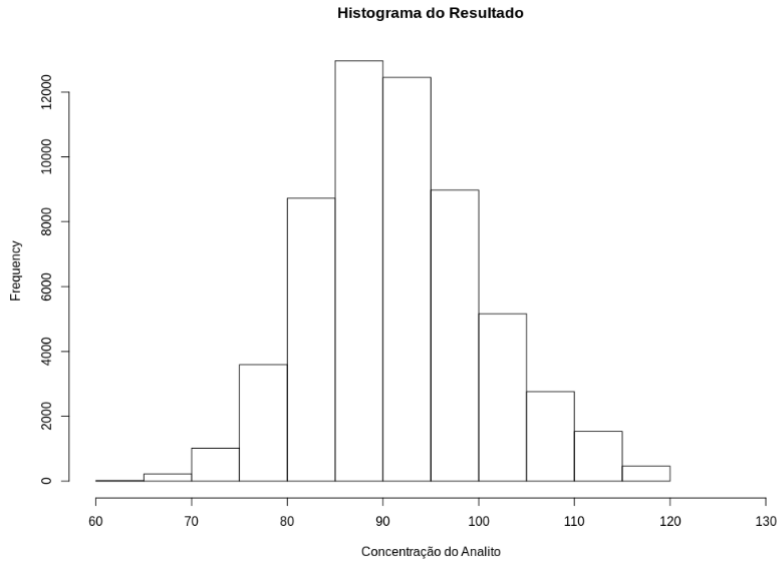
Após a reconstrução do dataset com expurgo dos outlires, o histograma, segundo o melhor produto normalizado, ficou demonstrado conforme os Gráficos 5 e 6.

Gráfico 5 – Histograma dos dados de insulina após exclusão dos *outliers*




Fonte: O autor, 2020.

Gráfico 6 – Histograma dos dados de glicose após exclusão dos *outliers*




Fonte: O autor, 2020.

## 4.1 **Distribuição Total das Ocorrências**

Os dados para distribuição total das amostras foram catalogados e curados até o momento, onde está livre de víeis e dentro da distribuição normal, para Glicose utilizaremos a distribuição Gauseana, para a Insulina precisaremos utilizar a distribuição Gamma pois esta se adequou melhor a referência.

Para a todo teste que seja utilizada a distribuição Gamma, aplicaremos os testes de Skewness ou assimetria e de Kurtosis ou curtose onde avaliaremos se a distribuição, no analito Insulina, o histograma é uma amostra de uma distribuição Gama com o parâmetro de forma 1.5. A distribuição Gama é uma distribuição inclinada com a quantidade de inclinação, dependendo do valor do parâmetro de forma. O grau de decaimento à medida que nos afastamos do centro também depende do valor do parâmetro informado. Para esse conjunto de dados, a assimetria é 0,7653 e a curtose é 2,91154, o que indica assimetria e curtose moderadas.

Para distribuição total, não foi levado em consideração o teste de Shapiro, pois este está limitado a n=5000, então, não foi levado em observação o teste para avaliação de normalidade.

Neste foi utilizado somente o teste Kolmogorov-Smirnov para avaliação de distribuições de probabilidade unidimensionais que podem ser usadas para comparar uma

amostra com uma distribuição de probabilidade de referência, então foi utilizado o modelo gráfico para interpretação dos resultados.

Este, para Insulina, foi encontrado o valor D=0.087257 e p-value < 2.2216, não foi aprofundado o estudo sobre este analito e está em aberto, a intensão é demonstrar o poder de decisão sobre os valores demonstrados, possivelmente o analito possui dois lados.

Este, para Glicose, foi encontrado o valor de D=0.05557 e p-value < 2.2216, não foi aprofundado o estudo sobre este analito, a intensão é demonstrar o poder de decisão sobre os valores demonstrados, possivelmente o analito possui dois lados.

É importante observar que os dados biológicos tendem a ser distorcidos. É afirmado por vários autores que a distorção positiva é comum em dados biológicos e, portanto, pode-se esperar estar presente na distribuição de indivíduos de referência Clemson et al.[35], Solberg(37), Wright e Royston [58], Schork et al.[62] afirmaram que "a distorção pode ser uma parte integrante parte de uma característica biológica e pode, de fato, ter um significado biológico ".

Com os dados avaliados e redimensionados, foi possível aplicar a equação de bhat para determinação do intervalo de referência sugerido e foi encontrado para Glicose 73 a 108 mg/dL e para insulina 1.2 a 21 mcU/mL demonstrados na Tabela 3.

## 4.2 **Distribuição Aletória de 5000 Analitos**

Os dados para distribuição com 5000 das amostras foram catalogados e curados até o momento, onde está livre de víeis e dentro da distribuição normal, para Glicose utilizaremos a distribuição Gauseana, para a Insulina precisaremos utilizar a distribuição Gamma pois esta se adequou melhor a referência.

O N de 5000 foi utilizado para variar significantemente da distribuição total e também conseguir encaixar no teste de Shapiro, onde este é limitado a amostragem de 5000 unidades.

A função em R utilizado para randomizar as amostras de Insulina foi:

```
myfkInsCorrigido <- sample(InsCorrigido,size=5000, replace=FALSE)
```

A função em R utilizado para randomizar as amostas de Glicose foi:

myfkGlCorrigido <- sample(GlCorrigido,size=5000, replace=FALSE)

Com o número de amostras aleatoriamente sorteadas, limitado ao n=5000, foi possível a aplicação do teste de Shapiro-Wilk para determinação de normalidade, onde para glicose foi encontrado o p<2.216 e W=0.99042 e para insulina foi encontrado p<2.216 e W=94508.

Com os dados avaliados e redimensionados, foi possível aplicar a equação de bhat para determinação do intervalo de referência sugerido e foi encontrado para Glicose 73 a 109 mg/dL e para insulina 1.2 a 22.2 mcU/mL demonstrados na Tabela 3.

# 5 **DISCUSSÃO**

A interpretação do resultado de um teste bioquímico requer responder à pergunta se o resultado do teste pertence ou não à distribuição de valores encontrados em uma população específica. A composição dessa população-alvo pode variar de acordo com o interesse do usuário pretendido. Isso requer a definição precisa dos membros da população-alvo.

A coleção de dados resultado da pesquisa aponta para a independência dos analitos quando comparado a glicose com a insulina pelo teste de Spearman com resultado de 0,292 onde aponta uma fraca relação de amplitude dos dois analitos, quando comparado a glicose com a hemoglobina glicada reflete uma média relação de amplitude dos analitos pelo teste de Spearman com o resultado de 0,477 e a comparação da Insulina com a hemoglobina glicada reflete fraca relação de amplitude dos dois analitos pelo teste de Spearman com resultado de 0,170 – explicar com mais detalhes como foi utilizada a hemoglobina glicada. Todos os testes foram realizados na linguagem R 3.5.1 conectado ao banco de dados PostgreSQL contendo a extração dos dados, essa fraca relação denota a independência dos analitos quando utilizados para caracterizar a população desejada.

Com a população atual foi possível utilizar o aplicativo em R para realizar uma mensuração parcial dos IR, histograma e avaliar o cálculo estatístico e Bhattacharya para previsão do intervalo de referência, o modelo dinâmico onde reflete a pesquisa temporal, poderá ser agendado nas ações automáticas do sistema operacional desejado, podendo submeter relatórios por e-mail ou em pasta compartilhada.

O objetivo deste estudo foi preencher esta lacuna, mesmo que realizado com uma população brasileira regional, porém, as características demográficas desta população a tornam, se não ideal, pelo menos apropriada. Isto se deve ao fato de que a população avaliada se caracteriza por um intenso movimento migratório, o que pode representar uma amostragem de diferentes regiões do Brasil.

Neste estudo foi dada ênfase à distribuição dos valores em percentis, tendo em vista os parâmetros estudados serem preferencialmente tratados como limites de decisão[25]. Entretanto serão discutidos, também os achados em termos de média e desvio-padrão, bem como o IR proposto.

Além disso, as condições sob as quais as amostras analíticas são coletadas e processadas precisam ser rigorosamente padronizadas. Apesar das críticas levantadas sobre o

uso dos chamados métodos indiretos para determinar os IR[3,36-38]), pensamos que continuava sendo uma abordagem alternativa atraente, em vista das dificuldades práticas mencionadas.

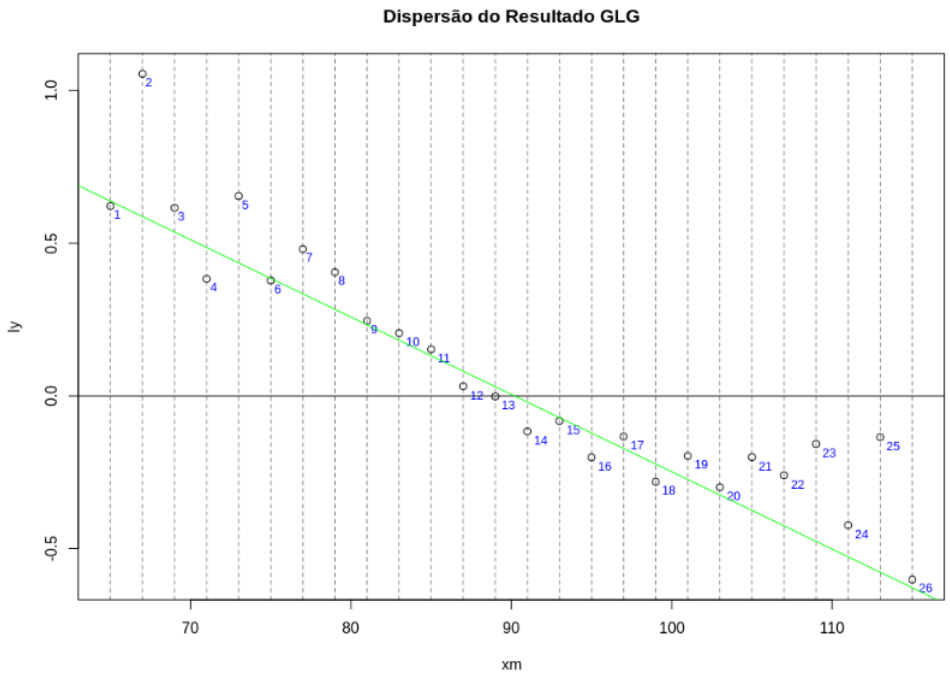
A aplicação da técnica de convolução Bhattacharya de distribuições gaussianas parcialmente sobrepostas é dificultada pela restrição de que o método se torna inválido quando as distribuições compostas estão muito próximas umas das outras, impedindo assim o reconhecimento da parte linear na primeira função derivada.

Em nossa opinião, essa desvantagem não impedirá a aplicação desta técnica quando ela for usada corretamente. Portanto, incluímos a restrição de que os arquivos de resultados só podem ser processados quando uma parte linear grande o suficiente da primeira função derivada, em termos da porcentagem da população total, pode ser encontrada. Na prática, sempre definimos esse limite restritivo em 40%, segundo Baadenhuijsen e Smit[60].

Outro argumento a favor dessa técnica é o de concordar com as constatações relatadas por outros autores sobre a dependência de idade e sexo dos IR. É preciso ter certeza de que, frequentemente, esses relacionamentos são relativamente fracos. No entanto, esta técnica parece ser capaz de resolver relacionamentos tão fracos.

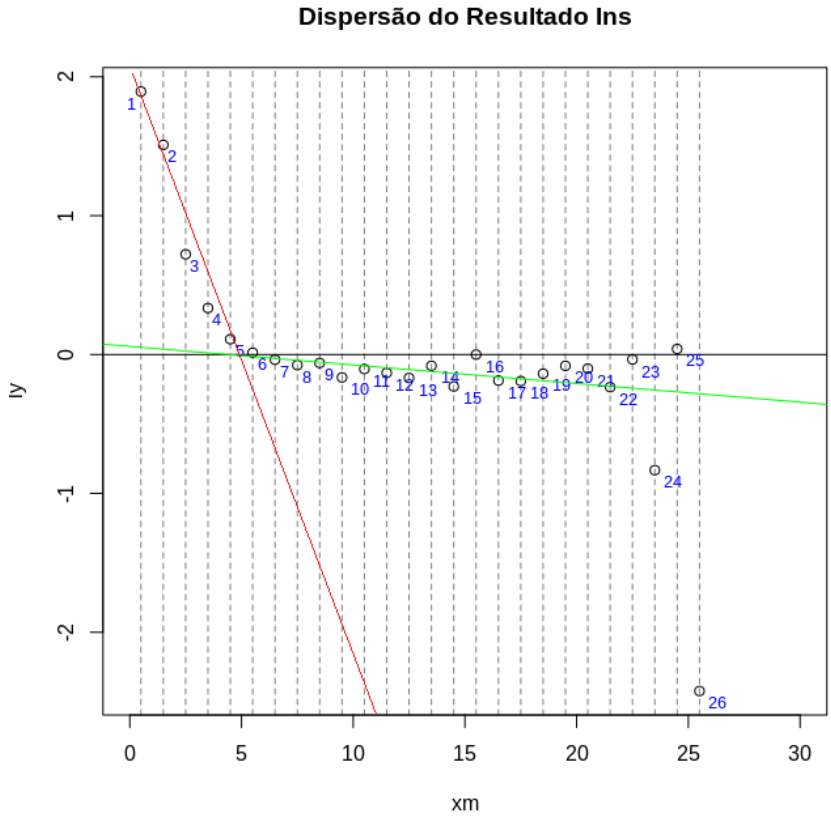
Foi possível demonstrar a interferência de idade e sexo quando comparamos o resultado absoluto encontrado, a faixa da insulina ficou mais alargada do que a faixa preconizada pelos laboratórios, ao plotar o gráfico de dispersão, foi possível evidenciar a interferência, segundo o Gráfico 7 e a faixa da glicose se demonstrou linearmente coerente, conforme Gráfico 8.

Gráfico 7 – Gráfico de dispersão dos resultados de Glicose




Fonte: O autor, 2020.

Gráfico 8 – Gráfico de dispersão dos resultados de Insulina




Fonte: O autor, 2020. Estes gráficos sugiro estarem nos resultados.

## 5.1 **Comparação dos Resultados**

Para validar esses procedimentos para obter limites de referência indiretos a partir dos resultados dos pacientes, usamos uma amostra de referência de acordo com as recomendações do IFCC (5). Como as frações 0,025 e 0,975 são geralmente selecionados como limites de referência, estabeleceu-se que qualquer limite de referência inferior ou superior obtido por um procedimento não-IFCC será válido se não for estatisticamente diferente (P.0.05) dos frágeis 0.025 ou 0.975, respectivamente, na amostra de referência. Assim, foi utilizado o teste estatístico comparando uma fração observada com um dado teórico.

A aplicação do modelo estatístico de Bhattacharya permitiu o entendimento, comparação e sugestionar um intervalo de referência a partir das populações distintas, seja ela uma população aleatória ou de volume total, conforme demonstrado na Tabela 3.

Tabela 3 – Resumo dos dados levantados após aplicação estatística

| Exame | Valor de Referência atual | 5000 Aleatório | Distribuição total |
|---|---|---|---|
| **Glicose** | 70 a 99 mg/dL | 73 a 109 mg/dL | 73 a 108 mg/dL |
| **Insulina** | 2 a 13 mcU/mL | 1.2 a 22.2 mcU/mL | 1.2 a 21 mcU/mL |

Fonte: O autor, 2020.

Sugiro que os valores encontrados por você desta tabela estejam no resultados e aqui na discussão só a comparação com os valores utilizados atuais

## CONCLUSÃO

Foi levantado a hipótese de que os resultados derivados deste estudo criarão uma base para definir IR confiáveis de laboratório clínico para a população do Rio de Janeiro. Embora este estudo tenha provado que isso poderia ser feito, muito mais trabalho precisa ser feito preparar e selecionar o conjunto de dados para definir IR confiáveis de laboratório clínico;

Foi levantado a hipótese de que os resultados derivados deste estudo são comparáveis aos em uso pelo CLSI. O estudo demonstrou que dados laboratoriais pré-testados podem ser usados para determinar IR, desde que:

Os objetivos específicos são:

*a)* Verificar a viabilidade da utilização de um banco de dados laboratorial;

*b)* Extrair dados laboratoriais clínicos existentes e pré-testados do LIS em um banco de dados em um formato adequado para análise estatística;

*c)* Produção de algoritmos para extração de dados específicos para pesquisa cientifica, *Data Mining*, em um formato adequado para uso estatístico;

*d)* Produção de algoritmos para filtro e limpeza do banco de dados, *Data Cleaning;*

*e)* Produção de algoritmo para determinação do IR por metodologia indireta.

Discussão:

**a)** o conjunto de dados seja representativo, ou seja, todos os resultados de maior número de regiões abordáveis possíveis. As amostras de laboratório analisadas por empresas privadas de laboratório provêm de pessoas predominantemente saudáveis, porque muitos exames laboratoriais realizados por médicos particulares são principalmente rotineiros;

**b)** se possível, deve-se tomar mais cuidado para evitar a inclusão de pacientes não saudáveis. Afirma-se que, se fosse possível eliminar os resultados dos pacientes hospitalizados e também encontrar maneiras de garantir o registro de dados clínicos precisos dos pacientes testados, mais resultados de pacientes não saudáveis poderiam ser removidos e impedir a inclusão de aqueles no conjunto de dados. Não foi avaliado a possibilidade e critério de

exclusão de pacientes hospitalizados por estes constituírem o grupo disperso na sociedade;

**c)** uma limitação pode ser considerada é o não particionamento de resultados em faixas etárias. Embora existissem dados sobre a idade, isso não foi considerado preciso o suficiente para ser incluídos no estudo e, portanto, não foi realizada a partição etária. Sugere-se que que, se os resultados pudessem ser divididos em categorias etárias[63], consideráveis melhora da sensibilidade em relação a faixas inespecíficas, mesmo quando diferenças de idade e sexo são significativamente diferente, seria alcançado.

# REFERÊNCIAS


1. Ferre-masferrer M, Fuentes-arderiu X. Indirect reference limits estimated from patients ’ results by three mathematical procedures. 1999;279:97–105.
2. Forsman RW. Why is the laboratory an afterthought for managed care organizations? Clin Chem. 1996 May;42(5):813–6.
3. Geffré A, Friedrichs K, Harr K, Concordet D, Trumel C, Braun JP. Reference values: A review. Vet Clin Pathol. 2009;38(3):288–98.
4. Arderiu XF. Intervalos de referencia biológicos. 2011;46–51.
5. Henny J, Vassault A, Boursier G, Vukasovic I, Mesko Brguljan P, Lohmander M, et al. Recommendation for the review of biological reference intervals in medical laboratories. Clin Chem Lab Med. 2016 Jan 1;54(12).
6. Colantonio DA, Kyriakopoulou L, Chan MK, Daly CH, Brinc D, Venner AA, et al. Closing the gaps in pediatric laboratory reference intervals: A caliper database of 40 biochemical markers in a healthy and multiethnic population of children. Clin Chem. 2012;58(5):854–68.
7. Ferreira CEDS, Andriolo A. Intervalos de referência no laboratório clínico. Vol. 44, Jornal Brasileiro de Patologia e Medicina Laboratorial. 2008. p. 11–6.
8. Jones GRD, Haeckel R, Loh TP, Sikaris K, Streichert T, Katayev A, et al. Indirect methods for reference interval determination - Review and recommendations. Clin Chem Lab Med. 2019 Apr 19;57(1):20–9.
9. Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth BD. Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database. AMIA . Annu Symp proceedings AMIA Symp. 2017;2017:384–92.
10. Bellazzi R. Big Data and Biomedical Informatics: A Challenging Opportunity. IMIA Yearb. 2014 May 22;9(1):8–13.
11. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage. 2015;35(2):137–44.
12. Instituto Brasileiro de Geografia e Estatística - IBGE. Projeção da população do Brasil por sexo e idade 1980-2050. Vol. 39, Ministério do planejamento, orçamento e gestão. 2008. 1–63 p.
13. Instituto Brasileiro de Geografia e Estatística . Características étnico-raciais da população: um estudo das categorias de classificação de cor ou raça (2008). Vol. 41, Ibge. 2011. 1–95 p.
14. Reis PP dos. A miscigenação e a etnia brasileira. Rev Hist (Costa Rica). 1961;23(48):323.
15. Andrade JHF de. O Brasil e a organização internacional para os refugiados (1946-1952). Rev Bras Política Int. 2005;48(1):60–96.
16. Penalva DQF. Síndrome metabólica: diagnóstico e tratamento. Rev Med. 2008 Dec 18;87(4):245.
17. Almeida APF, Moura L, Chaves FR, Romaldini JH. Dislipidemias e diabetes mellitus : fisiopatologia e tratamento. Rev Ciências Médicas. 2007;16:267–77.
18. Avezedo S, Victor EG, Oliveira DC de. Diabetes mellitus e aterosclerose: noções básicas da fisiopatologia para o clínico geral: [revisão]. Rev Soc Bras Clín Méd. 2010;8(6).
19. Executiva E. Diretrizes - Consenso Da Sociedade Brasileira De Diabetes Sobre O Diagnóstico E Classificação Do Diabetes Melito E Tratamento Do Diabetes Tipo 2. Rev Bras Estud Pedagógicos [Internet]. 2020 Aug 31;101(258). Available from:

http://rbepold.inep.gov.br/index.php/rbep/article/view/4559
20. Rydén L, Standl E, Małgorzata B, Van Den Berghe G, Betteridge J, De Boer MJ, et al. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: Executive summary. The task force on diabetes and cardiovascular diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). Eur Heart J. 2007;28(1):88–136.
21. Gräsbeck R. The evolution of the reference value concept. Clin Chem Lab Med. 2004 Jan 5;42(7):692–7.
22. Schneider AJ, Ph D. Some thoughts on Normal , or Standard , Values in Clinical medicine. Am Acad Pediatr [Internet]. 1960;26(6). Available from: https://pediatrics.aappublications.org/content/26/6/973
23. Barth JH. Reference ranges still need further clarity. Ann Clin Biochem. 2009;46(1):1–2.
24. Friedberg RC. The Origin of Reference Intervals: A College of American Pathologists Q-Probes Study of "Normal Ranges" Used in 163 Clinical Laboratories. Yearb Pathol Lab Med. 2008 Jan;2008(3):3–4.
25. Clinical and Laboratory Standards Institute (CLSI). Defining , Establishing , and Verifying Reference Intervals in the Clinical Laboratory ; Approved Guideline — Third Edition. Vol. 28. 2008.
26. Amador E, Hsi BP, Massod MF. Indirect Methods for Estimating the Normal Range. Am J Clin Pathol [Internet]. 1969 Nov 1;52(5):538–46. Available from: https://academic.oup.com/ajcp/article-lookup/doi/10.1093/ajcp/52.5.538
27. Bock BJ, Dolan CT, Miller GC, Fitter WF, Hartsell BD, Crowson AN, et al. The Data Warehouse as a Foundation for Population-Based Reference Intervals. Am J Clin Pathol. 2003;120(5):662–70.
28. Aytekin M, Emerk K. Accurate Reference Intervals are Required for Accurate Diagnosis and Monitoring of Patients. Ejifcc. 2008;19(2):137–41.
29. Organização Mundial da Saúde. Constituição da Organização Mundial da Saúde adotada pela Conferência Internacional de Saúde. 2nd ed. Nova York; 1946.
30. Petitclerc C. Normality: The unreachable star? Clin Chem Lab Med. 2004;42(7):698–701.
31. Adeli K. eJIFCC Special Issue on Laboratory Reference Intervals. EJIFCC [Internet]. 2008 Oct;19(2):94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27683303
32. Gorender J. Globalização, tecnologia e relações de trabalho. Estud Avançados [Internet]. 1997 Apr;11(29):311–61. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40141997000100017&lng=pt&tlng=pt
33. Chiavegatto Filho ADP. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol e Serviços Saúde. 2015;24(2):325–32.
34. Grossi E. The REALAB Project: A New Method for the Formulation of Reference Intervals Based on Current Data. Clin Chem. 2005 Jul 1;51(7):1232–40.
35. Clemson L, Turner JR, Turner JR, Jacquez F, Raglin W, Reed G, et al. Fasting Glucose. Encycl Behav Med. 2013;784–5.
36. Tate JR, Yen T, Jones GRD. Transference and validation of reference intervals. Clin Chem. 2015;61(8):1012–5.
37. Solberg HE. Using a Hospitalized Population to Establish Reference Intervals : Pros and Cons. 1994;2206:2205–6.
38. Arzideh F. Estimation of Medical Reference Limits by Truncated Gaussian and Truncated Power Normal Distributions. Bremen Univ [Internet]. 2008;3(September). Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.1345

39. Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate of normal range. Clin Chem. 1971;17(4):275–84.
40. Bhattacharya CG. A Simple Method of Resolution of a Distribution into Gaussian Components. Biometrics. 1967;23(1):115.
41. Hemel JB, Hindriks FR, van der Slik W. Critical discussion on a method for derivation of reference limits in clinical chemistry from a patient population. J Automat Chem [Internet]. 1985 Mar;7(1):20–30. Available from: https://www.jstor.org/stable/2528285?origin=crossref
42. Dasgupta A, Wahed A. Laboratory Statistics and Quality Control. In: Clinical Chemistry, Immunology and Laboratory Quality Control [Internet]. Elsevier; 2014. p. 47–66. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780124078215000048
43. Hager HW, Bain LJ. Inferential procedures for the generalized gamma distribution. J Am Stat Assoc. 1970;65(332):1601–9.
44. Tadikamalla PR, Penn P. Random sampling from the generalized gamma distribution. Computing. 1979;23(2):199–203.
45. Huang PH, Hwang TY. On new moment estimation of parameters of the generalized gamma distribution using it's characterization. Taiwan J Math. 2006;10(4):1083–93.
46. Dadpay A, Soofi ES, Soyer R. Information measures for generalized gamma family. J Econom. 2007;138(2):568–85.
47. Nadarajah S, Gupta AK. A generalized gamma distribution with application to drought data. Math Comput Simul. 2007;74(1):1–7.
48. Nayeban S, Rezaei Roknabadi AH, Mohtashami Borzadaran GR, Khorashadizadeh M. Comparing Bhattacharyya and Kshirsagar bounds with bootstrap method. Hacettepe J Math Stat. 2018;48(2):564–79.
49. Park SY, Bera AK. Maximum entropy autoregressive conditional heteroskedasticity model. J Econom [Internet]. 2009;150(2):219–30. Available from: http://dx.doi.org/10.1016/j.jeconom.2008.12.014
50. Hoffmann RG. Statistics in the Practice of Medicine. JAMA J Am Med Assoc. 1963;185(11):864–73.
51. Shapiro ASS, Wilk MB. An Analysis of Variance Test for Normality ( Complete Samples ) Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2333709. Biometrika Trust. 1965;52(3/4):591–611.
52. Royston JP. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. Appl Stat. 1982;31(2):115.
53. Yap BW, Sim CH. Comparisons of various types of normality tests. J Stat Comput Simul. 2011;81(12):2141–55.
54. White CA, Kennedy JF. Methods of enzymatic analysis, 3rd edition, volume VI: Metabolites 1: Carbohydrates edited by H. U. Bergmeyer, J. Bergmeyer and M. GraRl, Verlag Chemie, Weinheim, 1984. pp. xxix+701. Br Polym J. 1985 Dec;17(4):379–379.
55. Lang DA, Matthews DR, Peto J, Turner RC. Cyclic Oscillations of Basal Plasma Glucose and Insulin Concentrations in Human Beings. N Engl J Med. 1979 Nov 8;301(19):1023–7.
56. Steiner DF. Adventures with insulin in the islets of langerhans. J Biol Chem. 2011;286(20):17399–421.
57. Shine B. Use of routine clinical laboratory data to define reference intervals. Ann Clin Biochem [Internet]. 2008 Sep 1;45(5):467–75. Available from: http://acb.sagepub.com/lookup/doi/10.1258/acb.2008.008028
58. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. Stat Methods Med Res [Internet]. 1999 Apr 2;8(2):93–112. Available from:

http://journals.sagepub.com/doi/10.1177/096228029900800202
59. Karson M. Handbook of Methods of Applied Statistics. Volume I: Techniques of Computation Descriptive Methods, and Statistical Inference. Volume II: Planning of Surveys and Experiments. I. M. Chakravarti, R. G. Laha, and J. Roy, New York, John Wiley; 1967, $9.00. J Am Stat Assoc. 1968;63(323):1047–9.
60. Baadenhuijsen BH, Smit JC. Indirect Estimation of Clinical Chemical Reference Intervals from Total Hospital Patient Data : Application of a Modified Bhattacharya Procedure. J Clin Chcm Clin Biochem. 1985;23(12):829–39.
61. Baadenhuijsen H, Smit JC. Indirect Estimation of Clinical Chemical Reference Intervals from Total Hospital Patient Data: Application of a Modified Bhattacharya Procedure. Clin Chem Lab Med. 1985;23(12):829–40.
62. Schork NJ, Weder AB, Schork MA. On the asymmetry of biological frequency distributions. Genet Epidemiol. 1990;7(6):427–46.
63. Ortolá J, Fuentes-Arderiu X. Intra- and interindividual biological variation of the serum concentration of phospholipids. Clin Chem [Internet]. 1991 Apr 1;37(4):583–583. Available from: https://academic.oup.com/clinchem/article/37/4/583/5649373