Identificação de Grupos de Risco para Insuficiência Renal Aguda em Pacientes Hospitalizados: Uma Análise Baseada em Descoberta de Subgrupos

Rodrigo S. Nascimento
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
rodrigosalesn@ufmg.br

Renato Vimieiro
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
rvimieiro@dcc.ufmg.br

***Resumo*—Esta pesquisa investigou os subgrupos de pacientes hospitalizados com diferentes perfis de risco no desenvolvimento da Insuficiência Renal Aguda (IRA) utilizando técnicas de descoberta de subgrupos (SD) aplicadas a dados retrospectivos de 76.957 internações do Kansas Medical Center, onde 9,4% dos casos apresentaram quadros clínicos de IRA. Foram analisadas 233 variáveis distribuídas em sete categorias. O total de 205 variáveis clínicas foram selecionadas para processamento por meio do algoritmo SSDP+, que identificou 15 subgrupos com distribuições de risco não usuais para IRA, todos validados estatisticamente. Os padrões encontrados foram estratificados com relação a taxa de incidência de IRA em três categorias: alto risco ($\geq$40% de incidência), risco moderado-alto (25–39%) e risco moderado (10–24%). Alguns marcadores comuns encontrados nos padrões identificados incluem quadros clínicos com a presença de Diabetes Mellitus, fibrose cística e infecções respiratórias. A insuficiência cardíaca emergiu como um fator de risco central, presente em 7 subgrupos. Observou-se que interações entre comorbidades, como hipertensão e insuficiência cardíaca, geram efeitos de amplificação do risco que não são capturados em análises univariadas. Esses achados permitem um melhor entendimento da necessidade de intervenções preventivas direcionadas a populações específicas.**

***Palavras-chave*—Insuficiência Renal Aguda, Descoberta de Subgrupos, Grupos de Risco**

I. Introdução

A Insuficiência Renal Aguda (IRA) representa uma das complicações mais críticas no ambiente hospitalar, caracterizada pela perda rápida da capacidade de filtração renal e acúmulo de substâncias tóxicas no organismo. Com uma incidência significativa e alta taxa de letalidade, a IRA está associada ao aumento do tempo de internação e dos custos relacionados a assistência à saúde. Nesse contexto, é importante realizar estudos para a identificação precoce de pacientes em risco dessa condição para implementação de medidas preventivas eficazes.

Este trabalho foi motivado pelo estudo desenvolvido por He et al. [1], que apresentou uma abordagem de modelagem preditiva para IRA utilizando dados de prontuários eletrônicos de 76.957 internações em um hospital acadêmico terciário. Os autores propuseram uma metodologia multi-perspectiva, contemplando quatro cenários clínicos distintos, incluindo predição no momento da admissão e predição com janelas de tempo variáveis antes do início da IRA. Utilizando cinco métodos de aprendizado de máquina, incluindo Regressão Logística e Random Forest, alcançaram desempenho considerado adequado para a prática clínica, com valores de AUC entre 0,720 e 0,764. Embora o estudo realizado represente um avanço significativo nas metodologias de predição de IRA, sua abordagem não fornece descrições claras e interpretáveis dos diferentes perfis de pacientes associados ao risco dessa condição, lacuna que o presente trabalho busca preencher através da técnica de mineração de dados conhecida como Descoberta de Subgrupos.

A. Objetivos

A pergunta de pesquisa que norteia este estudo é: Quais subgrupos de pacientes hospitalizados apresentam perfis de risco diferenciados para o desenvolvimento de insuficiência renal aguda (IRA), considerando variáveis clínicas, demográficas e laboratoriais? Assim, o objetivo principal é identificar e caracterizar esses subgrupos através da aplicação de técnicas de descoberta de subgrupos.

Os objetivos específicos incluem: (1) realizar análise exploratória abrangente de dados clínicos retrospectivos para caracterização da população estudada e identificação de padrões preliminares; (2) aplicar uma algoritmo de descoberta de subgrupos para identificação de combinações específicas de características correlacionadas ao risco elevado ou reduzido para desenvolvimento de IRA; (3) avaliar a qualidade e a significância dos subgrupos identificados através de métricas estatísticas apropriadas; e (4) interpretar os resultados no contexto clínico.

B. Estrutura do Trabalho

Este trabalho está organizado em seis seções principais, além das referências bibliográficas. A presente seção apresenta a introdução ao tema, caracterização do problema, motivação e objetivos da pesquisa. A seção II apresenta os fundamentos teóricos sobre insuficiência renal aguda e descoberta de subgrupos. A seção III é dedicada à revisão

da literatura, abordando fatores de risco associados à IRA, aplicações clínicas de descoberta de subgrupos e diferentes abordagens metodológicas existentes para essa técnica. A seção IV detalha a metodologia empregada, estruturada em duas etapas complementares: análise exploratória de dados e aplicação de técnicas de descoberta de subgrupos. A seção V apresenta os resultados obtidos, incluindo a caracterização da base de dados, geração e validação estatística dos subgrupos identificados, interpretação clínica dos padrões descobertos e discussão sobre limitações metodológicas. Por fim, a seção VI apresenta as conclusões do estudo, destacando as principais contribuições para a área e sugerindo direcionamentos para pesquisas futuras.

II. Fundamentos Teóricos

A seguir, serão apresentados os principais conceitos que estão relacionados à temática.

A. Insuficiência Renal Aguda

A Insuficiência Renal Aguda (IRA) é uma condição em que o sistema renal perde rapidamente sua capacidade de filtrar o sangue, levando ao acúmulo de substâncias tóxicas e desequilíbrios no organismo [2]. É considerada uma das complicações mais importantes de atenção em pacientes hospitalizados, estando associada ao aumento significativo das taxas de mortalidade, tempo de internação e custos totais relacionados à saúde [3]. Um estudo realizado por Susantitaphong et al. [4] mostrou em meta-análise que a IRA afeta cerca de 1 em cada 5 pacientes hospitalizados, com mortalidade variando entre 20-50% dependendo da gravidade. Turgut et al. [5] demonstrou que existe um aumento progressivo dessa incidência na última década, tanto em pacientes que evoluem para terapia renal substitutiva quanto naqueles com formas menos graves da doença.

Historicamente, a falta de consenso em relação à detecção de IRA levou à existência de mais de 30 critérios diagnósticos diferentes. Atualmente, o critério KDIGO é o que mais se destaca nas práticas clínicas pela sua abrangência de detecção [6].

Critério de diagnóstico KDIGO (*Kidney Disease: Improving Global Outcomes*): Define IRA como uma elevação da creatinina (marcador de função renal no sangue) igual ou superior a 0,3 mg/dL em 48 horas OU igual ou superior a 1,5 vezes o valor basal em um intervalo de até sete dias [7].

B. Descoberta de Subgrupos

A Descoberta de Subgrupos (*Subgroup Discovery* - SD) pode ser definida como uma técnica de mineração de dados que visa encontrar subgrupos de uma população que são estatisticamente 'mais interessantes', ou seja, são tão grandes quanto possível e apresentam características estatísticas (distribucionais) incomuns com relação a uma propriedade de interesse [8].

As técnicas de SD estão posicionadas na interseção entre indução preditiva e descritiva, combinando características de ambas as abordagens [9]. Enquanto métodos de classificação (indução preditiva) buscam prever valores de variáveis, as técnicas descritivas buscam padrões gerais nos dados sem alvos específicos, a tarefa da descoberta de subgrupos integra elementos dessas duas vertentes para extrair conhecimento que seja interpretável a partir de dados rotulados.

Outras técnicas semelhantes à SD incluem Mineração de Conjuntos Contrastantes (*Contrast Set Mining*) e Mineração de Padrões Emergentes (Emerging Pattern Mining), mas a principal diferença é que enquanto a SD tenta descrever distribuições incomuns no espaço de busca em relação a um valor da variável alvo, as outras duas técnicas buscam relações dos dados com respeito aos possíveis valores da variável alvo com base principalmente em medidas de cobertura e precisão [10].

C. Definição do problema de pesquisa

Para entender como as técnicas de SD serão empregadas na solução do problema desta pesquisa, é necessário compreender a definição de subgrupo e medidas de qualidade.

Um subgrupo S é formalmente definido como $S = \{\, e \in D \mid \text{cond}(e) \,\}$, onde D representa o conjunto de dados e e são as instâncias que satisfazem a condição $\text{cond}(e)$ (também chamada de descrição). A estrutura final do subgrupo segue o padrão "SE condição, ENTÃO propriedade-alvo", associada a uma medida de qualidade que quantifica sua significância estatística [10].

As medidas de qualidade são essenciais nesse tipo de tarefa, pois determinam quais padrões são considerados relevantes. A escolha de quais medidas usar depende diretamente do interesse específico da análise [11]. A seguir estão algumas das medidas de qualidade frequentemente usadas em algoritmos de SD [10]:

1) Suporte: Mede a frequência de exemplos corretamente classificados cobertos por uma regra. É calculado como

$$\text{Sup} = \frac{n(\text{Cond} \wedge \text{Alvo})}{N}, \tag{1}$$

onde $n(\text{Cond} \wedge \text{Alvo})$ representa o número de instâncias que satisfazem a condição Cond e também pertencem ao valor da variável alvo Alvo na regra, e N é o número total de instâncias no conjunto de dados.

2) Confiança: A confiança mede a frequência relativa de instâncias que satisfazem uma regra completa entre aquelas que satisfazem apenas o antecedente. É calculada como

$$\text{Conf} = \frac{n(\text{Cond} \wedge \text{Alvo})}{n(\text{Cond})}, \tag{2}$$

onde $n(\text{Cond})$ representa o número de instâncias que satisfazem as condições determinadas pela parte antecedente da regra.

*3) Qg (*Quality Generalization*):* Avalia a qualidade de um subgrupo ao balancear dois aspectos essenciais: a cobertura de exemplos que satisfazem tanto a condição do subgrupo quanto o alvo (precisão) e a generalização do padrão encontrado. É calculada como

$$\text{Qg} = \frac{n(\text{Cond} \wedge \text{Alvo})}{n(\text{Cond} \wedge \neg\text{Alvo}) + g}, \quad (3)$$

onde $n(\text{Cond} \wedge \neg\text{Alvo})$ contabiliza os exemplos que satisfazem a condição mas não pertencem ao alvo (ruído), e g é um parâmetro de generalização.

4) WRAcc (Acurácia Relativa Ponderada): Define-se como a precisão relativa de uma regra ponderada por sua cobertura. É calculada por

$$\text{WRAcc} = \frac{n(\text{Cond})}{N} \left(\frac{n(\text{Cond} \wedge \text{Alvo})}{n(\text{Cond})} - \frac{n(\text{Alvo})}{N} \right), \quad (4)$$

em que N é o número total de instâncias no conjunto de dados.

No contexto deste trabalho, buscamos identificar subgrupos descritivos que sejam significativos e com distribuições incomuns de ocorrências de IRA em relação à distribuição geral. Esses subgrupos são caracterizados por combinações de variáveis clínicas, laboratoriais e demográficas relacionadas aos pacientes (condição), que apresentam divergências significativas na distribuição da propriedade-alvo (variável que determina a presença ou ausência de IRA).

A Tabela I ilustra exemplos fictícios de subgrupos identificados para essa tarefa, destacando não apenas sua condição descritiva, mas também a divergência na distribuição do atributo alvo em relação à população geral. Nesses exemplos, é possível observar que os padrões descobertos evidenciam associações relevantes entre características dos pacientes e a ocorrência de IRA, permitindo identificar tanto grupos de risco quanto grupos protegidos.

TABELA I
EXEMPLOS DE ESTRUTURAS, MÉTRICAS DE QUALIDADE E DISTRIBUIÇÃO DA VARIÁVEL-ALVO PARA SUBGRUPOS IDENTIFICADOS.

Descrição do Subgrupo	**Métrica de Qualidade**	**IRA +**	**IRA -**
Idade $> 64 \wedge$ Insuficiência Cardíaca Crônica	Suporte: 6%	25%	75%
Idade 18–25 $\wedge$ Ausência de Comorbidades	Suporte: 5%	2%	98%

III. REVISÃO DA LITERATURA

Antes de adentrar na análise proposta por este trabalho, é fundamental situar a pesquisa no contexto do debate acadêmico já estabelecido. A revisão da literatura que se segue tem como propósito sistematizar o conhecimento produzido, primeiro, sobre os fatores de risco associados à insuficiência renal aguda e, em seguida, sobre as diferentes abordagens metodológicas para a descoberta de subgrupos e suas aplicações em contextos clínicos.

A. Fatores de risco associados a Insuficiência Renal Aguda

Para este estudo, um maior entendimento dos fatores de risco associados à IRA é fundamental para contextualizar os subgrupos identificados nos resultados da metodologia e analisar as suas significâncias clínicas. Existem diversas pesquisas científicas da área médica que revelam fatores de risco associados ao desenvolvimento da IRA, abrangendo características demográficas, condições clínicas pré-existentes e intervenções terapêuticas.

1) Fatores Demográficos: A idade avançada é um dos principais fatores de risco para IRA, uma vez que alterações morfofuncionais decorrentes do envelhecimento tornam os rins mais vulneráveis a disfunções [12]. Além disso, o gênero também influencia a incidência: homens apresentam maior predisposição do que mulheres, mesmo quando ajustados fatores como status socioeconômico, etnia e hábitos nocivos à saúde [13].

2) Fatores Relacionados a Comorbidades e Histórico Médico: Diversas condições clínicas elevam o risco de IRA. A hipertensão arterial sistêmica (HAS) duplica as chances de desenvolvimento da doença [12], enquanto o diabetes mellitus (DM) contribui significativamente devido a alterações microvasculares que prejudicam a circulação renal [14]. Pacientes com doença renal crônica (DRC) prévia têm risco mais que dobrado de evoluir para quadros graves de insuficiência, dada sua reduzida reserva funcional [14]. Doenças hepáticas crônicas, como cirrose, também aumentam a vulnerabilidade devido a complicações hepatorrenais e a falta de volume sanguíneo adequado no organismo [15]. A insuficiência cardíaca é outro fator crítico, elevando o risco em mais de cinco vezes ao comprometer o débito cardíaco e, consequentemente, a circulação de sangue nos rins [12].

Intervenções cirúrgicas também desempenham um papel relevante. Grandes cirurgias, especialmente cardíacas ou abdominais (como as envolvendo fígado ou intestino), sobrecarregam os rins e podem levar a falta de circulação sanguínea suficiente para o funcionamento adequado do sistema renal [16]. Da mesma forma, pacientes submetidos a transplantes de órgãos como coração, fígado e pulmão têm risco aumentado devido ao uso de imunossupressores e à maior incidência de infecções pós-operatórias [17].

3) Fatores de Risco Relacionados a Intervenções Terapêuticas: Algumas classes de medicamentos estão diretamente associadas à IRA. Os anti-inflamatórios não esteroides (AINEs), por exemplo, inibem a síntese de hormônios essenciais para a vasodilatação renal, podendo causar lesões especialmente em idosos, pacientes com quadro de desidratação ou com alguma doença renal prévia [5]. Drogas vasoativas, como vasopressores, também representam um risco significativo, pois reduzem o fluxo sanguíneo renal por meio da vasoconstrição, levando à isquemia [14]. Além disso, o uso de antibióticos, principalmente em combinações de diferentes classes, pode quadruplicar o risco de IRA devido ao efeito nefrotóxico de certos fármacos [12].

Além da compreensão dos marcadores associados a IRA, é

importante entender como a descoberta de subgrupos pode ser útil na área médica e suas principais abordagens.

B. Aplicações da Descoberta de Subgrupos em contextos clínicos

O uso de técnicas de SD no contexto clínico tem demonstrado resultados promissores em diversas áreas. No campo da oncologia, Gómez-Bravo et al. [18] aplicaram uma nova proposta de algoritmo, denominado IGSD (InfoGained Subgroup Discovery), para analisar padrões de tratamento em pacientes com câncer de pulmão. O método combinou Information Gain e Odds Ratio para identificar padrões com maior aceitação clínica e relevância estatística. Os resultados demonstraram que o IGSD superou métodos tradicionais, como FSSD [19] e SSD++ [20], na descoberta de padrões relacionados às características dos pacientes, tratamentos prescritos e desfechos clínicos. O resultado desse trabalho evidencia o valor da customização de algoritmos SD para contextos clínicos específicos, fornecendo um possível caminho metodológico para para diversas abordagens da área médica.

Em um estudo com foco em COVID-19, Vagliano et al. [21] aplicaram três métodos de descoberta de subgrupos (SSD++, PRIM e APRIORI-SD) a dados de 14.548 pacientes internados em UTIs holandesas para compreender a heterogeneidade do risco de mortalidade hospitalar. Os resultados indicaram ampla variação no número de subgrupos identificados (de 5 a 62), com o SSD++ se destacando tanto nas métricas quantitativas quanto na aceitação clínica, com 91% dos subgrupos considerados relevantes pelos especialistas da área. Esses resultados demonstram a aplicabilidade de técnicas SD na estratificação de risco em ambientes críticos.

Yang et al. [22] apresentaram o algoritmo SDLD (Subgroup Discovery for Longitudinal Data), desenvolvido especificamente para dados longitudinais em prontuários eletrônicos. O método combinou árvores de interação generalizada com estimadores de máxima verossimilhança direcionada para identificar pacientes com HIV que apresentavam maior risco de ganho de peso ao receber terapias antirretrovirais contendo um princípio ativo específico. A abordagem identificou o gênero dos pacientes como principal modificador de efeito, com mulheres apresentando ganho de peso significativamente maior.

A revisão desses trabalhos demonstra que os algoritmos SD são eficazes em identificar padrões clinicamente relevantes em tratamentos e fatores de risco associados a condições como câncer, COVID-19 e HIV. Os resultados observados sugerem um potencial para uma aplicação similar no contexto deste trabalho, uma vez que também buscamos padrões associados ao desenvolvimento de IRA sem depender de hipóteses pré-especificadas.

C. Abordagens de Descoberta de Subgrupos

Em SD, o espaço de busca dos subgrupos é definido pela linguagem de descrição adotada, que especifica quais atributos e operadores podem ser combinados nas descrições de um subgrupo (exemplo: apenas conjunções ‘E’ entre atributos binários ou discretizados) e impõe restrições de complexidade (exemplo: máximo de 2 atributos por condição) [8] [11]. A dimensão desse espaço tem uma relação exponencial com o número de características e valores considerados. Isso torna necessária a aplicação de estratégias de busca eficientes para lidar com as limitações de poder computacional [24].

Para percorrer esse espaço de forma eficiente, duas abordagens principais são utilizadas nos algoritmos de SD: abordagem exaustiva e abordagem heurística.

D. Métodos Exaustivos de Descoberta de Subgrupos

Os métodos exaustivos garantem encontrar todos os padrões de subgrupos com algum grau de relevância contidos em um conjunto de dados, percorrendo sistematicamente todo o espaço de busca [25]. Dessa forma, esses métodos realizam uma busca completa de padrões possíveis, evitando assim a presença de ”pontos cegos” que ocorrem em abordagens heurísticas.

Atzmueller [8] destaca que essas abordagens são especialmente úteis para métodos exploratórios e descritivos, pois asseguram que todos os padrões verdadeiramente interessantes serão identificados. A importância disso ocorre em casos onde o usuário está interessado em compreender completamente o espaço de soluções e não deseja perder nenhum padrão potencialmente valioso.

Para lidar com o espaço de busca exponencial, os algoritmos exaustivos empregam técnicas de poda baseadas em propriedades como suporte mínimo e limites superiores em medidas de qualidade como a acurácia relativa ponderada (WRAcc) [26]. Estas estratégias permitem reduzir significativamente o espaço de busca sem perder algum padrão.

Alguns exemplos de algoritmos clássicos para essa abordagem incluem:

1) Apriori-SD: Desenvolvido por Kavšek e Lavrač [27], o Apriori-SD é uma adaptação do algoritmo Apriori para descoberta de subgrupos. Ele começa gerando subgrupos com um único seletor e verifica se atendem a um suporte mínimo pré-definido. Em seguida, combina iterativamente esses subgrupos válidos para formar candidatos mais complexos, avaliando sua qualidade e podando aqueles que não atingem o suporte mínimo (evitando assim a expansão desnecessária de ramos inúteis). O processo se repete até que nenhum novo candidato possa ser gerado, e um pós-processamento remove subgrupos redundantes, mantendo apenas os mais relevantes.

2) SD-Map: Proposto por Atzmueller e Puppe [25], o SD-Map adapta o método FP-growth para descoberta de subgrupos. Ele constrói uma árvore de padrões frequentes (FP-tree) que armazena seletores relevantes de forma compacta e hierárquica, eliminando a necessidade de gerar candidatos repetidamente. A mineração é feita de forma recursiva, calculando a qualidade dos subgrupos diretamente durante o processo, sem etapas intermediárias de geração e teste. O algoritmo também lida com valores ausentes, garantindo robustez em dados reais, e finaliza com uma seleção que filtra redundâncias, retendo apenas os subgrupos mais significativos.

E. Métodos Heurísticos de Descoberta de Subgrupos

As abordagens heurísticas visam uma exploração mais direcionada do espaço de busca, percorrendo os caminhos potencialmente mais promissores. Esses métodos são particularmente úteis quando a busca exaustiva é computacionalmente inviável, como em conjuntos de dados grandes ou de alta dimensionalidade [10]. Dessa maneira, esses métodos sacrificam a garantia de encontrar todos os padrões possíveis em favor de soluções aproximadas que podem ser obtidas em tempo computacional razoável [26].

Flexibilidade e adaptabilidade a diferentes tipos de dados e restrições de recursos são outras características dos algoritmos desse grupo. Eles frequentemente incorporam mecanismos para promover a diversidade entre os subgrupos descobertos, o que é importante para muitos cenários de aplicação. Outra característica frequente é a capacidade de interromper a busca a qualquer momento e fornecer os melhores subgrupos encontrados até então, conhecida como propriedade ”any-time”. Isso é particularmente útil em aplicações com restrições de tempo ou quando os usuários desejam feedbacks iterativos durante o processo de descoberta [28].

Alguns exemplos de algoritmos representativos desse tipo abordagem incluem:

*1) Busca em Feixe (*Beam Search*):* Este método heurístico mantém apenas os k melhores subgrupos (definidos pela largura do feixe) em cada iteração, expandindo-os de forma gulosa [10]. Ele começa com subgrupos simples e avança gradualmente para combinações mais complexas, usando uma função de qualidade para selecionar quais serão mantidos. Embora seja rápido para feixes estreitos, tende a gerar redundâncias e pode perder ótimos locais mais profundos no espaço de busca devido à sua natureza gulosa [28].

2) PRIM: O PRIM (*Patient Rule Induction Method*), proposto por Friedman e Fisher [29], é um algoritmo que refina subgrupos em forma de ”caixas” através de fases de encolhimento e expansão. Na etapa de encolhimento, remove-se pequenas porções das bordas para aumentar a média da variável alvo dentro do subgrupo. Já na expansão, tenta-se ampliar a caixa sem prejudicar significativamente a qualidade. É eficaz para atributos numéricos e permite um controle intuitivo entre cobertura e precisão, mas pode ser computacionalmente custoso em grandes conjuntos de dados [8].

*3) SSDP (*Simple Search Discriminative Patterns*):* Projetado para alta dimensionalidade, o SSDP [30] é uma abordagem evolutiva que representa padrões candidatos como conjuntos de inteiros, usando tabelas hash para evitar duplicações. Inicia com todos os padrões unidimensionais e aplica operadores genéticos adaptativos (seleção por torneio, mutação e crossover), com taxas que se ajustam dinamicamente. O critério de parada ocorre quando os melhores k padrões se estabilizam após reinicializações da população. O método é simples, exige poucos parâmetros e evita redundâncias.

4) MCTS (Monte Carlo Tree Search): Adaptado para descoberta de subgrupos por Bosc et al. [28], o MCTS combina busca em árvore com simulações aleatórias, usando o critério UCB (Upper Confidence Bound) para balancear exploração e explotação. A cada iteração, seleciona nós promissores, expande a árvore de forma assimétrica e retropropaga os resultados das simulações. Diferente de métodos gulosos, o MCTS é eficaz em encontrar múltiplos ótimos locais distribuídos no espaço de busca.

IV. METODOLOGIA

Compreendendo os conceitos envolvidos para a tarefa de estudo, a metodologia foi estruturada em duas partes complementares, sendo a primeira voltada para a identificação e análise de dados existentes sobre a ocorrência de insuficiência renal aguda em hospitalizações, e a segunda, destinada à aplicação de técnicas de descoberta de subgrupos e à avaliação qualitativa de seus resultados, com o intuito de atingir o objetivo principal da pesquisa.

A. Análise de Dados

Conforme mencionado na seção I, a base de dados utilizada para responder a pergunta de pesquisa elaborada foi usada previamente em estudos de predição da IRA via aprendizado supervisionado.

Com essa fonte de dados, foi conduzida uma análise exploratória para compreender a estrutura e características do seu conteúdo. Essa etapa incluiu: (1) estudo de quantificação e categorização das variáveis; (2) análise de proporção de valores para as variávies identificadas; (3) e verificação das ocorrências mais comuns para variáveis clínicas que caracterizam o estado clínico de um paciente e sua relação direta com o quadro de IRA na população estudada.

Com base nos resultados da análise exploratória, foram definidas as estratégias de pré-processamento necessárias, incluindo tratamento de valores ausentes e codificação de variáveis categóricas. A definição do desfecho de interesse (desenvolvimento de IRA) foi estabelecida de acordo com os critérios utilizados no estudo original que disponibilizou os dados, garantindo consistência com os padrões clínicos reconhecidos.

B. Aplicação da Descoberta de Subgrupos

A segunda parte deste estudo é composta da aplicação efetiva da técnica de descoberta de subgrupos aos dados explorados na etapa anterior. Essa etapa incluiu: (1) geração dos subgrupos; (2) análise da significância estatística dos padrões gerados; (3) e análise interpretativa dos padrões em contexto clínico.

Para a geração dos subgrupos, foi utilizado o algoritmo SSDP+ (*Simple Search Discriminative Patterns Plus*) [31], selecionado por sua capacidade de lidar com a alta dimensionalidade da base de dados em análise (233 variáveis). Conforme apresentado na seção III.E, a abordagem evolutiva do SSDP emprega estratégias de busca dirigida para explorar o espaço de soluções de forma eficiente, mantendo a capacidade de identificar padrões relevantes.

Uma característica distintiva da versão plus em relação à versão original do SSDP é a incorporação de mecanismos de análise de similaridade entre subgrupos. Essa funcionalidade

é útil para identificar e eliminar subgrupos redundantes, isto é, aqueles que, embora apresentem descrições diferentes, cobrem essencialmente as mesmas populações de pacientes ou revelam padrões equivalentes. No contexto da pesquisa, esse recurso permite concentrar a análise em subgrupos distintos, evitando interpretações repetitivas de padrões semelhantes.

A aplicação do SSDP+ foi configurada com os seguintes parâmetros:

Parâmetro k: Define o número de subgrupos a serem retornados pelo algoritmo com base em uma ordem descendente de qualidade. Foi estabelecido o valor k = 15 visando equilibrar a diversidade de padrões identificados com a viabilidade de análise clínica detalhada dos resultados sem gerar um volume excessivo de padrões.

Métrica de qualidade: A métrica Qg (seção II.C) foi selecionada para medir a qualidade dos subgrupos por sua capacidade de balancear precisão e generalização. Após alguns testes iniciais para o escolher o valor ideal do parâmetro de generalização g, o valor de 75 foi o mais adequado para favorecer subgrupos que apresentam não apenas alta concentração de casos positivos de IRA, mas também cobertura populacional significativa.

Variável-alvo: O rótulo de Insuficiência Renal Aguda com valor positivo, representando pacientes que desenvolveram IRA durante a hospitalização.

Além dos parâmetros de configuração básica do algoritmo, para avaliação de subgrupos redundantes, foram configurados os seguintes parâmetros no módulo de similaridade do SSDP+:

Medida de similaridade: Utilizado o índice de Jaccard para quantificar a sobreposição entre populações de pacientes cobertas por diferentes subgrupos.

Tamanho da cache: Define a quantidade de subgrupos na memória para comparações de similaridade durante o processo evolutivo. Para a análise feita, foi considerado o tamanho 10.

Limiar de similaridade: Define o quanto dois subgrupos podem ser parecidos antes de serem considerados redundantes. Se sua similaridade (calculada pela cobertura de exemplos positivos) for maior que esse limiar, apenas o melhor subgrupo é mantido na lista principal dos top-k subgrupos, enquanto versões similares são armazenadas no cache secundário. Isso garante diversidade nos resultados sem descartar informações potencialmente relevantes. Foi estabelecido o valor de 0,60 para permitir certa variação nas descrições dos subgrupos enquanto evita redundâncias significativas na cobertura populacional.

Considerando que o SSDP+ foi projetado para trabalhar com dados categóricos, foi necessária uma etapa de transformação dos valores das variáveis para adequação ao formato exigido pelo algoritmo. As variáveis originalmente binárias foram convertidas em categorias explícitas [TRUE] e [FALSE]. Os rótulos da variável-alvo foram codificados como ”p” (positivo) para pacientes que desenvolveram IRA e ”n” (negativo) para aqueles que não apresentaram a condição durante a internação. Variáveis demográficas, sinais vitais e exames laboratoriais mantiveram seus valores com as categorizações originais.

Um aspecto de decisão metodológica importante refere-se à exclusão das variáveis relacionadas a medicamentos de todas as análises apresentadas. Essa decisão foi motivada por incertezas identificadas no processo de marcação dessas variáveis na base de dados original. Especificamente, foi verificado que o registro temporal de administração de medicamentos em pacientes que não desenvolveram IRA apresenta ambiguidades quanto ao período de referência considerado, o que poderia introduzir um erro significativo em padrões que possam apontar essa categoria de variável.

A significância estatística dos subgrupos identificados pelo SSDP+ foi avaliada por meio de um teste de permutação. Esse procedimento verifica se os padrões descobertos são, de fato, relevantes ou se poderiam ter ocorrido simplesmente ao acaso.

O teste de permutação é baseado no princípio de que, se os padrões identificados são reais, isto é, se existe associação verdadeira entre as características do subgrupo e o desenvolvimento de IRA, esses padrões não devem emergir quando a relação entre variáveis clínicas e desfecho é quebrada de forma artificial. Conforme o estudo prévio realizado por Duivesteijn et al. (2011) [32], essa análise foi conduzida com os seguintes passos:

Permutação dos rótulos: Os rótulos da variável-alvo foram aleatoriamente permutados m vezes mantendo todas as demais variáveis fixas. Este processo remove qualquer relação real entre as características dos pacientes e o desfecho, criando um cenário sob a hipótese nula de ausência de associação. Foram realizadas 5000 permutações.

Recálculo das métricas: Para cada permutação, a métrica Qg foi recalculada para todos os subgrupos originalmente identificados, utilizando suas descrições (conjuntos de condições) fixas mas aplicadas aos rótulos permutados.

Comparação com valores originais: Foi realizado um registro de n vezes onde o valor de Qg obtido com os dados permutados (Qg permutado) foi igual ou superior ao valor Qg original de cada subgrupo.

Cálculo do p-valor: O p-valor representa a probabilidade de observar um valor de qualidade igual ou superior ao obtido originalmente caso não houvesse associação real entre as características do subgrupo e o desfecho. Subgrupos com p-valores baixos (tipicamente p ¡ 0,05) indicam que é improvável obter valores de Qg tão elevados por acaso, fornecendo evidência estatística de que o padrão identificado é genuíno. Para cada subgrupo, o p-valor foi calculado como

$$\text{p-valor} = \frac{n(\text{Qg}_{\text{permutado}} \geq \text{Qg}_{\text{original}}) + 1}{m + 1} \tag{5}$$

Após a identificação e validação estatística dos subgrupos, foi feita uma análise da relevância clínica dos padrões descobertos. Um dos diferenciais da abordagem adotada nessa metodologia é a identificação de combinações de fatores que, em conjunto, impactam o risco de IRA de modo não detectável por análises univariadas tradicionais. A detecção dessas interações multivariadas fundamentou uma estratificação de risco populacional, distinguindo grupos de pacientes de alto, moderado-alto e moderado risco comparado a taxa global de

incidência de IRA na população estudada. Além disso, os padrões desses perfis foram confrontados com os fatores de risco descritos na literatura médica para verificar se estão de acordo com o conhecimento já estabelecido ou revelam associações que merecem investigações adicionais.

V. RESULTADOS

A. Análise de Dados

Conforme descrito na seção I, a base de dados selecionada foi descrita no estudo realizado por He et al. [1]. Esses dados foram obtidos do repositório clínico HERON (Health Enterprise Repository for Ontological Narration). Os dados são retrospectivos de pacientes internados na universidade University of Kansas Medical Center (KUMC), localizada em Kansas City, EUA, coletados entre novembro de 2007 e dezembro de 2016. A população de estudo consiste em adultos ($>$18 anos) com internação hospitalar de pelo menos 2 dias. A coorte inicial contemplou 179.370 internações correspondentes a 96.590 pacientes.

Enquanto He et al. [1] focou em modelos preditivos para IRA, este estudo adota uma abordagem descritiva, caracterizando os pacientes e identificando padrões nos dados. Assim, complementa trabalhos anteriores ao revelar fatores associados à IRA que apoiam a interpretação de modelos preditivos e a formulação de hipóteses clínicas.

He et al. [1] analisou e classificou os dados em diferentes perspectivas temporais para estudar a predição da IRA com o uso de aprendizado de máquina supervisionado. Para este estudo, utilizamos o conjunto correspondente à primeira perspectiva de análise descrita pelos autores, compreendendo dados coletados até um dia antes do desenvolvimento de IRA em um paciente. Para os pacientes que não apresentaram o quadro de IRA, os dados coletados correspondem ao total do seu tempo de internação.

Os dados dessas internações apresentam 233 variáveis. O conteúdo original apresentava 1.287 variáveis, mas existiu um processo de filtragem para anonimizar os pacientes da amostra visando o compartilhamento público. Além disso, esse conjunto final de dados passou por critérios de exclusão elaborados pelos autores para eliminar as internações que não atendem aos requisitos para avaliar o desenvolvimento de IRA durante a hospitalização. Após aplicação desses critérios, a coorte final analisada passou a ter 76.957 internações, sendo que destas, 7.259 (9,4%) apresentaram episódios de insuficiência renal aguda.

Foram excluídas:

1) Internações sem dados suficientes para determinação do desfecho (se o paciente desenvolveu ou não IRA).

2) Internações com evidência de disfunção renal moderada ou grave pré-existente. Esse critério de exclusão visou focar apenas nos casos de IRA adquiridas apenas durante a hospitalização.

O desfecho de interesse (incidência de IRA) foi definido utilizando o critério KDIGO baseados nos valores de creatinina sérica (SCr) de exames clínicos. O valor basal de SCr foi definido como a última medição disponível na janela de até 2 dias antes da admissão ou a primeira medição após a admissão dos pacientes. Todos os valores de SCr entre a admissão e a alta foram avaliados para determinar a ocorrência de IRA.

As variáveis registradas foram organizadas em 7 categorias, conforme apresentado na Tabela II.

TABELA II
DETALHES DAS VARIÁVEIS POR CATEGORIA

Categoria	Número de variáveis	Detalhes
Demográfica	3	Idade, Gênero e Etnia
Sinais Vitais	5	IMC, pressão arterial diastólica, pressão arterial sistólica, pulso e temperatura.
Exames Laboratoriais	14	Albumina, ALT, AST, amônia, bilirrubina, ureia, cálcio, CK-MB, CK, glicose, lipase, plaquetas, troponina e leucócitos
Comorbidades	26	Condições pré-existentes baseadas no núcleo de comorbidades da United Health Care (UHC)
Diagnósticos de Admissão	12	Baseadas no sistema APR-DRG (All Patient Refined Diagnosis Related Group)
Medicações	28	Mapeadas para princípios ativos RxNorm (padrão mantido pela U.S. National Library of Medicine)
Histórico Médico	144	Códigos de diagnóstico ICD-9 (Classificação Internacional de Doenças, 9ª versão) agrupados

As variáveis demográficas têm seus valores categorizados da seguinte forma:

1) Idade: Categorizada em intervalos de 18-25, 26-35, 36-45, 46-55 e 56-64 e $>$64 anos.

2) Gênero: Masculino ou feminino.

3) Etnia: Branco, Afro-americano, Asiático ou Outros.

As distribuições das variáveis demográficas Idade e Etnia, apresentadas na Figura 1 e Figura 2, indicam predominância de indivíduos mais velhos, com maior proporção na faixa acima de 64 anos (27%), seguida por 56–64 anos (23%) e 46–55 anos (20%). A amostra é majoritariamente composta por indivíduos brancos (76%). Em relação ao gênero, apresenta leve predominância feminina, com 54,8% doss pacientes identificados como mulheres.

As variáveis de contexto clínico foram representadas pelo último valor registrado antes do evento de IRA. Sinais vitais foram agrupados em faixas de medição (Tabela III), e exames laboratoriais foram categorizados como “presente e normal”, “presente e anormal” ou “desconhecido”, seguindo intervalos de referência padrão. Valores ausentes foram classificados como “desconhecidos”.

As variáveis das categorias comorbidades, histórico médico e diagnósticos de admissão foram simplificadas para valores binários (”Sim/Não”), dada a natureza dos dados. A categoria

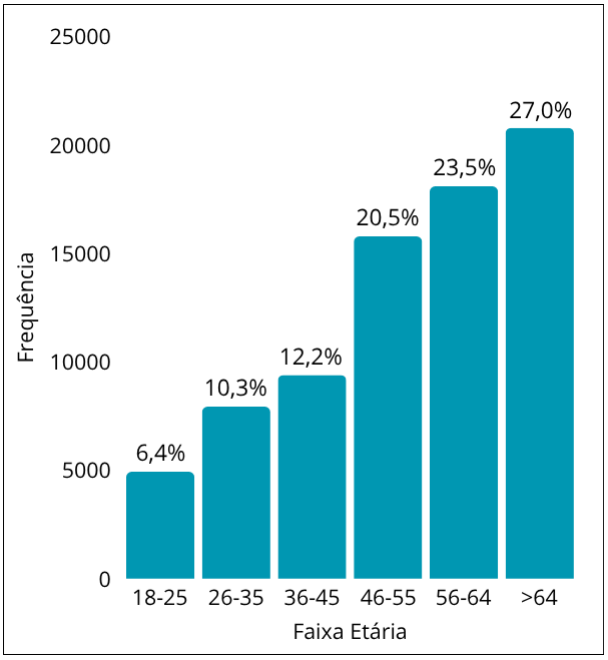

Figura 1. Distribuição da variável Idade na base de dados.

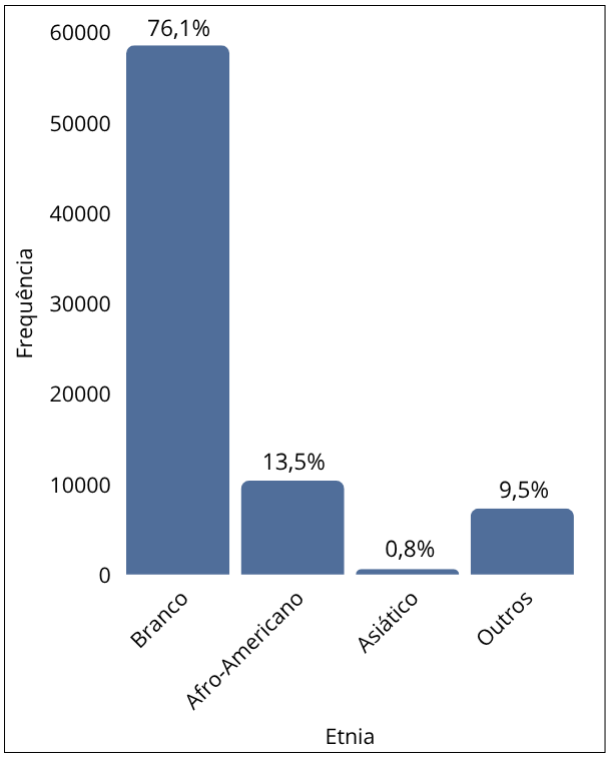

Figura 2. Distribuição da variável Etnia na base de dados.

TABELA III
FAIXAS DE MEDIÇÕES PARA OS SINAIS VITAIS NO CONJUNTO DE DADOS. A TEMPERATURA, QUE ESTAVA ORIGINALMENTE EM FAHRENHEIT, FOI CONVERTIDA PARA CELSIUS.

Sinal Vital	Intervalos (Valores)
IMC	$<$18.5, [18.5–24.9], [25.0–29.9], $\geq$30.0
Pressão Arterial Diastólica (mmHg)	$<$80, [80–89], [90–99], $\geq$100
Pressão Arterial Sistólica (mmHg)	$<$120, [120–139], [140–159], $\geq$160
Pulso (bpm)	$<$50, [50–65], [66–80], [81–100], $>$100
Temperatura (°C)	$<$35.0, [35.0–36.4], [36.5–37.5], [37.6–40.0], $>$40.0

de medicamentos foi excluída das análises conforme apontado na seção IV.B.

Uma outra perspectiva importante de análise é a busca da compreensão de possíveis relações dessas variáveis com um desfecho positivo ou negativo de IRA. As tabelas IV, V, e VI mostram a proporção das cinco variáveis de maior ocorrência na base de dados para cada categoria analisada. Os gráficos das figuras 3, 4 e 5 apresentam as proporções de casos positivos e negativos de IRA para essas variáveis.

TABELA IV
VARIÁVEIS DE MAIOR OCORRÊNCIA NA CATEGORIA COMORBIDADES CONSIDERANDO A POPULAÇÃO GERAL.

Variável	Total de Ocorrências	Proporção
Hipertensão	35.794	46,5%
Diabetes c/ complicações crônicas	14.107	18,3%
Perda de peso	13.812	17,9%
Diabetes s/ complicações crônicas	13.608	17,7%
Doença pulmonar crônica	12.725	16,5%

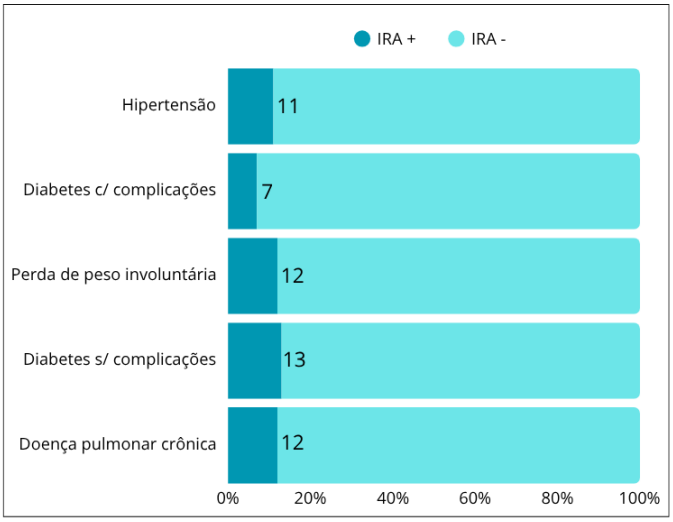

Figura 3. Distribuição dos casos de IRA para variáveis mais frequentes da categoria Comorbidades.

Apesar de analisar as variáveis de forma isolada, os resultados indicam tendências relevantes que podem apoiar futuras análises clínicas dos subgrupos identificados. No grupo das comorbidades, observa-se que, embora a hipertensão seja a mais comum, sua taxa de IRA se mantém próxima à média populacional (9,4%). Em contraste, condições como diabetes sem complicações crônicas e perda de peso involuntária apresentam taxas de IRA bem mais elevadas. Já o diabetes com complicações crônicas exibe uma taxa inferior à do diabetes sem complicações, sugerindo possíveis diferenças no manejo clínico e indicando a necessidade de investigação adicional.

Os diagnósticos de admissão revelam uma grande variabilidade no risco de IRA associado às variáveis. Infecções graves como septicemia e tratamentos oncológicos, como a quimioterapia parecem estar ligados a taxas mais elevadas de IRA, enquanto procedimentos eletivos, como cirurgias ortopédicas programadas, apresentam esse risco associado bastante reduzido.

TABELA V
VARIÁVEIS DE MAIOR OCORRÊNCIA NA CATEGORIA DIAGNÓSTICO DE ADMISSÃO CONSIDERANDO A POPULAÇÃO GERAL.

Variável	Total de Ocorrências	Proporção
Septicemia	2.845	3,7%
Reabilitação	1.919	2,5%
Quimioterapia	1.811	2,4%
Subst. articulação do joelho	1.747	2,3%
Distúrbios vesícula biliar	1.681	2,2%

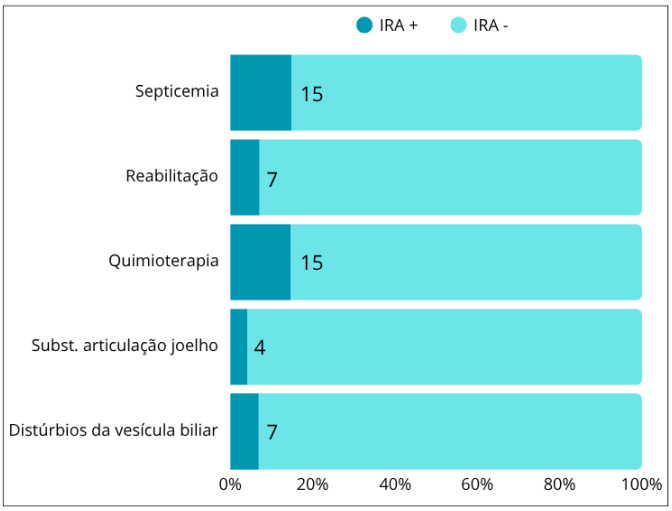

Figura 4. Distribuição dos casos de IRA para variáveis mais frequentes da categoria Diagnóstico de Admissão.

Por fim, o histórico médico mostra um padrão mais homogêneo, com taxas de IRA próximas da taxa geral, sugerindo que condições crônicas estáveis não são, por si só, possíveis fortes preditores de risco.

TABELA VI
VARIÁVEIS DE MAIOR OCORRÊNCIA NA CATEGORIA HISTÓRICO MÉDICO CONSIDERANDO A POPULAÇÃO GERAL.

Variável	Total de Ocorrências	Proporção
Hipertensão	24.033	31,2%
Outros cuidados pós-tratamento	23.761	30,9%
Distúrbios do metabolismo lipídico	15.689	20,4%
Outros distúrbios do sistema nervoso	14.178	18,4%
Outros distúrbios gastrointestinais	13.363	17,4%

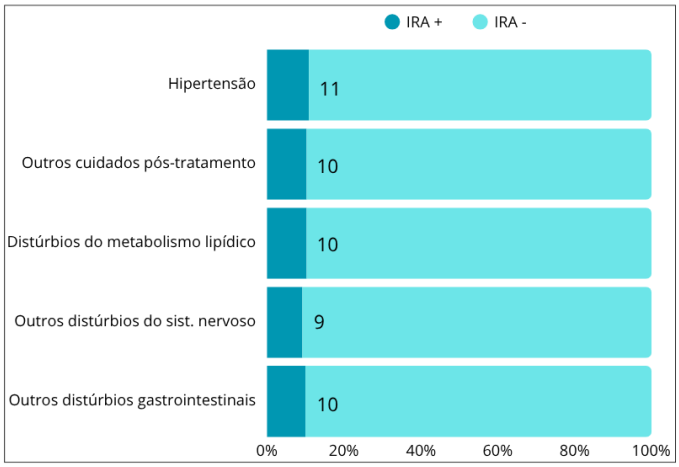

Figura 5. Distribuição dos casos de IRA para variáveis mais frequentes da categoria Histórico Médico.

B. Geração e Validação Estatística dos Subgrupos

A aplicação do algoritmo SSDP+ aos dados resultou na identificação de subgrupos com distribuições variadas e de alta incidência de IRA em relação à taxa global observada na população estudada. Os dados dos 15 subgrupos com os maiores valores de qualidade pela métrica Qg e os respectivos resultados do teste de permutação são descritos nos quadros Tabela VII e Tabela VIII.

A validação estatística por meio do teste de permutação demonstrou que os padrões descobertos são confiáveis. Todos os 15 subgrupos apresentaram p-valores de 0,0002, indicando que a probabilidade de obter os valores de qualidade Qg observados ao acaso é muito baixa.

O gráfico ilustrado na Figura 6 mostra a distribuição dos valores de Qg permutados em relação aos valores originais para cada subgrupo e reforça visualmente essa significância. É possível observar uma separação clara entre os valores originais de Qg e a distribuição dos valores permutados. Nenhuma das permutações apresentou um valor de Qg que superasse o valor original correspondente.

C. Interpretação Clínica dos Resultados

Os resultados que caracterizam a relação dos padrões identificados com a taxa de incidência da IRA permitiram a

TABELA VII
PRINCIPAIS SUBGRUPOS IDENTIFICADOS, COM SEUS PADRÕES, QUALIDADE (QG), COBERTURA E PROPORÇÃO DE CASOS DE IRA+.

Subgrupo	Padrão	Qg	Cobertura	IRA+ (%)
S1	Quimioterapia ∧ Câncer Metastático ∧ Transtornos de Coagulação e Hemorrágicos	0,505	83	64
S2	Perda de Peso ∧ Pneumonia ∧ Outras Infecções Respiratórias ∧ Fibrose Cística	0,361	87	49
S3	Idade 18 – 25 Anos ∧ Fibrose Cística ∧ Septicemia ∧ Diabetes Mellitus	0,333	97	44
S4	Etnia Afro-Americana ∧ Insuficiência Cardíaca ∧ Hipertensão Essencial	0,286	343	27
S5	Anemia ∧ Fibrose Cística ∧ Septicemia ∧ Diabetes Mellitus	0,278	118	36
S6	Pneumonia ∧ Anemias ∧ Insuficiência Cardíaca Congestiva	0,271	441	25
S7	Perda de Peso ∧ Anemia ∧ Outras Infecções, Incluindo Verminoses	0,270	137	33
S8	Sexo Feminino ∧ Diabetes Mellitus ∧ Insuficiência Cardíaca	0,258	76	41
S9	Transtornos Eletrolíticos ∧ Pneumonia ∧ Insuficiência Cardíaca	0,256	249	27
S10	Obesidade ∧ Pneumonia ∧ Insuficiência Cardíaca	0,251	204	27
S11	Choque ∧ Hipertensão Essencial ∧ Transtornos Eletrolíticos	0,251	324	25
S12	Insuficiência Cardíaca ∧ Doença Pulmonar Crônica	0,248	1.284	21
S13	Insuficiência Cardíaca ∧ Perda de Peso ∧ Insuficiência Respiratória	0,246	82	38
S14	Septicemia ∧ Outras Infecções Respiratórias ∧ Fibrose Cística	0,246	62	44
S15	Transtornos Esofágicos ∧ Septicemia ∧ Transtornos Pancreáticos (Não Diabetes)	0,223	260	23

TABELA VIII
RESULTADOS DO TESTE DE PERMUTAÇÃO PARA OS SUBGRUPOS GERADOS.

Subgrupo	Qg original	Qg permutado (média)	Qg permutado (máximo)	p-valor
S1	0,505	0,053	0,129	0,0002
S2	0,361	0,054	0,133	0,0002
S3	0,333	0,056	0,147	0,0002
S4	0,286	0,084	0,139	0,0002
S5	0,278	0,062	0,142	0,0002
S6	0,271	0,088	0,147	0,0002
S7	0,270	0,065	0,140	0,0002
S8	0,258	0,050	0,153	0,0002
S9	0,256	0,078	0,149	0,0002
S10	0,251	0,074	0,139	0,0002
S11	0,251	0,083	0,143	0,0002
S12	0,248	0,098	0,133	0,0002
S13	0,246	0,052	0,163	0,0002
S14	0,246	0,045	0,114	0,0002
S15	0,223	0,080	0,151	0,0002

classificação dos pacientes em diferentes categorias de risco comparadas à taxa global de 9,4%. Os padrões revelaram uma amplitude considerável de risco, com taxas de IRA positiva (IRA^{+}) variando de 21% a 64%. A estratificação dos pacientes foi feita em três categorias adaptadas com base em um estudo realizado por Erfurt et al. (2023) [33]: alto risco (taxa de IRA+ $\geq 40\%$), risco moderado alto (entre $25\% \leq$ IRA+ $\leq 39\%$) e risco moderado (entre $10\% \leq$ IRA+ $\leq 24\%$).

D. Subgrupos de alto risco (IRA+ ≥ 40%)

Cinco subgrupos apresentaram taxas superiores a 40%, caracterizando perfis de vulnerabilidade alta que podem ser considerados de atenção clínica prioritária.

O subgrupo S1 (Tabela VII) é o de maior risco identificado (IRA+ = 64%) e é caracterizado pela combinação de diagnóstico de admissão por quimioterapia, presença de câncer metastático e histórico de distúrbios hematológicos. A quimioterapia, especialmente em pacientes oncológicos, é reconhecidamente nefrotóxica, podendo causar lesões graves e diretas nos rins [34]. Pacientes com câncer metastático já apresentam maior vulnerabilidade devido ao estado inflamatório sistêmico, desnutrição e possível comprometimento da oxigenação do tecido renal [35]. Os episódios hemorrágicos nos históricos médicos desses pacientes também indicam a possibilidade de comprometimento da saúde renal [36].

O subgrupo S2 (IRA+ = 49%) combina perda de peso involuntária, pneumonia, outras infecções do trato respiratório superior e fibrose cística. A perda de peso involuntária é um marcador de fragilidade clínica e desnutrição que, conforme identificado na análise exploratória do estudo, está associada

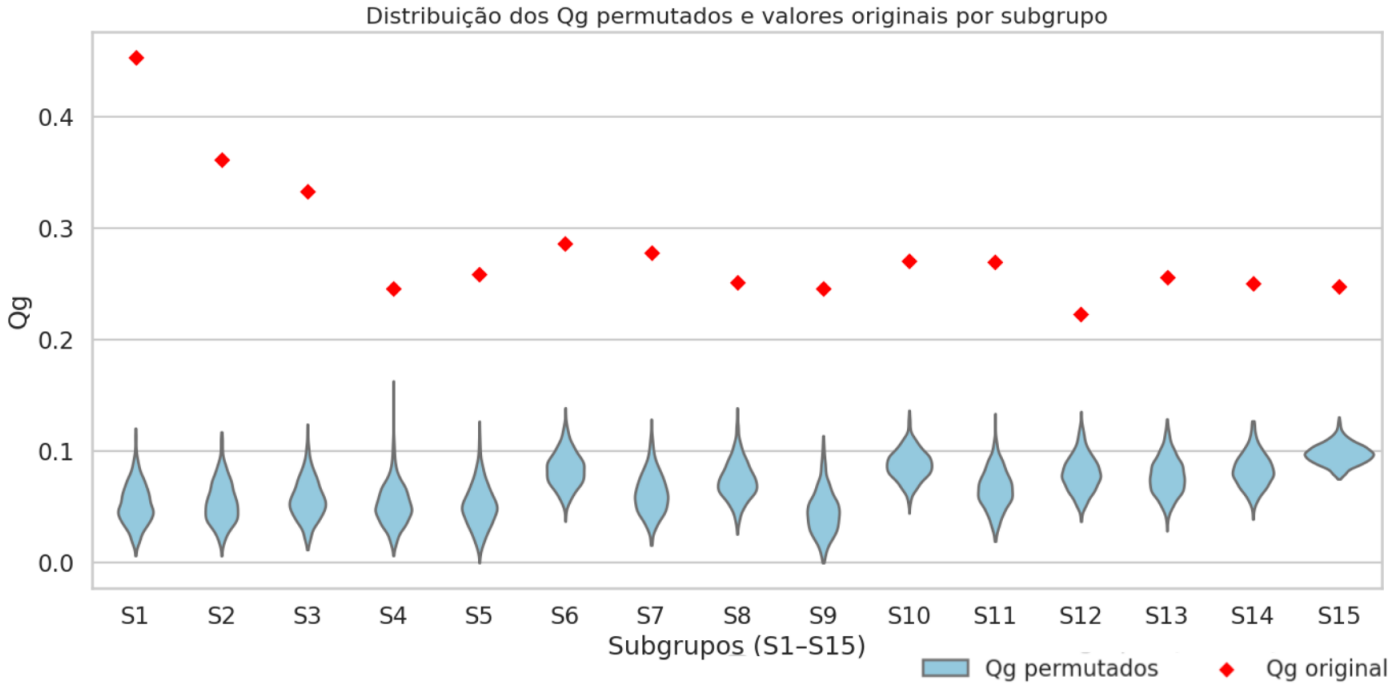

Figura 6. Resultados das distribuições da métrica Qg para os testes de permutação.

a 12,4% de incidência de IRA de forma isolada. Quando combinada com processos infecciosos respiratórios, o risco se elevou consideravelmente. A pneumonia é conhecida por aumentar o risco de IRA através de mecanismos como resposta inflamatória sistêmica e uso de antibióticos nefrotóxicos [37]. A presença da doença genética fibrose cística não é um fator de risco direto da IRA mas adiciona complexidade ao quadro, uma vez que esses pacientes sofrem de perda crônica de sal no organismo levando a um estado frequente de desidratação e baixo volume sanguíneo. Isso coloca os rins sob estresse constante para reter água e sal, podendo, a longo prazo, favorecer uma lesão do tecido renal [38].

No subgrupo S3 (IRA+ = 44%), o padrão é marcado por pacientes jovens (entre 18 e 25 anos) com fibrose cística, diagnósticos de septicemia e diabetes mellitus. O padrão é particularmente interessante pela faixa etária jovem, que normalmente seria considerada protetora conforme apresentado na seção III.A. A presença de fibrose cística e septicemia sugerem episódios de infecções graves que estão frequentemente associadas a lesão renal [39]. Esses marcadores também estão combinados no padrão encontrado para o subgrupo S14 (IRA+ = 44%), reforçando a relação com a alta incidência da IRA. Conforme descrito na revisão da literatura, o diabetes mellitus presente em S3 e outros subgrupos gerados já é estabelecido como um fator de risco que contribui com alterações metabólicas e microvasculares que tornam os rins mais vulneráveis. A identificação desse subgrupo demonstra que, mesmo em pacientes jovens, a combinação de múltiplos fatores de risco pode elevar a vulnerabilidade para IRA de forma considerável.

A relação de diabetes e IRA volta a aparecer no subgrupo S8 (IRA+ = 41%) atrelada a pacientes do sexo feminino com insuficiência cardíaca congestiva. Esse padrão evidencia a síndrome cardiorrenal, um conjunto de condições onde a disfunção cardíaca e renal interagem de forma bidirecional [40]. A especificidade para o sexo feminino nesse subgrupo sugere possíveis diferenças na apresentação clínica ou na resposta fisiológicas à combinação desses fatores, aspecto que merece investigação adicional.

E. Subgrupos de risco moderado-alto ($25\% \leq$ *IRA+* $\leq 39\%$*)*

Sete subgrupos apresentaram taxas classificadas de risco moderado-alto, representando perfis que também requerem alguma forma de monitoramento mais intensivo, uma vez que a taxa de incidência de IRA fica, em média, três vezes maior que a taxa global.

Um padrão de destaque presente nessa categoria é a combinação de etnia afro-americana com insuficiência cardíaca congestiva e hipertensão essencial (27% de incidência). Esse é o subgrupo com maior cobertura populacional entre os identificados para essa categoria, sugerindo um perfil de risco prevalente. Fatores sociais e estruturais (incluindo acesso desigual a cuidados, atraso no diagnóstico e diferenças na qualidade do atendimento) são conhecidos como amplificadores de risco para a população afro-americana nos Estados Unidos e podem levar à detecção tardia e menor prevenção da progressão para IRA nos hospitais [41]. A hipertensão, embora apresente incidência de IRA próxima à média populacional quando analisada isoladamente (10,8%), quando combinada com insuficiência cardíaca, cria um contexto de vulnerabilidade maior. Esse resultado ressalta

a importância de considerar fatores demográficos e socioeconômicos na estratificação de risco.

Os subgrupos envolvendo comorbidades como anemias e obesidade demonstram mais uma vez a importância das interações multifatoriais. A anemia está associada a pior prognóstico em pacientes hospitalizados, reduzindo a capacidade de transporte de oxigênio e sobrecarga renal [42]. A obesidade, por sua vez, está relacionada a inflamação crônica de baixo grau, resistência insulínica e alterações hemodinâmicas que podem comprometer a função renal. Essas comorbidades são especialmente fatores de impulsionamento de riscos quando associados a processos infecciosos agudos como pneumonias (subgrupo S6) ou insuficiência cardíaca (subgrupos S9 e S10) [43].

Outro marcador interessante encontrado nos padrões desses subgrupos é o histórico de outros tipos de infecções, incluindo verminoses (subgrupo S7). Apesar de não haver estudos diretos de relações com a incidência de IRA, é possível associar aos efeitos sistêmicos secundários das verminoses como a anemia, desnutrição, inflamação crônica, desidratação ou sepse secundária que já são estabelecidos como impulsionadores do desfecho [44].

F. Subgrupos de risco moderado ($10\% \leq$ *IRA+* $\leq 24\%$*)*

Três subgrupos foram classificados como risco moderado, representando o dobro da taxa global populacional. Esses subgrupos foram os que apresentaram a maior cobertura populacional média quando comparados aos subgrupos de outras categorias.

O subgrupo S12 é definido pela presença de insuficiência cardíaca congestiva combinada com doença pulmonar crônica (IRA+ = 21%) e é o mais abrangente em termos de cobertura populacional, incluindo 1.284 pacientes. Esse padrão representa uma condição clínica não incomum em hospitalizações, particularmente em população idosa. Doenças pulmonares crônicas, como doença pulmonar obstrutiva crônica (DPOC), frequentemente coexiste com insuficiência cardíaca, configurando um fenômeno conhecido como "comorbidade cardiopulmonar" [45]. A falta prolongada de oxigênio no sangue, causada por essa combinação de problemas nos pulmões e coração, pode fazer os vasos dos rins se contraírem e reduzir o fluxo de sangue para esse órgão. Mesmo que o risco de lesão renal seja menor nesse grupo, a grande quantidade de pessoas afetadas torna o problema importante para ações de monitoramento.

Outro padrão de destaque nessa categoria é a presença de pacientes com histórico de choques hospitalares (subgrupo S11). O choque hospitalar é classificados conforme sua causa principal: hipovolêmico (perda de sangue ou líquidos), cardiogênico (falha na função do coração), distributivo (alteração na distribuição do fluxo sanguíneo, como no choque séptico ou anafilático) e obstrutivo (bloqueio físico do fluxo sanguíneo, como em embolia pulmonar ou tamponamento cardíaco). Sua ocorrência pode apresentar sequelas decorrentes da redução prolongada da circulação de sangue nos tecidos durante o evento, especialmente nos rins [46], resultando em perda parcial e permanente da função renal. Dessa forma, esse histórico pode indicar indivíduos mais suscetíveis ao desenvolvimento de insuficiência renal aguda, mesmo combinado com fatores de risco mais fracos.

G. Interação de Fatores de Risco para IRA

Conforme evidenciado na análise exploratória, algumas variáveis se associam a incidências de IRA próximas à taxa global quando analisadas isoladamente. Por exemplo, a hipertensão, presente em 35.794 internações, mostrou associação a 10,8% de incidência de IRA de forma isolada, valor apenas ligeiramente superior à taxa populacional (9,4%). No entanto, quando a hipertensão é combinada com insuficiência cardíaca congestiva (como o padrão do subgrupo S4), a taxa de IRA eleva-se para 27%, sugerindo que o efeito combinado desses fatores supera significativamente o risco individual.

Esse fenômeno é observado de forma consistente em múltiplos subgrupos identificados. As doenças pulmonares, o diabetes mellitus e a fibrose cística, por exemplo, embora não apresentem valores extremos de associação de risco quando considerados isoladamente, marcam presença em grande parte dos subgrupos identificados. Essa recorrência sugere que essas variáveis clínicas atuam como fatores amplificadores de risco comuns.

Similarmente, a insuficiência cardíaca congestiva aparece em 7 dos 15 subgrupos, evidenciando seu papel central na fisiopatologia da IRA hospitalar. A identificação repetida desse fator em múltiplas combinações reforça a importância do monitoramento em pacientes com disfunção cardíaca para o desfecho da IRA, principalmente quando há comorbidades adicionais.

Outro ponto observado é a presença de septicemia em quatro subgrupos, incluindo dois dos cinco subgrupos de alto risco. A sepse é reconhecida como uma das principais causas de IRA em pacientes críticos, com taxas de incidência que podem alcançar 50% em pacientes sépticos [47]. Dessa forma, é possível presumir que a identificação de combinações envolvendo septicemia indicam um risco elevado imediato de desenvolvimento da IRA.

H. Limitações e Considerações Metodológicas

É importante contextualizar os resultados dentro das limitações metodológicas do estudo. A exclusão das 28 variáveis relacionadas a medicamentos, motivada por incertezas no processo de marcação temporal dos dados, pode ter impedido a identificação de padrões envolvendo interações medicinais específicas que são reconhecidamente nefrotóxicas.

Adicionalmente, a análise foi baseada em dados retrospectivos de um único centro, o que pode limitar a generalização dos achados para outras populações com características demográficas e práticas clínicas diferentes. Estudos em múltiplos centros seriam importantes para validar os subgrupos identificados e verificar se os mesmos padrões emergem em diferentes contextos assistenciais.

Por fim, embora os resultados indiquem padrões de acordo com o que se encontra na literatura, seria importante que essas interpretações fossem validadas por profissionais da área médica, a fim de confirmar sua relevância clínica e aprimorar a compreensão dos achados nesse cenário.

VI. CONCLUSÃO

Este estudo apresentou uma abordagem sistemática de identificação e caracterização de grupos de risco para insuficiência renal aguda em pacientes hospitalizados, utilizando técnicas de descoberta de subgrupos aplicadas a dados clínicos retrospectivos. A pesquisa foi estruturada como uma análise complementar ao trabalho preditivo desenvolvido por He et al. (2019) [1], oferecendo uma perspectiva descritiva para os perfis de pacientes na distribuição de ocorrências dos quadros de IRA.

A análise exploratória dos dados das 76.957 internações hospitalares, das quais 9,4% apresentaram episódios de IRA, revelou padrões importantes nas associações entre características clínicas e o desenvolvimento dessa condição. Os resultados preliminares identificaram indícios de fatores de risco que vão além das análises univariadas, destacando a importância de considerar combinações específicas de variáveis para uma compreensão mais abrangente do risco individual.

A aplicação do algoritmo SSDP+ resultou na identificação de 15 subgrupos com significância estatística validada. Os padrões descobertos revelaram uma estratificação clara de risco quando comparado a incidência global, com taxas de IRA variando de 21% a 64%, permitindo classificar os pacientes em três categorias principais: alto risco ($\geq$40%), risco moderado-alto (25-39%) e risco moderado (20-24%).

Entre os principais achados, destacam-se: (1) a identificação do perfil de mais alto risco, superior a 6 vezes a taxa global, caracterizado pela combinação de quimioterapia, câncer metastático e distúrbios hematológicos, com 64% de incidência de IRA; (2) a presença recorrente da insuficiência cardíaca congestiva em 7 dos 15 subgrupos, evidenciando seu papel como fator amplificador de risco; (3) a demonstração de efeitos sinérgicos entre fatores que individualmente apresentam risco moderado, como a hipertensão que, quando combinada com insuficiência cardíaca congestiva, superam suas taxas de risco individuais; e (4) a identificação de padrões multivariados envolvendo septicemia, diabetes mellitus, fibrose cística e doenças pulmonares como amplificadores de risco quando presentes em combinação com outras comorbidades.

As limitações do estudo incluem o uso de dados retrospectivos de uma única instituição, o que pode limitar a generalização dos resultados para outras populações, e a exclusão de variáveis dos medicamentos ministrados durante as internações devido a inconsistências na marcação temporal da base de dados original.

É esperado que o trabalho realizado possa contribuir para o avanço da precisão no contexto da prevenção da IRA, oferecendo uma ferramenta complementar aos modelos preditivos existentes. A identificação de subgrupos específicos pode facilitar o desenvolvimento de estratégias preventivas personalizadas, permitindo intervenções mais direcionadas e eficazes para populações de alto risco.

REFERÊNCIAS

[1] J. He, Y. Hu, X. Zhang, L. Wu, L. R. Waitman, and M. Liu, "Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records," *JAMIA Open*, vol. 2, no. 1, pp. 115-124, 2019.

[2] D. Ponce, C. P. F. Zorzenon, N. Y. Santos, and A. L. Balbi, "Acute kidney injury in intensive care unit patients: a prospective study on incidence, risk factors and mortality," *Revista Brasileira de Terapia Intensiva*, vol. 23, no. 3, pp. 321-326, 2011.

[3] S. Pozzoli, M. Simonini, and P. Manunta, "Predicting acute kidney injury: current status and future challenges," *Journal of Nephrology*, vol. 31, no. 2, pp. 209-223, 2018.

[4] P. Susantitaphong, M. Siribamrungwong, K. Doi, T. Noiri, N. Terrin, and B. L. Jaber, "Performance of urinary liver-type fatty acid–binding protein in acute kidney injury: a meta-analysis," *American Journal of Kidney Diseases*, vol. 61, no. 3, pp. 430-439, 2013.

[5] F. Turgut, A. S. Awad, and E. M. Abdel-Rahman, "Acute kidney injury: medical causes and pathogenesis," *Journal of Clinical Medicine*, vol. 12, no. 1, pp. 375, 2023.

[6] M. C. Sampaio, C. A. G. Máximo, C. M. Montenegro, V. C. C. Mota, O. P. Fernandes, R. M. Bianco, and A. P. M. Amorim, "Comparação de critérios diagnósticos de insuficiência renal aguda em cirurgia cardíaca," *Arquivos Brasileiros de Cardiologia*, vol. 101, no. 1, pp. 18-25, 2013.

[7] A. Khwaja, "KDIGO clinical practice guidelines for acute kidney injury," *Nephron Clinical Practice*, vol. 120, no. 4, pp. c179-c184, 2012.

[8] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35-49, 2015.

[9] P. Kralj-Novak, N. Lavrač, and G. I. Webb, "Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining," *Journal of Machine Learning Research*, vol. 10, pp. 377-403, 2009.

[10] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and Information Systems*, vol. 29, no. 3, pp. 495-525, 2011.

[11] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach, "Decision support through subgroup discovery: three case studies and the lessons learned," *Machine Learning*, vol. 57, no. 1-2, pp. 115-143, 2004.

[12] C. R. Benichel and S. Meneguin, "Risk factors for acute renal injury in intensive clinical patients," *Acta Paulista de Enfermagem*, vol. 33, pp. eAPE20190218, 2020.

[13] C. Loutradis, L. Pickup, J. P. Law, I. Dasgupta, C. J. Ferro, C. E. Hutchinson, S. Cockwell, and P. Cockwell, "Acute kidney injury is more common in men than women...", *Biology of Sex Differences*, vol. 12, no. 1, pp. 1-9, 2021.

[14] C. L. Arenas, A. C. P. Forero, D. C. V. Ángel, A. F. O. Lasso, D. F. P. Ibarra, and J. A. S. Restrepo, "Risk factors for acute kidney injury in COVID-19 patients," *Brazilian Journal of Nephrology*, vol. 45, no. 1, pp. 29-41, 2023.

[15] E. Campello, A. Zanetto, C. M. Radu, C. Bulato, C. Gavasso, M. Franceschet, A. Ferrarese, U. Cillo, P. Simioni, and S. Piano, "Acute kidney injury is associated with increased microvesicles...", *Digestive and Liver Disease*, vol. 53, no. 3, pp. 323-329, 2021.

[16] J. A. Kellum, J. W. O. van Till, and G. Mulligan, "Targeting acute kidney injury in COVID-19," *Nephrology Dialysis Transplantation*, vol. 35, no. 10, pp. 1652-1662, 2020.

[17] R. A. S. Santos, H. S. Ribeiro, E. A. Vieira, and A. P. Ferreira, "Epidemiological profile of acute kidney injury in critically ill patients...", *Brazilian Journal of Nephrology*, vol. 43, no. 2, pp. 184-194, 2021.

[18] D. Gómez-Bravo, A. García, G. Vigueras, B. Ríos, and J. L. Fernández-Alemán, "A new algorithm for Subgroup Set Discovery based on Information Gain," *arXiv:2307.15089*, 2023.

[19] A. Belfodil, A. Belfodil, A. Bendimerad, P. Lamarre, C. Robardet, M. Kaytoue, and M. Plantevit, "FSSD – a fast and efficient algorithm for subgroup set discovery," in *DSAA*, 2019.

[20] H. M. Proença, P. Grünwald, T. Bäck, and M. van Leeuwen, "Robust subgroup discovery," *Data Mining and Knowledge Discovery*, vol. 36, no. 5, pp. 1885–1970, 2022.

[21] I. Vagliano et al., "Automated identification of patient subgroups…", *Computers in Biology and Medicine*, vol. 163, 107146, 2023.

[22] J. Yang et al., "Tree-based subgroup discovery using electronic health records…", *Biostatistics*, vol. 25, no. 2, pp. 323-335, 2024.

[23] G. Bosc, J. F. Boulicaut, C. Raïssi, and M. Kaytoue, "Anytime discovery of a diverse set of patterns…", *DMKD*, vol. 32, no. 3, pp. 604-650, 2018.

[24] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data," *Statistics and Computing*, vol. 9, no. 2, pp. 123-143, 1999.

[25] T. Pontes, R. Vimieiro, and T. B. Ludermir, "SSDP: A Simple Evolutionary Approach…", in *BRACIS*, pp. 1-6, 2016.

[26] B. Kavšek and N. Lavrač, "APRIORI-SD…", *Applied Artificial Intelligence*, vol. 20, no. 7, pp. 543-583, 2006.

[27] S. Helal, "Subgroup discovery algorithms: a survey…", *Journal of Computer Science and Technology*, vol. 31, no. 3, pp. 561-576, 2016.

[28] G. I. Webb, "Discovering significant patterns," *Machine Learning*, vol. 68, no. 1, pp. 1-33, 2007.

[29] M. Atzmueller and F. Puppe, "SD-Map – A fast algorithm for exhaustive subgroup discovery," in *PKDD*, pp. 6-17, 2006.

[30] T. Pontes, R. Vimieiro, and T. Ludermir, "SSDP…", *BRACIS*, 2016.

[31] T. Lucas, R. Vimieiro, T. Ludermir, "SSDP+: a Diverse and More Informative Subgroup Discovery Approach for High Dimensional Data," *Anais do XXXVII Simpósio Brasileiro de Banco de Dados*, 2022.

[32] W. Duivesteijn, A. Knobbe, "Exploiting False Discoveries - Statistical Validation of Patterns and Quality Measures in Subgroup Discovery," *Proceedings of the 2011 IEEE International Conference on Data Mining*, pp. 151–160, 2011.

[33] S. Erfurt, R. Lehmann, I. Matyukhin, B. Marahrens, S. Patschan, and D. Patschan, "Stratification of Acute Kidney Injury Risk, Disease Severity, and Outcomes by Electrolyte Disturbances," *Journal of Clinical Medical Research*, vol. 15, no. 2, pp. 59–67, 2023.

[34] M. L. C. Santos, B. B. Brito, F. A. F. Silva, A. C. S. Botelho, and F. F. Melo, "Nephrotoxicity in cancer treatment: An overview," *World Journal of Clinical Oncology*, vol. 11, no. 4, pp. 190–204, 2020.

[35] Perazella, M. A., "Onco-nephrology: renal toxicities of chemotherapeutic agents," *Clinical Journal of the American Society of Nephrology*, vol. 7, no. 10, pp. 1713–1721, 2012.

[36] L.-N. Gao, D. Yan, X.-H. Liu, D. Chen, H. Guo, and J. Liu, "Association of Systemic Inflammatory Biomarkers (NLR, MLR, PLR, SII, SIRI) with Preeclampsia-Related Kidney Injury: A Retrospective Observational Study," *Journal of Inflammation Research*, 2025.

[37] Murugan, R., Kellum, J. A., "Acute kidney injury: what's the prognosis?," *Nature Reviews Nephrology*, vol. 7, no. 4, pp. 209–217, 2011.

[38] Southern, K. W., "Acute renal failure in people with cystic fibrosis," *Thorax*, vol. 62, no. 6, pp. 472–473, 2007.

[39] Zarbock, A., Weiss, R., Albert, F., Rutledge, K., Kellum, J. A., Bellomo, R., Grigoryev, E., Candela-Toha, A. M., Demir, Z. A., Legros, V., Rosenberger, P., Galán Menéndez, P., Garcia Alvarez, M., Peng, K., Léger, M., Khalel, W., Orhan-Sungur, M., Meersch, M., … (e col.) "Epidemiology of surgery-associated acute kidney injury (EPIS-AKI): a prospective international observational multi-center clinical study," *Intensive Care Medicine*, vol. 49, no. 12, pp. 1441–1455, 2023.

[40] D. Banerjee, M. A. Ali, A. Y.-M. Wang, and V. Jha, "Acute kidney injury in acute heart failure – when to worry and when not to worry?," *Nephrology Dialysis Transplantation*, vol. 40, no. 1, pp. 10–18, 2025.

[41] H. Choi, J. Y. Lee, Y. Sul, S. Kim, J. B. Ye, J. S. Lee, S. Yoon, J. Seok, J. Han, J. H. Choi, and H. R. Kim, "Comparing machine learning and logistic regression for acute kidney injury prediction in trauma patients: A retrospective observational study at a single tertiary medical center," *Medicine (Baltimore)*, vol. 102, no. 33, e34847, 2023.

[42] Y. Lombardi, C. Ridel, and M. Touzot, "Anaemia and acute kidney injury: the tip of the iceberg?," *Clinical Kidney Journal*, vol. 14, no. 2, pp. 471–473, 2020.

[43] Hsu, C. Y., McCulloch, C. E., Iribarren, C., Darbinian, J., Go, A. S., "Body mass index and risk for end-stage renal disease," *Annals of Internal Medicine*, vol. 144, no. 1, pp. 21–28, 2006.

[44] Herberg, J., Pahari, A., Walters, S., Levin, M., "Infectious Diseases and the Kidney," *Pediatric Nephrology*, pp. 1235–1273, 2009.

[45] Morgan, A. D., Zakeri, R., Quint, J. K., "Defining the relationship between COPD and CVD: what are the implications for clinical practice?," *Therapeutic Advances in Respiratory Disease*, vol. 12, 2018.

[46] Prowle, J. R., Kirwan, C. J., Bellomo, R., "Fluid management for the prevention and attenuation of acute kidney injury," *Nature Reviews Nephrology*, vol. 10, no. 1, pp. 37–47, 2014.

[47] Hoste, E. A. J., Bagshaw, S. M., Bellomo, R., Cely, C. M., Colman, R., Cruz, D. N., … Kellum, J. A., "Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study," *Intensive Care Medicine*, vol. 41, no. 8, pp. 1411–1423, 2015.