Capítulo 3

Modelagem, Mineração e Análise de Jornadas / Trajetórias de Pacientes

Caroline de Oliveira Costa Souza Rosa, Márcia Ito, Alex Borges Vieira, Antônio Tadeu Azevedo Gomes

Abstract

The sequence of visits and procedures performed by the patient in the health system, also known as the patient's pathway or trajectory, can reveal important information about the clinical treatment adopted and the health service provided. The rise of electronic health data availability made it possible to assess the pathways of a large number of patients. Nevertheless, some challenges also arose concerning how to synthesize these pathways and how to mine them from the data, fostering a new field of research. The objective of this mini-course is to present the area of patient pathway mining, highlighting representation models, mining techniques, methods of analysis and examples of case studies.

Resumo

A sequência de atendimentos e procedimentos realizados pelo paciente no sistema de saúde, a qual é denominada jornada ou trajetória do paciente, pode revelar informações importantes sobre o tratamento clínico adotado e o serviço de saúde prestado. Com o aumento no registro de dados de saúde de forma eletrônica, tornou-se possível o estudo das jornadas de um grande volume de pacientes. No entanto, surgiram também desafios sobre como sintetizar essas jornadas e de que modo extraí-las dos dados, fomentando um novo campo de pesquisa. O objetivo deste minicurso é apresentar a área de mineração de jornadas de pacientes, destacando modelos de representação, técnicas de mineração, métodos de análise e exemplos de estudos de caso.

3.1. Introdução

O aumento de casos de doenças crônicas e de doenças agudas que deixam sequelas como a COVID-19 adicionaram outras complicações na já complexa assistência que demandam estes pacientes. Geralmente esses pacientes possuem múltiplas condições crônicas ou

de longo termo e que podem incluir tanto doenças físicas quanto mentais; por exemplo, um paciente com diabetes que contraiu a COVID-19 depois de curado pode permanecer com uma sequela motora. Eles são classificados como pacientes com multimorbidade. Pacientes nessas condições têm a sua vida diária muito afetada, mais especificamente na sua qualidade de vida ou no acometimento de doenças mentais como ansiedade e depressão. Além disso, a multimorbidade aumenta a fragmentação do cuidado e afeta o relacionamento familiar. Dessa forma há um aumento no uso do serviço de saúde por pacientes nestas condições se comparado com os que não tem multimorbidade. A gestão desses pacientes no serviço de saúde se torna complexo, pois além da complexidade de lidar com várias doenças há o problema da intensidade de intervenções. Os serviços de saúde precisam se organizar de forma a comportar a intensidade e a integração necessária para pacientes com multimorbidade. Além disso os protocolos clínicos existentes nem sempre consideram a multimorbidade e assim é difícil para o profissional de saúde saber todos os cuidados necessários às múltiplas recomendações para o mesmo paciente. Para o paciente e seus cuidadores a decisão de qual recomendação seguir também se torna um problema. Uma forma de elaborar protocolos clínicos e organizar melhor o serviço de saúde é estudar e entender a trajetória ou jornada de pacientes nestas condições [Poitras et al. 2018, Ito 2020].

Entende-se por **Jornada/Trajetória de Paciente**[1] como uma sequência de eventos de saúde ou intervenções que o paciente realizou em sua passagem pelo sistema de saúde. Em alguns casos esses eventos podem ocorrer fora ou dentro da instituição de saúde, como por exemplo o uso de medicamentos que pode acontecer na casa do paciente. Assim, estes eventos podem envolver tanto atividades de cuidado em saúde como consultas, administração de medicamentos, internações e cirurgias, quanto atividades relacionadas ao serviço de saúde prestado, como as unidades e departamentos visitados e as especialidades consultadas.

Embora a compreensão da jornada percorrida por um único paciente possa trazer informações relevantes para amparar seu processo de tratamento, a avaliação de jornadas se torna especialmente vantajosa quando informações de coortes maiores de pacientes estão disponíveis. Nesse caso, é possível, por exemplo, avaliar o desempenho de diferentes unidades dentro do sistema de saúde [Sato et al. 2020, Villamil et al. 2017]. Pode-se também avaliar se diretrizes de tratamento ou diagnóstico estão sendo seguidas [Rinner et al. 2018, Durojaiye et al. 2018], ou acompanhar a progressão de doenças que antecederam um evento crítico [de Oliveira et al. 2020b], de modo que futuros pacientes que apresentem os mesmos padrões sejam acompanhados com mais ênfase. Outro exemplo de aplicação é a otimização de *layouts* hospitalares [Rismanchian and Lee 2017, Arnolds and Gartner 2018], minimizando distâncias percorridas a partir do entendimento das jornadas mais frequentes. Entretanto, quando se trata de dados na área da saúde, há desafios a serem superados com relação aos dados em si e aos processos neles contidos.

[1]Em tradução livre de *Patient Pathway*. Como não há um termo oficial em português para *Patient Pathway*, os autores decidiram que para efeito deste capítulo usarão os termos trajetória e jornada como sinônimos.

3.1.1. Dados em Saúde

Dentre as características que dificultam a análise de dados de saúde atualmente, destacam-se o volume, o reuso e a qualidade dos dados.

A adoção de meios eletrônicos para registrar dados da área de saúde tem permitido que um **volume** cada vez maior de dados seja coletado [Pastorino et al. 2019]. Esse volume está intimamente ligado ao conceito de *Big Data* [Baro et al. 2015]. Apesar do valioso conhecimento implícito nesses dados [Dahlem et al. 2015], sua externalização depende da disponibilidade de recursos computacionais e do desenvolvimento de técnicas e algoritmos para manejá-los.

O volume e a variedade crescentes de dados de saúde torna possível seu **reuso** para fins outros que não os para que foram originalmente coletados [Baro et al. 2015]. Embora o reuso de dados facilite a realização de pesquisas, evitando os obstáculos e os custos da coleta de dados primários, a qualidade e a adequação destes dados para os novos fins devem ser analisadas com cautela [Weiskopf and Weng 2013, Kurniati et al. 2019].

Por fim, questões ligadas à **qualidade** dos dados, especialmente secundários, envolvem a completude de informações sobre um paciente, a validade do que foi registrado, a concordância e plausibilidade entre os elementos dos registros e outros elementos ou outras fontes, e a contemporaneidade dos dados [Weiskopf and Weng 2013]. Para contornar tais desafios é necessário, por exemplo, utilizar atributos diferentes do de interesse para estimar informações faltantes [Zaballa et al. 2020] ou comparar diferentes atributos no intuito de testar a validade dos dados [Bettencourt-Silva et al. 2015].

3.1.2. Processos em Saúde

Os processos em saúde centrados no paciente apresentam variabilidade alta. Isso se dá devido a múltiplos fatores:

- **Especificidade dos pacientes.** Cada paciente reage de uma forma aos procedimentos e medicações a que é submetido. Embora o tratamento de primeira linha se mostre eficiente para uma parcela dos pacientes, nem todos obterão resultados satisfatórios, enquanto outros poderão apresentar reações adversas graves. Há ainda o risco de haver incompatibilidade entre o protocolo clínico de uma doença e as diretrizes de tratamento das comorbidades do paciente.
- **Decisões médicas.** A jornada de um paciente é fortemente influenciada pelas decisões tomadas pela equipe clínica envolvida no seu acompanhamento. Diferentes graus de conhecimento médico e até mesmo a indisponibilidade de recursos podem afetar as decisões tomadas e, consequentemente, a trajetória do paciente. Além disso, em determinadas situações, pode não haver um protocolo clínico estabelecido, de modo que a experiência e a percepção do especialista prevalecerão.
- **Decisões dos pacientes.** Em diversas situações, mas especialmente na atenção primária, o paciente tem grande influência sobre sua jornada. Por exemplo, ele pode ou não tomar a medicação prescrita ou retornar para uma consulta de revisão.
- **Dinamismo da área da saúde.** A área da saúde é marcada por mudanças e inovações que incluem, por exemplo, o desenvolvimento de novas medicações e tra-

tamentos, o estabelecimento de diretrizes mais confiáveis, e a descoberta de novas doenças [Rebuge and Ferreira 2012].

- **Complexidade de dados.** Como discutido anteriormente (Seção 3.1.1), os dados da saúde são complexos, e essa complexidade propaga-se para a análise das jornadas, que podem envolver a identificação de sequências temporais nesses dados.
- **Multidisciplinaridade.** Cada vez mais os pacientes têm sido acompanhados por times formados por diferentes especialidades e, possivelmente, distribuídos entre múltiplas unidades de saúde [Rebuge and Ferreira 2012]. Além da variabilidade advinda das decisões de múltiplos especialistas, percebe-se que a constituição dos times de cuidado varia para cada paciente [Conca et al. 2018].

3.1.3. Mineração de Jornadas de Pacientes

Por conta da complexidade e da quantidade das variáveis existentes na condução da saúde de um paciente, o tipo de informação utilizada direciona o tipo de análise e que aqui será denominada como **Perspectiva** [Manktelow et al. 2022]. As perspectivas possíveis incluem o diagnóstico, especialidades, departamentos, atividades assistenciais, entre outras; é possível também fazer a combinação entre essas perspectivas gerando uma terceira análise [de Oliveira et al. 2020b, Conca et al. 2018, Arnolds and Gartner 2018, Rebuge and Ferreira 2012, Najjar et al. 2018, Zhang et al. 2015a]. O estudo da jornada do paciente não é uma novidade, mas a viabilidade de análise vem crescendo à medida que o acesso a dados de saúde – sejam eles administrativos ou clínicos – vem sendo possível [Rotondi et al. 1997].

Mineração da Trajetória do Paciente[2] é a denominação de métodos e técnicas automáticas que existem para minerar os dados de saúde com a finalidade de encontrar jornadas de grupos de pacientes. Para a montagem das jornadas são selecionados eventos e comportamentos relevantes. Procura-se representar essas jornadas de uma forma que especialistas do domínio consigam analisá-las. São inúmeras as possibilidades de análise [Rinner et al. 2018, Chen et al. 2018, Conca et al. 2018, Dahlin and Raharjo 2019, Arnolds and Gartner 2018]; dentre elas pode-se destacar:

- Aderência a protocolos clínicos;
- Busca por anomalias nos protocolos clínicos padrões, ao comparar o protocolo clínico padrão com o encontrado no mundo real;
- Busca por novos protocolos clínicos resultados de diagnósticos ou tratamentos de sucesso no mundo real;
- Análise de custo de procedimentos, tratamentos e diagnósticos;
- Busca por anomalias e gargalos nos processos do serviço de saúde;
- Proposição de melhorias ou readequação do *layout* e dos processos no serviço de saúde.

[2]Em tradução livre de *Patient Pathway Mining*.

3.1.4. Objetivos e Estrutura deste Capítulo

Devido à complexidade e quantidade de variáveis existentes nos dados clínicos e administrativos de um paciente [Baro et al. 2015, Pastorino et al. 2019], definir as variáveis que serão utilizadas para projetar a jornada do paciente de forma compreensível e de fácil análise por seres humanos é um dos grandes desafios. Como resultado, há vários modelos para representar jornadas de pacientes, algoritmos para minerá-las e métodos para analisá-las [Najjar et al. 2018, de Oliveira et al. 2020b, Rebuge and Ferreira 2012].

Assim, o objetivo deste capítulo é apresentar e discutir sobre os principais métodos e técnicas existentes para a descoberta automática e análise das jornadas de pacientes. O restante do capítulo está organizado da seguinte forma: na Seção 3.2 são descritas as principais técnicas de representação e mineração da modelagem de jornadas dos pacientes. Além disso, as principais técnicas de mitigação da complexidade de dados para lidar com o volume e a variabilidade dos dados em saúde são discutidas naquela seção. Na Seção 3.3 algumas técnicas para a análise da jornada do paciente são descritas, abordando o agrupamento de jornadas, a análise de conformidade de jornadas e a análise da temporalidade de eventos em jornadas. Os estudos de casos sob as perspectivas clínicas e de serviço de saúde são discutidos na Seção 3.4, a fim de ilustrar como a modelagem e a análise de jornadas de pacientes têm sido aplicadas nas diferentes áreas da saúde. Finalizando, na Seção 3.5 são feitas as considerações finais, incluindo um resumo do conteúdo apresentado e uma discussão sobre os desafios e as possibilidades de pesquisa na área.

3.2. Técnicas de Modelagem

As jornadas de pacientes, ao serem mineradas, podem ser representadas na forma de uma lista de padrões relevantes ou através de um único modelo que sumarize seus resultados. Além da forma de representação, também é importante decidir o nível de detalhe que os resultados da mineração das jornadas terão e os tipos de informações fornecidas (p.ex. a ordem dos eventos, o tempo entre as atividades e os recursos envolvidos). Essas escolhas dependem do modelo matemático utilizado para representar as jornadas e do algoritmo escolhido para descobri-las. Nesta seção, discutimos exemplos de modelos (Subseção 3.2.1) e algoritmos (Subseção 3.2.2) encontrados na literatura. Além disso, também apresentamos algumas estratégias adotadas para lidar com os desafios impostos pela variabilidade dos dados de saúde (Subseção 3.2.3).

3.2.1. Representação de Jornadas de Pacientes

Na literatura, há diversas formas para se modelar jornadas de pacientes. A escolha do modelo de representação depende do tipo de informação contida na trajetória, do objetivo do estudo e das relações buscadas entre as atividades.

Uma maneira direta de representar jornadas é usando sequências de eventos [Antonelli et al. 2012, Aspland et al. 2021, Defossez et al. 2014, Egho et al. 2014, Fei and Meskens 2013, Huang et al. 2012, Hur et al. 2020, Le Meur et al. 2015, Perer et al. 2015]. Nesse caso, os resultados são apresentados em forma de lista, como exemplificado na Figura 3.1. Essa abordagem viabiliza a análise de características das diferentes jornadas (variantes) e dos pacientes que as seguem, além de se adequar bem a estudos que pretendem utilizar as jornadas de pacientes como entrada para algoritmos de otimiza-

ção [Arnolds and Gartner 2018] ou ferramentas preditivas [Hur et al. 2020, Kempa-Liehr et al. 2020]. Além disso, rótulos de atividades podem se repetir em diferentes momentos de uma mesma variante, de forma que a noção exata do caminho percorrido é preservada.

Outra forma de se modelar uma jornada de paciente é utilizando uma sequência de intervalos de tempo, cada um dos quais possuindo um conjunto de atividades frequentes. Os autores que utilizaram esta estratégia estavam principalmente interessados em descobrir jornadas clínicas realmente adotadas [Cho et al. 2020, Huang et al. 2013, Wang et al. 2017]. A tarefa pode envolver não apenas a determinação do conjunto de ações, mas também os intervalos de tempo entre as ações realizadas.

Em contrapartida, o ajuste extremo aos dados faz com que essa forma de representação de jornadas como sequências de eventos não tenha uma **generalização** alta. Generalização diz respeito à capacidade do modelo de gerar padrões (jornadas) não registrados nos dados usados para aprendê-lo. Essa característica pode ser desejável, pois os dados usados na construção do modelo correspondem a uma amostra da realidade, e portanto, é razoável que padrões não observados na amostra ocorram na realidade.

Jornada	Frequência
A B C	10
B C D	10
B C	8
A D	5
A B A	2
A B C D	2
C D	1
D	1
B D	1
A A B	1

Figura 3.1. Exemplo de representação de jornadas de pacientes por meio de sequências. Neste caso, A, B, C e D representam as atividades seguidas pelos pacientes (p.ex. procedimentos realizados). A primeira coluna do quadro apresenta as variantes de jornadas, enquanto a segunda coluna apresenta a frequência absoluta de cada variante.
Fonte: Os Autores (2022)

Apesar de simples e direta, essa abordagem também pode ser usada de forma auxiliar em trabalhos que utilizam modelos mais sofisticados de jornadas de pacientes, como álgebras de processo, pois permite a inspeção das variantes mais frequentes [Kempa-Liehr et al. 2020, Gonzalez-Garcia et al. 2020, Kim et al. 2013, Durojaiye et al. 2018, Andrews et al. 2020, Kurniati et al. 2019]. As sequências podem representar tanto as jornadas exatamente como foram registradas nos dados, quanto subsequências frequentes obtidas, por exemplo, com algoritmos de mineração de sequências. Independente de se utilizar as jornadas originais ou os padrões mais frequentes, haverá uma lista de sequências a serem analisadas. Em geral, quando a variabilidade das jornadas for baixa, isto é, o número de jornadas diferentes é baixo quando comparado com o número de pacientes, é viável que os especialistas analisem todas as jornadas listadas para extração de conhecimento. No entanto, à medida que a variabilidade das jornadas aumenta, a tendência é que a maio-

ria delas seja seguida por poucos ou mesmo por um único paciente. Como resultado, a listagem de sequências será longa e a inspeção dos resultados poderá ser impraticável.

Alternativamente, um modelo que resuma múltiplas trajetórias individuais, ou seja, um modelo capaz de replicar as trajetórias observadas ou a maioria delas, pode representar as jornadas dos pacientes de forma mais concisa. Exemplos de tais modelos genéricos são grafos, extensões probabilísticas de autômatos de estados finitos e álgebras de processo.

Grafos. Um **grafo** é uma estrutura matemática que representa um conjunto de objetos (nós) e as relações entre eles (arestas). No domínio das jornadas de pacientes, os nós podem denotar os encontros, enquanto as arestas podem indicar quais são as direções seguidas. Alguns algoritmos de mineração de processos, como o *Fuzzy Miner* [Günther and van der Aalst 2007], geram um grafo como seu modelo de processo. Os algoritmos dessa linha, frequentemente, incluem nós virtuais para indicar o início e o fim das jornadas [Lin et al. 2001, Sato et al. 2020, Najjar et al. 2018]. A Figura 3.2 apresenta um exemplo desse tipo de modelo.

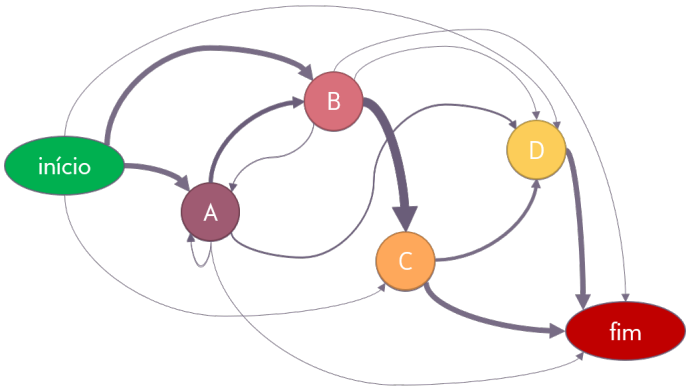

Figura 3.2. Exemplo de representação de jornadas de pacientes por meio de um grafo. Neste caso, os vértices A, B, C e D representam as atividades seguidas pelos pacientes (p.ex. procedimentos realizados), e as arestas direcionadas indicam que houve uma sucessão direta observada entre o nó de origem e o de destino. A espessura das arestas indica a sua frequência de ocorrência.
Fonte: Os Autores (2022)

Os grafos podem representar todo o conjunto de dados [Basole et al. 2015] ou, quando há muitas variantes de jornadas, podem apresentar apenas comportamentos frequentes ou relevantes [Lin et al. 2001, Zhang et al. 2015a, Prodel et al. 2018]. Além disso, as arestas podem conter informações sobre o tempo ou o fluxo de pacientes entre os eventos [Gonzalez-Garcia et al. 2020, Arias et al. 2020].

Tradicionalmente, os nós de um grafo têm rótulos exclusivos. Assim, quando uma atividade, representada por um nó, aparece repetidamente nas jornadas, o grafo terá ciclos que podem dificultar a interpretação do modelo, principalmente a noção de quais nós anteriores foram visitados antes do atual. Como alternativa, pode-se utilizar grafos acíclicos (também conhecidos como grafos árvore) [Dagliati et al. 2017]. A Figura 3.3 apresenta um exemplo de grafo acíclico elaborado com os mesmos dados do grafo cíclico da Figura 3.2. Nos grafos acíclicos, a repetição de rótulos no modelo faz com que as jornadas sejam apresentadas exatamente como observadas ou mineradas. Em contrapartida,

enquanto grafos cíclicos têm capacidade de generalização, grafos acíclicos reproduzem apenas jornadas observadas nos dados.

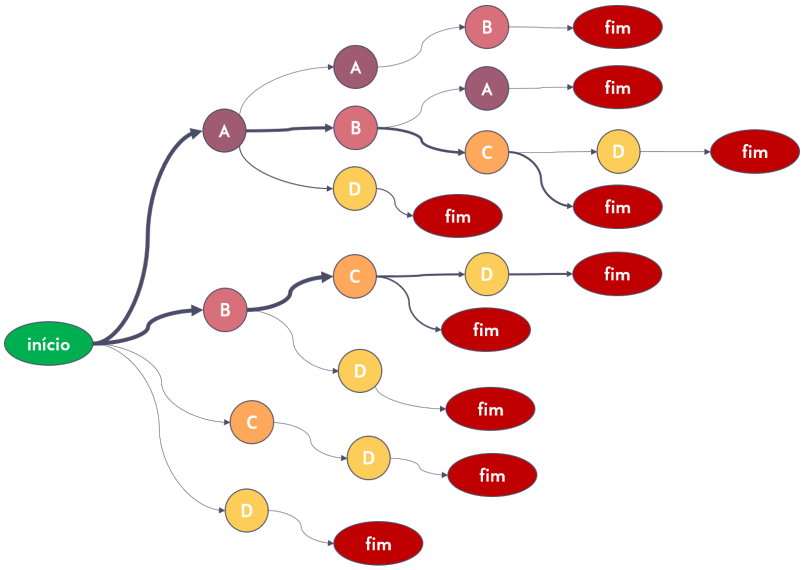

Figura 3.3. Exemplo de representação de jornadas de pacientes por meio de um grafo acíclico. Neste caso, os vértices A, B, C e D representam as atividades seguidas pelos pacientes (e.g. procedimentos realizados), e as arestas direcionadas indicam que houve uma sucessão direta observada entre o nó de origem e o de destino. A espessura das arestas indica a sua frequência de ocorrência. Fonte: Os Autores (2022)

Como alternativa, o modelo proposto por Duma and Aringhieri 2020, denominado *Hybrid Activity Tree*, começa adicionando todas as jornadas observadas em um grafo árvore direcionado (isto é, acíclico). Posteriormente o algoritmo converte ramificações infrequentes em um grafo cíclico direcionado sem rótulos repetidos. O objetivo dos autores desse trabalho foi o de balancear a preservação de memória dos grafos acíclicos com a capacidade de generalização dos grafos cíclicos.

Um grafo com múltiplas camadas em que cada camada corresponde a uma posição em uma jornada foi proposto por de Oliveira et al. 2020a. Como resultado, uma atividade pode aparecer várias vezes no modelo, desde que as ocorrências estejam em camadas diferentes. Além disso, entre um mesmo par de nós pode haver múltiplas arestas indicando que duas jornadas com a mesma sequência de atividades podem não ter o mesmo significado se o tempo entre as atividades for significativamente diferente.

Extensões probabilísticas de autômatos de estados finitos. Um autômato de estados finitos é um modelo matemático que descreve uma máquina abstrata com um número finito de estados. Essa máquina está em apenas um estado por vez. Uma transição indica uma mudança de estado. **Cadeias de Markov** e **Autômatos de Estados Finitos Probabilísticos** são exemplos de extensões probabilísticas desses autômatos, onde as transições de estado são associadas a probabilidades. A principal diferença entre esses dois modelos é que, nos autômatos probabilísticos, cada transição está associada a uma condição que precisa ser satisfeita para que a transição ocorra com certa probabilidade; nas cadeias de Markov, as transições ocorrem imediatamente com certa probabilidade. Em geral, tais

modelos são similares a grafos (cíclicos) cujas arestas contêm a probabilidade de transição entre os vértices que as definem. (Uma outra interpretação, matricial, para cadeias de Markov é apresentada na Subseção 3.3.2.) Alguns autores modelaram jornadas de pacientes como Modelos de Markov [Baker et al. 2017, Garg et al. 2009, Villamil et al. 2017], Modelos de Markov Ocultos [Zhang and Padman 2015] ou Autômatos de Estados Finitos Probabilísticos [Arnolds and Gartner 2018].

Álgebra de processo. Quando se espera que as jornadas dos pacientes sejam o resultado de um processo subjacente bem estruturado, modelos apoiados em **álgebra de processo** (p.ex. com operações como junções e divisões AND/OR/XOR) são adequados. Nessa linha, são exemplos de modelos aplicados na literatura associada a jornadas de pacientes o BPMN [Kurniati et al. 2019, Lu et al. 2016, Stefanini et al. 2020, Tamburis and Esposito 2020], *HeuristicNet* [Caron et al. 2014, Durojaiye et al. 2018], *Inductive Visual Model* [Andrews et al. 2020, Kempa-Liehr et al. 2020, Marazza et al. 2020], as redes de Petri [Durojaiye et al. 2018, Marazza et al. 2020, Rebuge and Ferreira 2012, Stefanini et al. 2020, Tamburis and Esposito 2020], *Fork/Join Network* [Senderovich et al. 2016] e modelos declarativos de processo [Mertens et al. 2018, Mertens et al. 2018]. Um exemplo desse tipo de modelo, usando a notação BPMN, é apresentado na Figura 3.4. Nele é indicado que as jornadas se iniciam com a atividade **B** sozinha ou em paralelo à atividade **A**. Em seguida, a atividade **C** pode ou não ocorrer e, por fim, a atividade **D** pode ou não ocorrer; se ambas **C** e **D** ocorrerem, então **C** tem que ocorrer antes de **D**.

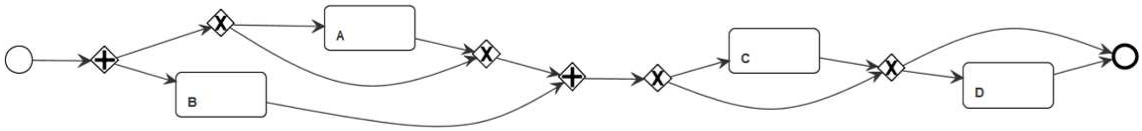

Figura 3.4. Exemplo de representação de jornadas de pacientes por meio de um modelo BPMN. Neste caso, A, B, C e D representam as atividades seguidas pelos pacientes (p.ex. procedimentos realizados).
Fonte: Os Autores (2022)

Os modelos dessa categoria trazem um ganho na capacidade de representação com relação aos anteriores, pois conseguem expressar mais tipos de dependência entre as atividades, como pontos em que atividades ocorrem de forma paralela ou em que deve-se escolher entre uma ou outra atividade. De forma semelhante ao caso dos grafos, as arestas podem conter informações sobre o número de pacientes ou o tempo médio, por exemplo. No entanto, deve-se notar que, quando a variabilidade de jornadas é alta, esses modelos podem se tornar complexos de serem lidos. Além disso, geralmente essas notações não admitem repetição de rótulos de atividades, o que dificulta a interpretação de modelos com muitos ciclos, como no caso dos grafos cíclicos.

3.2.2. Mineração de Jornadas

Nesta subseção, apresentamos métodos que podem ser usados para mineração (descoberta) de jornadas de pacientes. Nós introduzimos diferentes técnicas identificadas na literatura relacionada, agrupando-as em seis categorias, descritas a seguir.

3.2.2.1. Algoritmos baseados em atividades/sucessões frequentes

Os trabalhos classificados neste primeiro grupo utilizam diferentes métodos, mas têm em comum o fato de que eventos e transições só são selecionados para o modelo final se forem suficientemente frequentes.

Os algoritmos de mineração de sequências, também chamados de algoritmos de mineração de padrões frequentes ou sequenciais, visam encontrar subsequências frequentes nos dados de acordo com um limite de suporte mínimo escolhido. Eles são baseados principalmente na estratégia do algoritmo Apriori [Agrawal et al. 1994], proposto para descobrir conjuntos frequentes de itens e posteriormente adaptado para padrões sequenciais [Agrawal and Srikant 1995]. Com esta abordagem, em vez de se verificar a prevalência de todas as sequências possíveis, são consideradas apenas aquelas cujas subsequências são frequentes. Por exemplo, dadas três atividades – A, B e C – o padrão ABC só pode ser frequente se ambas as subsequências AB e BC forem frequentes. Há várias implementações de algoritmos de mineração de sequências na literatura, e alguns deles foram adotados por autores da área de mineração de jornadas de pacientes. Por exemplo, pode-se citar o algoritmo BIDE [Antonelli et al. 2012] e o algoritmo CloFAST [Hur et al. 2020]. Além destes, Perer et al. 2015 propôs uma modificação do algoritmo SPAM.

Outros autores, no entanto, optaram por implementar um novo algoritmo de mineração de sequências para lidar com jornadas de pacientes. Por exemplo, Egho et al. 2014 considerou que cada encontro na trajetória de um paciente pode conter vários registros de diferentes tipos (p.ex. conjunto de procedimentos, diagnósticos, especialidade médica). Ele propôs um novo algoritmo de mineração de sequência para descobrir padrões multidimensionais e, em trabalhos posteriores, ele também propôs uma métrica de similaridade para comparar tais sequências [Egho et al. 2015] (vide Subseção 3.3.2). Se por um lado esse trabalho destacou como as jornadas de pacientes podem envolver múltiplas perspectivas, por outro, Huang et al. 2012 chamaram a atenção para a incorporação de informações de tempo nos padrões minerados. Os autores propuseram um novo algoritmo de mineração de sequências, cujo último passo converte os padrões frequentes em *chronicles*. Em outras palavras, além de minerar subsequências frequentes, os autores também mineram intervalos de tempo chave entre duas atividades que não necessariamente ocorrerem uma imediatamente após a outra no modelo.

Em [Huang et al. 2013] e [Wang et al. 2017], os autores incluíram informações temporais nos modelos de forma diferente. Eles propuseram um método para minerar jornadas de pacientes expressas como uma sequência de intervalos de tempo, cada um com um conjunto de ações frequentes. Eles usaram otimização para definir os melhores intervalos e, em seguida, aplicaram algoritmos baseados nos algoritmos Apriori [Huang et al. 2013, Wang et al. 2017] e FP-growth [Wang et al. 2017] para descobrir padrões frequentes. Estas jornadas de pacientes intervaladas assemelham-se à forma como alguns protocolos clínicos são definidos e, de fato, ambos os estudos tiveram como objetivo descobrir os protocolos clínicos efetivamente adotados na prática.

Uma desvantagem dos algoritmos de mineração de sequências é que as sequências têm poder expressivo limitado, pois só podem representar sucessões diretas de atividades. Alternativamente, Lin et al. 2001 e Yang and Hwang 2006 propuseram uma modificação

dos algoritmos de mineração de padrões sequenciais para descobrir subgrafos frequentes. A utilização de grafos permite a representação de atividades concorrentes no modelo de jornada. Lin et al. 2001 também incluiu uma opção para definir janelas de tempo e minerar os padrões de jornada em cada uma.

O algoritmo proposto por Chen et al. 2018 inclui tanto sequências de intervalos como grafos. O método de descoberta da jornada começa com a definição dos intervalos de tempo de acordo com especialistas. Em seguida, um algoritmo identifica as atividades realizadas em cada intervalo e constrói, para cada paciente, matrizes de transição entre pares de intervalos, formando uma cadeia de Markov. Essa cadeia representa a jornada do paciente e é sobre ela que uma métrica de similaridade com outras jornadas é proposta (vide Subseção 3.3.2). O agrupamento das jornadas similares permitiram aos autores identificar transições frequentes ("núcleo denso") e incluí-las em um grafo para representar o comportamento frequente do grupo.

Essa abordagem de agrupar pacientes e identificar atividades e transições frequentes também tem sido utilizada por outros autores. Najjar et al. 2018 propuseram o uso de duas etapas de agrupamento de pacientes. Na primeira, eles utilizaram modelos ocultos de Markov para agrupar as jornadas, e na segunda, utilizaram um algoritmo hierárquico (vide Subseção 3.3.1) para agrupar os *clusters* da primeira fase. Em seguida, eles construíram um grafo para descrever o comportamento frequente de cada *cluster* final. Por outro lado, Zhang et al. 2015a e Zhang et al. 2015b começaram agrupando pacientes com agrupamentos hierárquicos, usando uma medida de similaridade baseada na subsequência comum mais longa (vide Subseção 3.3.2). Eles construíram um modelo de Markov usando transições de visitas como estados ("super par") para cada *cluster*. Os estados e transições cuja frequência foi superior a um limiar escolhido foram usados para construir um grafo, ou seja, o modelo final da jornada (vide estudo de caso na Subseção 3.4.2.2).

Embora os grafos possam representar atividades concorrentes, tradicionalmente não permitem que os nós tenham rótulos repetidos e, consequentemente, se as jornadas tiverem muitas atividades repetidas, os grafos acabam tendo muitos *loops* e ciclos, tornado-os difíceis de ler. Alternativamente, o método proposto por Dagliati et al. 2017 começa com a seleção das primeiras atividades frequentes e constrói uma árvore adicionando ramos cuja frequência é superior a um limite pré-determinado. Os autores adicionam informações temporais (p.ex. tempo médio) a nós e transições. Eles comparam os valores de biomarcadores selecionados de pacientes que seguiram diferentes jornadas. Em um artigo posterior [Dagliati et al. 2018], os autores propuseram uma extensão do método para identificar atividades concorrentes.

3.2.2.2. Algoritmos baseados na descoberta de dependências entre atividades

O desenvolvimento da mineração de processos, uma área de pesquisa que envolve algoritmos e métodos para descobrir e avaliar processos usando dados de registro (*log*) de eventos, inovou a área de mineração de jornadas. Uma das principais vantagens dos algoritmos voltados à mineração de processos é a capacidade de descobrir e representar mais tipos de relações entre atividades, como dependências de longo prazo, atividades mutuamente exclusivas e atividades paralelas.

O *Heuristics Miner* [Weijters et al. 2006], um dos primeiros algoritmos de mineração de processos, constrói um modelo de processo identificando relações de dependência entre atividades, levando em consideração uma medida de sua frequência. Vários autores têm utilizado este algoritmo para descobrir jornadas de pacientes [Rebuge and Ferreira 2012, Caron et al. 2014] (vide estudo de caso na Subseção 3.4.1.1). Ele foi usado, por exemplo, para avaliar a qualidade do banco de dados MIMIC-III para estudos de mineração de processos [Kurniati et al. 2019]; para comparar as jornadas clínicas adotadas por quatro hospitais [Partington et al. 2015]; para descobrir protocolos clínicos efetivamente seguidos [Yoo et al. 2015]; para comparar as jornadas de pacientes antes e depois da construção de um novo prédio em um hospital [Yoo et al. 2016]; e para mapear as jornadas de departamentos visitados por pacientes com trauma pediátrico [Durojaiye et al. 2018].

Mais recentemente, Leemans et al. 2013 propuseram o algoritmo do *Inductive Miner*. Esse algoritmo decompõe as sequências observadas em subconjuntos, que também são particionados para revelar a estrutura subjacente do processo. O algoritmo organiza os resultados em uma árvore de processos que pode ser convertida em outras notações de modelagem de processos. Este algoritmo foi usado, por exemplo, para descobrir as jornadas de pacientes que sofreram acidentes rodoviários [Andrews et al. 2020] e para descobrir jornadas de pacientes com câncer de pulmão [Stefanini et al. 2020].

Os modelos de jornada de pacientes não apenas proporcionam uma melhor compreensão de um grupo específico de pacientes, mas permitem compará-los com outros grupos. Marazza et al. 2020 propuseram um método para comparar jornadas de duas subpopulações de pacientes com câncer que consiste em obter um modelo de processo para cada grupo, converter os modelos de processo em gráficos direcionados e compará-los com Distância de Edição de Grafos (vide Subseção 3.3.2), ou com um vetor de atributos do grafo. Os autores utilizaram o *Inductive Miner* para obter jornadas representadas como redes de Petri, e também uma extensão do algoritmo – *o Inductive Miner Infrequent* – para minerar jornadas representadas com o *Inductive Visual Model*. Este último algoritmo foi também o que o Tamburis and Esposito 2020 escolheram para descobrir as jornadas de tratamento da catarata. Eles usaram o modelo de processo para criar e executar um modelo de simulação. Da mesma forma, Kempa-Liehr et al. 2020 aplicaram o *Inductive Miner Infrequent* para descobrir as jornadas de pacientes com apendicite (vide estudo de caso na Subseção 3.4.1.2). Eles utilizaram as jornadas mais frequentes como atributos para prever o tempo de permanência no pós-operatório.

A maioria dos autores nessa categoria usou algoritmos de mineração de processos estabelecidos para minerar jornadas de pacientes. No entanto, alguns autores propuseram novos algoritmos com foco nas características dos dados de saúde. Por exemplo, Lu et al. 2016 propuseram um novo algoritmo capaz de descobrir Dependências Cíclicas Decomponíveis, que acontecem, por exemplo, quando duas atividades parecem ser concorrentes, mas o valor de um atributo determina qual delas deve acontecer primeiro. Eles realizaram um estudo de caso considerando um processo organizacional de saúde. Mertens et al. 2018 apresentaram um algoritmo para minerar modelos de processos declarativos. Eles se concentram na descoberta de regras que externalizam o conhecimento tácito, em vez de produzir um modelo de processo estruturado. Eles mineram as regras em duas etapas. A primeira é baseada no algoritmo Apriori, e a segunda utiliza uma abordagem genética. Os autores realizaram um estudo de caso considerando fraturas relacionadas ao braço. Em

uma publicação posterior [Mertens et al. 2020], eles propuseram uma metodologia para aplicar mineração declarativa de processos em saúde com base em seu algoritmo.

3.2.2.3. Algoritmos baseados na simplificação de um modelo geral

A mineração de processos, tradicionalmente, visa descobrir modelos de processo a partir de dados de registros de eventos, assumindo que há um processo subjacente bem estruturado. No entanto, essa suposição nem sempre é válida. As jornadas seguidas pelos pacientes dependem de características pessoais, conhecimentos médicos e disponibilidade de recursos. Portanto, os modelos descobertos representam tantos comportamentos que se tornam difíceis de ler e interpretar. Alguns autores propuseram algoritmos alternativos para evitar esses modelos ilegíveis, também conhecidos como modelos "espaguete" [Günther and van der Aalst 2007]. O algoritmo *Fuzzy Miner* [Günther and van der Aalst 2007] inicialmente traduz atividades e transições entre elas em nós e arestas de um grafo direcionado. Em seguida, as arestas são filtradas e os nós são mantidos, agrupados ou removidos. A seleção de nós e arestas não se baseia exclusivamente em sua frequência, mas outras medidas podem ser usadas, como a importância relativa de uma aresta para um nó e a posição do nó na rede (p.ex. um nó que representa uma bifurcação no modelo pode ter importância na estrutura do processo). Na área de mineração de jornadas de pacientes, além do *Fuzzy Miner* em si [Lismont et al. 2016, Kim et al. 2013, Partington et al. 2015], algoritmos baseados nele também foram adotados [Xu et al. 2016]. O algoritmo de descoberta de processos implementado no software Disco©, por exemplo, é baseado no *Fuzzy Miner*, e vários autores o utilizaram para minerar jornadas de pacientes [Benevento et al. 2019, Dahlin and Raharjo 2019, Erdogan and Tarhan 2018, Rismanchian and Lee 2017, Sato et al. 2020, Rinner et al. 2018, Stefanini et al. 2018, Sawhney et al. 2021]. Da mesma forma, outra ferramenta comercial de mineração de processos, desenvolvida pela empresa Celonis©, também é baseada no*Fuzzy Miner* [Lira et al. 2019] e tem sido aplicada na área Arias et al. 2020. Além disso, o bupaR© [Janssenswillen et al. 2019], um pacote de mineração de processos para R, possui uma função para traçar um grafo (mapa de processo) com a possibilidade de aplicar vários filtros para simplificá-lo. Esse pacote também foi usado por alguns autores para minerar trajetórias de pacientes [Prokofyeva and Zaytsev 2020, Gonzalez-Garcia et al. 2020].

Além das abordagens baseadas no algoritmo *Fuzzy Miner*, Duma and Aringhieri 2020 propuseram um novo modelo de jornada (*Hybrid Activity Tree*) e o algoritmo de descoberta correspondente. Ele se inicia construindo um grafo acíclico (grafo árvore) cujos nós representam as atividades nas jornadas. Como nenhum filtro é aplicado inicialmente, todo comportamento observado é incluído no grafo, que não permite nenhum comportamento não observado. Em seguida, o algoritmo converte ramificações infrequentes em grafos direcionados para aumentar a generalização do modelo. Portanto, o modelo inicial é simplificado e se torna mais geral. Além desse trabalho, Arnolds and Gartner 2018 usou uma estrutura em forma de árvore para representar o comportamento geral (*Probabilistic Prefix Tree Acceptor*). Nesse modelo, os nós representam estados e os ramos representam atividades. O algoritmo adiciona todo comportamento observado ao modelo e então mescla os estados de acordo com a similaridade da distribuição de frequência de seus ramos de saída. Da mesma forma, o algoritmo PALIA [Fernandez and Benedí 2008] usado

por Conca et al. 2018 inicia a construção de um modelo (*Parallel Acceptor Tree*) com as atividades, enquanto identifica seus paralelismos; posteriormente, os nós e ramos podem ser fundidos ou excluídos para simplificar o modelo.

3.2.2.4. Algoritmos baseados em otimização

O quarto grupo de técnicas de mineração de jornadas de pacientes é semelhante ao grupo de algoritmos baseados na simplificação de um modelo inicial. De fato, algoritmos baseados em otimização também escolhem os componentes mais importantes para o modelo final. No entanto, ao invés de simplificar um modelo inicial, um problema de otimização é definido para obter o modelo de jornadas de modo a otimizar a função objetivo definida.

Em 2018, Prodel et al. 2018 propuseram um novo algoritmo de mineração de jornadas baseado na maximização da métrica de *replayability* do modelo descoberto. Essa métrica mede a capacidade do modelo de reproduzir o comportamento observado, e ela pode ser definida de diferentes maneiras. Particularmente, os autores usaram oito métricas diferentes. A saída do problema de otimização é um grafo direcionado e o algoritmo de busca tabu utilizado para resolvê-lo decide quais atividades serão representadas/agrupadas em um nó e quais arestas serão incluídas no modelo. Enquanto a função objetivo visa maximizar a adequação do modelo, um limiar é estabelecido para limitar o número total de nós e arestas para evitar um modelo excessivamente complexo. Posteriormente, o método foi utilizado em outros estudos de caso [de Oliveira et al. 2020b].

Em [de Oliveira et al. 2020a] o método baseado na maximização de *replayability* do modelo descoberto foi estendido para minerar grafos com camadas e arestas caracterizadas por intervalos de tempo. A extensão fornece um modelo mais representativo e os grafos que retornam permitem que eventos recorrentes apareçam em diferentes camadas do modelo. Além disso, eles suportam múltiplas transições entre o mesmo par de nós para representar um conjunto de intervalos de tempo distintos. A métrica de *replayability* considera a porcentagem de eventos que puderam ser reproduzidos, a proporção de arestas faltantes ou intervalos de tempo inadequados, e se houve eventos "pulados". Existem restrições quanto ao número de nós, arestas e camadas. Em outra publicação, De Oliveira et al. 2020 propuseram uma etapa de pré-processamento para agrupar eventos, e utilizaram o método para minerar as jornadas.

Cho et al. 2020 também desenvolveram um algoritmo de mineração de jornadas baseado em otimização. Seu objetivo é decidir quais ordens, ou seja, sucessões diretas entre duas atividades, devem ser incluídas no modelo para maximizar sua adequação e precisão, de acordo com duas taxas de conformidade propostas. O algoritmo obtém o modelo ótimo verificando iterativamente se a inclusão de uma ordem aumenta ou não a taxa média de conformidade.

3.2.2.5. Extração de jornadas de pacientes como elas são

Na literatura, há também métodos que não usam nem algoritmos de mineração de jornada, nem métodos de agrupamento específicos. Isso ocorre principalmente quando o estudo de

caso ou a construção da jornada resultam em poucas variantes, ou quando há uma etapa de pré-processamento que simplifica significativamente as jornadas.

Exemplos de jornadas simples são as consideradas por Zhang et al. 2018. As jornadas consideradas por eles referem-se a trocas de medicamentos para tratamento de condições específicas e podem ser representadas por gráficos. Le Meur et al. 2015 também apresentam jornadas simples, modeladas exatamente por três eventos, cada um resumindo a assistência pré-natal do trimestre correspondente de gestantes.

Quando se trata de pré-processamento para simplificar jornadas, Huang et al. 2018 propôs um método para agrupar eventos de jornada em tópicos de tratamento. Embora tenham fornecido uma representação visual de apenas algumas jornadas selecionadas no estudo, eles mencionaram a intenção de processar as trajetórias gerais do tópico em trabalhos futuros. Outro exemplo é o trabalho de Defossez et al. 2014, no qual converteram grupos de eventos de trajetórias de câncer de mama em estados de acordo com um conjunto de regras.

Bettencourt-Silva et al. 2015 propuseram várias etapas de filtragem e agrupamentos de eventos. Eles desenvolveram um sistema para especialistas avaliarem os resultados em jornadas simples. Esse sistema inclui, por exemplo, um gráfico mostrando a jornada de atividades clínicas versus a evolução de um biomarcador medido para um paciente. Basole et al. 2015 também se concentraram em fornecer uma representação visual das jornadas dos pacientes. Eles usaram um grafo e organizaram estrategicamente os nós para facilitar sua interpretação.

3.2.2.6. Outros

Os cinco trabalhos a seguir não se encaixam bem em nenhuma das classificações propostas anteriormente para técnicas de mineração de jornada de paciente. Também, por si só, eles não constituem um outro grupo, com características específicas.

Dois estudos compartilham o fato de começarem com um modelo de jornada predefinido. Um deles atualiza o modelo e o enriquece enquanto revisa a cadeia de eventos (registro de eventos) [Baker et al. 2017]. O segundo calcula os parâmetros do modelo com os dados observados [Garg et al. 2009].

Mais precisamente, Baker et al. 2017 começam com um modelo de Markov proposto por especialistas para retratar as jornadas dos pacientes. Eles obtêm as probabilidades de transição com os dados das jornadas observadas e atualizaram os estados do modelo caso um evento imprevisto apareça nos dados. Além disso, identificam quantos pacientes tem em cada cada estado, como parte de sua subjornada mais grave. Garg et al. 2009 também começam com um modelo de Markov e estimam as probabilidades de transição com os dados disponíveis. Identificam, entre outras, as jornadas mais prováveis, mais caras e mais curtas.

Villamil et al. 2017 escolhem uma abordagem diferente. Eles usam um algoritmo de agrupamento originalmente como seu método de mineração de jornada, definindo o número de clusters para um. O método começa gerando um número predefinido de modelos de Markov. Em cada iteração, ele atribui cada jornada de paciente ao modelo com

maior probabilidade de reproduzi-lo e atualiza os parâmetros dos modelos. Ele repete essas etapas até que os parâmetros dos modelos convirjam. Ao final, obtém-se um único modelo de Markov para representar cada grupo.

Com foco nas jornadas de diagnóstico, Khan et al. 2018 utilizou redes complexas para representar a evolução das comorbidades em pacientes diabéticos. Eles prepararam uma rede de doenças para pacientes não diabéticos e outra para pacientes diabéticos. Compararam a prevalência de doenças (nós) em cada rede e identificaram aquelas que foram mais prevalentes para pacientes diabéticos.

Por fim, para avaliar a conformidade das jornadas de pacientes com câncer, Senderovich et al. 2016 utilizam uma abordagem baseada em álgebra intervalar e probabilidades Markovianas para descobrir a estrutura do modelo e caracterizar sua dinâmica para construir uma rede *Fork-Join* que descreva o processo.

3.2.3. Mitigação da Complexidade de Dados

Em geral, dados provenientes da área da saúde trazem consigo desafios relacionados ao número elevado de variáveis disponíveis, além da variabilidade de comportamentos observados, já que as jornadas dos pacientes tendem a variar de acordo com características pessoais e presença de comorbidades, além do conhecimento tácito da equipe médica e dos recursos disponíveis [Rebuge and Ferreira 2012, Günther and van der Aalst 2007]. Ao serem confrontados com tais dificuldades, muitos autores optaram por utilizar métodos auxiliares para mitigar a variabilidade dos dados. Um resumo das técnicas usadas é apresentado nesta seção.

Em primeiro lugar, pode-se focar na simplificação do número de atividades, seja por meio de filtragem ou agrupamento. A filtragem de atividades é um dos métodos relatados com mais frequência pelos autores e, em muitos casos, ela é guiada pela opinião de especialistas. Nesses casos, os autores obtêm, junto a técnicos da área médica, uma listagem das atividades consideradas relevantes (ou irrelevantes) para o estudo de caso. A filtragem também pode ser feita com base na frequência das atividades [Wang et al. 2017, De Oliveira et al. 2020], seja removendo atividades infrequentes ou atividades com frequência muito alta, como exames laboratoriais de rotina. Outras perspectivas podem ser usadas durante o processo de seleção; em [Zaballa et al. 2020], por exemplo, apesar da perspectiva adotada ser a de intervenções (procedimentos), os eventos foram filtrados de acordo com o diagnóstico correspondente ou, na falta dele, com a especialidade responsável pelo atendimento (vide estudo de caso na Subseção 3.4.2.1). É importante destacar que o processo de filtragem pode remover padrões infrequentes, porém relevantes da análise, como influência de comorbidades, novas alternativas de tratamento ou jornadas com indícios de fraude no serviço prestado. Alguns autores optaram propositalmente por não filtrar atividades do seu conjunto de dados. Nestes casos, o agrupamento de eventos e pacientes [Najjar et al. 2018] ou a seleção de atividades por meio de otimização [Prodel et al. 2018, de Oliveira et al. 2020a] foram usadas para lidar com a variabilidade dos dados.

O agrupamento de eventos também foi utilizado com frequência e, como a filtragem, pode ser feito de acordo com a direção de especialistas [Chen et al. 2018, Dagliati et al. 2018], ou selecionando um nível de granularidade maior quando as atividades pos-

suem códigos que respeitam uma hierarquia [Marazza et al. 2020, Villamil et al. 2017], como a Classificação Internacional de Doenças. De forma alternativa, técnicas para agrupar de forma automática os eventos podem ser adotadas, como *autoencoding* [De Oliveira et al. 2020],[3] modelos de tópicos [Chiudinelli et al. 2020, Xu et al. 2017, Huang et al. 2018], clusterização [Antonelli et al. 2012, Najjar et al. 2018], mineração de conjuntos de itens [Perer et al. 2015], e ainda, o uso de uma ontologia para identificar eventos ligados a uma mesma ação [Leonardi et al. 2018]. Mais detalhes sobre algumas dessas técnicas de agrupamento e como podem ser usadas na análise de trajetórias são apresentados na Seção 3.3.1.

Além da simplificação do número de atividades, pode-se filtrar ou agrupar jornadas. No caso da filtragem, o critério de seleção pode ser a frequência [Rismanchian and Lee 2017, Tamburis and Esposito 2020] ou a duração (tempo de permanência) [Wang et al. 2017] das jornadas. Já no que diz respeito ao agrupamento, há relatos do uso de diferentes técnicas, como árvores de decisão [Duma and Aringhieri 2020], *k-medoids* [Aspland et al. 2021, Zaballa et al. 2020], agrupamento hierárquico [Zhang and Padman 2015, Zhang et al. 2015b, Zhang et al. 2015a], e agrupamento baseado em modelos de Markov [Rebuge and Ferreira 2012].

Outra estratégia de simplificação é a edição manual do modelo de jornadas descoberto [Andrews et al. 2020, Stefanini et al. 2020], geralmente conduzida sob direção da equipe médica.

Alguns estudos não mencionam o uso de métodos auxiliares para lidar com a complexidade dos dados, mas, em geral, há alguma característica no estudo de caso conduzido ou nas técnicas de descoberta que podem atuar de forma compensatória. Por exemplo, estudos de caso envolvendo o tratamento de casos agudos [Cho et al. 2020, Kempa-Liehr et al. 2020, Yoo et al. 2015, Gonzalez-Garcia et al. 2020, Sato et al. 2020, Sawhney et al. 2021], de emergência [Basole et al. 2015, Benevento et al. 2019, Partington et al. 2015], ou de alternância de medicamentos [Zhang et al. 2018], tendem a resultar em jornadas mais curtas do que outros casos, como o acompanhamento de pacientes com doenças crônicas. Por outro lado, o uso de métodos de mineração de jornadas envolvendo etapas de filtragem ou agrupamento, como nos algoritmos de mineração de sequência [Huang et al. 2012, Huang et al. 2013, Hur et al. 2020] ou otimização [Prodel et al. 2018, de Oliveira et al. 2020a], lidam com a variabilidade dos dados durante a etapa de descoberta do modelo. Alguns estudos que envolveram departamento/recurso [Arnolds and Gartner 2018, Durojaiye et al. 2018, Senderovich et al. 2016] ou jornadas organizacionais [Arias et al. 2020, Kim et al. 2013, Lu et al. 2016] também não mencionaram o uso de métodos auxiliares.

3.3. Técnicas de Análise

Nesta subseção, apresentamos métodos que podem ser usados para análise de jornadas de pacientes. Nós introduzimos diferentes técnicas identificadas na literatura relacionada, agrupando-as em três categorias, descritas a seguir.

[3]Um *autoencoder* é um tipo de rede neural utilizado no aprendizado não-supervisionado de representações eficientes de dados, objetivando geralmente a redução de dimensionalidade desses dados.

3.3.1. Agrupamento (*clustering*)

As técnicas apresentadas nesta subseção permitem identificar classes (grupos, ou *clusters*) de jornadas de pacientes.

Técnicas de agrupamento objetivam particionar elementos de um conjunto em grupos distintos de modo que observações sobre elementos de cada grupo sejam semelhantes e sobre elementos de grupos diferentes sejam distintas. O conceito de similaridade ou diferença entre observações sobre elementos depende de conhecimento de domínio e da técnica empregada. Tal conhecimento é tipicamente expresso por meio de um ou mais parâmetros da técnica de agrupamento, associados a um espaço de atributos que definem as observações sobre os elementos. Há dois tipos principais de algoritmos de agrupamento: algoritmos *k*-particionados e algoritmos hierárquicos.

Algoritmos *k*-particionados dividem elementos de um conjunto em um número finito (e parametrizável) k de grupos. O parâmetro k é o principal ingrediente desses algoritmos no que se refere ao estabelecimento do conhecimento de domínio. Exemplos de algoritmos desse tipo incluem o *k-means* [MacQueen 1967], em que cada grupo é representado pelo seu **centroide** no espaço de atributos, e o *k-medoids* [Kaufman and Rousseeuw 1990], em que cada grupo é representado pelo seu elemento mais próximo do centroide do grupo. Esses algoritmos iniciam com uma atribuição aleatória de cada elemento do conjunto a um dos k grupos, seguida de um processo iterativo em dois passos: (i) computar o centroide de cada um dos k grupos; (ii) re-atribuir cada elemento do conjunto ao grupo cujo centroide (ou elemento representante do grupo) esteja mais próximo (ou seja, com menor distância Euclidiana no espaço de atributos). Zaballa et al. 2020 utilizaram o *k-medoids* e optaram pela distância de edição para medir a similaridade entre as jornadas (vide estudo de caso na Subseção 3.4.2.1). Já Fei and Meskens 2013 usaram duas etapas de agrupamento baseadas no *k-medoids*. A primeira é baseada em características de jornada, enquanto a segunda considera a sequência de ações.

Algoritmos hierárquicos propõem-se a eliminar a limitação dos algoritmos *k*-particionados de predeterminar o valor de k. Esses algoritmos constroem representações baseadas em árvores (**dendrogramas**) de maneira *bottom-up*, partindo inicialmente de n grupos, um para cada elemento do conjunto. Segue-se então um processo de, no máximo, $n-1$ passos, onde em cada passo computa-se a **dissimilaridade** entre todos os grupos e funde-se os dois grupos menos dissimilares em um único grupo. A medida de dissimilaridade adotada depende do domínio de aplicação e da referência do grupo: exemplos de medidas de dissimilaridade incluem distância Euclidiana e distância baseada em correlação. Essas medidas podem ainda usar como referência (*linkage*, no jargão do algoritmo) o centroide dos grupos, a média das similaridades par-a-par entre elementos de cada grupo, entre outras possibilidades. O número de passos do processo determina a altura do dendrograma, e tem um efeito prático sobre o nível de agrupamento semelhante ao parâmetro k dos algoritmos *k*-particionados. Essa técnica foi aplicada em [Zhang et al. 2015a] para agrupar pacientes com doença renal crônica de uma clínica comunitária na Pensilvânia em grupos cujas trajetórias identificaram progressões típicas da doença e práticas consistentes com as diretrizes médicas vigentes, com consequentes melhorias na condição de saúde dos pacientes (vide estudo de caso na Subseção 3.4.2.2).

Outros tipos de algoritmos de agrupamento tentam mitigar limitações dos algorit-

mos acima com respeito ao conhecimento de domínio e à regularidade da "geometria" dos grupos formados. Um dos algoritmos nessa linha é o de **propagação de afinidade** [Frey and Dueck 2007]. Esse algoritmo parte do princípio de que dois elementos do conjunto podem ser comparados quanto à sua similaridade (vide Subseção 3.3.2, para o caso de comparação entre sequências). Assim como o *k-medoids*, esse algoritmo escolhe representantes para cada grupo a ser formado, com a diferença de que o número de grupos não precisa ser informado de início. Esse algoritmo foi usado em [Chen et al. 2018] para agrupar sequências de 99.393 prescrições médicas (de 138 medicamentos diferentes) relacionadas ao atendimento de 8.287 pacientes com infarto cerebral em 20 hospitais na China. Os resultados da análise deste estudo permitiram identificar os melhores regimes de tratamento para pacientes em diferentes situações de criticalidade.

Outro algoritmo nessa linha é o DBSCAN [Ester et al. 1996], que se baseia na ideia de particionar os elementos do conjunto em células, de modo que células com maior número de elementos (células mais **densas**) têm maior probabilidade de se tornarem centroides de grupos, e células menos densas de se encontrarem nas fronteiras entre grupos. O principal parâmetro desse algoritmo é a máxima distância permitida entre dois elementos de um mesmo conjunto. Em [Antonelli et al. 2012], o DBSCAN foi empregado no agrupamento de exames feitos em instantes de tempo próximos (com no máximo 12 dias de espaçamento) por 134 pacientes diagnosticados com câncer de cólon em um hospital na Itália. Os resultados do agrupamento do DBSCAN, combinados a outras técnicas de análise de trajetórias, permitiram aos autores do estudo identificar que o protocolo de testagem era raramente seguido e trajetórias alternativas eram frequentemente adotadas, incluindo exames diagnósticos diferentes daqueles orientados pelo protocolo. Segundo os autores, esses resultados sugerem que o protocolo de testagem não se aplicava a todos os pacientes e, no caso de sintomas menos comuns e similares a de outras doenças, investigações alternativas foram prescritas.

3.3.2. Conformidade

As técnicas apresentadas nesta subseção permitem comparar jornadas de pacientes entre si ou com modelos teóricos.

Comparações entre jornadas de pacientes são um exemplo do problema clássico em computação de **comparação de sequências**. Um dos primeiros e mais conhecidos trabalhos na área [Levenshtein et al. 1966] propõe o conceito de **distância de edição** entre sequências. A distância de edição corresponde ao número mínimo de operações de edição (p.ex. inserção, remoção, ou substituição de itens) necessárias para transformar uma sequência em outra. Essa técnica é explorada em [Defossez et al. 2014] como parte da análise das trajetórias de 159 pacientes com câncer de mama, com o objetivo de indicar se tais trajetórias atendiam as diretrizes sugeridas de pontualidade quanto às etapas de atendimento. Os resultados demonstraram ser possível desenvolver a partir dessa técnica um sistema de informação que produza indicadores de pontualidade no atendimento.

Outro trabalho clássico na área é apresentado em [Hirschberg 1975]. Esse trabalho propõe uma métrica baseada no cômputo da **subsequência mais longa em comum** (*Longest Common Subsequence* – LCS)[4] entre duas sequências. Outros trabalhos,

[4]Uma subsequência é uma sequência que pode ser derivada de outra por meio da remoção de itens sem

como [Sequeira and Zaki 2002], propuseram métricas baseadas também na subsequência mais longa em comum.

O trabalho apresentado em [Wang and Lin 2007] argumenta que métricas baseadas na subsequência mais longa em comum ignoram informações importantes eventualmente presentes em outras subsequências, e propõe o **número de subsequências em comum** entre as sequências como a base para a definição de métricas melhores. Em [Egho et al. 2015] é proposta uma métrica de similaridade entre duas sequências ordenadas de conjuntos de itens (*itemsets*) baseada no cômputo eficiente do número de subsequências em comum. A métrica é definida como a razão entre o número dessas subsequências em comum entre as duas sequências e o número máximo de subsequências distintas nas duas sequências. Essa métrica é empregada na análise da jornada de 828 pacientes com câncer de pulmão na região leste da França. Os resultados obtidos com essa métrica mostraram-se bastante próximos do conhecimento que especialistas na área médica tinham a respeito das jornadas de pacientes naquela região. Esses resultados também enfatizaram características geográficas importantes nessas jornadas, sugerindo que a variabilidade na organização do tratamento de câncer estaria relacionada a fatores locais.

Sequências ordenadas apresentam como principal limitação o fato de que nem sempre apresentam correspondência direta com a forma como se dão os processos centrados em pacientes, em que eventos podem ocorrer sem uma ordem particular. Nesse sentido, outras técnicas de medição de similaridade têm sido propostas.

Especificamente no caso de sequências rotuladas com estampas de tempo (**séries temporais**), uma técnica bem conhecida de comparação é a de **distorção dinâmica do tempo** (*Dynamic Time Warping* – DTW) proposta em [Berndt and Clifford 1994]. A técnica DTW se baseia no alinhamento dos elementos das duas séries temporais de forma a minimizar a soma das distâncias entre esses elementos. Essa técnica foi empregada em [Forestier et al. 2012] para computar uma métrica de similaridade entre processos cirúrgicos, focando nos diferentes tipos de atividades conduzidas durante a cirurgia, sua sequência e seu espaçamento temporal. Experimentos conduzidos com dados de 24 cirurgias de hérnia de disco ocorridas em um hospital universitário na Alemanha demonstraram que a técnica foi capaz de identificar automaticamente grupos de cirurgiões de acordo com seu nível de expertise (sênior ou júnior).

Quando a similaridade entre sequências precisa ser definida a partir de um número grande ou indeterminado de métricas, técnicas mais complexas são necessárias.

Uma estratégia para computar a similaridade entre sequências baseada em um número de métricas determinável a priori é por meio de modelos probabilísticos. Um desses modelos é o de **alocação de Dirichlet latente** (*Latent Dirichlet Allocation* – LDA), proposto em [Blei et al. 2003]. O modelo LDA parte da ideia de que uma sequência pode ser vista como um coleção de variáveis aleatórias intercambiáveis, representável por um modelo de mistura de distribuições. No modelo LDA, essas variáveis aleatórias são dependentes de uma ou mais variáveis latentes (chamadas no LDA de **tópicos**) com distribuição de Dirichlet. A principal aplicação do modelo LDA é na identificação de tópicos comuns a conjuntos de sequências. Em [Huang et al. 2014, Huang et al. 2015, Huang et al. 2016]

alterar a ordem dos itens no conjunto.

o modelo LDA é empregado na descoberta de comportamentos de tratamento latentes – os tópicos do modelo LDA – em jornadas de pacientes de tal modo que as similaridades entre duas jornadas podem ser medidas com base nos tópicos. Esses trabalhos apresentam três possíveis aplicações da técnica (recuperação de jornadas, agrupamento de jornadas, e detecção de anomalias em jornadas), que são avaliadas em conjuntos de dados de jornadas de pacientes de um hospital chinês.

Quando o número de métricas não é possível de ser determinado a priori, técnicas de **aprendizado de representação** (*Representation Learning* – RL) podem ser empregadas [Bengio et al. 2013]. Técnicas de RL permitem que atributos (**características**, no jargão de aprendizado de máquina) que servem de entrada em processos de classificação ou regressão possam ser aprendidos, em vez de definidos a priori. Por exemplo, em [Zhu et al. 2016], mapas de características são extraídos das jornadas dos pacientes como parte do processo de treinamento de uma **rede neural convolucional** (*Convolutional Neural Network* – CNN). O resultado do treinamento é um classificador baseado na similaridade de jornadas. Esse trabalho mostra a capacidade do preditor obtido de agrupar, com boa precisão, jornadas de 7.135 pacientes em quatro coortes distintos: doença pulmonar obstrutiva crônica, diabetes, problemas cardíacos, e obesidade.

Outra técnica passível de ser aplicada para medir a similaridade de sequências, empregada em [Chen et al. 2018], é baseada na teoria de cadeias de Markov. Essa teoria estabelece a base para a modelagem e análise de processos estocásticos que apresentam como característica o fato de que a distribuição condicional do estado futuro do processo estocástico (dito Markoviano) depende somente do seu estado atual. Em termos matriciais, e para processos discretos, pode-se representar o estado futuro $i+1$ de um processo Markoviano como $S_{i+1} = T_{(i \to i+1)} S_i$, onde as componentes de S denotam as variáveis aleatórias do processo, e $T_{(i \to i+1)}$ denota a matriz de transição do processo do estado i para o estado $i+1$. Em [Chen et al. 2018] sequências de matrizes de transição são obtidas a partir de sequências de conjuntos de prescrições médicas, com cada componente de S representando a prescrição de um certo medicamento em um certo momento do tratamento. Com essas sequências de matrizes de transição, uma métrica de similaridade entre duas sequências de prescrições médicas é então obtida por meio do cômputo de uma média ponderada das distâncias de Manhattan calculadas par-a-par entre cada matriz de transição de cada uma das duas sequências de prescrições.

Todas as técnicas apresentadas até o momento nesta subseção partem do pressuposto que as trajetórias de pacientes são modeladas como sequências de eventos. Quando um modelo de grafos é empregado para essa representação, métricas de similaridade específicas são necessárias. O trabalho de [Sanfeliu and Fu 1983] foi o primeiro a formalizar o problema do **casamento inexato de grafos**, definindo o conceito de **distância de edição de grafos** em analogia ao termo usado para comparação de sequências. Trata-se de um problema *HP-hard*, com algoritmos aproximativos de custo computacional alto (p.ex. [Zeng et al. 2009]). Mais recentemente, técnicas de aprendizado de máquina supervisionado têm sido empregadas para resolver o problema [Li et al. 2019], mas demandam a rotulação dos dados, o que nem sempre está disponível. Em [Marazza et al. 2020] uma técnica de aprendizado não-supervisionado é empregada por meio da definição manual de características dos grafos representantes das trajetórias (p.ex. número de nós, número de arestas, grau médio, etc.), características essas sobre as quais métricas de

similaridade como as vistas anteriormente nesta subseção podem ser computadas. Essa técnica foi comparada à inspeção visual de especialistas sobre casos de câncer presentes na base pública MIMIC[5] e na base de dados de pacientes de um hospital holandês, tendo como resultado uma correlação positiva particularmente alta para a base do hospital, o que os autores consideraram como ilustrativo do potencial da técnica.

3.3.3. Temporalidade

As técnicas apresentadas nesta subseção permitem analisar características temporais de jornadas de pacientes. Em particular, o foco é no problema da análise e estimação do tempo de espera de pacientes em filas de atendimento.

A técnica mais simples de análise de tempo de espera é a de **médias móveis**. O objetivo das médias móveis é identificar tendências de longo prazo. Elas são calculadas pela média de um grupo de observações de uma variável de interesse em um período de tempo específico. Essa média é tomada como representativa desse período em uma linha de tendência. Diz-se que essas médias baseadas em períodos "se movem" porque quando uma nova observação é coletada ao longo do tempo, a observação mais antiga do conjunto sendo calculada é eliminada e a observação mais recente é incluída na média. Em [Dong et al. 2018] esta técnica é citada como o estado-da-prática em hospitais, tanto para a análise como para a estimação de tempo de espera.

Diversas outras técnicas de análise e estimação de tempo de espera são apresentadas na literatura médica; em [Pianykh and Rosenthal 2015] algumas das técnicas mais conhecidas são analisadas e, no caso de técnicas de regressão, um estudo das variáveis preditoras mais relevantes é conduzido. Segundo esse trabalho, uma técnica de análise bastante empregada na área é a de simulação discreta de eventos, particularmente no que se refere à prevenção de superlotação de salas de emergência [Duguay and Chetouane 2007, Connelly and Bair 2004]. Essa técnica permite avaliar o comportamento de um modelo de operação de um sistema como uma sequência de eventos.[6] Cada evento está associado a um instante de tempo particular e demarca uma mudança de estado no sistema; assume-se que mudanças no sistema não ocorrem entre eventos consecutivos, daí a natureza discreta dessa técnica. Trata-se de uma abordagem adequada para lidar com com cenários "*what if*", mas dificilmente ajusta-se bem para o gerenciamento de filas de espera em tempo real.

Para a estimação de tempo de espera, técnicas de regressão são as mais exploradas. Essas técnicas estabelecem um modelo matemático (tipicamente linear, na forma $y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$) que descreve o relacionamento entre variáveis **preditoras** $\{x_i\}_{i=1..p}$ e uma variável **dependente** y (ou **alvo**, **resposta**). Em modelos de regressão linear, o ajuste dos coeficientes $\{\beta_i\}_{i=0..p}$ é feito por meio de um procedimento denominado de **mínimos quadrados**. Ainda que amplamente empregado, o ajuste por mínimos quadrados apresenta limitações, particularmente quando o número de observações não é muito maior que o número de variáveis preditoras, já que se baseia nas médias das variáveis preditoras e da variável dependente (no caso em que o número de variáveis preditoras é maior que o

[5] `https://mimic.physionet.org`

[6] A simulação em si não é o modelo! O modelo a ser avaliado por simulação pode ser expresso, por exemplo, como um sistema de filas multiservidor não estacionário, cuja avaliação analítica é complexa.

número de observações, o ajuste por mínimos quadrados não pode ser usado).

Alguns artigos na literatura médica apresentam a regressão **quantílica** como o estado-da-arte na área, com diretrizes bem claras para seu uso [Ding et al. 2010, Sun et al. 2012]. Esta técnica de regressão apresenta maior robustez contra *outliers* comparativamente à regressão linear, pois estima medianas (ou quaisquer outros quantis de interesse) em vez de médias. Em [Sun et al. 2012] essa técnica foi aplicada a 13.200 visitas feitas a um departamento de emergência médica em Singapura. Utilizou-se como variáveis preditoras nesse trabalho o nível de acuidade do paciente, o tamanho das filas de espera, e as taxas de atendimento. Foram obtidas com esse estudo medianas de erros de predição absolutos variando entre 9.2 e 15.1 minutos, o que foi considerado pelos autores como uma indicação da boa acurácia da técnica.

O trabalho de [Ang et al. 2016] propõe uma técnica chamada Q-Lasso para a predição de tempo de espera que combina a técnica de regressão **Lasso** (*least absolute shrinkage and selection operator*) com estimadores baseados em modelos de fluidos. A regressão Lasso adiciona ao ajuste por mínimos quadrados um termo de penalização $\lambda \sum_{i=1}^{p} |\beta_i|$ que tenta "encolher" os coeficientes $\{\beta_i\}_{i=1..p}$ para zero, potencialmente selecionando as variáveis preditoras efetivamente importantes. A determinação do parâmetro λ tem impacto grande na eficácia da técnica, de modo que estratégias de validação cruzada são recomendadas. Já os estimadores baseados em modelos de fluidos aproximam um modelo estocástico de operação baseado em filas por um modelo determinístico (descrito, por exemplo, como um conjunto de equações diferenciais ordinárias). A técnica Q-Lasso acrescenta dois estimadores baseados em modelos de fluidos às variáveis preditoras candidatas na regressão Lasso. Essa combinação mitiga o problema de subestimação de comprimento das filas presente tanto nos estimadores baseados em modelos de fluidos como nos estimadores baseados em regressão quantílica. Essa combinação foi aplicada a dados históricos de 4 hospitais diferentes nos EUA, tendo obtido maior precisão na estimação do tempo de espera de pacientes na emergência desses hospitais do que as técnicas de média móvel, de modelos de fluidos e de regressão quantílica.

3.4. Estudos de Caso

Embora as jornadas representem os encontros dos pacientes com o sistema de saúde, elas podem ser construídas a partir de diferentes pontos de vista, dependendo do tipo de dado escolhido para a caracterização dos eventos. Por exemplo, pode-se definir uma jornada como a sucessão de procedimentos médicos a que um paciente foi submetido, ou então, como a sequência de departamentos (ou unidades de saúde) visitados. Em ambos os casos, o histórico dos pacientes foi utilizado para obtenção dos dados, porém perspectivas diferentes foram adotadas para a construção das jornadas.

Nesta seção, destacamos sete perspectivas registradas na literatura. De forma geral, elas se relacionam ou com o tratamento médico provido ou com o serviço de saúde prestado, conforme apresentado na Figura 3.5.

Uma das perspectivas mais comuns ocorre no contexto clínico e é denotada aqui como "**Intervenção**". Ela abrange procedimentos médicos, administração de medicamentos e outras atividades relacionadas (p.ex. admissão e alta). É comumente usada para avaliar protocolos clínicos [Cho et al. 2020, Gonzalez-Garcia et al. 2020], identificar e/ou

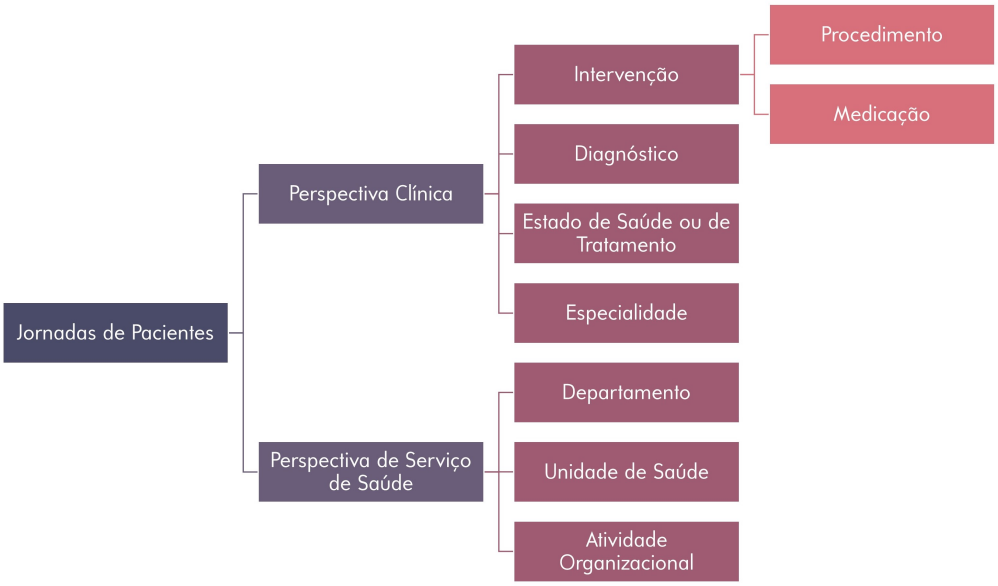

Figura 3.5. Tipos de perspectivas adotadas para construção de jornadas de pacientes.

acompanhar grupos de pacientes [Andrews et al. 2020, Chiudinelli et al. 2020], e comparar protocolos usados em diferentes unidades de saúde [Dahlin and Raharjo 2019, Leonardi et al. 2018].

A perspectiva de "**Diagnóstico**" contempla jornadas que representam as doenças que os pacientes contraíram ou desenvolveram ao longo do tempo [Prodel et al. 2018, de Oliveira et al. 2020b, Khan et al. 2018], amparando o processo de identificação de comorbidades frequentes e de padrões de evolução de doenças associados a maior risco. Exemplos da perspectiva "**Estado de Saúde ou Fase do Tratamento**" envolvem a evolução das fases de tratamento de pacientes idosos [Garg et al. 2009] ou jornadas formadas por indicadores da qualidade do atendimento trimestral de gestantes [Le Meur et al. 2015]. Ainda sob a categoria de perspectiva clínica, as jornadas de pacientes podem conter as "**Especialidades**" que atenderam os pacientes [Conca et al. 2018], permitindo avaliar os impactos de times uni ou multidisciplinares no acompanhamento dos pacientes.

Sob o olhar do serviço em saúde, a perspectiva "**Departamento**" revela não apenas como os pacientes circulam num hospital [Rismanchian and Lee 2017], mas permite interpretar os departamentos como reflexo do tratamento fornecido [Durojaiye et al. 2018]. Nos casos em que múltiplas unidades de saúde estão envolvidas, em vez de departamentos pode-se avaliar a trajetória de "**Unidades**" [Egho et al. 2015]. Há ainda jornadas de "**Atividades Administrativas**" (ou "organizacionais"), como agendamento, cadastro e pagamento [Arias et al. 2020, Erdogan and Tarhan 2018, Kim et al. 2013, Lu et al. 2016, Rebuge and Ferreira 2012, Yoo et al. 2016], que podem auxiliar na estimativa ou avaliação de eficiência do uso dos recursos das unidades de saúde.

A maioria dos autores apresentam jornadas de pacientes considerando uma única perspectiva. No entanto, alguns optam por construir jornadas usando múltiplas perspectivas (vide Tabela 3.1). Por exemplo, dados de internações hospitalares geralmente contêm informações de diagnóstico, intervenção e unidade de saúde, e todas essas perspectivas

podem ser mantidas na análise [Egho et al. 2015]. O objetivo de Egho et al. 2014 foi justamente o de fornecer um método para mineração de sequências de conjuntos de itens multidimensionais, como os dados de internações. Foi observado o nível hierárquico de cada categoria de dados para descobrir padrões frequentes. Outras abordagens incluem o agrupamento de dados formando eventos multi-perspectiva [Najjar et al. 2018], e a rotulagem de eventos, seja por meio da concatenação das informações [Zhang et al. 2015a,Zhang and Padman 2015,Zhang et al. 2015b], ou do uso de *autoencoding* [De Oliveira et al. 2020].

Tabela 3.1. Exemplos de Jornadas de Pacientes construídas sob múltiplas perspectivas

Referência	**Perspectivas**
[Dagliati et al. 2018, de Oliveira et al. 2020a,De Oliveira et al. 2020]	intervenção e diagnóstico
[Egho et al. 2014,Egho et al. 2015]	intervenção, diagnóstico e unidade/ departamento
[Najjar et al. 2018]	intervenção, diagnóstico, unidade/ departamento e especialidade
[Zhang et al. 2015a, Zhang and Padman 2015, Zhang et al. 2015b]	intervenção, diagnóstico e tipo de visita
[Fei and Meskens 2013]	unidade/ departamento especialidade

Outra opção do uso de várias perspectivas em um único estudo é a de modelar as jornadas de pacientes com apenas uma perspectiva, porém utilizar outras de forma auxiliar. Por exemplo, Fei and Meskens 2013 estudaram jornadas de especialidade, mas, ao comparar duas sequências, utilizaram os departamentos onde ocorreram as visitas para medir a distância entre elas. Já Zaballa et al. 2020 usaram informações sobre especialidades médicas para estimar se um evento com diagnóstico ausente provavelmente estaria relacionado ao estudo de caso considerado.

Alguns autores trabalharam com eventos qualitativos e quantitativos. Bettencourt-Silva et al. 2015 construíram jornadas de intervenções e as compararam com séries temporais de um biomarcador. Como espera-se que mudanças na tendência dos biomarcadores estejam associadas a mudanças no tratamento, eles destacaram que as séries de biomarcadores podem ser usadas para avaliar a qualidade dos dados de intervenção. No trabalho de Conca et al. 2018, os autores rotularam os pacientes de acordo com a evolução do resultado de um exame laboratorial e avaliaram se esses rótulos se correlacionavam com os grupos de pacientes categorizados de acordo com jornadas de especialidades. Dagliati et al. 2017 também compararam diferentes grupos de jornadas de pacientes, mas utilizando medições de biomarcadores em pontos estratégicos das jornadas.

Nas próximas seções, apresentaremos exemplos de estudos de caso que realizam a modelagem e análise das trajetórias de pacientes em duas categorias de perspectivas:

- Clínica – trabalhos que contemplam jornadas de procedimentos, diagnósticos e outras atividades relacionadas diretamente ao tratamento do paciente;
- Serviço de Saúde – trabalhos que tratam de jornadas de atividades organizacionais, departamentos hospitalares ou outros eventos relacionados ao serviço de saúde.

3.4.1. Perspectiva Serviço de Saúde

Inicialmente, técnicas para encontrar trajetórias críticas na indústria foram desenvolvidas para identificar e gerenciar a limitação dos processos de produção. Na indústria, a produção é afetada por qualquer variação no processo. Dessa forma, ao definir os processos e o tempo deles, os gerentes podem encontrar as áreas críticas, medir as variações e tentar implementar melhorias. Quando as melhorias são aplicadas, as variações e o tempo para completar a trajetória diminuem, e, por consequência, os custos caem e a qualidade da produção aumenta. Quando o mesmo princípio é aplicado na indústria de saúde, adaptações são necessárias, sendo que a trajetória crítica se traduz na trajetória do paciente no sistema de saúde. Desta forma, se tem que, inicialmente, a trajetória de paciente foi criada sem fins assistenciais, pois foi projetada como uma ferramenta para o apoio à redução de custos e melhorias da qualidade da assistência nos estabelecimentos de saúde. Para este tipo de análise é preciso: (a) definir os processos e os tempos de execução das etapas do processo (nesse ponto é feita a extração e modelagem da trajetória do paciente); (b) analisar o processo procurando por gargalos e anormalidades; (c) priorizar os pontos críticos do processo que precisam de melhorias; (d) medir as variações dos pontos críticos no processo de negócio e com a análise dos dados; e (e) propor e implementar as melhorias no processo [Partington et al. 2015, Ursoniu et al. 2012].

De acordo com o NHS England e o NHS Improvement,[7] reduzir o tempo de espera ao longo da jornada do paciente é desejável, pois demonstra que o sistema de saúde é mais eficiente, as variações são reduzidas e a experiência do paciente é melhorada. Contudo, é preciso ter cuidado ao diminuir o tempo de espera, pois a tendência é agrupar os trabalhos semelhantes e processá-los em "lote", o que pode gerar esperas desnecessárias ao longo da jornada de atendimento do paciente. Por exemplo, um grupo de radiologistas que revisam os laudos de tomografia uma vez por semana, fazem isto, pois acham que agrupando os exames irão gastar um tempo menor em cada laudo, e com isto imaginam que estão sendo eficientes. No entanto, dessa forma, os pacientes esperaram pelo resultado de seus exames de 1 a 9 dias. Após o estudo da jornada do paciente [NHS England and NHS Improvement 2022], tem-se que quando o laudo é feito ao mesmo tempo em que o exame, o tempo de espera do paciente é menor que 2,5 dias. Assim, o ideal é que se faça o laudo no mesmo dia do exame. Isso demonstra que a análise da jornada de pacientes é importante para que o planejamento do processo não seja feito de modo empírico.

O estudo da jornada do paciente permite customizar o tempo de atendimento do paciente de acordo com a sua necessidade, assim como gerenciar e analisar as informações ao longo da trajetória do paciente permitem melhorar a assistência, otimizar o uso dos recursos e reduzir os custos operacionais. É neste contexto que foram selecionados 2 estudos de casos na perspectiva de serviço de saúde para apresentar como se extraiu a jornada do paciente, e como ela foi analisada a fim de entender o processo real das organizações de saúde.

[7]NHS England e NHS Improvement lideram o *National Health Service* (NHS) na Inglaterra com o objetivo de prestar os melhores cuidados aos pacientes.

3.4.1.1. Estudo de caso 1: Análise dos processos do setor de emergência da radiologia [Rebuge and Ferreira 2012]

O Hospital São Sebastião (HSS) é um hospital público com aproximadamente 300 leitos, localizado em Santa Maria da Feira em Portugal. Os objetivos estratégicos do hospital são prestar serviços centrados no paciente, melhorar a eficiência e qualidade dos processos mais importantes e reduzir os custos dos serviços. O hospital desenvolveu o próprio sistema de informação, chamado Medtrix, que apoia a maioria das atividades realizadas. Esse sistema tem uma visão integrada das informações clínicas nos diferentes departamentos em que o paciente foi atendido. Desta forma, o Medtrix é uma boa fonte de dados para extrair a trajetória do paciente no hospital.

Após uma discussão com os *stakeholders* do hospital, decidiu-se o escopo do projeto. Assim, ficou decidido que seriam modeladas trajetórias de pacientes da emergência, envolvendo as atividades de triagem, tratamento, diagnóstico, exames médicos e encaminhamentos. Essa decisão foi baseada em: (a) a qualidade percebida dos serviços é baseada principalmente na opinião dos pacientes, portanto entender a sua trajetória é uma prioridade; (b) o comportamento dos fluxos da emergência e as interações requeridas entre os médicos é um dos mais complexos de entender; (c) há um grande interesse em conhecer o desempenho dos serviços de emergência e toda análise que possa esclarecer é bem-vindo; e (d) desde que o Medtrix foi integrado ao sistema de radiologia da emergência, os processos se tornaram maduros, e portanto há a facilidade em conseguir o registro de eventos.

A extração dos dados não foi algo simples, pois o Medtrix possui mais de 400 tabelas que não estão documentadas. Inicialmente, foi feita uma análise exploratória dos dados, e ao final optou-se por criar um ambiente com os dados relevantes do domínio de estudo. A nova base de dados contém os eventos de emergência de janeiro a julho de 2009. Assim, a tabela de episódios contém os pacientes que passaram pela emergência e as demais tabelas representam os procedimentos realizados com a situação do exame para cada paciente (agendado, cancelado, a ser laudado, etc.). Em cada tabela tem-se a origem e a estampa de tempo do evento o qual permite explorar a organização e o desempenho das trajetórias. O campo "denominação" possui uma descrição do evento, possibilitando explorar o comportamento da trajetória nos diferentes níveis de abstração, dando maior flexibilidade às análises. Esse formato permite criar tipos diferentes de registro de eventos a partir da mesma base de dados, tornando mais fácil extrair, modelar e analisar vários tipos de trajetórias.

Para a extração e análise das trajetórias foi desenvolvido um ambiente baseado no *toolkit* ProM [Verbeek et al. 2010] chamado MPMS (*Medtrix Process Mining Studio*), pois foi necessário automatizar a exportação dos registros de eventos e oferecer um ambiente para a visualização dos resultados que atendesse as necessidades dos gestores do hospital. O componente ***Log Preparation*** é responsável por criar os registros de eventos que são usados para construir a trajetória do paciente. A extração desses dados é feita por meio de consultas SQL e filtros podem ser aplicados caso não se queiram todos os registros. O componente ***Log Inspector*** traz informações estatísticas sobre o registro de evento. O componente ***Sequence Clustering*** é um conjunto de técnicas que agrupam os registros de eventos e fazem uma pré-análise do processo. O objetivo é fornecer mais do

que um simples modelo de trajetória ao sistematizar o processo de análise das várias versões e das anomalias do comportamento.[8] O componente ***Performance Analysis*** fornece o tempo máximo, mínimo e médio entre as tarefas, assim como o rendimento dos processos. O resultado é apresentado em gráficos. Finalmente, o componente ***Social Network Analysis*** consegue descobrir a rede social de acordo com duas métricas: "*handover of work*" e "*working together*" que são as mesmas encontradas no ProM.

Tendo o ambiente e os dados preparados, iniciou-se o estudo analisando o fluxo do departamento de emergência da radiologia. Neste estudo de caso são apresentados os resultados da análise da sequência de agrupamento, da trajetória em si e do desempenho. Os dirigentes do hospital acreditam que os médicos não sigam o processo estabelecido para o atendimento na emergência da radiologia, em que um mesmo médico deve dar o diagnóstico, tratamento, solicitar os exames e encaminhar o paciente. Dessa forma, essa foi a trajetória analisada neste estudo de caso.

Para extrair os registros de eventos da radiologia obtidos, o MPMS selecionou os que continham o exame radiológico e a evolução do exame. Foram obtidos 27.930 processos (trajetórias), com 179.354 eventos e 12 atividades diferentes (os exames permitem 11 situações possíveis). Desses processos foram encontrados 2.296 tipos diferentes de trajetórias. Para encontrar a trajetória usou-se o *Heuristic Miner* do ProM e o resultado convertido para redes de Petri (vide Figura 3.6). O esperado era encontrar uma trajetória que se destacaria como padrão, mas não foi o que aconteceu. A presença de tantas transições (diferentes escolhas) após cada passo sugere que na verdade há muitas variantes deste processo.

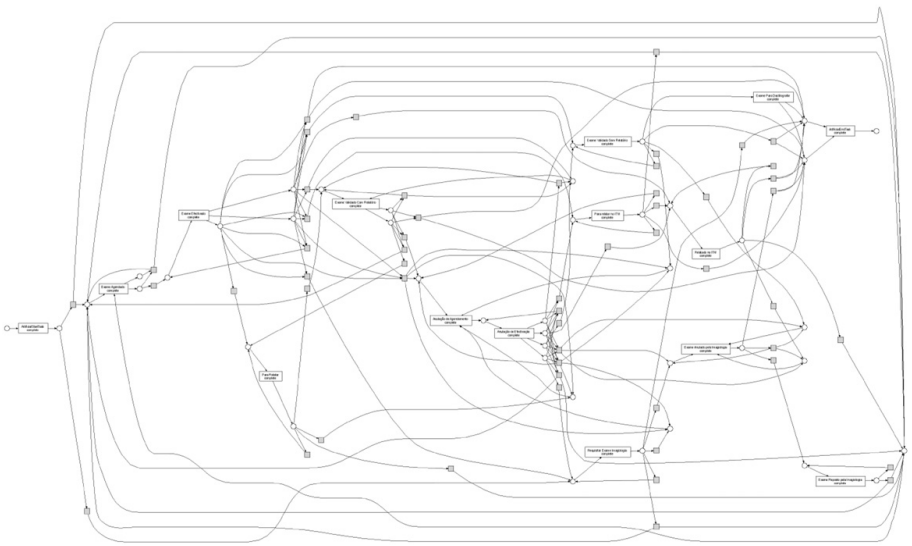

Figura 3.6. Trajetória global da radiologia em redes de Petri, convertido do *Heuristic Miner* do ProM.
(Fonte: [Rebuge and Ferreira 2012])

Em seguida, o algoritmo de agrupamento de sequências foi aplicado para separar o comportamento regular das variantes e infrequentes. Inicialmente, o agrupamento foi

[8]Para entender os detalhes do método recomendamos acessar o artigo de Rebuge e Ferreira [Rebuge and Ferreira 2012].

feito de forma automática utilizando o *Microsoft Sequence Clustering*, mas o resultado continuou alto (22 agrupamentos). Após alguns experimentos chegaram a 8 agrupamentos que pareceram resultados mais intuitivos e que eram passíveis de análise. Neste caso é preciso encontrar um meio termo, pois muitos agrupamentos têm variações tão pequenas que as trajetórias são muito semelhantes entre si, e assim difíceis de analisar. Por outro lado, um pequeno número de agrupamentos agrega variações tão discrepantes que geram modelos muito complexos para interpretar e analisar.

Com os 8 agrupamentos, o MPMS criou um diagrama em forma de grafo, em que cada agrupamento é representado por um vértice, e as arestas apresentam a distância entre eles. Desse modo, foi possível identificar os agrupamentos mais similares. Também foi possível encontrar o agrupamento com a trajetória dominante, pois um atributo dos vértices indicava o número de trajetórias nele contido. A partir do agrupamento mais frequente, construiu-se um modelo de trajetória para representar o comportamento regular. A Figura 3.7 apresenta a trajetória encontrada como regular e apresenta a seguinte sequência: (1) o exame é solicitado; (2) o exame é agendado; (3) o exame é realizado; (4) o exame é validado sem o laudo.

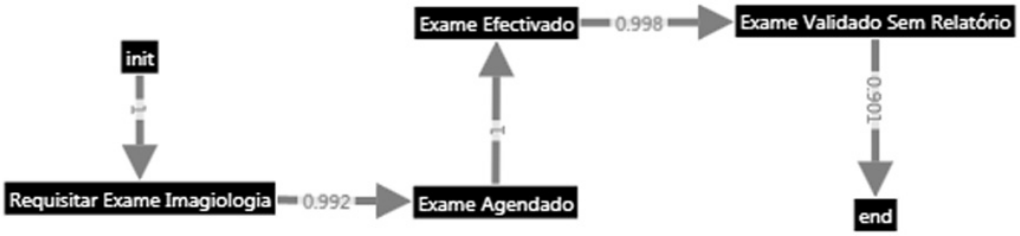

Figura 3.7. Comportamento regulara da trajetória da radiologia. (Fonte: [Rebuge and Ferreira 2012])

A fim de entender as variações, fez-se uma análise utilizando a árvore geradora mínima do diagrama, através da qual é possível encontrar o grau de similaridade entre os agrupamentos. Posteriormente, foi possível comparar as trajetórias resultantes dos agrupamentos para encontrar as diferenças e entender o comportamento divergente, ou investigar a causa da diferença no comportamento. Por exemplo, ao comparar o agrupamento regular com um não regular, foi possível definir que, uma vez o exame solicitado, existia a probabilidade de 0,073 que ele fosse cancelado; e, uma vez cancelado, o processo terminava. Foram também feitas análises em relação ao tempo de espera entre uma atividade a outra, além do estudo de gargalos no processo.

Para responder aos gestores se os médicos seguem a recomendação de atenderem o paciente desde o início até o encaminhamento dele, foi usado o componente ***Social Network Analysis*** do MPMS. Com este componente uma rede social foi criada usando a técnica "*handover of work*". Nela foi possível detectar os médicos que por algum motivo transferiram a assistência de seus pacientes a outros médicos, mas não é possível saber o motivo, pois neste caso o médico pode ter transferido o paciente por não se tratar da especialidade dele, por exemplo, e não por negligência.

É importante destacar que a modelagem da trajetória do paciente é um retrato do que aconteceu, porém sem dados adicionais é difícil interpretar as razões pelas quais houve um desvio da trajetória recomendada ou do padrão encontrado.

3.4.1.2. Estudo de caso 2: Predição de novos eventos em Cirurgias de Apendicite [Kempa-Liehr et al. 2020]

Este estudo de caso faz a análise e mineração da jornada de pacientes que fizeram a retirada do apêndice (apendicectomia) a partir de registros eletrônicos médicos (EMR). Após a modelagem e análise da jornada é feita a aplicação de técnicas probabilísticas de aprendizado de máquina para prever o tempo de recuperação de um paciente após a apendicectomia. Essa cirurgia foi escolhida por ser geralmente simples e com início e término bem definidos.

O ProM [Verbeek et al. 2010] foi a ferramenta escolhida para a extração e análise da trajetória, pois não é necessário conhecimento prévio sobre mineração de processos. Além disso, essa ferramenta parece apresentar facilidade de interpretação clínica e a possibilidade de ser combinada com modelos de aprendizado de máquina.

O método utilizado para a mineração do processo, sua análise e proposta de análises preditivas compreende os seguintes passos:

- Processamento clínico dos dados:
  - identificar e sintetizar os padrões repetitivos de atividades;
  - agrupar os procedimentos pelos departamentos.
- Visualização das trajetórias:
  - importar os dados processados para o ProM;
  - usar as funções do ProM para descobrir e revisar as trajetórias dos pacientes.
- Análise de conformidade:
  - identificar desvios das trajetórias do paciente;
  - ajustar as trajetórias de acordo com os desvios.
- Avaliação do desempenho das trajetórias:
  - identificar indicadores de desempenho críticos (tempo de hospitalização, taxa de readmissões, etc.);
  - avaliar e analisar o desempenho da trajetória baseado em suas variações.
- Análise do desempenho das trajetórias
  - desenvolver modelos de aprendizado de máquina para identificar fatores de risco;
  - identificar possíveis formas para melhorar o desempenho da trajetória.

Para o estudo foram coletados dados de 448 pacientes com apendicite do North Shore Hospital de Auckland, Nova Zelândia, no período de 2015 a 2017. Os dados estavam armazenados no Sistema de Informação de Radiologia (SIR) e no Sistema Administrativo do Hospital. O conjunto de dados compreendia os procedimentos que foram

categorizados pelos departamentos clínicos do hospital em que foram executados (farmácia, enfermaria, centro cirúrgico, entre outros). Além disso, havia a indicação se o procedimento foi feito antes ou após a cirurgia.

Após o processamento dos dados clínicos, as trajetórias foram geradas pelo ProM sem os procedimentos relacionados aos medicamentos, pois o uso de antibióticos pelos pacientes antes e depois da cirurgia é padrão. Foram encontradas 13 variações de jornadas (vide Figura 3.8), sendo que as quatro mais comuns compreendem aproximadamente 88% dos pacientes. Várias análises comparativas entre as jornadas foram feitas em relação ao tempo de internação após a cirurgia. Observou-se que há variações no tempo de internação, e as jornadas em que o paciente ficou mais tempo internado merecem uma investigação para entender o que aconteceu.

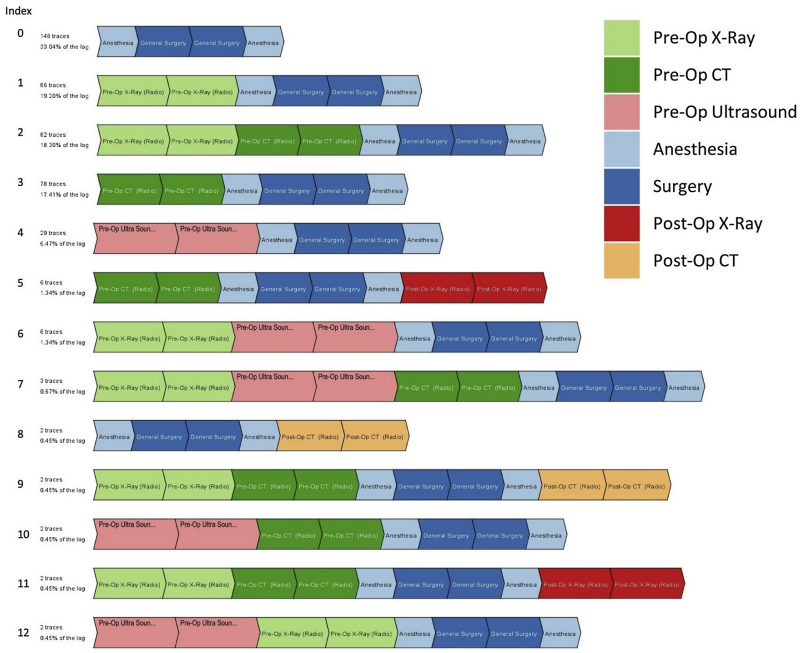

Figura 3.8. Variações das trajetórias de apendicite geradas pelo *plug-in Explore Event Log* do ProM. As quatro primeiras correspondem a 88% dos pacientes. (Fonte: [Kempa-Liehr et al. 2020])

Com o auxílio de médicos especialistas chegou-se a um modelo de trajetória com variações significativas clinicamente (vide Figura 3.9) e que foram usadas como base para investigar, por meio de modelos probabilísticos de aprendizado de máquina, os fatores que influenciam no tempo de internação após a cirurgia. A hipótese é que a idade e a severidade da inflamação do apêndice seriam os fatores que influenciariam no tempo de internação pós-operatório. Como não há uma forma objetiva de conhecer a severidade da inflamação, considerou-se para este estudo o tempo de cirurgia. A análise do comportamento das trajetórias não é a mesma ao se considerar a idade e o tempo de cirurgia. Os estudos sugerem que o tempo de internação é menor em pacientes mais novos, porém quando o idoso não possui comorbidades (outras doenças além da apendicite) o tempo de internação é basicamente o mesmo que o dos mais jovens.

Assim, conclui-se que analisar as trajetórias de pacientes utilizando técnicas de mineração de processo aliadas a modelos probabilísticos de aprendizado de máquina pa-

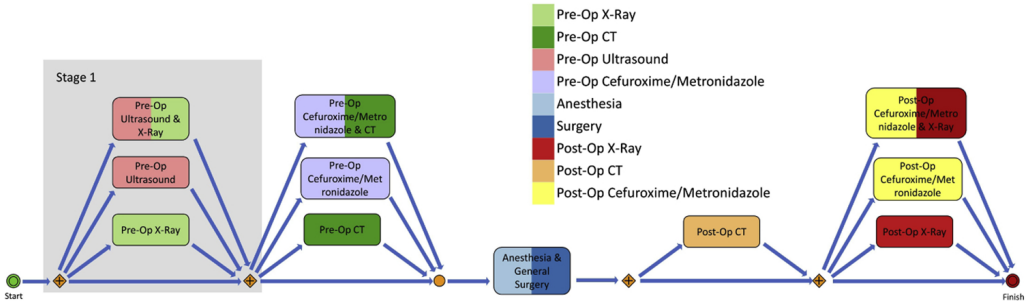

Figura 3.9. Modelo de trajetória de apendicite criado após análise de especialistas das 13 variações. (Fonte: [Kempa-Liehr et al. 2020])

recem ser promissores para entender as não conformidades de trajetórias de paciente. Para análises detalhadas do processo, a visualização fornecida pelo ProM é suficiente, porém para tornar as trajetórias mais fáceis para uma interpretação clínica é necessário que especialistas façam uma reformulação das trajetórias de forma manual.

3.4.2. Perspectiva Clinica

O monitoramento e gerenciamento da saúde de pacientes crônicos com ou sem multimorbidade é um desafio. São pacientes atendidos por várias especialidades médicas, além de outros profissionais que influenciam nos hábitos de cada um, como nutricionistas, fisioterapeutas, entre outros. Coordenar as atividades de todos de tal forma que suas orientações não entrem em conflito ou sejam contraditórias é essencial para o sucesso no tratamento e no uso dos recursos [Ito 2020].

Os profissionais que atendem a um mesmo paciente formam o time de cuidado. Coordenação do cuidado é a definição da coordenação que é preciso ter entre esses profissionais na assistência ao paciente. A coordenação de cuidado tem como objetivo reduzir a fragmentação do cuidado e melhorar a qualidade do serviço de saúde prestado ao paciente e sua família. Desde os anos 1960, o conceito de gerenciamento da saúde por meio da coordenação do cuidado vem sendo estudado e implementado [McDonald et al. 2007].

Nos anos 2000 com o aumento das doenças crônicas, notou-se que tratar os pacientes de forma padronizada de acordo com a sua doença não estava surtindo o efeito desejado. A mesma doença, devido a características diferentes entre os pacientes, não tinha um comportamento padrão. Assim, foi proposto o cuidado centrado no paciente no qual, ao invés de se ter linhas de cuidados baseadas na doença, essas seriam propostas de acordo com as condições clínicas e socioeconômicas dos pacientes. Porém, individualizar o tratamento pode causar confusão nas condutas clínicas. Assim, apesar de todo paciente ser único, existem grupos que possuem características comuns que permitem definir protocolos clínicos padrões de acordo com cada grupo. Assim, conhecer, analisar e avaliar a jornada de pacientes comuns é algo desejável para proporcionar um cuidado adequado aos pacientes [Ito 2020, Rogers et al. 2008].

Nesse contexto, modelar a jornada do paciente se torna um instrumento para implementar melhorias no cuidado centrado no paciente, pois permite aumentar a organização dos processos de cuidado e melhora o trabalho em equipe dos profissionais de saúde. Dessa forma, a jornada clínica do paciente é considerada uma ferramenta baseada

em evidências para orientar os cuidados de saúde. É composta por atendimentos multidisciplinares estruturados que detalham etapas essenciais no tratamento de doentes com problemas clínicos específicos. Desde os anos de 1980 são definidos protocolos clínicos padrões que são elaborados a partir de estudos de jornadas clínicas de pacientes encontradas em pesquisas clínicas científicas. Esses protocolos clínicos são usados para entender a evolução de doenças e otimizar o tratamento [Kinsman et al. 2010, Ursoniu et al. 2012].

Tendo em vista a necessidade de extração da jornada clínica de pacientes comuns para incrementar os protocolos clínicos, a mineração de jornada de pacientes parece ser um meio possível para comparar os protocolos clínicos padrão com o que ocorre no mundo real e, com isso, propor adaptações e melhorias. Além disso, é possível encontrar novos agrupamentos de pacientes para que cada vez mais os protocolos clínicos possam abranger um número maior de pacientes com cuidados adequados às suas condições clínicas e socioeconômicas.

3.4.2.1. Estudo de caso 1: Identificação de tratamentos comuns em câncer de mama [Zaballa et al. 2020]

Neste estudo de caso, pretende-se analisar a sequência de procedimentos médicos de um sistema de saúde no tratamento de câncer de mama. O método utilizado para modelar a jornada do paciente seguiu os seguintes passos: (1) identificar os procedimentos no sistema de saúde associado à doença; (2) identificar os pacientes com o tratamento completo da doença; (3) descobrir as jornadas de tratamentos comuns de pacientes com aquele diagnóstico.

Como a jornada está relacionada com o tratamento realizado, os dados contêm o serviço médico realizado, a especialidade médica e o diagnóstico. As jornadas geradas não estão relacionadas a apenas um diagnóstico, pois isso significa que o paciente possui multimorbidade (outras doenças, além do câncer de mama) e são denominadas como jornada do tratamento. Verifica-se que, com a associação desses dados, é possível ter a jornada do tratamento da doença, caso jornadas por doença sejam modeladas. Nesse caso somente os registros que possuem o diagnóstico servem para criar a jornada de tratamento por doença. Porém, como registros administrativos nem sempre possuem o diagnóstico, técnicas para contornar esse problema são propostas neste estudo.

O método detalhado para a modelagem das jornadas encontra-se na Figura 3.10 e cada etapa é detalhada a seguir:

- **Criação da jornada de tratamentos a partir de registros administrativos do sistema de saúde** – É preciso converter os registros numa jornada de tratamento. Então, é preciso criar uma sequência de procedimentos de tal forma que cada paciente esteja associado a uma jornada. Foram usadas sequências discretas, variadas em tamanho por conta da heterogeneidade das histórias dos pacientes.
- **Extração completa dos tratamentos associados aos diagnósticos** – Após a criação das jornadas, subsequências de procedimentos associados com o diagnóstico de interesse foram extraídos das jornadas encontradas. As etapas para encontrar essas subsequências:

  - **Identificar os pacientes com os diagnósticos de interesse e os seus procedimentos** – Como existiam registros que não tinham o diagnóstico, não foi possível extrair a subsequência pelo diagnóstico inicial. Então foi proposta uma métrica por relevância considerando a especialidade médica para identificar quais os procedimentos típicos estão relacionados ao diagnóstico de interesse.[9]
  - **Identificar os pacientes com alta probabilidade de ter o tratamento completo** – Encontrar as subsequências somente pelo diagnóstico inicial do paciente não é o suficiente, pois há um mesmo procedimentos para mais de um tipo de diagnóstico, assim critérios de seleção foram definidos e aplicados: (a) certificar-se que o tratamento está todo relacionado com o diagnóstico alvo – pacientes com diagnósticos semelhantes são desconsiderados; (b) retirar os pacientes em que não há o início e/ou termino do tratamento na jornada criada; (c) permitir somente pacientes que tenham um acompanhamento mínimo no diagnóstico alvo.

- **Agrupar as jornadas de tratamento de doenças** – Após criar as jornadas das doenças foi possível identificar os grupos mais significantes e agrupá-los. A ideia é agrupar os tratamentos similares entre si, mas que não são similares ao tratamento dado a outros grupos. A similaridade foi calculada pelo método da distância de Levenshtein (vide Subseção 3.3.2) e agrupados pelo método *k-medoids* (vide Subseção 3.3.1).

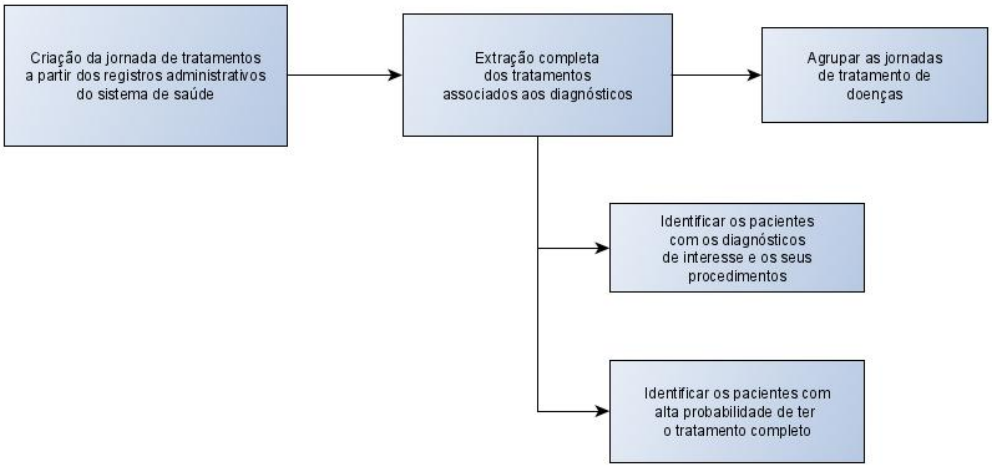

Figura 3.10. Método para modelar a jornada de tratamento por diagnóstico. (Fonte: adaptado de [Zaballa et al. 2020])

Para modelar a jornada foram utilizados registros administrativos do sistema público de saúde do país Basco (Espanha) chamado Osakidetza. Essa base tinha os registros de 579.798 pacientes pertencentes a vários níveis do sistema de saúde (1 hospital, 11 ambulatórios e serviços de emergência) de janeiro de 2016 a dezembro de 2017. Destes, 1.456 pacientes foram diagnosticados com câncer de mama. Após aplicar o método

[9]Para detalhes do método recomenda-se a leitura de [Zaballa et al. 2020].

para identificar os procedimentos relacionados ao câncer de mama, 21 pacientes foram retirados do estudo.

Para encontrar os pacientes com tratamento completo foram definidos critérios de seleção: (a) incluir pacientes com diagnóstico de câncer de mama; (b) ter pelo menos um exame de biopsia que atestasse a localização do câncer; (c) o diagnóstico tem que ter ocorrido entre 1/02/2016 a 30/09/2017; (d) a radioterapia é feita a cada 2 dias e a quimioterapia a cada 1-3 semanas, assim se tiver alguma radioterapia ou quimioterapia nas últimas 3 semanas de 2017, significa que o tratamento não acabou em 2017; e (e) o período mínimo de acompanhamento do paciente deve ser de pelo menos 3 meses, e a quantidade de procedimentos tem que ser pelo menos de 15. Após aplicados os critérios de seleção, restaram 440 pacientes.

O algoritmo *k-medoids* foi empregado, e ao final encontrou-se 5 grupos de jornadas que estão representadas na Figura 3.11, na qual as linhas horizontais representam as jornadas de tratamento da doença e as verticais, os serviços hospitalares em que o paciente compareceu para executar o procedimento. A fim de validar os resultados foi feita uma comparação com protocolos clínicos de câncer de mama da *European Society for Medical Oncology* e validados por médicos especialistas.

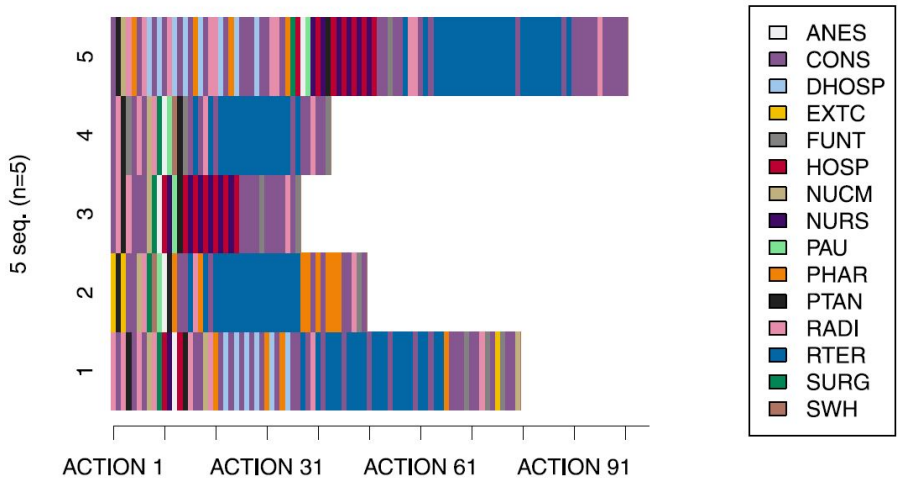

Figura 3.11. Resultado dos agrupamentos, foram encontrados 5 grupos. Neste gráfico *ACTION* são os procedimentos. A nomenclatura dos serviços é: ANES: Anestesia; CONS: Consulta; DHOSP: Hospital Dia; EXTC: Consulta externa; FUNT: Teste funcional; HOSP: Internação; NUCM: Medicina nuclear; NURS: Enfermaria; PAU: Unidade de cuidado pós-operatório; PHAR: Farmácia; PTAN: Anatomia patológica; RADI: Radiologia; RTER: Radioterapia; SURG: Cirurgia; SWH: Cirurgia sem internação.
(Fonte: [Zaballa et al. 2020])

Para cada grupo foi definido uma jornada padrão:

- Grupo 1: Cirurgia + Quimioterapia + Radioterapia (66 pacientes, 15%);
- Grupo 2: Cirurgia + Radioterapia + Terapia Hormonal (89 pacientes, 20,3%);
- Grupo 3: Cirurgia + Internação (108 pacientes, 24,6%);

- Grupo 4: Cirurgia + Quimioterapia (137 pacientes, 31,2%);
- Grupo 5: Quimioterapia + Cirurgia + Internação + Radioterapia (40 pacientes, 9,1%).

As análises realizadas encontraram achados interessantes, como, por exemplo, de que a mamografia bilateral foi realizada em todos os cinco grupos, parecendo ser um exame recomendado no tratamento do câncer de mama. As consultas regulares por estes pacientes seguiram o recomendado (a cada 3-4 meses). No geral, o método adotado parece promissor em extrair jornadas de registros administrativos, mas ainda é difícil modelar as jornadas na qual nem todos os registros possuem o diagnóstico devido a existência de multimorbidade. Foi possível verificar a utilidade de se encontrar jornadas reais e confrontá-las com os protocolos clínicos padrão para identificar desvios no tratamento. Encontrar os desvios não significa que houve erro ou negligência, mas pontos que precisam ser investigados, e, se for o caso, mudanças nos protocolos clínicos levando em conta a regionalização devem ser definidas.

Há limitações no método proposto, como aplicá-lo em diagnósticos de doenças em que os procedimentos realizados são comuns a várias doenças, de forma que é praticamente impossível associá-lo a uma única doença (p.ex. a sinusite aguda não possui um ou mais procedimentos específicos para que se possa extrair da base de dados uma jornada de tratamento). Outro ponto é com relação a doenças de longa duração, tão longas que ultrapassam os registros existentes na base de dados. Por exemplo, no caso do tratamento de câncer de mama, tratamentos com terapia hormonal duram de 5 a 10 anos, de modo que bases de dados com registros com menos de 5 anos não irão possuir o tratamento completo.

3.4.2.2. Estudo de caso 2: Extração e visualização da jornada clínica de doentes renais crônicos [Zhang et al. 2015b]

Medicina baseada em evidência é uma abordagem sistemática para proporcionar uma assistência de saúde consistente, segura e confiável. Contudo, modelos validados, métodos e ferramentas necessários para aplicar a medicina baseada em evidências no ponto de cuidado, principalmente aqueles que evidenciem as particularidades regionais e a prática real inexistem. Seguir protocolos padrões que não considerem o regionalismo e as condições socioeconômicas dos pacientes pode, ao invés de melhorar a assistência, torná-la impossível na prática; ao mesmo tempo, não é possível praticar a medicina baseada no empirismo dos profissionais [Djulbegovic and Guyatt 2017].

Uma fonte importante para evidências clínicas é a jornada clínica do paciente ao indicar a sequência mais provável de intervenções no tratamento para grupos específicos de pacientes. Ao mesmo tempo, com o uso intenso da tecnologia da informação na área da saúde, é possível obter dados que permitem realizar as análises e revisões dos fluxos de trabalho, e, por consequência, melhorar a prática da medicina baseada em evidências por meio do aprendizado de jornadas clínicas que ocorrem no mundo real (jornadas baseadas em práticas clínicas[10]).

[10]Tradução livre de *practice-based clinical pathways*

É nesse contexto que os autores utilizam a tecnologia para extrair, modelar e analisar a jornada clínica de pacientes renais crônicos. Para o processo de desenvolvimento da jornada clínica, os seguintes passos são propostos: (a) Obter e tratar o dado do registro de saúde (EHR); (b) processar o dado e modelar o problema; (c) identificar os subgrupos de pacientes relevantes para o estudo; (d) minerar os tratamentos padrões mais comuns; (e) passar por avaliações de especialistas; e (f) modelar as jornadas baseadas nas práticas clínicas. Visualizações dos modelos são criadas para que possam ser usadas pelas instituições de saúde para a revisão da prática clínica e o apoio à decisão e pelos pacientes para comunicação na decisão compartilhada e educação.

Para este estudo de caso, os dados foram extraídos de uma comunidade clínica de nefrologia da Pensilvânia. Foram encontrados 1.576 pacientes com doenças crônicas no período de 01/01/2009 a 30/06/2013. Os pacientes tinham pelo menos 5 consultas realizadas. As consultas poderiam ser no consultório, no hospital ou com fins educacionais. Esses pacientes realizaram um total de 17.358 consultas.

Neste estudo, inicialmente cada consulta, contendo quatro tipos de informação (motivo da consulta, procedimento, medicação e o diagnóstico), é representada por um nó na jornada clínica do paciente. Em seguida, são definidos super-nós formados pela combinação dessas informações (os valores dos quatro campos são concatenados, e cada combinação única recebe um rótulo). A sequência desses super-nós ordenados pela data das consultas forma a jornada clínica do paciente. Cada paciente tem uma única sequência, iniciando com a primeira consulta registrado no sistema de informação e termina com a última consulta. Com os dados obtidos, após o processamento, foram encontrados 804 super-nós.

As jornadas criadas a partir da sequência dos super-nós geram uma jornada de paciente distinta para cada paciente. Neste caso, foram geradas 1.576 jornadas, então é preciso agrupá-las a fim de encontrar jornadas padrões de acordo com cada grupo. A proposta de agrupamento é por similaridade entre as sequências e foi usado o LCS. Ao aplicar o algoritmo de agrupamento proposto foram identificados 31 subgrupos.

O próximo passo é extrair a jornada clínica do paciente padrão dos subgrupos. Para cada grupo, identificou-se as sucessões diretas entre super-nós existentes nas jornadas, e cada uma delas foi traduzida em um estado (chamado super-par) de um modelo de Markov. Esse procedimento foi feito para compensar a perda de memória dos modelos de Markov. Em seguida, cada modelo de Markov foi examinado e estados e transições frequentes foram selecionados para formar um grafo, tornando o modelo mais simples e fácil de entender clinicamente. Neste caso, os super-pares são os nós, e as arestas representam a transição entre os super-pares (vide Figura 3.12).

Ao final foram encontrados 3.505 super-pares e para alguns subgrupos a trajetória foi construída. Foi feito um estudo comparativo da trajetória de um subgrupo com a que foi gerada pelo *Heuristic Miner* do ProM [Verbeek et al. 2010] do mesmo subgrupo. Foram encontradas vantagens do algoritmo proposto em relação ao *Heuristic Miner*, uma delas é a habilidade de mostrar a associação temporal correta e a informação codificada. Outro ponto a favor é que a trajetória é mais fiel com a realidade, pois o *Heuristic Miner* simplifica demais os processos e fatos são omitidos, o que pode tornar a trajetória algo sem sentido para os especialistas. Assim, parece que o algoritmo proposto permite que os

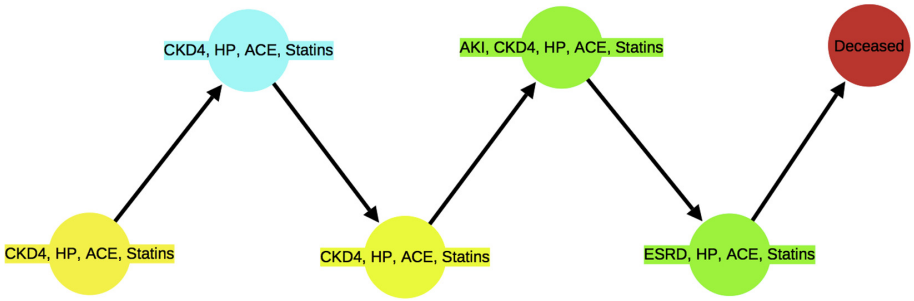

Figura 3.12. Visualização da trajetória clínica dos pacientes do subgrupo 29. Nó amarelo: consulta, nó verde: internação, nó azul: visita educacional, nó vermelho: óbito, CKD4: doença crônica no estágio 4, HP: hipertensão, Statins: estatina (medicamento), ACE: inibidor da enzima conversora de angiotensina, AKI: diagnóstico de doença crônica, ESRD: estágio final da doença renal crônica. (Fonte: [Zhang et al. 2015b])

tomadores de decisões possam ter uma análise melhor e mais condizente com a realidade. Visualizar as trajetórias obtidas pelo algoritmo pode auxiliar na comunicação entre médicos e pacientes e permitir a tomada de decisão compartilhada. Vale ressaltar que este tipo de análise depende da completude da base de dados, pois, caso alguma informação esteja faltando, não é possível extrair corretamente a trajetória clínica do paciente.

3.5. Sumário e Perspectivas

As primeiras tentativas de descobrir jornadas de pacientes usavam principalmente técnicas de mineração de sequências ou outros métodos de descoberta baseados no algoritmo Apriori. Os modelos mais comuns para representar as jornadas incluíam sequências, grafos e conjuntos de intervalos temporais. Dessa forma, eles geralmente se restringiam à representação de sucessões diretas entre as atividades. No entanto, parte deles permitia a repetição de rótulos de atividades, favorecendo a legibilidade. A noção temporal geralmente se restringia à apresentação da ordem dos acontecimentos. Embora tais métodos fossem capazes de descobrir padrões frequentes de forma eficaz, eles estavam sujeitos ao risco de negligenciar comportamentos pouco frequentes porém relevantes.

Com o desenvolvimento da mineração de processos como área de pesquisa e sua aplicação na área da saúde, a expressividade dos modelos de jornadas de pacientes aumentou, graças à identificação de divisões/junções dos processos e de dependências de longo prazo, por exemplo. Além disso, a maioria das ferramentas de mineração de processos fornece um resumo do tempo decorrido entre as atividades. Se por um lado os modelos se tornaram mais expressivos, por outro, o desafio de lidar com a alta variabilidade de percursos e atividades se tornou mais evidente. Os algoritmos tradicionais de mineração de processos se concentram na descoberta de modelos de processos estruturados e, portanto, frequentemente se defrontam com processos "espaguete" ao lidar com dados de saúde. Além disso, vários deles não suportam rótulos de atividades repetidos, o que pode dificultar a leitura do modelo. Alternativamente, métodos auxiliares, como agrupamento de eventos ou pacientes, e filtragem de eventos ou jornadas foram adotados.

O agrupamento de pacientes tem sido utilizado não apenas como método auxiliar, mas também como a principal estratégia de avaliação das jornadas dos pacientes, seja

tomando uma jornada central como representativa do cluster ou avaliando a distribuição temporal das atividades dentro de cada cluster. Embora essa estratégia lide bem com a variabilidade dos dados, ela não gera um modelo de jornadas explícito. Outras abordagens utilizadas focaram em casos específicos, como cálculo de parâmetros de modelos de Markov ou a extração direta de jornadas como se apresentam nos dados em casos mais simples.

Enquanto isso, no âmbito da mineração de processos, outros algoritmos de descoberta foram propostos, com foco em processos menos estruturados, como os encontrados na área da saúde. Um exemplo é o algoritmo *Fuzzy Miner* que usa um grafo como modelo de processo e o simplifica priorizando nós e arestas relevantes. Deste modo, é possível reduzir a variabilidade das jornadas sem necessariamente remover comportamentos potencialmente significativos, embora pouco frequentes. Muitos softwares, inclusive comerciais, implementaram essa abordagem de simplificação, tornando-a popular para a mineração de jornadas de pacientes. No entanto, o uso de um grafo que não permite repetição de rótulos de atividades e que mostra apenas sucessões diretas entre elas, aliado a falhas de implementações de algoritmos de descoberta, pode levar a modelos pouco claros e até enganosos, como [van der Aalst 2019] discutiu recentemente. Além disso, embora a ideia proposta inicialmente no algoritmo *Fuzzy Miner* fosse de selecionar nós e arestas relevantes para a estrutura do modelo de processo, aparentemente a maioria dos autores que utilizaram algoritmos de simplificação limitaram-se a identificar atividades e transições frequentes.

Mais recentemente, problemas de otimização têm sido propostos para a obtenção de modelos de jornadas de pacientes. Essa estratégia é um pouco semelhante à da simplificação, pois também busca por nós e arestas chaves para serem incluídos no modelo final, porém é guiada pela otimização de uma função objetivo (p.ex. valor de uma métrica). Dependendo do modelo matemático utilizado, esses algoritmos podem enfrentar os mesmos problemas que os de simplificação, mas representações alternativas foram propostas, como o *Time Grid Process Model* de [de Oliveira et al. 2020a], que distingue atividades que ocorrem em diferentes momentos de um caminho. Os autores que usaram essa abordagem focaram principalmente em maximizar a medida de *replayability* do modelo. Em outras palavras, seus algoritmos escolhem atividades e transições de modo que o modelo de jornadas reproduza o máximo possível o comportamento observado, o que tende a favorecer padrões frequentes.

Logo, a revisão de modelos e métodos de mineração de jornadas de pacientes sugere que o desenvolvimento de um método de mineração que lide com a variabilidade de dados selecionando atividades e transições relevantes com base em sua importância para o modelo de jornadas, como nos casos de simplificação, mas levando em conta também a influência do tempo e do contexto seria vantajoso. Ele também deve fornecer um modelo de jornadas que seja fácil de ler e interpretar. Trata-se de um tema desafiador e ainda em aberto na literatura.

Referências

[Agrawal and Srikant 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering*,

pages 3–14. IEEE.

[Agrawal et al. 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Citeseer.

[Andrews et al. 2020] Andrews, R., Wynn, M. T., Vallmuur, K., Ter Hofstede, A. H., and Bosley, E. (2020). A comparative process mining analysis of road trauma patient pathways. *International Journal of Environmental Research and Public Health*, 17(10).

[Ang et al. 2016] Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., and Aratow, M. (2016). Accurate emergency department wait time prediction. *Manuf. Serv. Oper. Manag.*, 18(1):141–156.

[Antonelli et al. 2012] Antonelli, D., Baralis, E., Bruno, G., Chiusano, S., Mahoto, N. A., and Petrigni, C. (2012). Analysis of diagnostic pathways for colon cancer. *Flexible Services and Manufacturing Journal*, 24(4):379–399.

[Arias et al. 2020] Arias, M., Rojas, E., Aguirre, S., Cornejo, F., Munoz-Gama, J., Sepúlveda, M., and Capurro, D. (2020). Mapping the patient's journey in healthcare through process mining. *International Journal of Environmental Research and Public Health*, 17(18):1–16.

[Arnolds and Gartner 2018] Arnolds, I. V. and Gartner, D. (2018). Improving hospital layout planning through clinical pathway mining. *Annals of Operations Research*, 263(1-2):453–477.

[Aspland et al. 2021] Aspland, E., Harper, P. R., Gartner, D., Webb, P., and Barrett-Lee, P. (2021). Modified Needleman–Wunsch algorithm for clinical pathway clustering. *Journal of Biomedical Informatics*, 115.

[Baker et al. 2017] Baker, K., Dunwoodie, E., Jones, R. G., Newsham, A., Johnson, O., Price, C. P., Wolstenholme, J., Leal, J., McGinley, P., Twelves, C., and Hall, G. (2017). Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *International Journal of Medical Informatics*, 103:32–41.

[Baro et al. 2015] Baro, E., Degoul, S., Beuscart, R., and Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed research international*, 2015.

[Basole et al. 2015] Basole, R. C., Braunstein, M. L., Kumar, V., Park, H., Kahng, M., Chau, D. H., Tamersoy, A., Hirsh, D. A., Serban, N., Bost, J., Lesnick, B., Schissel, B. L., and Thompson, M. (2015). Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. *Journal of the American Medical Informatics Association*, 22(2):318–323.

[Benevento et al. 2019] Benevento, E., Aloini, D., Squicciarini, N., Dulmin, R., and Mininno, V. (2019). Queue-based features for dynamic waiting time prediction in emergency department. *Measuring Business Excellence*, 23(4):458–471.

[Bengio et al. 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[Berndt and Clifford 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, page 359–370. AAAI Press.

[Bettencourt-Silva et al. 2015] Bettencourt-Silva, J. H., Clark, J., Cooper, C. S., Mills, R., Rayward-Smith, V. J., and Iglesia, B. D. L. (2015). Building Data-Driven Pathways from Routinely Collected Hospital Data: A Case Study on Prostate Cancer. *JMIR Medical Informatics*, 3(3).

[Blei et al. 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

[Caron et al. 2014] Caron, F., Vanthienen, J., Vanhaecht, K., Limbergen, E. V., De Weerdt, J., and Baesens, B. (2014). Monitoring care processes in the gynecologic oncology department. *Computers in Biology and Medicine*, 44(1):88–96.

[Chen et al. 2018] Chen, J., Sun, L., Guo, C., Wei, W., and Xie, Y. (2018). A data-driven framework of typical treatment process extraction and evaluation. *Journal of Biomedical Informatics*, 83:178–195.

[Chiudinelli et al. 2020] Chiudinelli, L., Dagliati, A., Tibollo, V., Albasini, S., Geifman, N., Peek, N., Holmes, J. H., Corsi, F., Bellazzi, R., and Sacchi, L. (2020). Mining post-surgical care processes in breast cancer patients. *Artificial Intelligence in Medicine*, 105.

[Cho et al. 2020] Cho, M., Kim, K., Lim, J., Baek, H., Kim, S., Hwang, H., Song, M., and Yoo, S. (2020). Developing data-driven clinical pathways using electronic health records: The cases of total laparoscopic hysterectomy and rotator cuff tears. *International Journal of Medical Informatics*, 133.

[Conca et al. 2018] Conca, T., Saint-Pierre, C., Herskovic, V., Sepúlveda, M., Capurro, D., Prieto, F., and Fernandez-Llatas, C. (2018). Multidisciplinary collaboration in the treatment of patients with type 2 diabetes in primary care: Analysis using process mining. *Journal of Medical Internet Research*, 20(4).

[Connelly and Bair 2004] Connelly, L. G. and Bair, A. E. (2004). Discrete event simulation of emergency department activity: A platform for system-level operations research. *Academic Emergency Medicine*, 11(11):1177–1185. Cited By :159.

[Dagliati et al. 2017] Dagliati, A., Sacchi, L., Zambelli, A., Tibollo, V., Pavesi, L., Holmes, J. H., and Bellazzi, R. (2017). Temporal electronic phenotyping by mining careflows of breast cancer patients. *Journal of Biomedical Informatics*, 66:136–147.

[Dagliati et al. 2018] Dagliati, A., Tibollo, V., Cogni, G., Chiovato, L., Bellazzi, R., and Sacchi, L. (2018). Careflow Mining Techniques to Explore Type 2 Diabetes Evolution. *Journal of Diabetes Science and Technology*, 12(2):251–259.

[Dahlem et al. 2015] Dahlem, D., Maniloff, D., and Ratti, C. (2015). Predictability bounds of electronic health records. *Scientific Reports*, 5:1–9.

[Dahlin and Raharjo 2019] Dahlin, S. and Raharjo, H. (2019). Relationship between patient costs and patient pathways. *International Journal of Health Care Quality Assurance*, 32(1):246–261.

[De Oliveira et al. 2020] De Oliveira, H., Augusto, V., Jouaneton, B., Lamarsalle, L., Prodel, M., and Xie, X. (2020). Automatic and explainable labeling of medical event logs with autoencoding. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3076–3084.

[de Oliveira et al. 2020a] de Oliveira, H., Augusto, V., Jouaneton, B., Lamarsalle, L., Prodel, M., and Xie, X. (2020a). Optimal process mining of timed event logs. *Information Sciences*, 528:58–78.

[de Oliveira et al. 2020b] de Oliveira, H., Prodel, M., Lamarsalle, L., Inada-Kim, M., Ajayi, K., Wilkins, J., Sekelj, S., Beecroft, S., Snow, S., Slater, R., and Orlowski, A. (2020b). "Bow-tie" optimal pathway discovery analysis of sepsis hospital admissions using the Hospital Episode Statistics database in England. *JAMIA Open*, 3(3):439–448.

[Defossez et al. 2014] Defossez, G., Rollet, A., Dameron, O., and Ingrand, P. (2014). Temporal representation of care trajectories of cancer patients using data from a regional information system: An application in breast cancer. *BMC Medical Informatics and Decision Making*, 14(1):24.

[Ding et al. 2010] Ding, R., McCarthy, M. L., Desmond, J. S., Lee, J. S., Aronsky, D., and Zeger, S. L. (2010). Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Academic Emergency Medicine*, 17(8):813–823.

[Djulbegovic and Guyatt 2017] Djulbegovic, B. and Guyatt, G. H. (2017). Progress in evidence-based medicine: a quarter century on. *The lancet*, 390(10092):415–423.

[Dong et al. 2018] Dong, J., Yom-Tov, G., and Yom-Tov, E. (2018). The impact of delay announcements on hospital network coordination and waiting times. *Management Science*.

[Duguay and Chetouane 2007] Duguay, C. and Chetouane, F. (2007). Modeling and improving emergency department systems using discrete event simulation. *Simulation*, 83(4):311–320. Cited By :192.

[Duma and Aringhieri 2020] Duma, D. and Aringhieri, R. (2020). An ad hoc process mining approach to discover patient paths of an Emergency Department. *Flexible Services and Manufacturing Journal*, 32(1):6–34.

[Durojaiye et al. 2018] Durojaiye, A. B., McGeorge, N. M., Puett, L. L., Stewart, D., Fackler, J. C., Hoonakker, P. L., Lehmann, H. P., and Gurses, A. P. (2018). Mapping the Flow of Pediatric Trauma Patients Using Process Mining. *Applied Clinical Informatics*, 9(3):654–666.

[Egho et al. 2014] Egho, E., Jay, N., Raïssi, C., Ienco, D., Poncelet, P., Teisseire, M., and Napoli, A. (2014). A contribution to the discovery of multidimensional patterns in healthcare trajectories. *Journal of Intelligent Information Systems*, 42(2):283–305.

[Egho et al. 2015] Egho, E., Raïssi, C., Calders, T., Jay, N., and Napoli, A. (2015). On measuring similarity for sequences of itemsets. *Data Mining and Knowledge Discovery*, 29(3):732–764.

[Erdogan and Tarhan 2018] Erdogan, T. G. and Tarhan, A. (2018). A goal-driven evaluation method based on process mining for healthcare processes. *Applied Sciences (Switzerland)*, 8(6).

[Ester et al. 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

[Fei and Meskens 2013] Fei, H. and Meskens, N. (2013). Clustering of Patients' Trajectories with an Auto-Stopped Bisecting K-Medoids Algorithm. *Journal of Mathematical Modelling and Algorithms*, 12(2):135–154.

[Fernandez and Benedí 2008] Fernandez, C. and Benedí, J. M. (2008). Timed Parallel Automaton learning in Workflow Mining problems. In *1er. Congreso Internacional de Mecatrónica y 2do. Congreso Nacional UP*, pages 1–8, Tuxtla Gutiérrez, Mexico.

[Forestier et al. 2012] Forestier, G., Lalys, F., Riffaud, L., Trelhu, B., and Jannin, P. (2012). Classification of surgical processes using dynamic time warping. *Journal of Biomedical Informatics*, 45(2):255–264.

[Frey and Dueck 2007] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814):972–976.

[Garg et al. 2009] Garg, L., McClean, S., Meenan, B., and Millard, P. (2009). Non-homogeneous Markov models for sequential pattern mining of healthcare data. *IMA Journal of Management Mathematics*, 20(4):327–344.

[Gonzalez-Garcia et al. 2020] Gonzalez-Garcia, J., Telleria-Orriols, C., Estupinan-Romero, F., and Bernal-Delgado, E. (2020). Construction of Empirical Care Pathways Process Models from Multiple Real-World Datasets. *IEEE Journal of Biomedical and Health Informatics*, 24(9):2671–2680.

[Günther and van der Aalst 2007] Günther, C. W. and van der Aalst, W. M. P. (2007). Fuzzy mining – adaptive process simplification based on multi-perspective metrics. In Alonso, G., Dadam, P., and Rosemann, M., editors, *Business Process Management*, pages 328–343, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Hirschberg 1975] Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343.

[Huang et al. 2014] Huang, Z., Dong, W., Duan, H., and Li, H. (2014). Similarity measure between patient traces for clinical pathway analysis: Problem, method, and applications. *IEEE Journal of Biomedical and Health Informatics*, 18(1):4–14.

[Huang et al. 2016] Huang, Z., Dong, W., Ji, L., He, C., and Duan, H. (2016). Incorporating comorbidities into latent treatment pattern mining for clinical pathways. *Journal of Biomedical Informatics*, 59:227–239.

[Huang et al. 2015] Huang, Z., Dong, W., Ji, L., Yin, L., and Duan, H. (2015). On local anomaly detection and analysis for clinical pathways. *Artificial Intelligence in Medicine*, 65(3):167–177.

[Huang et al. 2018] Huang, Z., Ge, Z., Dong, W., He, K., and Duan, H. (2018). Probabilistic modeling personalized treatment pathways using electronic health records. *Journal of Biomedical Informatics*, 86:33–48.

[Huang et al. 2012] Huang, Z., Lu, X., and Duan, H. (2012). On mining clinical pathway patterns from medical behaviors. *Artificial Intelligence in Medicine*, 56(1):35–50.

[Huang et al. 2013] Huang, Z., Lu, X., Duan, H., and Fan, W. (2013). Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1):111–127.

[Hur et al. 2020] Hur, C., Wi, J., and Kim, Y. (2020). Facilitating the development of deep learning models with visual analytics for electronic health records. *International Journal of Environmental Research and Public Health*, 17(22):1–14.

[Ito 2020] Ito, M. (2020). Chapter 6 - patient-centered care. In Gogia, S., editor, *Fundamentals of Telemedicine and Telehealth*, pages 115–126. Academic Press.

[Janssenswillen et al. 2019] Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., and Vanhoof, K. (2019). bupar: Enabling reproducible business process analysis. *Knowledge-Based Systems*, 163:927–930.

[Kaufman and Rousseeuw 1990] Kaufman, L. and Rousseeuw, P. (1990). *Partitioning Around Medoids (Program PAM)*, chapter 2, pages 68–125. John Wiley & Sons, Ltd.

[Kempa-Liehr et al. 2020] Kempa-Liehr, A. W., Lin, C. Y. C., Britten, R., Armstrong, D., Wallace, J., Mordaunt, D., and O'Sullivan, M. (2020). Healthcare pathway discovery and probabilistic machine learning. *International Journal of Medical Informatics*, 137.

[Khan et al. 2018] Khan, A., Uddin, S., and Srinivasan, U. (2018). Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression. *International Journal of Medical Informatics*, 115:1–9.

[Kim et al. 2013] Kim, E., Kim, S., Song, M., Kim, S., Yoo, D., Hwang, H., and Yoo, S. (2013). Discovery of outpatient care process of a tertiary university hospital using process mining. *Healthcare Informatics Research*, 19(1):42–49.

[Kinsman et al. 2010] Kinsman, L., Rotter, T., James, E., Snow, P., and Willis, J. (2010). What is a clinical pathway? development of a definition to inform the debate. *BMC medicine*, 8.

[Kurniati et al. 2019] Kurniati, A. P., Rojas, E., Hogg, D., Hall, G., and Johnson, O. A. (2019). The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. *Health Informatics Journal*, 25(4):1878–1893.

[Le Meur et al. 2015] Le Meur, N., Gao, F., and Bayat, S. (2015). Mining care trajectories using health administrative information systems: The use of state sequence analysis to assess disparities in prenatal care consumption. *BMC Health Services Research*, 15(1).

[Leemans et al. 2013] Leemans, S. J., Fahland, D., and van der Aalst, W. M. (2013). Discovering block-structured process models from event logs-a constructive approach. In *International conference on applications and theory of Petri nets and concurrency*, pages 311–329. Springer.

[Leonardi et al. 2018] Leonardi, G., Striani, M., Quaglini, S., Cavallini, A., and Montani, S. (2018). Leveraging semantic labels for multi-level abstraction in medical process mining and trace comparison. *Journal of Biomedical Informatics*, 83:10–24.

[Levenshtein et al. 1966] Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

[Li et al. 2019] Li, Y., Gu, C., Dullien, T., Vinyals, O., and Kohli, P. (2019). Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning*, pages 3835–3845. PMLR.

[Lin et al. 2001] Lin, F.-r., Chou, S.-c., Pan, S.-m., and Chen, Y.-m. (2001). Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, 62(1):11–25.

[Lira et al. 2019] Lira, R., Salas-Morales, J., Leiva, L., de la Fuente, R., Fuentes, R., Delfino, A., Nazal, C. H., Sepúlveda, M., Arias, M., Herskovic, V., and Munoz-Gama, J. (2019). Process-Oriented Feedback through Process Mining for Surgical Procedures in Medical Training: The Ultrasound-Guided Central Venous Catheter Placement Case. *International Journal of Environmental Research and Public Health*, 16(11):1877.

[Lismont et al. 2016] Lismont, J., Janssens, A. S., Odnoletkova, I., vanden Broucke, S., Caron, F., and Vanthienen, J. (2016). A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways. *Computers in Biology and Medicine*, 77:125–134.

[Lu et al. 2016] Lu, F., Zeng, Q., and Duan, H. (2016). Synchronization-core-based discovery of processes with decomposable cyclic dependencies. *ACM Transactions on Knowledge Discovery from Data*, 10(3).

[MacQueen 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967).*

[Manktelow et al. 2022] Manktelow, M., Iftikhar, A., Bucholc, M., McCann, M., and O'Kane, M. (2022). Clinical and operational insights from data-driven care pathway mapping: a systematic review. *BMC Medical Informatics and Decision Making*, 22(1):43.

[Marazza et al. 2020] Marazza, F., Bukhsh, F. A., Geerdink, J., Vijlbrief, O., Pathak, S., van Keulen, M., and Seifert, C. (2020). Automatic process comparison for subpopulations: Application in cancer care. *International Journal of Environmental Research and Public Health*, 17(16):1–23.

[McDonald et al. 2007] McDonald, K. M., Sundaram, V., Bravata, D. M., Lewis, R., Lin, N., Kraft, S. A., McKinnon, M., Paguntalan, H., and Owens, D. K. (2007). Closing the quality gap: a critical analysis of quality improvement strategies (vol. 7: Care coordination). Technical Report 04(07)-0051-7, Agency for Healthcare Research and Quality. Technical Review 9 (Preparedby the Stanford University-UCSF Evidence-based Practice Center under contract 290-02-0017).

[Mertens et al. 2018] Mertens, S., Gailly, F., and Poels, G. (2018). Discovering healthcare processes using DeciClareMiner. *Health Systems*, 7(3):195–211.

[Mertens et al. 2020] Mertens, S., Gailly, F., Van Sassenbroeck, D., and Poels, G. (2020). Integrated Declarative Process and Decision Discovery of the Emergency Care Process. *Information Systems Frontiers*.

[Najjar et al. 2018] Najjar, A., Reinharz, D., Girouard, C., and Gagné, C. (2018). A two-step approach for mining patient treatment pathways in administrative healthcare databases. *Artificial Intelligence in Medicine*, 87:34–48.

[NHS England and NHS Improvement 2022] NHS England and NHS Improvement (2022). Flow – reduce unnecessary waits. Disponível em `https://www.england.nhs.uk/sustainableimprovement/qsir-programme/qsir-tools/quality-service-improvement-and-redesign-qsir-tools-by-stage-of-the-patient-pathway/`. Acesso em 28-05-2022.

[Partington et al. 2015] Partington, A., Wynn, M., Suriadi, S., Ouyang, C., and Karnon, J. (2015). Process mining for clinical processes: A comparative analysis of four australian hospitals. *ACM Transactions on Management Information Systems*, 5(4).

[Pastorino et al. 2019] Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., and Boccia, S. (2019). Benefits and challenges of big data in healthcare: an overview of the european initiatives. *European journal of public health*, 29(Supplement_3):23–27.

[Perer et al. 2015] Perer, A., Wang, F., and Hu, J. (2015). Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 56:369–378.

[Pianykh and Rosenthal 2015] Pianykh, O. S. and Rosenthal, D. I. (2015). Can we predict patient wait time? *Journal of the American College of Radiology*, 12(10):1058–1066.

[Poitras et al. 2018] Poitras, M.-E., Maltais, M.-E., Bestard-Denommé, L., Stewart, M., and Fortin, M. (2018). What are the effective elements in patient-centered and multimorbidity care? A scoping review. *BMC health services research*, 18(1):1–9.

[Prodel et al. 2018] Prodel, M., Augusto, V., Jouaneton, B., Lamarsalle, L., and Xie, X. (2018). Optimal Process Mining for Large and Complex Event Logs. *IEEE Transactions on Automation Science and Engineering*, 15(3):1309–1325.

[Prokofyeva and Zaytsev 2020] Prokofyeva, E. S. and Zaytsev, R. D. (2020). Clinical pathways analysis of patients in medical institutions based on hard and fuzzy clustering methods. *Business Informatics*, 14(1):19–31.

[Rebuge and Ferreira 2012] Rebuge, Á. and Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2):99–116.

[Rinner et al. 2018] Rinner, C., Helm, E., Dunkl, R., Kittler, H., and Rinderle-Ma, S. (2018). Process mining and conformance checking of long running processes in the context of melanoma surveillance. *International Journal of Environmental Research and Public Health*, 15(12).

[Rismanchian and Lee 2017] Rismanchian, F. and Lee, Y. H. (2017). Process Mining–Based Method of Designing and Optimizing the Layouts of Emergency Departments in Hospitals. *Health Environments Research and Design Journal*, 10(4):105–120.

[Rogers et al. 2008] Rogers, H., Maher, L., and Plsek, P. E. (2008). New rules for driving innovation in access to secondary care in the NHS.

[Rotondi et al. 1997] Rotondi, A. J., Brindis, C., Cantees, K. K., Deriso, B. M., Ilkin, H. M., Palmer, J. S., Gunnerson, H. B., and Watkins, W. D. (1997). Benchmarking the perioperative process. I. Patient routing systems: A method for continual improvement of patient flow and resource utilization. *Journal of Clinical Anesthesia*, 9(2):159–169.

[Sanfeliu and Fu 1983] Sanfeliu, A. and Fu, K.-S. (1983). A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362.

[Sato et al. 2020] Sato, D. M., Mantovani, L. K., Safanelli, J., Guesser, V., Nagel, V., Moro, C. H., Cabral, N. L., Scalabrin, E. E., Moro, C., and Santos, E. A. (2020). Ischemic stroke: Process perspective, clinical and profile characteristics, and external factors. *Journal of Biomedical Informatics*, 111(1):103582.

[Sawhney et al. 2021] Sawhney, S., Tan, Z., Black, C., Marks, A., Mclernon, D. J., Ronksley, P., and James, M. T. (2021). Validation of Risk Prediction Models to Inform Clinical Decisions After Acute Kidney Injury. *American Journal of Kidney Diseases*, 78(1):28–37.

[Senderovich et al. 2016] Senderovich, A., Weidlich, M., Yedidsion, L., Gal, A., Mandelbaum, A., Kadish, S., and Bunnell, C. A. (2016). Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. *Information Systems*, 62:185–206.

[Sequeira and Zaki 2002] Sequeira, K. and Zaki, M. (2002). Admit: Anomaly-based data mining for intrusions. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 386–395, New York, NY, USA. Association for Computing Machinery.

[Stefanini et al. 2018] Stefanini, A., Aloini, D., Benevento, E., Dulmin, R., and Mininno, V. (2018). Performance analysis in emergency departments: a data-driven approach. *Measuring Business Excellence*, 22(2):130–145.

[Stefanini et al. 2020] Stefanini, A., Aloini, D., Benevento, E., Dulmin, R., and Mininno, V. (2020). A data-driven methodology for supporting resource planning of health services. *Socio-Economic Planning Sciences*, 70(October 2019):100744.

[Sun et al. 2012] Sun, Y., Teow, K. L., Heng, B. H., Ooi, C. K., and Tay, S. Y. (2012). Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine*, 60(3):299–308.

[Tamburis and Esposito 2020] Tamburis, O. and Esposito, C. (2020). Process mining as support to simulation modeling: A hospital-based case study. *Simulation Modelling Practice and Theory*, 104(June):102149.

[Ursoniu et al. 2012] Ursoniu, S., Vernic, C., Muntean, C., and Timar, B. (2012). Nursing case management: Identifying, coordinating and monitoring the implementation of care services for patients. *Annals. Computer Science Series*, 10(2).

[van der Aalst 2019] van der Aalst, W. M. (2019). A practitioner's guide to process mining: Limitations of the directly-follows graph. *Procedia Computer Science*, 164:321–328.

[Verbeek et al. 2010] Verbeek, H., Buijs, J., Van Dongen, B., and van der Aalst, W. M. (2010). Prom 6: The process mining toolkit. *Proc. of BPM Demonstration Track*, 615:34–39.

[Villamil et al. 2017] Villamil, M. D. P., Barrera, D., Velasco, N., Bernal, O., Fajardo, E., Urango, C., and Buitrago, S. (2017). Strategies for the quality assessment of the health care service providers in the treatment of Gastric Cancer in Colombia. *BMC Health Services Research*, 17(1).

[Wang and Lin 2007] Wang, H. and Lin, Z. (2007). A novel algorithm for counting all common subsequences. In *Proceedings of the 2007 IEEE International Conference on Granular Computing*, GRC '07, page 502, USA. IEEE Computer Society.

[Wang et al. 2017] Wang, T., Tian, X., Yu, M., Qi, X., and Yang, L. (2017). Stage division and pattern discovery of complex patient care processes. *Journal of Systems Science and Complexity*, 30(5):1136–1159.

[Weijters et al. 2006] Weijters, A., van Der Aalst, W. M., and De Medeiros, A. A. (2006). Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP*, 166:1–34.

[Weiskopf and Weng 2013] Weiskopf, N. G. and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151.

[Xu et al. 2016] Xu, X., Jin, T., Wei, Z., Lv, C., and Wang, J. (2016). TCPM: Topic-Based Clinical Pathway Mining. *Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2016*, pages 292–301.

[Xu et al. 2017] Xu, X., Jin, T., Wei, Z., and Wang, J. (2017). Incorporating Topic Assignment Constraint and Topic Correlation Limitation into Clinical Goal Discovering for Clinical Pathway Mining. *Journal of Healthcare Engineering*, 2017.

[Yang and Hwang 2006] Yang, W. S. and Hwang, S. Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31(1):56–68.

[Yoo et al. 2016] Yoo, S., Cho, M., Kim, E., Kim, S., Sim, Y., Yoo, D., Hwang, H., and Song, M. (2016). Assessment of hospital processes using a process mining technique: Outpatient process analysis at a tertiary hospital. *International Journal of Medical Informatics*, 88:34–43.

[Yoo et al. 2015] Yoo, S., Cho, M., Kim, S., Kim, E., Park, S. M., Kim, K., Hwang, H., and Song, M. (2015). Conformance analysis of clinical pathway using electronic health record data. *Healthcare Informatics Research*, 21(3):161–166.

[Zaballa et al. 2020] Zaballa, O., Pérez, A., Inhiesto, E. G., Ayesta, T. A., and Lozano, J. A. (2020). Identifying common treatments from Electronic Health Records with missing information. An application to breast cancer. *PLoS ONE*, 15(12 December).

[Zeng et al. 2009] Zeng, Z., Tung, A. K., Wang, J., Feng, J., and Zhou, L. (2009). Comparing stars: On approximating graph edit distance. *Proceedings of the VLDB Endowment*, 2(1):25–36.

[Zhang et al. 2018] Zhang, X., Wang, L., Miao, S., Xu, H., Yin, Y., Zhu, Y., Dai, Z., Shan, T., Jing, S., Wang, J., Zhang, X., Huang, Z., Wang, Z., Guo, J., and Liu, Y. (2018). Analysis of treatment pathways for three chronic diseases using OMOP CDM. *Journal of Medical Systems*, 42(12).

[Zhang and Padman 2015] Zhang, Y. and Padman, R. (2015). Innovations in Chronic Care Delivery Using Data-Driven Clinical Pathways. *The American journal of managed care*, 21(12):661–668.

[Zhang et al. 2015a] Zhang, Y., Padman, R., and Patel, N. (2015a). Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data. *Journal of Biomedical Informatics*, 58:186–197.

[Zhang et al. 2015b] Zhang, Y., Padman, R., Wasserman, L., Patel, N., Teredesai, P., and Xie, Q. (2015b). On clinical pathway discovery from electronic health record data. *IEEE Intelligent Systems*, 30(1):70–75.

[Zhu et al. 2016] Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., and Wang, F. (2016). Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 749–758.

