

**Universidade Federal do Maranhão Agência de Inovação,
Empreendedorismo, Pesquisa, Pós-Graduação e
Internacionalização Programa de Pós-Graduação em Educação
Física Mestrado Acadêmico**

**UTILIZAÇÃO DE MÍDIAS SOCIAIS NA VIGILÂNCIA DE
COMPORTAMENTOS RELACIONADOS À ATIVIDADE FÍSICA E
AO SEDENTARISMO: UMA REVISÃO SISTEMÁTICA.**

Sonny Állan Silva Bezerra

**São Luís
2023**

SONNY ÁLLAN SILVA BEZERRA

**UTILIZAÇÃO DE MÍDIAS SOCIAIS NA VIGILÂNCIA DE
COMPORTAMENTOS RELACIONADOS À ATIVIDADE FÍSICA E
AO SEDENTARISMO: UMA REVISÃO SISTEMÁTICA.**

Dissertação apresentada ao Programa de Pós-Graduação em Educação Física da Universidade Federal do Maranhão, para a obtenção do Título de Mestre em Educação Física.

Área de Concentração: Biodinâmica do Movimento Humano

Linha de Pesquisa: Atividade Física relacionada a Saúde Humana

Orientador: Prof. Dr. Emanuel Péricles Salvador

Coorientador: Prof. Dr. Davi Viana dos Santos

São Luís

2023

SONNY ÁLLAN SILVA BEZERRA

**UTILIZAÇÃO DE MÍDIAS SOCIAIS NA VIGILÂNCIA DE
COMPORTAMENTOS RELACIONADOS À ATIVIDADE FÍSICA E
AO SEDENTARISMO: UMA REVISÃO SISTEMÁTICA.**

Dissertação apresentada ao Programa de Pós-Graduação em Educação Física da Universidade Federal do Maranhão, para a obtenção do Título de Mestre em Educação Física.

A Banca Examinadora da defesa da Dissertação de Mestrado apresentada em sessão pública, considerou o candidato aprovado em: ____/____/____.

Prof. Dr. Emanuel Péricles Salvador (Orientador)

Universidade Federal do Maranhão, PPGEF

Prof. Dr. Davi Viana dos Santos (Coorientador)

Universidade Federal do Maranhão, PPGCC e DCCMAPI

Prof. Dr. Bruno Feres de Souza (Examinador-Externo)

Universidade Federal do Maranhão, PPGSC e PPGCC

Prof^a. Dr^a. Andréa Dias Reis (Examinadora-Interno)

Universidade Federal do Maranhão, PPGEF

Prof. Dr. Geraldo Braz Junior (Examinador-externo)
Universidade Federal do Maranhão, PPGCC e DCCMAPI

Prof. Dr. Almir Vieira Dibai Filho (Examinador suplente)
Universidade Federal do Maranhão, PPGEF

São Luís

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Silva Bezerra, Sonny Állan.

Utilização de mídias sociais na vigilância de comportamentos relacionados à atividade física e ao sedentarismo : uma revisão sistemática / Sonny Állan Silva Bezerra. - 2023.

66 p.

Coorientador(a): Davi Viana dos Santos.

Orientador(a): Emanuel Péricles Salvador.

Dissertação (Mestrado) - Programa de Pós-graduação em Educação Física/ccbs, Universidade Federal do Maranhão, São Luís, 2023.

1. Comportamento motor. 2. Infodemia. 3. Infodemiologia. 4. Mídias sociais. 5. Monitoramento epidemiológico. I. Péricles Salvador, Emanuel. II. Viana dos Santos, Davi. III. Título.

AGRADECIMENTOS

Primeiramente, agradeço a Deus pela oportunidade de poder concluir esta etapa de minha vida, com saúde mental e corporal, e por colocar em meu caminho tão nobres colegas, amigos, professores e orientadores. Agradeço aos meus pais por todo o esforço feito por eles para que eu pudesse me desenvolver intelectualmente ao ponto atual, sempre me apoiando e orientando para o caminho mais correto e justo.

Aos meus orientadores, Professor Dr. Emanuel Salvador e Prof. Dr. Davi Viana, agradeço a confiança, apoio e orientação, que, nem de longe, foi conflituosa, mas, do contrário, amplamente harmônica. Agradeço também aos meus colegas de laboratório e de produção, Michele e Denílson, que além de tudo, também são grandes amigos e contribuíram primorosamente neste trabalho, assim como em tantos outros.

Em especial, agradeço a minha esposa, Every Monteiro, também bacharel em Educação Física e pesquisadora em formação, cujo apoio diário se traduz em afeto, provocações intelectuais e nos mais puros momentos de alegria com cada passo dado por nós dois rumo ao desconhecido. Iniciamos essa aventura de fazer uma segunda graduação e avançamos juntos na pós-graduação, e que assim possamos continuar.

RESUMO

Objetivo: esta pesquisa teve como intuito principal descrever o estado da arte da utilização de dados obtidos em mídias sociais na vigilância de desfechos relacionados à atividade física e sedentarismo. **Metodologia:** trata-se de uma revisão sistemática de artigos publicados entre os anos de 2012 e 2022 e presentes nas bases de dados PubMed/MEDLINE, Embase e El Compendex, analisados por equipe composta por um pesquisador principal, dois pesquisadores auxiliares e 2 pesquisadores coordenadores com larga experiência. As buscas foram feitas entre 19 de maio de 2022 e 14 de junho de 2022 e o processo seguiu as etapas de busca sistemática, filtragem de artigos por título e resumo, seleção dos artigos após leitura completa, extração de dados e análise da qualidade metodológica. Foram incluídos na revisão artigos originais revisados por pares, de natureza observacional, onde a atividade física ou o sedentarismo fossem um dos desfechos primários ou secundários e os dados utilizados fossem obtidos em mídias sociais. **Resultados:** Um total de 6447 registros foram encontrados e 20 publicações foram incluídas na revisão. O Twitter foi a mídia social investigada em 17 artigos. A análise de conteúdo foi empregada em todos os artigos e em 13 publicações foram identificadas técnicas de análise e classificação de sentimentos. O geoprocessamento de dados foi observado em 14 estudos e a aprendizagem de máquina foi utilizada em nove estudos para coletar e analisar os dados. A maioria dos estudos apresentou qualidade moderada-bom, com média de oito pontos em escala de zero a 10. Seis estudos avaliaram a viabilidade do uso dos dados e de técnicas de tratamento ou extração desses dados, todos com resultados positivos e qualidade metodológica superior ou igual a média. **Conclusão:** a utilização de dados das mídias sociais para investigação de desfechos relacionados a atividade física se mostra viável e potencialmente útil na vigilância de comportamentos e sentimentos. Contudo, limitações sociodemográficas apresentadas podem comprometer a validade externa dos resultados e questões éticas relacionados ao uso desse tipo de dados ainda devem ser discutidas.

Palavras-chave: Comportamento motor. Mídias sociais. Monitoramento epidemiológico. Infodemia. Infodemiologia.

ABSTRACT

Objective: this research aimed to describe the state of the art of the use of data obtained in social media in the surveillance of outcomes related to physical activity and sedentary lifestyle. **Methodology:** this is a systematic review of articles published between 2012 and 2022 and present in the PubMed/MEDLINE, Embase and El Compendex databases, analyzed by a team composed of a principal researcher, 2 auxiliary researchers and 2 coordinating researchers with extensive experience. The searches were made between May 19, 2022 and June 14, 2022 and the process followed the stages of systematic search, filtering of articles by title and abstract, selection of articles after full reading, data extraction and analysis of the risk of bias. The review included original peer-reviewed articles of an observational nature, where physical activity or sedentary lifestyle was one of the primary or secondary outcomes and the data used were obtained on social media. **Results:** A total of 6447 records were found and 20 publications were included in the review. Twitter was the social media investigated in 17 articles. Content analysis was used in all articles and in 13 publications analysis techniques and sentiment classification were identified. Data geoprocessing was observed in 14 studies and machine learning was used in nine studies to collect and analyze data. Most of the studies presented moderate-good quality, with an average of eight points on a scale from zero to 10. Six studies evaluated the feasibility of using data and techniques for processing or extracting these data, all with positive results and superior or superior methodological quality. equal to average. **Conclusion:**The use of data from social media to investigate outcomes related to physical activity is feasible and potentially useful in monitoring behaviors and feelings. However, presented sociodemographic limitations may compromise the external validity of the results and ethical issues related to the use of this type of data still need to be discussed.

Keywords: Physical Activity. Social Media. Epidemiological Monitoring. Infodemic. Infodemiology.

LISTA DE FIGURAS

Figura 1. Síntese das perguntas norteadoras.....	13
Figura 2. Fluxograma do processo de revisão em acordo com as diretrizes PRISMA.	21

LISTA DE QUADROS

Quadro 1. Relação de sintaxes de buscas por base de dados	18
Quadro 2. Formulário para extração de dados	19
Quadro 3. Natureza dos estudos: características temporais e desenho de estudo	22
Quadro 4. Natureza dos dados utilizados nos estudos	23
Quadro 5. Objetivos declarados e análises de dados realizados	26

LISTA DE TABELAS

Tabela 1. Resultado da análise da qualidade metodológica dos artigos revisados, com base no checklist de Kodf-Petersen.....	29
--	----

Sumário

1. Introdução	11
2. Objetivos e hipóteses	12
2.1. Objetivo geral	13
2.2. Objetivo específico	13
2.3. Hipóteses.....	13
3. Revisão de literatura.....	14
3.1. Propostas conceituais para caracterização das pesquisas com dados obtidos na internet na área da saúde.	14
4. Metodologia.....	16
4.1. Critérios de inclusão de artigos.	18
4.2. Extração dos dados e análise da qualidade metodológica.	20
5. Resultados	20
5.1. Visão geral	20
5.2. Sobre a natureza dos dados utilizados: dados primários, secundários, fonte e tipos de dados.	22
5.3. Sobre objetivos, variáveis, tipos de análise e principais resultados dos estudos revisados.....	24
5.4. Análise metodológica.....	28
5.5. Requisitos éticos e limitações sociodemográficas dos estudos.....	29
5.6. Resultados práticos desta revisão	30
6. Discussão.....	30
6.1. Principais mídias sociais e objetivos investigados nos estudos	30
6.2. Natureza dos dados e mensuração da atividade física.....	31
6.3. Viabilidade de uso e qualidade metodológica.....	33
6.4. Limitações	34
7. Conclusão e sugestões para novas pesquisas	36
REFERÊNCIAS.....	37
APÊNDICES	41
ANEXOS	66

1. Introdução

A Organização Mundial da Saúde recomenda a prática mínima de 150 minutos de atividade física (AF) moderada por semana (WHO, 2020), o equivalente a 20 minutos de atividade simples e de baixo impacto, como a caminhada. De forma anedótica, supostamente essa seria uma meta realista, pensando no ato de caminhar. Mas o sedentarismo e a inatividade física ainda são um dos maiores problemas do século XXI (WHO, 2015), associadas com diversas doenças crônicas, mortalidade precoce, com altos custos sociais e econômicos (KRUK, 2014; PRATT et al., 2014). Assim, é possível argumentar que iniciar e manter o hábito de realizar AF regularmente seja um fenômeno complexo e que o monitoramento dos níveis populacionais de AF e fatores associados seja uma necessidade de grande importância (GUTHOLD et al., 2018).

Algumas formas já tradicionais de monitoramento dos níveis populacionais de AF são baseadas em questionários como o IPAQ ou pelo uso de acelerometria e pedometria (BAUMAN et al., 2009; SKENDER et al., 2016). Contudo, propostas alternativas de monitoramentos epidemiológicos se tornaram uma nova realidade em diversas áreas da saúde nas últimas duas décadas, incluindo o uso de big data, que Eysenbach (2002) nomeou como infodemiologia.

Dentro dessa linha de pesquisas, a aplicação das mídias sociais como fonte de dados para o rastreamento epidemiológico de diversos desfechos é comum, com destaque para vigilância de doenças infectocontagiosas (CHEW; EYSENBACH, 2010; EYSENBACH, 2009; MOWERY, 2016), incluindo a COVID-19 (SPRINGER; ZIEGER; STRZELECKI, 2021).

Uma característica dessas pesquisas é o uso de técnicas de análise de sentimentos e de algoritmos para identificação de linguagem natural como forma de tratamento dos dados obtidos, o que pode ser interessante para a análise de fenômenos complexos que envolvem interações sociais e até mesmo sentimentais (ALESSA; FAEZIPOUR, 2019; CHEW; EYSENBACH, 2010).

Dentro desse contexto, a aplicação das mídias sociais como fonte de dados no estudo de fenômenos relacionados ao estilo de vida e a prática de AF com

características epidemiológicas parece ser uma possibilidade com grande potencial (CESARE et al., 2019b; LIU; CHEN; KUO, 2019).

Esse potencial de uso e exemplos de metodologias aplicáveis para utilização desses dados no monitoramento em tempo real ou predição dos níveis de AF foram abordados por Liu e Young (2019) em uma revisão narrativa e argumentativa. No entanto, após leitura prévia de outros artigos sobre o tema, alguns desafios ficaram evidentes em relação a padronização, a viabilidade e a qualidade metodológica desses estudos, sintetizados na Figura 1, como perguntas norteadoras deste trabalho.

Desta forma, entendendo que o desenvolvimento de uma revisão sistemática seria a principal opção metodológica para sintetizar informações e avaliar a qualidade de estudos que fizeram uso de dados obtidos nas mídias sociais para vigilância de desfechos relacionados a AF, justifica-se a realização deste estudo.

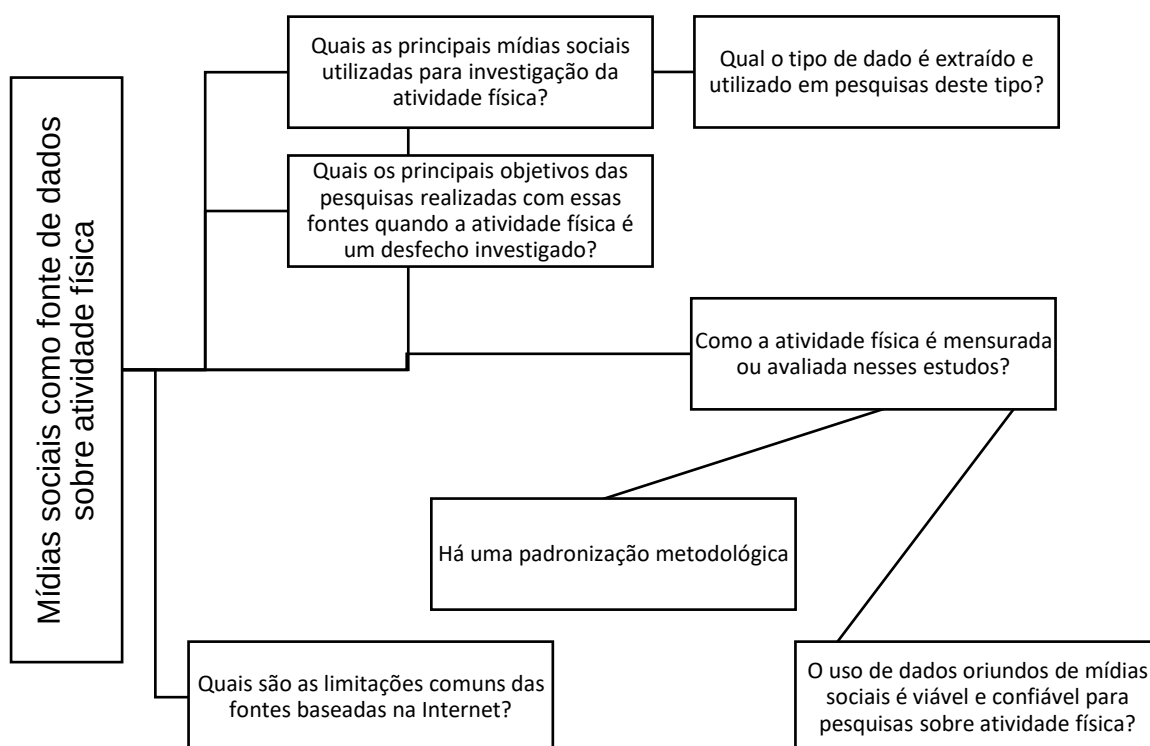


Figura 1. Síntese das perguntas norteadoras

2. Objetivos e hipóteses

2.1. Objetivo geral

Analisar o estado da arte sobre o uso de mídias sociais na vigilância de desfechos relacionados à AF e sedentarismo.

2.2. Objetivo específico

- Sintetizar informações sobre os principais métodos de coleta e análise de dados dos estudos de vigilância de desfechos relacionados à AF e ao sedentarismo que utilizaram dados obtidos nas mídias sociais digitais como fonte.
- Investigar a qualidade metodológica de estudos de vigilância (*infodemiology* e *infoveillance*) com desfechos relacionados à AF e ao sedentarismo que utilizaram dados oriundos das mídias sociais digitais como fonte.

2.3. Hipóteses

Na revisão sistemática publicada por Rebholz-Schuhmann (2020) a aplicação de dados obtidos em mídias sociais e em outras fontes baseadas na internet no rastreamento de doenças infectocontagiosas, os artigos analisados pelos autores atestaram a viabilidade do uso desses dados e uma qualidade metodológica satisfatória. Na revisão narrativa proposta por Liu e Young (2019), os autores descreveram diferentes possibilidades metodológicas já utilizadas para uso desses dados em pesquisas onde a AF era o desfecho. Em ambos os casos, o Twitter surge como principal mídia social utilizada com estes objetivos e uma falta de padronização nessas pesquisas também é evidente e mencionada pelos autores. Levando em consideração esses achados e as características em comum com a proposta do nosso estudo, as hipóteses a seguir foram formuladas:

- Hipótese 1: O uso de dados obtidos nas mídias sociais digitais aplicados para vigilância epidemiológica de comportamentos ligados à AF é viável.
- Hipótese 2: O uso de mídias sociais como fonte de dados para pesquisa epidemiológica em AF e saúde ainda carece de padronização metodológica.

- Hipótese 3: O *Twitter* é a principal fonte de dados utilizada nas pesquisas até o momento.

3. Revisão de literatura

O acesso à informação facilitado pela internet pode ser considerada uma das maiores revoluções tecnológicas do final do século 20. O uso cada vez mais massivo dessa tecnologia por usuários comuns, bem como a disponibilização de informações geradas pelo consumo digital (compras, pesquisas no Google, uso de mídias sociais e aplicativos móveis etc.) abre uma possibilidade de investigação epidemiológica sobre o comportamento dos indivíduos por meio da coleta desses dados para fins diversos (MAVRAGANI, 2020; ZERAATKAR; AHMADI, 2018).

Outras áreas do saber, como o marketing, economia e a comunicação social utilizam essa massa de dados para investigar, prever e influenciar comportamentos, realizar estudos econométricos ou identificar opiniões de eleitores, por exemplo (VOSEN; SCHMIDT, 2011).

É neste contexto que o uso dos dados digitais disponíveis em *Big Data* para pesquisas epidemiológicas emergiu como linha de pesquisa em meados dos anos 2000 e vem ganhando força, com uma ampliação no escopo dessas investigações, na variedade de métodos e de desenhos de estudo. Essa linha de pesquisa foi definida em 2002 como *infodemiology* ou infodemiologia, como já é possível identificar nas listas de descritores da Biblioteca Virtual em Saúde (BVS)¹ e que abrange também o conceito de *infoveillance*, que pode ser traduzido como infovigilância (EYSENBACH, 2002, 2006, 2009), listada como sinônimo de infodemiologia na BVS.

3.1. Propostas conceituais para caracterização das pesquisas com dados obtidos na internet na área da saúde.

¹ https://pesquisa.bvsalud.org/portal/decs-locator/?lang=pt&mode=&tree_id=L01.313.750.500

Em 2021, a Organização Mundial da Saúde (OMS) promoveu um evento online mundial, a *3rd WHO Infodemic Management Conference*, para discutir o impacto da chamada “epidemia de informações” durante a pandemia de COVID-19. Esse é um conceito relativamente simples de ser entendido, definido pela própria OMS como uma superabundância de informações digitais sobre um determinado assunto que provoca um “tsunami” capaz de confundir o público sobre o que é verdadeiro e confiável (WORLD HEALTH ORGANIZATION, 2020). Paradoxalmente, esse conceito, apesar de simples e de dialogar com os já mencionados *infodemiology* e *infoveillance*, pode levar o público a uma confusão conceitual.

Como forma de tornar a leitura posterior deste documento mais didática, faremos um apanhado geral e didático sobre esses três conceitos, iniciando pela *infodemiology*: uma vertente da epidemiologia baseada na coleta massiva de dados disponíveis em plataformas digitais e capaz de identificar e prever comportamentos humanos relacionados a saúde com as ferramentas da recente e propagada ciência de dados (EYSENBACH, 2002).

Os primeiros estudos enquadrados nessa linha utilizavam majoritariamente dados hospitalares digitalizados para identificar ou prever situações de característica epidemiológica. Posteriormente, o uso das mídias sociais como o *Twitter*, dados de localização por GPS, compartilhamento de informações auto reportadas em sites e blogs, bem como padrões de busca no Google Search ganharam destaque como fontes de informação utilizadas nesses estudos (EYSENBACH, 2009; MAVRAGANI, 2020).

Em outra pesquisa dessa primeira leva de estudos, os autores identificaram uma associação entre o relato de sintomas gripais no *Twitter* e os dados da pandemia de H1N1 em 2009, no Canadá, ilustrando o poder preditivo desses dados na identificação de surtos de doenças e seu potencial uso para identificação e controle de epidemias (CHEW; EYSENBACH, 2010), reforçando o que já havia sido defendido por um dos autores ainda em 2006 com base em achados relacionados a um surto gripal no Canadá em 2004-2005 e o padrão de buscas por sintomas no Google Search (EYSENBACH, 2006).

Nesse contexto, outro conceito abordado é o de *infoveillance*, que se refere ao uso de dados digitais de larga escala para vigilância epidemiológica (CHEW; EYSENBACH, 2010; MAVRAGANI, 2020; SPRINGER; ZIEGER; STRZELECKI, 2021). Desta forma, este trabalho de revisão pode ser entendido dentro do conceito geral de infodemiologia, abordando a infovigilância de comportamentos relacionados a AF e saúde e que fazem uso de dados obtidos a partir das mídias sociais, que se inserem no conceito de infodemia.

4. Metodologia

Este é um estudo bibliográfico caracterizado como revisão sistemática realizada de acordo com as diretrizes propostas pelo *Preferred Reporting Items for Systematic Review* - PRISMA (PAGE et al., 2021). Sendo uma revisão sistemática com foco na análise metodológica e não em intervenções, podendo ser classificada como revisão sistemática de metodologia (*methology sistematic review*) (MUNN et al., 2018), não houve a necessidade de seu registro na plataforma PROSPERO, conforme informações da própria plataforma e também não foram observados pressupostos para realização de meta-análises.

No processo de busca e filtragem dos artigos, foram considerados apenas estudos com dados originais, com dados primários, publicados entre os anos de 2012 e 2022. As bases de dados exploradas foram PubMed/MEDLINE, Embase e El Compendex, escolhidas após busca preliminar, sendo aqueles repositórios que retornaram um maior volume de documentos quando comparados a outras bases.

De acordo com o objetivo de analisar a qualidade metodológica dos artigos revisados, optou-se por priorizar manuscritos avaliados por pares e oriundos de bases bibliográficas de grande reconhecimento internacional, não utilizando a denominada literatura cinza.

A primeira rodada de buscas sistemáticas ocorreu no dia 19 de maio de 2022. Após identificação de necessidade, uma nova busca foi feita, no dia 14 de junho de 2022, adicionando um segundo conjunto de descritores, conforme apresentado no quadro 1.

No total, cinco pesquisadores participaram dos processos de planejamento, seleção preliminar de artigos, filtragem, extração de informações, cabendo ao pesquisador principal conduzir a escrita e argumentação final deste manuscrito. Todo o processo foi organizado com a ajuda da plataforma online Parsifal².

A equipe responsável foi composta pelo pesquisador principal, discente do Programa de Pós Graduação em Educação Física da Universidade Federal do Maranhão (PPGEF-UFMA); dois coordenadores, um docente dos programas de Pós Graduação em Educação Física e Saúde Coletiva da UFMA, com experiência em epidemiologia da AF e fatores associados, e outro docente do programa de Pós Graduação em Ciência da Computação da UFMA, com experiência na produção de revisões sistemáticas, desenvolvimento de softwares e computação aplicada à saúde; e dois pesquisadores auxiliares graduados em educação física, que já desenvolveram pesquisas em epidemiologia da AF e geoprocessamento de dados.

A dupla de pesquisadores auxiliares foi responsável pela busca paralela e independente dos artigos, além de auxiliarem na extração dos dados via formulário. O pesquisador principal ficou responsável por realizar a segunda camada de filtragem dos artigos segundo, com leitura completa, pautando-se nos critérios de elegibilidade e exclusão, além de comandar a extração dos dados e a análise metodológica dos artigos. Os dois coordenadores, pesquisadores mais experientes da equipe, revisaram todo o processo.

Após testarmos isoladamente diversos descritores, dois conjuntos foram utilizados para realizar a busca sistemática associados aos operadores booleanos (AND e OR):

- *(((((Physical activity) OR (Exercise)) OR (Sedentary Behaviour)) AND (Infodemiology))).*
- *(((((Physical activity) OR (sedentary behavior)) OR (exercise)) AND (social media)).*

Dependendo da base, algumas variações na construção da sintaxe ocorreram, conforme apresentado no Quadro 1.

² <https://parsif.al/>

Quadro 1. Relação de sintaxes de buscas por base de dados

Base de dados	Sintaxes de busca
El Compendex	#1 (((Physical activity) OR (Exercise)) OR (Sedentary Behaviour)) AND (Infodemiology)) #2 ((((((Physical activity) OR (sedentary behavior)) OR (exercise)) AND (social media)) WN ALL)) AND ({ja} WN DT))
EMBASE	#1 ('physical activity'/exp OR 'physical activity' OR (physical AND ('activity'/exp OR activity)) OR 'exercise'/exp OR exercise OR 'sedentary behaviour'/exp OR 'sedentary behaviour' OR (sedentary AND ('behaviour'/exp OR behaviour))) AND ('infodemiology'/exp OR infodemiology) #2 ('physical activity'/exp OR 'physical activity' OR (physical AND ('activity'/exp OR activity)) OR 'sedentary behavior'/exp OR 'sedentary behavior' OR (sedentary AND ('behavior'/exp OR behavior)) OR 'exercise'/exp OR exercise) AND ('social media'/exp OR 'social media' OR (('social'/exp OR social) AND ('media'/exp OR media)))
MEDLINE/Pubmed	#1 (((Physical activity) OR (Exercise)) OR (Sedentary Behaviour)) AND (Infodemiology)) #2 (((Physical activity) OR (sedentary behavior)) OR (exercise)) AND (social media)

4.1. Critérios de inclusão de artigos.

Os critérios de inclusão e exclusão foram definidos após duas reuniões entre os integrantes da equipe que realizou esta revisão, com base na leitura prévia de outros artigos de revisão e nos objetivos deste trabalho, sendo os critérios de inclusão:

- Artigos com dados originais e revisados por pares
- Estudos observacionais, cujo desfecho analisado estivesse relacionado à AF e seu pressuposto antagonista, o sedentarismo, e que tenham tido como fonte de dados utilizada as mídias sociais

digitais, utilizando autorrelatos, análise de sentimentos e geoprocessamento, com objetivos de vigilância epidemiológica.

Quadro 2. Formulário para extração de dados

Item	Classificações
Ano de publicação	-
Natureza do estudo em relação a temporalidade	Transversal, longitudinal
Desenho de estudo (múltipla resposta)	Análise de conteúdo, Projeção de comportamento, Viabilidade de uso (da mídia social), Vigilância de comportamento
Mídia social investigada	Blogs, Facebook, Fóruns, <i>Instagram</i> , Tik Tok, <i>Twitter</i> , Weibo
Tipo de dados coletados (das mídias sociais)	Autorrelato, Dados de wearable, Fotografias, Geolocalização
Desfechos	Comportamento sedentário, Fatores associados e comorbidades, Níveis de AF, Percepções sobre prática de atividade física e exercício físico (dos usuários da mídia social)
Disponibilidades dos dados	Privados-gratuitos (dados obtidos de empresas privadas, mas disponibilizados gratuitamente para pesquisadores) Privados-pagos (dados obtidos de empresas privadas e que necessitam de pagamento para serem acessados), Públicos-sem restrição (dados de origem estatal, com acesso livre e gratuito) , Públicos-restritos (dados disponíveis em bancos de dados públicos, de origem estatal, mas com acesso restrito)
Outras fontes/tipos de dados	Prontuários e/ou inquéritos públicos, Dados socioeconômicos.
Métodos de tratamento e análise dos dados	Análise de linguagem, Análise de sentimentos, Análise manual (de conteúdo), Correlações, Deep Learning, Machine Learning, Modelos de regressão, Projeções estatísticas (forecasting), Técnicas de classificação (de conteúdo)
Quesitos éticos	Completamente relatados, Parcialmente relatados, Não relatados
Principais resultados	
Limitações	

Já os critérios adotados para exclusão de estudos foram:

- Não utilizaram dados coletados em mídias sociais, ainda que tenham sido caracterizados como de *infodemiology*. Por exemplo, estudos que utilizaram dados do Google Trends.

- Ensaios clínicos
- Artigos de opinião, cartas ao editor.
- Outras revisões sistemáticas ou de escopo.
- Livros e capítulos de livro.
- Resumos expandidos para congressos.
- Artigos *pré-prints* e artigos publicados em revistas consideradas predatórias.

4.2. Extração dos dados e análise da qualidade metodológica.

A extração das informações dos artigos selecionados foi feita com um formulário (Quadro 2) inserido na plataforma Parsifal e utilizado pelo pesquisador principal, com participação dos dois pesquisadores auxiliares. Para análise de qualidade metodológica, adotamos um *checklist* de 10 critérios, valendo um ponto cada, desenvolvido por Kofod-Petersen (2018), já utilizado no estudo de Rebholz-Schuhmann (2020), uma revisão sistemática com metodologia similar a deste estudo, mas com objetivo de sintetizar informações sobre a viabilidade da aplicação de dados obtidos na internet na vigilância de doenças infectocontagiosas, permitindo uma comparação entre os resultados obtidos nesta etapa e uma referência metodologicamente próxima. Esse instrumento está disponibilizado integralmente no Anexo 1 deste documento. A análise da qualidade metodológica foi realizada primeiramente pelo pesquisador principal, sendo revisada por ambos os pesquisadores de maior experiência e coordenadores do projeto, de forma cega.

5. Resultados

5.1. Visão geral

Como apresentado na Figura 1, as buscas retornaram 6447 documentos, sendo a maioria dos documentos obtidos na base de dados EMBASE (60,8%), 40% dos documentos foram obtidos na Medline/Pubmed e 8% na EI Compendex. Após a primeira filtragem desses documentos, 5000 foram rejeitados e 1382 artigos duplicados foram encontrados. Para segunda fase de filtragem, 65 artigos foram

eleitos e 45 foram excluídos após a leitura aprofundada dos artigos inteiros. Em ambas as etapas, foram seguidos os critérios de inclusão e exclusão para filtragem dos documentos.

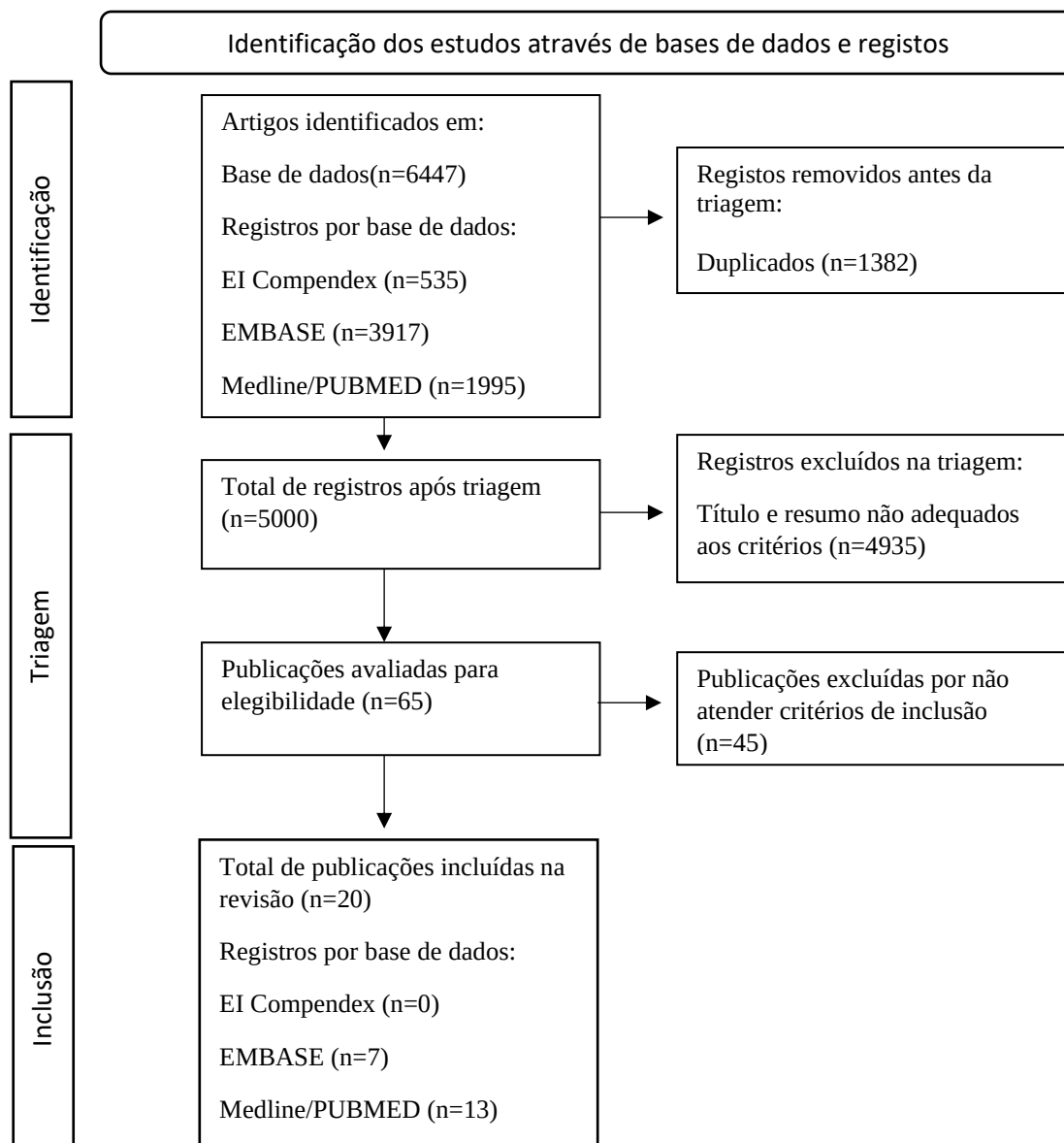


Figura 1. Fluxograma do processo de revisão em acordo com as diretrizes PRISMA.

Ao final do processo, 20 artigos foram incluídos na revisão final, tiveram seus dados extraídos e submetidos as avaliações de qualidade metodológica. Nenhum dos artigos incluídos na revisão teve registro oriundo da base EI Compendex. A base de dados com maior número de registros incluídos foi a Medline/PUBMED, com 13 publicações.

Quadro 3. Natureza dos estudos: características temporais e desenho de estudo.

Artigo (referência)	Desenho (recorte temporal)	Desenho (tipo de estudo)
(SHAKERI HOSSEIN ABAD et al., 2022)	Longitudinal	Analítico, Descritivo - Exploratório
(LIU; CHEN; KUO, 2019)	Transversal	Analítico, Descritivo - Exploratório
(ZHANG et al., 2013)	Transversal	Descritivo - Exploratório
(VICKEY; BRESLIN, 2015)	Transversal	Analítico
(MEAH et al., 2020)	Transversal	Descritivo - Exploratório
(LIU et al., 2021)	Longitudinal	Analítico
(NGUYEN et al., 2017a)	Transversal	Analítico, Descritivo - Exploratório, Preditivo.
(PAUL; DREDZE, 2014)	Transversal	Descritivo - Exploratório
(HARRIS et al., 2018)	Transversal	Analítico, Descritivo - Exploratório
(CESARE et al., 2019b)	Transversal	Analítico
(NGUYEN et al., 2016)	Transversal	Analítico
(MURALIDHARA; PAUL, 2018)	Transversal	Analítico, Descritivo - Exploratório
(BUNNEY et al., 2017)	Transversal	Analítico, Descritivo - Exploratório
(LYDECKER et al., 2021)	Transversal	Analítico, Descritivo - Exploratório
(HALLWARD; DUNCAN, 2021)	Transversal	Analítico, Descritivo - Exploratório
(OSER et al., 2019)	Transversal	Descritivo - Exploratório
(HUANG et al., 2020)	Transversal	Descritivo - Exploratório
(NGUYEN et al., 2017)	Transversal	Analítico, Descritivo - Exploratório
(GORE; DIALLO; PADILLA, 2015)	Transversal	Descritivo - Exploratório
(TURNER-MCGRIEVY; BEETS, 2015)	Transversal	Analítico, Descritivo - Exploratório

5.2. Sobre a natureza dos dados utilizados: dados primários, secundários, fonte e tipos de dados.

A maioria dos estudos utilizou exclusivamente o *Twitter* como mídia social para extração de dados (n=16), dois utilizaram exclusivamente o *Instagram* e outros dois utilizaram duas ou mais mídias sociais. Quanto ao tipo de dado utilizado, sete estudos utilizaram apenas autorrelatos (tweets ou legendas) em suas análises, dez estudos utilizaram autorrelatos e geolocalização juntos, dois utilizaram fotografias

e autorrelatos. Apenas um estudo combinou autorrelatos com dados de *wearables*, que são equipamentos vestíveis capazes de coletar e até interpretar dados, como pulseiras e relógios inteligentes.

Além da utilização de dados oriundos das mídias sociais, a maioria dos artigos também utilizou outras fontes para caracterização da amostra ou compor os modelos de análise estatística (n=14). Entre estes, quatro artigos apresentavam dados de saúde coletados por meio de inquéritos ou prontuários públicos e três utilizaram dados socioeconômicos, enquanto outros sete estudos mesclaram o uso de dados socioeconômicos e de saúde obtidos em bancos públicos de dados.

Quanto a disponibilidade dos dados, todos os estudos foram classificados como privados-gratuitos, por pertencerem a grupos empresariais privados e estarem disponíveis gratuitamente para pesquisadores. Os artigos que fizeram uso de dados obtidos em bancos públicos foram todos classificados como públicos-sem restrição (n=12), não sendo descritas restrições de acesso para pesquisadores.

Quadro 4. Natureza dos dados utilizados nos estudos.

Artigo (referência)	Origem – mídia social utilizada	Tipo de dados minerados	Outras fontes de dados	Disponibilidade dos dados
(SHAKERI HOSSEIN ABAD et al., 2022)	<i>Twitter</i>	Autorrelatos, Geolocalização	Socioeconômicos	Privados-gratuitos, Públicos-sem restrição
(LIU; CHEN; KUO, 2019)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários/ inquéritos públicos, Socioeconômicos	Privados-gratuitos, Públicos-sem restrição
(ZHANG et al., 2013)	<i>Twitter</i>	Autorrelatos	Socioeconômicos	Privados-gratuitos
(VICKEY; BRESLIN, 2015)	<i>Twitter</i>	Autorrelatos, Dados de wearables	Dados de prontuários ou inquéritos públicos	Privados-gratuitos, Públicos-sem restrição
(MEAH et al., 2020)	<i>Twitter</i>	Autorrelatos		Privados-gratuitos

Quadro 4. Continuação: natureza dos dados utilizados nos estudos.

Artigo (referência)	Origem – mídia social utilizada	Tipo de dados minerados	Outras fontes de dados	Disponibilidade dos dados
---------------------	---------------------------------	-------------------------	------------------------	---------------------------

(LIU et al., 2021)	<i>Instagram</i>	Autorrelatos, Fotografias	Socioeconômicos	Privados-gratuitos
(OSER et al., 2019)	Facebook, <i>Twitter</i>	Autorrelatos		Públicos-gratuitos
(HUANG et al., 2020)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Socioeconômicos	Públicos-gratuitos
(NGUYEN et al., 2017)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Socioeconômicos	Privados-gratuitos
(GORE; DIALLO; PADILLA, 2015)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos	Privados-gratuitos
(TURNER-MCGRIEVY; BEETS, 2015)	<i>Twitter</i>	Autorrelatos	Socioeconômicos	Privados-gratuitos
(NGUYEN et al., 2017a)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Socioeconômicos	Privados-gratuitos
(PAUL; DREDZE, 2014)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Geolocalização	Privados-gratuitos
(HARRIS et al., 2018)	<i>Twitter</i>	Autorrelatos, Fotografias		Privados-gratuitos
(CESARE et al., 2019b)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Socioeconômicos	Privados-gratuitos, Públicos-sem restrição
(NGUYEN et al., 2016)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Socioeconômicos	Privados-gratuitos, Públicos-sem restrição
(MURALIDHARA; PAUL, 2018)	<i>Instagram</i>	Autorrelatos		Privados-gratuitos
(BUNNEY et al., 2017)	<i>Twitter</i>	Autorrelatos, Geolocalização	Dados de prontuários ou inquéritos públicos, Socioeconômicos	Privados-gratuitos
(LYDECKER et al., 2021)	<i>Twitter</i>	Autorrelatos	Dados de prontuários ou inquéritos públicos	Privados-gratuitos
(HALLWARD; DUNCAN, 2021)	Blogs/Fóruns, <i>Instagram</i> , <i>Twitter</i>	Autorrelatos		Públicos-gratuitos

5.3. Sobre objetivos, variáveis, tipos de análise e principais resultados dos estudos revisados.

Foram identificados dez estudos incluídos na fase final de revisão com a vigilância dos comportamentos associados à prática de AF como objetivo principal (Quadro 5). Foi identificada análise de conteúdo em 13 artigos. Quatro estudos investigaram a viabilidade do uso desses dados para fins de vigilância da AF.

Quanto as variáveis, nove artigos mensuraram os níveis de AF populacional e em oito deles os níveis de AF foram investigados conjuntamente com a presença de fatores associados: DCNTs, fatores de risco cardiovascular, qualidade do sono e dor lombar crônica. A outra parte dos estudos investigou apenas as percepções e sentimentos sobre AF e exercício físico, associados ou não à outras variáveis.

Todos os estudos fizeram algum tipo de análise de conteúdo e 13 estudos utilizaram análise e classificação de sentimentos. Em nove artigos foi identificado o uso de *machine learning*, ou aprendizado de máquina, um método computacional caracterizado pelo uso de algoritmos cujo objetivo principal é gerar modelos preditivos, detecção de informações e de classificação de forma autônoma com base em um banco de dados pré-estabelecido (PAIXÃO et al., 2022).

Nos estudos revisados, o *machine learning* foi utilizado para identificar, analisar e rotular (classificar) sentimentos com base nos autorrelatos feitos nas mídias sociais estudadas. Em 13 estudos a rotulagem de sentimentos foi feita manualmente pelos pesquisadores ou revisada para identificar casos em que o algoritmo não foi eficiente.

Três artigos descreveram a criação de bancos de dados (*datasets*) utilizando as mídias sociais para posterior vigilância de comportamento. Em 12 estudos foram identificadas associações estatisticamente significantes entre dados derivados das mídias sociais e dados públicos oficiais, indicadores de risco cardiovascular ou Doenças Crônicas Não Transmissíveis (DCNTs).

Dos estudos que tiveram como objetivo identificar a viabilidade de uso dos dados oriundos de mídias sociais na investigação de desfechos relacionados à AF, quatro estudos utilizaram o *Twitter* e dois investigaram dados do *Instagram*. Em todos eles, o desfecho foi positivo para viabilidade de uso.

Quadro 5. Objetivos declarados e análises de dados realizados.

Artigo (referência)	Objetivos	Tipo de análises
(SHAKERI HOSSEIN ABAD et al., 2022)	Análise de conteúdo, Vigilância de comportamento	AL, ML, Rotulagem não automatizada
(LIU; CHEN; KUO, 2019)	Análise de conteúdo, Viabilidade de uso	AL, Análise de sentimentos, ML, Modelos de regressão, Rotulagem não automatizada
(ZHANG et al., 2013)	Análise de conteúdo, Vigilância de comportamento	AL, Análise de sentimentos, ML, Modelos de regressão, Rotulagem não automatizada
(VICKEY; BRESLIN, 2015)	Análise de conteúdo, Vigilância de comportamento	AL, Análise de sentimentos, ML, Rotulagem não automatizada
(MEAH et al., 2020)	Análise de conteúdo, Vigilância de comportamento	AL, Modelos de regressão, Projeções estatísticas (forecasting), Técnicas de rotulagem
(LIU et al., 2021)	Análise de conteúdo, Viabilidade de uso	AL, Modelos de regressão
(NGUYEN et al., 2017a)	Análise de conteúdo, Projeção de comportamento, Vigilância de comportamento	AL, Análise de sentimentos, Modelos de regressão, Projeções estatísticas (forecasting), Rotulagem não automatizada
(PAUL; DREDZE, 2014)	Análise de conteúdo, Viabilidade de uso, Vigilância de comportamento	AL, Análise de sentimentos, ML, Modelos de regressão, Rotulagem não automatizada
(HARRIS et al., 2018)	Análise de conteúdo, Vigilância de comportamento	AL, Análise de sentimentos, Rotulagem não automatizada
(CESARE et al., 2019b)	Projeção de comportamento, Vigilância de comportamento	AL, Análise de sentimentos, Rotulagem não automatizada, ML, Modelos de regressão
(NGUYEN et al., 2016)	Análise de conteúdo, Viabilidade de uso	AL, Análise de sentimentos, Rotulagem não automatizada ou revisão de rotulagem automatizada, ML, Modelos de regressão
(MURALIDHARA; PAUL, 2018)	Análise de conteúdo, Viabilidade de uso	AL, Correlações
(BUNNEY et al., 2017)	Análise de conteúdo, Vigilância de comportamento	AL, Análise de sentimentos, Modelos de regressão, Rotulagem não automatizada
(LYDECKER et al., 2021)	Análise de conteúdo, Viabilidade de uso, Vigilância de comportamento	AL, Análise de sentimentos, Rotulagem não automatizada, ML, Modelos de regressão.
(HALLWARD; DUNCAN, 2021)	Análise de conteúdo, Projeção de comportamento	AL, Rotulagem não automatizada

AL: análise de linguagem. ML: *machine learning*.

A viabilidade do uso desses dados foi analisada de diversas formas, com resultados que indicaram resultado favorável, mas com metodologias diferentes. No estudo conduzido por Liu, Chen e Kuo (2019), o modelo de regressão que cruzava dados públicos sobre nível de AF e o volume de tweets no período investigado apresentou associação significativa entre a porcentagem de tweets relacionados à AF e os níveis de AF já identificados na população dos EUA

(beta=0,11; SE 0,2; $P<0,001$) e também quando a AF foi ajustada à idade (beta=0,10; SE 0,20; $P<0,001$), por condado, nível de instrução e pelo índice de Gini. Contudo, o coeficiente de determinação do modelo de regressão foi baixo ($r^2=0,11$). Os autores concluíram que o uso de dados do *Twitter* como instrumento de vigilância dos níveis de AF populacional é viável, mas sugerem que novos estudos devem refinar a metodologia utilizada.

Outros estudos investigaram a viabilidade do uso de algoritmos para identificação e projeção de diversos tópicos em saúde, incluindo a AF ou o exercício físico. Em um desses estudos, Paul e Dredze (2014) testaram o algoritmo *Ailment Topic Aspect Model* (ATAM) e encontraram correlação positiva e alta para os tópicos relacionados com gripe sazonal ($r = 0,689$), alergias ($r = 0,810$), exercício físico ($r = 0,534$) e obesidade ($r = 0,631$), concluindo que os resultados apresentados demonstram que o ATAM é viável para descobrir tópicos relevantes para vigilância em saúde, utilizando o mínimo de recursos humanos para supervisão e nenhum dado histórico para treinar o modelo.

Em outro estudo, onde os autores investigaram a viabilidade de uso de diversos algoritmos para identificação de prevalência de dor lombar crônica inespecífica e fatores associados com base no autorrelato de usuários do Twitter, nos EUA, o algoritmo *Latent Dirichlet Allocation* (LDA) superou todos os outros, com pontuação de coerência mais alta e identificando 60 tópicos, com pontuação de coerência média de 0,562. Os autores concluíram que o algoritmo LDA apresenta alta viabilidade de uso para identificação de tópicos diversos em saúde (LYDECKER et al., 2021).

Outro estudo testou a viabilidade de dados obtidos no *Instagram* para prever os níveis de AF dos usuários, com base em uma análise de mediação entre níveis auto reportados de AF e comportamento dos usuários na respectiva mídia social (publicações, curtidas, perfis seguidos), identificando uma associação significativa e influência entre conteúdos publicados e seguidos pelos usuários e os níveis de AF reportados. Desta forma, os autores consideraram que o compartilhamento e o consumo de conteúdos relacionados ao exercício físico no *Instagram* está associado com uma propensão ao exercício físico e ao aumento dos níveis de AF (LIU et al., 2021).

A conclusão encontrada por Muralidhara e Paul (2018), que também investigaram o de uso dados da plataforma *Instagram* também foi positiva em relação a viabilidade. Mais precisamente, eles investigaram a viabilidade de uso de modelagem de tópicos (*polylingual topic model*) para identificar e rotular diversos tópicos relacionados a saúde e bem-estar, sendo os temas mais prevalentes aqueles relacionados à dieta (8.293/96.426; 8,6% das postagens) e ao exercícios (7.328/96.426; 7,6% das postagens).

5.4. Análise metodológica

Os estudos avaliados nesta revisão apresentaram uma pontuação média de oito pontos, em uma escala que vai de zero a dez, no instrumento proposto por Kodof-Petersen (2018). Mais da metade dos artigos avaliados ficaram com pontuação superior a sete, indicando boa qualidade metodológica. Destes, destacamos seis estudos que pontuaram acima da média (CESARE et al., 2019b; GORE; DIALLO; PADILLA, 2015; HUANG et al., 2020; LIU; CHEN; KUO, 2019; MEAH et al., 2020; MURALIDHARA; PAUL, 2018; NGUYEN et al., 2017a, 2017b; PAUL; DREDZE, 2014; SHAKERI HOSSEIN ABAD et al., 2022).

Por outro lado, seis estudos tiveram pontuação abaixo do valor médio, com destaque para dois artigos que ficaram com as pontuações mais baixas, inferiores a cinco pontos (HALLWARD; DUNCAN, 2021; MORROW et al., 2021a). Esses estudos possuem como característica comum um desenho de nuances qualitativas, que parecem não pontuar bem dentro do instrumento utilizado, sendo uma explicação plausível para o resultado apresentado.

Dois estudos foram classificados com nota máxima. Um deles avaliou a viabilidade de um algoritmo para identificação de tópicos relacionados a saúde no *instagram*, utilizando como dados para busca as legendas das fotos e *hashtags*, o podendo ser utilizado na identificação em tempo real e rotulagem de discussões sobre saúde, incluindo AF, nessa mídia social (MURALIDHARA; PAUL, 2018). O outro estudo com nota máxima foi o de Cesare et al (2019), que fez extração e tratamento de dados utilizando o algoritmo MALLET no Twitter, para identificar autorrelatos e categorizar sentimentos, combinou com análises de busca pelo

Google Trends e dados de geolocalização para identificar prevalência de obesidade e fatores associados, entre eles a AF, nos Estados Unidos da América.

Tabela 1. Resultado da análise da qualidade metodológica dos artigos revisados, com base no *checklist de Kodf-Petersen*

Artigo	Kodof-Petersen checklist
(SHAKERI HOSSEIN ABAD et al., 2022)	9,5
(LIU; CHEN; KUO, 2019)	9
(ZHANG et al., 2013)	6
(VICKEY; BRESLIN, 2015)	6
(MEAH et al., 2020)	8,5
(LIU et al., 2021)	7,5
(NGUYEN et al., 2017a)	9,5
(PAUL; DREDZE, 2014)	9
(HARRIS et al., 2018)	6,5
(CESARE et al., 2019b)	10
(NGUYEN et al., 2016)	8
(MURALIDHARA; PAUL, 2018)	10
(BUNNEY, P. E. et al, 2017)	8
(LYDECKER et al., 2021)	7,5
(HALLWARD; DUNCAN, 2021)	4,5
(MORROW et al., 2021b)	4
(HUANG et al., 2020)	9
(NGUYEN et al., 2017b)	9
(GORE; DIALLO; PADILLA, 2015)	9,5
(TURNER-MCGRIEVY; BEETS, 2015)	6,5
Média	8

5.5. Requisitos éticos e limitações sociodemográficas dos estudos

Em metade dos artigos incluídos na fase final desta revisão não foi possível identificar considerações explícitas sobre questões éticas e em outros 4 artigos, um relato parcial foi identificado, em geral apenas informando que o estudo foi aprovado por alguma instituição universitária, mas sem outros detalhes. Em 6 documentos, uma descrição detalhada dos procedimentos éticos foi identificada.

Em todos os que utilizaram como fonte dados do *Twitter*, uma problemática demográfica foi apresentada como principal limitação. A argumentação é de que a maior parte dos usuários dessa rede social são adultos jovens, dificultando assim a representatividade da amostra. Nos estudos que utilizaram geoprocessamento, outra limitação comum foi a quantidade média de *tweets* que permitem acesso à geolocalização do usuário. Nos artigos que utilizaram como fonte de dados o *Instagram*, a principal limitação diz respeito a dificuldade em coletar os dados, pois o acesso aos dados é bem mais restrito nessa rede.

5.6. Resultados práticos desta revisão

Por fim, relatamos como resultado prático deste estudo a produção de artigo científico a ser submetido inicialmente no periódico internacional *Informatics for Health and Social Care*, da editora Taylor & Francis, cujo fator de impacto atual é 3,082. O artigo pode ser visualizado na seção de apêndices deste documento (Apêndice 2), juntamente com o relatório das reuniões de planejamento entre os pesquisadores (Apêndice 1).

6. Discussão

O objetivo central deste estudo foi revisar a aplicação de dados obtidos em mídias sociais na vigilância de desfechos relacionados à prática de AF, especificamente os níveis de AF e percepções sobre sua prática autorrelatados por usuários dessas redes, avaliando a qualidade metodológica dos estudos atuais e descrevendo seus principais pontos. Assim, buscamos responder algumas questões norteadoras, discutidas a seguir.

6.1. Principais mídias sociais e objetivos investigados nos estudos

Como demonstram os resultados, o *Twitter* é a principal fonte de dados utilizada nos estudos incluídos nesta revisão. Isso provavelmente se deve pelas características específicas dessa mídia social, descrita comumente como um micro blog onde o autorrelato sobre a rotina pessoal é frequentemente feito pelos usuários, além de possuir uma API de acesso aberto para pesquisadores, ainda que atualmente possua algumas restrições (TORRES, 2020).

Em todos os estudos revisados, a análise de conteúdo era um objetivo em si. Isso decorre da natureza das próprias mídias sociais, já que há uma clara necessidade de interpretar os relatos dos usuários dessas mídias. A vigilância de comportamentos relacionados a prática de AF foi o outro objetivo de maior prevalência entre os estudos incluídos na revisão final. Esses resultados são muito próximos dos encontrados por Barros; Duggan; Rebholz-Schuhmann (2020) quando analisaram a aplicação desse tipo de dados em vigilância epidemiológica de doenças infectocontagiosas.

Portanto, até o presente momento, as pesquisas onde a AF se encontra entre os desfechos investigados utilizando dados obtidos em mídias sociais tem como principal objetivo a identificação de sentimentos relacionados a prática de atividade física ou um autorrelato de frequência.

6.2. Natureza dos dados e mensuração da atividade física

As mídias sociais digitais são espaços abertos para o usuário expor sentimentos, opiniões diversas, rotinas, fotografias, entre outras formas de conteúdos, mas ainda assim são produtos pertencentes a grupos empresariais, ou seja, os dados disponíveis nessas plataformas são privados, ainda que gratuitos e expostos publicamente. Nos estudos incluídos nesta revisão, os tipos de dados obtidos nessas plataformas foram majoritariamente o autorrelato e a geolocalização, combinados ou não com dados públicos, além de conteúdo secundário de inquéritos previamente realizados.

Essa combinação entre dados oriundos das mídias sociais e de outras fontes são evidentes em estudos cujo objetivos foram investigar associações entre

desfechos diversos e naqueles em que a AF surge como variável secundária. Em geral, nesses estudos, os desfechos primários foram obesidade (CESARE et al., 2019b; GORE; DIALLO; PADILLA, 2015; NGUYEN et al., 2017a, 2017b; PAUL; DREDZE, 2014; TURNER-MCGRIEVEY; BEETS, 2015), estilo de vida e felicidade (BUNNEY, P. E., ZINK, A. N., HOLM, A. A., BILLINGTON, C. J., & KOTZ, 2017; NGUYEN et al., 2016) ou doenças cardiovasculares (HUANG et al., 2020), mas também um dos estudos investigou o uso das mídias sociais na vigilância de dores lombares crônicas, sendo o relato de AF uma das variáveis associadas (LYDECKER et al., 2021).

Os resultados encontrados demonstram que a mensuração da AF, apesar de variada, pode se enquadrar em dois grandes grupos: volume e frequência de publicações sobre AF ou a utilização de técnicas de análise de sentimentos e rotulagem, seja manualmente ou por *machine learning*. Esses resultados corroboram os defendidos por Liu e Young (2019).

Outra metodologia explorada com frequência foi o geoprocessamento, incluindo os estudos que tiveram como objetivo criar bancos de dados para vigilância e análise de vizinhanças específicas (BUNNEY, P. E., ZINK, A. N., HOLM, A. A., BILLINGTON, C. J., & KOTZ, 2017; HUANG et al., 2020; NGUYEN et al., 2016, 2017b). Essa metodologia representa uma evolução nas observações sistemáticas da AF e sua relação com os espaços urbanos, sendo possível graças aos avanços tecnológicos e acessibilidade dos dados do Sistema de Informação Geográfica (SIG), cada vez mais utilizado em pesquisas na área da saúde e também na epidemiologia da AF, com grande amplitude de objetivos, que vão desde o monitoramento da locomoção populacional até mudanças ambientais do espaço urbano e suas interações com a AF (LOPES et al., 2019).

A combinação entre diferentes métodos de pesquisa denota a multidisciplinaridade dessas pesquisas e um potencial evolutivo para as investigações sobre a AF e fatores associados em níveis populacionais, com a utilização de tecnologias já bastante presentes em outras áreas do conhecimento, mas ainda pouco exploradas na área da Educação Física. Novas pesquisas podem testar o uso combinado de dados obtidos em mídias sociais com o uso dos tradicionais questionários e acelerômetros.

6.3. Viabilidade de uso e qualidade metodológica

A viabilidade de uso das mídias sociais foi pesquisada confirmada pelos estudos relacionados nesta revisão com variações metodológicas entre cada um deles. Contudo, percebemos que a viabilidade diretamente analisada foi a do uso de técnicas específicas de coleta e tratamento dos dados, sejam algoritmos como o LDA e o ATAM (LYDECKER et al., 2021; PAUL; DREDZE, 2014) ou modelagem de tópicos e análise de linguagem natural (LIU et al., 2021; MURALIDHARA; PAUL, 2018) dentro do ambiente das mídias sociais. Portanto, a aplicação dos dados de mídias sociais em pesquisas infodemiológicas sobre atividade física é indiretamente viável.

Confirmamos ainda a hipótese de que não haveria uma padronização metodológica para investigações de vigilância da AF com dados obtidos em mídias sociais. Contudo, não é correto traduzir este fato como uma advertência para o uso dessas plataformas como fonte de dados para as investigações, sendo que mesmo estudos com diferentes metodologias atestaram a viabilidade do uso desses dados. Além disto, a maioria dos estudos analisados apresentou boa qualidade metodológica, incluindo os seis estudos que investigaram especificamente a viabilidade.

Vários estudos combinaram diversas fontes de dados com os autorrelatos obtidos nas mídias sociais. Dos estudos que alcançaram pontuação igual ou superior à média na análise de qualidade, somente dois descreveram uso exclusivo de autorrelatos em suas análises (MEAH et al., 2020; MURALIDHARA; PAUL, 2018). Todos os outros combinaram autorrelatos com geolocalização. Esse fato é interessante, pois demonstra que o uso da geolocalização e do geoprocessamento parecem ser importantes na construção de uma vigilância da AF por meio das mídias sociais, com alguma associação com a qualidade desses estudos. Mas essa associação pode ser apenas o resultado de um viés do instrumento adotado para avaliação da qualidade metodológica, que parece priorizar estudos quantitativos.

Levando em consideração a qualidade metodológica média dos estudos avaliados e os resultados dos estudos que investigaram viabilidade indireta das

mídias sociais como fonte de informação, ganha força a hipótese de que as mídias sociais podem ser uma importante ferramenta no monitoramento de desfechos relacionados à AF, permitindo a implementação de diversas metodologias. No entanto, algumas observações devem ser feitas sobre as limitações desses estudos, conforme será tratado no próximo tópico.

6.4. Limitações

A limitação abordada com maior frequência nos estudos incluídos nesta revisão chama a atenção por ser considerada um ponto crítico em pesquisas: a amostra coletada pode não ser representativa da população. Isso se deve ao fato de as pesquisas com dados captados em mídias sociais sofrerem uma forte influência das características sociodemográficas dos usuários dessas plataformas de comunicação. Essa limitação também foi encontrada revisão sistemática feita por Barros; Duggan e Rebholz-Schuhmann (2020), onde analisaram a aplicabilidade de dados da internet na vigilância de doenças contagiosas e não contagiosas.

Ainda que não pareça ser conclusiva, essa limitação é um ponto crucial a ser discutido, pois confronta uma das principais vantagens apontadas nessas pesquisas: o volume de dados obtidos. Desta forma, ainda que o volume de dados seja expressivo, se não é possível confirmar uma representatividade dessa amostra, o resultado fica comprometido. Além disto, outra limitação importante se refere ao geoprocessamento. Apenas percentual menor dos usuários permite o acesso aos dados de geolocalização quando fazem uso das mídias sociais e isto também compromete a representatividade desses resultados (CESARE et al., 2019b; NGUYEN et al., 2017b). Uma sugestão para futuras pesquisas é utilizar algoritmos capazes de mensurar a localização dos usuários por processamento de texto ou mesmo de imagens.

Além desta limitação técnica, há um conflito quanto aos limites éticos do uso desses dados. A maioria dos artigos não apresentou uma discussão elaborada sobre o tema, o que representa um possível conflito ético que parece estar relacionado com a natureza desses dados. Uma das questões discutíveis é o uso

de autorrelatos feitos pelos usuários dessas mídias sociais podem sem os clássicos instrumentos de consentimento em pesquisa científica.

É possível considerar que esses dados disponibilizados pelas plataformas de mídias sociais são produtos comercializáveis, mas produzidos sem nenhum tipo de remuneração pelos usuários. Portanto, o uso deles em pesquisas acadêmicas poderia ser alvo sim de questionamentos éticos. Outro debate que poderia ser posto é a responsabilidade dos pesquisadores pelos critérios de anonimato em relação aos dados, autorrelatos e localização desses usuários. Como exemplo, o *Twitter*, principal plataforma utilizada nesses estudos, informa aos usuários sobre a possibilidade de uso desses dados em pesquisas, mas ainda é possível questionar se isso é suficiente para garantir o cumprimento dos parâmetros éticos em pesquisas científicas. Nenhum dos artigos incluídos nesta revisão discutiu esses temas.

Por outro lado, esta 'revisão também apresenta limitações em diversos aspectos, ainda que tenha cumprido seu papel investigativo e as etapas de sistematização de busca e seleção das publicações revisadas. A primeira de suas limitações é a quantidade e tipo de bases de dados utilizadas, que pode ter restringido o número de registros encontrados. A opção por não utilizar literatura cinza, *pré-prints* e revistas predatórias e o fato de terem sido incluídos apenas estudos em língua inglesa caracterizam um tipo de viés de seleção. Portanto, não é possível garantir que toda a evidência atual sobre o tema tenha sido sintetizada neste estudo.

Outra limitação diz respeito a possibilidade de que percepções individuais e enviesadas possam interferir na seleção dos documentos durante a fase de elegibilidade. Contudo, entendemos que a presença dos tradicionais protocolos de busca pareada e cega dos artigos, além das etapas de revisão dos processos acompanhadas por dois pesquisadores experientes minimiza esta limitação.

A heterogeneidade metodológica dos artigos revisados impacta diretamente nos resultados desta revisão, dificultando uma comparação mais justa entre os estudos. No entanto, a análise da qualidade oferece direcionamentos sobre quais

técnicas e desenhos de estudo podem ser mais interessantes para reproduzir por futuros pesquisadores.

Por fim, ressaltamos que quatro dos cinco pesquisadores que compõem a equipe possuem experiência em pesquisas quantitativas na área da epidemiologia da atividade física e saúde, não dominando conhecimentos específicos e aprofundados na área de ciências tecnológicas e da computação, como *machine learning*, por exemplo. Por este motivo, um quinto pesquisador, com vasta experiência nas ciências tecnológicas, foi convidado para coordenar o processo, no papel de coorientador.

7. Conclusão e sugestões para novas pesquisas

As evidências coletadas e sintetizadas nesta revisão apresentaram qualidade média moderada e podem trazer informações quanti-qualitativas importantes principalmente na identificação de sentimentos dos usuários dessas mídias sociais sobre a prática de AF.

Com base nos estudos revisados, sugerimos que a utilização dos autorrelatos, a análise de conteúdo e rotulagem de sentimentos sejam combinados com o geoprocessamento, quando possível. Além disso, visualizamos a possibilidade de combinação entre dados de mídias sociais e aqueles obtidos por meio de instrumentos tradicionais de pesquisas na área da epidemiologia da AF, como questionários e acelerômetros.

Outra sugestão é sobre a temporalidade de futuros estudos. Estudos de acompanhamento, do tipo “sobrevivência”, onde o objetivo seja identificar o padrão de utilização e a permanência dos usuários nessas mídias sociais reportando ou não a prática de atividade física. Este tipo de desenho permitiria inferir sobre fatores associados à comportamento de usuários que reportam suas atividades físicas rotineiramente em mídias sociais e por quanto tempo esse fenômeno pode ocorrer.

Até o momento, o *Twitter* segue como a principal mídia social digital utilizada com esse objetivo, mas a disponibilidade dos dados que essa plataforma oferece sofreu algumas modificações ao longo do tempo, não sendo mais completamente

gratuita. Sugerimos que os pesquisadores interessados verifiquem sempre a política de uso dos dados de qualquer plataforma.

Por outro lado, a validade externa desses resultados pode ser comprometida em razão das características sociodemográficas dos usuários da mídia social utilizada como fonte de dados e a geolocalização estava desabilitada pela maioria dos usuários, na maioria dos estudos avaliados, o que limita a utilização dos dados de geoprocessamento. Essa é uma limitação severa e que deve sempre ser devidamente abordada em pesquisas desse tipo. Além disto, sugerimos a discussão epistemológica sobre a utilização desses dados do ponto de vista ético-científico.

Por fim, considerando a qualidade metodológica identificada, os resultados sobre viabilidade nos estudos analisados e as limitações observadas, consideramos que o uso das mídias sociais como fonte de dados nas pesquisas de vigilância da AF, especialmente quando cruzados com outras fontes de dados, como a geolocalização, podem oferecer um panorama em tempo real sobre os sentimentos relacionados a prática de AF em comunidades e faixas populacionais específicas: jovens adultos e localidades onde o acesso a internet seja democrático.

REFERÊNCIAS

ALESSA, A.; FAEZIPOUR, M. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. **Journal of Medical Internet Research**, v. 21, n. 6, 2019.

BARROS, J. M.; DUGGAN, J.; REBHOLZ-SCHUHMANN, D. **The application of internet-based sources for public health surveillance (infoveillance): Systematic review** *Journal of Medical Internet Research*, 2020.

BAUMAN, A. et al. The international prevalence study on physical activity: Results from 20 countries. **International Journal of Behavioral Nutrition and Physical Activity**, v. 6, p. 1–11, 2009.

BUNNEY, P. E., ZINK, A. N., HOLM, A. A., BILLINGTON, C. J., & KOTZ, C. M. Leveraging geotagged Twitter data to examine neighborhood happiness. **Physiology & behavior**, v. 176, n. 801, p. 139–148, 2017.

CESARE, N. et al. Social media captures demographic and regional physical activity. **BMJ Open Sport and Exercise Medicine**, v. 5, n. 1, 2019a.

CESARE, N. et al. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. **Palgrave Communications**, v. 5, n. 1, p. 106, 17 dez. 2019b.

CHEW, C.; EYSENBACH, G. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. **PLoS ONE**, v. 5, n. 11, p. 1–13, 2010.

EYSENBACH, G. Infodemiology: the epidemiology of (mis)information. **The American Journal of Medicine**, v. 113, n. 9, p. 763–765, dez. 2002.

EYSENBACH, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 244–248, 2006.

EYSENBACH, G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. **Journal of medical Internet research**, v. 11, n. 1, 2009.

GORE, R. J.; DIALLO, S.; PADILLA, J. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. **PLOS ONE**, v. 10, n. 9, p. e0133505, 2 set. 2015.

GUTHOLD, R. et al. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. **The Lancet Global Health**, v. 6, n. 10, p. e1077–e1086, 2018.

HALLWARD, L.; DUNCAN, L. R. "Compulsive exercise is a socially acceptable prison cell": Exploring experiences with compulsive exercise across social media. **International Journal of Eating Disorders**, v. 54, n. 9, p. 1663–1671, 13 set. 2021.

HARRIS, J. K. et al. Messengers and Messages for Tweets That Used #thinspo and #fitspo Hashtags in 2016. **Preventing Chronic Disease**, v. 15, p. 170309, 4 jan. 2018.

HUANG, D. et al. Twitter-Derived Social Neighborhood Characteristics and Individual-Level Cardiometabolic Outcomes: Cross-Sectional Study in a Nationally Representative Sample. **Scientific reports**, v. 6, n. 1, p. e17969, 2020.

KOFOD-PETERSEN, A. How to do a structured literature review in computer science. **Researchgate**, n. May 2015, p. 1–7, 2018.

KRUK, J. Health and Economic Costs of Physical Inactivity. **Asian Pacific Journal of Cancer Prevention**, v. 15, n. 18, p. 7499–7503, 11 out. 2014.

LIU, S. et al. The Feasibility of Using Instagram Data to Predict Exercise Identity and Physical Activity Levels: Cross-sectional Observational Study. **Journal of Medical Internet Research**, v. 23, n. 4, p. e20954, 19 abr. 2021.

LIU, S.; CHEN, B.; KUO, A. Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study. **Journal of Medical Internet Research**, v. 21, n. 6, p. e12394, 3 jun. 2019.

LIU, S.; YOUNG, S. D. A survey of social media data analysis for physical activity

surveillance. **J Forensic Leg Med.**, p. 33–36, 2019.

LOPES, A. A. DOS S. et al. O Sistema de Informação Geográfica em pesquisas sobre ambiente, atividade física e saúde. **Revista Brasileira de Atividade Física & Saúde**, v. 23, p. 1–11, 7 ago. 2019.

LYDECKER, J. A. et al. Investigating Individuals' Perceptions Regarding the Context around the Low Back Pain Experience: Topic Modeling Analysis of Twitter Data. **Journal of Medical Internet Research**, v. 5, n. 2, p. 564–568, 2021.

MAVRAGANI, A. Infodemiology and infoveillance: Scoping review. **Journal of Medical Internet Research**, v. 22, n. 4, 2020.

MEAH, V. L. et al. Knowledge translation and social media: Twitter data analysis of the 2019 Canadian Guideline for Physical Activity throughout Pregnancy. **Canadian Journal of Public Health**, v. 111, n. 6, p. 1049–1056, 9 dez. 2020.

MORROW, D. et al. Using social media to understand adults' experiences of physical activity as a child living with type 1 diabetes. **Practical Diabetes**, v. 38, n. 3, p. 8–13, 2021a.

MORROW, D. et al. Using social media to understand adults' experiences of physical activity as a child living with type 1 diabetes. **Practical Diabetes**, v. 38, n. 3, p. 8–13, 2 maio 2021b.

MOWERY, J. Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns. **Online Journal of Public Health Informatics**, v. 8, n. 3, p. 1–19, 28 dez. 2016.

MUNN, Z. et al. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. **BMC Medical Research Methodology**, v. 18, n. 1, p. 5, 10 dez. 2018.

MURALIDHARA, S.; PAUL, M. J. #Healthy Selfies: Exploration of Health Topics on Instagram. **JMIR Public Health and Surveillance**, v. 4, n. 2, p. e10150, 29 jun. 2018.

NGUYEN, Q. C. et al. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. **JMIR Public Health and Surveillance**, v. 2, n. 2, p. e158, 17 out. 2016.

NGUYEN, Q. C. et al. Geotagged US tweets as predictors of county-level health outcomes, 2015-2016. **American Journal of Public Health**, v. 107, n. 11, p. 1776–1782, 2017a.

NGUYEN, Q. C. et al. Twitter-derived neighborhood characteristics associated with obesity and diabetes. **Scientific Reports**, v. 7, n. 1, p. 16425, 27 dez. 2017b.

OSER, T. K. et al. Using Social Media to Broaden Understanding of the Barriers and Facilitators to Exercise in Adults With Type 1 Diabetes. **Journal of Diabetes Science and Technology**, v. 13, n. 3, p. 457–465, 12 maio 2019.

PAGE, M. J. et al. The PRISMA 2020 statement : an updated guideline for reporting systematic reviews. p. 1–11, 2021.

PAIXÃO, G. M. DE M. et al. Machine Learning na Medicina: Revisão e

Aplicabilidade. **Arquivos Brasileiros de Cardiologia**, v. 118, n. 1, p. 95–102, 10 jan. 2022.

PAUL, M. J.; DREDZE, M. Discovering health topics in social media using topic models. **PLoS ONE**, v. 9, n. 8, 2014.

PRATT, M. et al. The cost of physical inactivity: moving into the 21st century: Table 1. **British Journal of Sports Medicine**, v. 48, n. 3, p. 171–173, fev. 2014.

SHAKERI HOSSEIN ABAD, Z. et al. **Physical activity, sedentary behavior, and sleep on Twitter_Multicountry and fully labeled public data set for digitalpublic health surveillance research.pdf**JMIR Public Health Surveill, , 2022. Disponível em: <<https://publichealth.jmir.org/2022/2/e32355>>

SKENDER, S. et al. Accelerometry and physical activity questionnaires - a systematic review. **BMC Public Health**, v. 16, n. 1, p. 515, 16 dez. 2016.

SPRINGER, S.; ZIEGER, M.; STRZELECKI, A. The rise of infodemiology and infoveillance during COVID-19 crisis. **One Health**, v. 13, n. July, p. 100288, 2021.

TORRES, A. **Um novo passo para o futuro da pesquisa acadêmica com a API do Twitter**. Disponível em: <https://blog.twitter.com/pt_br/topics/product/2019/-um-novo-passo-para-o-futuro-da-pesquisa-academica-com-a-api-do->.

TURNER-MCGRIEVEY, G. M.; BEETS, M. W. Tweet for health: using an online social network to examine temporal trends in weight loss-related posts. **Translational Behavioral Medicine**, v. 5, n. 2, p. 160–166, 29 jun. 2015.

VICKEY, T.; BRESLIN, J. Do As I Tweet, Not As I Do: Comparing Physical Activity Data Between Fitness Tweets and Healthy People 2020. **mHealth**, v. 1, n. 9, p. 1–7, 2015.

VOSEN, S.; SCHMIDT, T. Forecasting private consumption: survey-based indicators vs. Google trends. **Journal of Forecasting**, v. 30, n. 6, p. 565–578, set. 2011.

WHO. **Physical Inactivity: A Global Public Health Problem**.

WHO. **WHO Guidelines on physical activity and sedentary behaviour**. [s.l.: s.n.].

WORLD HEALTH ORGANIZATION. **Managing the COVID-19 infodemic**. Geneva: [s.n.].

ZERAATKAR, K.; AHMADI, M. Trends of infodemiology studies: a scoping review. **Health Information and Libraries Journal**, v. 35, n. 2, p. 91–120, 2018.

ZHANG, N. et al. Electronic word of mouth on twitter about physical activity in the United States: Exploratory infodemiology study. **Journal of Medical Internet Research**, v. 15, n. 11, p. 1–15, 2013.

APÊNDICES

Apêndice 1. Relatório de reuniões

Foram realizadas quatro reuniões de planejamento anteriores ao início das buscas. Uma reunião presencial foi feita entre o pesquisador principal e os dois pesquisadores auxiliares responsáveis pelas buscas nos repositórios de artigos para um dia antes da primeira busca. Os dois pesquisadores auxiliares catalogaram os documentos encontrados, enviando cada pesquisador uma planilha com as informações referentes ao título, resumo, data de publicação e indicação sobre aceite ou não. Ambas as planilhas foram repassadas ao coordenador geral para que fossem identificadas e sanadas as divergências, sendo encontradas quatro.

Após a identificação da necessidade de uma segunda busca, com um novo conjunto de descritores capaz de retornar artigos com as características procuradas, mas sem a definição de *infodemiology* em sua descrição, foi feita uma nova rodada de buscas, com o mesmo processo de filtragem realizado anteriormente. não havendo divergências entre os documentos selecionados.

Na segunda fase de filtragem, coube ao pesquisador principal realizar a leitura e avaliação dos artigos, não havendo necessidade de reuniões. Na fase seguinte, de extração dos dados, uma nova reunião foi realizada entre o pesquisador principal e os dois pesquisadores auxiliares. O formulário de extração estava disponível na plataforma Parsifal e foi utilizada pelos três pesquisadores. Por fim, na fase de análise metodológica, ainda em fase de revisão pelos dois pesquisadores de maior experiência, não há previsão de reunião.

Apêndice 2. Artigo derivado da dissertação em formatação condizente para submissão no periódico

Uso de mídias sociais na vigilância de comportamentos relacionados à atividade física: uma revisão sistemática metodológica.

Use of social media in surveillance of physical activity-related behaviors: a systematic methodological review.

Sonny Állan Silva Bezerra^{a*}; Davi Viana^b, Michele Maria Oliveira^a; Denilson dos Santos e Emanuel Péricles Salvador^c.

^aPrograma de Pós-graduação em Educação Física, Universidade Federal do Maranhão, São Luís, MA, Brasil; ^bPrograma de Pós-graduação em Ciência da Computação, Universidade Federal do Maranhão/ Universidade Federal do Piauí, São Luís, Maranhão; Departamento de Educação Física, Universidade Federal do Maranhão, São Luís, MA, Brasil.

* Autor correspondente: sonny.bezerra@discente.ufma.br

S.A.S Bezerra possui titulação de bacharel em Educação Física e em Comunicação Social e mestrado em Educação Física, sendo o autor principal deste manuscrito, participando dos processos de concepção, planejamento, execução e escrita. Davi Viana é docente do Programa de Pós Graduação em Ciência da Computação (UFMA/UFPI), foi coorientador do autor principal durante o mestrado e participou ativamente de todo o processo de concepção, planejamento e revisão do manuscrito. Michele M. Oliveira e Denilson Santos são discentes do Programa de Pós-graduação em Educação Física da Universidade Federal do Maranhão e participaram do planejamento e da execução deste trabalho. Emanuel Péricles Salvador é docente do Programa de Pós-graduação em Educação Física e em Saúde Coletiva, ambos da Universidade Federal do Maranhão, sendo o orientador e organizador principal deste trabalho.

Uso de mídias sociais na vigilância de comportamentos relacionados à atividade física: uma revisão sistemática metodológica.

Use of social media in surveillance of physical activity-related behaviors: a systematic methodological review.

RESUMO

Esta é uma revisão sistemática de artigos publicados entre os anos de 2012 e 2022, extraídos das bases de dados PubMed/MEDLINE, EMBASE e El Compendex, analisados por equipe composta por um pesquisador principal, dois pesquisadores auxiliares e 2 pesquisadores coordenadores com larga experiência. O estudo teve como objetivo principal descrever o estado da arte da utilização de dados obtidos em mídias sociais na vigilância de desfechos relacionados à atividade física e sedentarismo. As buscas foram feitas entre 19 de maio de 2022 e 14 de junho de 2022. Foram incluídos na revisão artigos originais revisados por pares, de natureza observacional, onde a atividade física ou o sedentarismo fossem um dos desfechos primários ou secundários e os dados utilizados fossem obtidos em mídias sociais. Um total de 6447 registros foram encontrados e 20 publicações foram incluídas na revisão. O Twitter foi a mídia social investigada em 17 artigos. A análise de conteúdo foi empregada em todos os artigos e em 13 publicações foram identificadas técnicas de análise e classificação de sentimentos. O geoprocessamento de dados foi observado em 14 estudos e a aprendizagem de máquina foi utilizada em nove estudos para coletar e analisar os dados. A maioria dos estudos apresentou qualidade moderada-bom, com média de oito pontos em escala de zero a 10. Seis estudos avaliaram a viabilidade do uso dos dados e de técnicas de tratamento ou extração desses dados, todos com resultados positivos e qualidade metodológica superior ou igual a média. A utilização de dados das mídias sociais para investigação de desfechos relacionados a atividade física se mostra viável e potencialmente útil na vigilância de comportamentos e sentimentos. Contudo, limitações sociodemográficas apresentadas podem comprometer a validade externa dos resultados e questões éticas relacionados ao uso desse tipo de dados ainda devem ser discutidas.

Palavras-chave: Atividade física. Mídias sociais. Monitoramento epidemiológico. Infodemia. Infodemiologia

ABSTRACT

This is a systematic review of articles published between 2012 and 2022, extracted from the PubMed/MEDLINE, EMBASE and El Compendex databases, analyzed by a team composed of a main researcher, two auxiliary researchers and 2 coordinating researchers with extensive experience. The main objective of the study was to describe the state of the art in the use of data obtained from social media in the surveillance of outcomes related to physical activity and sedentary lifestyle. The searches were carried out between May 19, 2022 and June 14, 2022. Original peer-reviewed, observational articles were included in the review, where physical activity or sedentary lifestyle were one of the primary or secondary outcomes and the data used were obtained from social media. A total of 6447 records were found and 20 publications were included in the review. Twitter was the social media investigated in 17 articles. Content analysis was used in all articles and in 13 publications analysis techniques and sentiment classification were identified. Data geoprocessing was observed in 14 studies and machine learning was used in nine studies to collect and analyze data. Most of the studies presented moderate-good quality, with an average of eight points on a scale from zero to 10. Six studies evaluated the feasibility of using data and techniques for processing or extracting these data, all with positive results and superior or superior methodological quality. equal to average. The use of data from social media to investigate outcomes related to physical activity is feasible and potentially useful in monitoring behaviors and feelings. However, presented sociodemographic limitations may compromise the external validity of the results and ethical issues related to the use of this type of data still need to be discussed.

Keywords: Physical Activity. Social Media. Epidemiological Monitoring. Infodemic. Infodemiology.

Introdução

A Organização Mundial da Saúde recomenda a prática mínima de 150 minutos de atividade física moderada por semana (WHO, 2020), o equivalente a pouco mais de 20 minutos diários de uma simples caminhada mais rápida. Aparentemente, essa é uma meta realista. Mas o sedentarismo e a inatividade física ainda são um dos maiores problemas do século XXI (WHO, 2015), associadas com diversas doenças crônicas e com mortalidade precoce, com altos custos sociais e econômicos (KRUK, 2014; PRATT et al., 2014). Assim, é possível argumentar que iniciar e manter o hábito de realizar atividade física regularmente seja um fenômeno complexo e que o monitoramento dos níveis populacionais de atividade física e fatores associados seja uma necessidade de grande importância (GUTHOLD et al., 2018).

Algumas formas já tradicionais de monitoramento dos níveis populacionais de atividade física são baseadas em questionários como o IPAQ ou pelo uso de acelerometria e pedometria (BAUMAN et al., 2009; SKENDER et al., 2016). Contudo, propostas alternativas de monitoramentos epidemiológicos se tornaram uma nova realidade em diversas áreas da saúde nas últimas duas décadas, incluindo o uso de *big data*, que Eysenbach (2002) nomeou como infodemiologia.

Dentro dessa linha de pesquisas, destacamos a aplicação das mídias sociais como fonte valiosa de dados para o rastreamento epidemiológico de diversos desfechos destacando-se as doenças infecciosas (Chew & Eysenbach, 2010; Eysenbach, 2009; Mavragani, 2020; Mowery, 2016), ganhando força durante a recente pandemia de COVID-19 (SPRINGER; ZIEGER; STRZELECKI, 2021).

Uma característica dessas pesquisas é o uso de técnicas de análise de sentimentos e de algoritmos para identificação de linguagem natural como forma de tratamento dos dados obtidos, o que pode ser interessante para a análise de fenômenos complexos que envolvem

interações sociais e até mesmo sentimentais (ALESSA; FAEZIPOUR, 2019; CHEW; EYSENBACH, 2010).

Dentro desse contexto, a aplicação das mídias sociais como fonte de dados no estudo de fenômenos relacionados ao estilo de vida e a prática de atividade física com características epidemiológicas parece ser uma possibilidade com grande potencial (CESARE et al., 2019a; LIU; CHEN; KUO, 2019). No entanto, após investigação previa de diversos trabalhos, algumas dúvidas relacionadas a viabilidade, a qualidade e a padronização metodológica persistiram, justificando a realização desta pesquisa de revisão bibliográfica sistemática.

Perguntas norteadoras

- Quais as principais mídias sociais utilizadas para investigação da atividade física?
- Quais os principais objetivos das pesquisas realizadas com essas fontes quando a atividade física é um desfecho investigado?
- Qual o tipo de dado é extraído e utilizado em pesquisas deste tipo?
- Como a atividade física é mensurada ou avaliada nesses estudos?
- Quais são as limitações comuns das fontes baseadas na Internet?
- Há uma padronização metodológica para pesquisas que usam as mídias sociais como instrumento para monitoramento e vigilância de desfechos relacionados à atividade física e ao sedentarismo?
- O uso de dados oriundos de mídias sociais é viável e confiável para

pesquisas sobre atividade física?

Objetivos

O objetivo principal desta revisão é analisar o estado da arte sobre o uso de mídias sociais na vigilância de desfechos relacionados à atividade física e sedentarismo. Como objetivos secundários ou específicos, Descrever os principais métodos de coleta e análise de dados dos estudos de vigilância de desfechos relacionados à atividade física e ao sedentarismo que utilizaram dados obtidos nas mídias sociais digitais como fonte. Avaliar a qualidade metodológica de estudos de vigilância (infodemiology e infoveillance) com desfechos relacionados à atividade física e ao sedentarismo que utilizaram dados oriundos das mídias sociais digitais como fonte.

Materiais e métodos

Este manuscrito é resultado de um estudo bibliográfico, com foco na descrição e na análise da qualidade metodológica dos artigos revisados e não nos efeitos de intervenções descritas na literature. Portanto, este estudo se enquadra na definição específica de revisão sistemática de metodologias (MUNN et al., 2018), não correspondendo aos parametros exigidos para seu registro na plataforma PROSPERO, conforme informações da própria plataforma.

Foram revisados apenas estudos com dados originais, com dados primários e publicados entre os anos de 2012 e 2022 nas bases de dados PubMed/MEDLINE, Embase e El Compendex, escolhidas após busca preliminar de reconhecimento, onde foi identificado que essas três bases concentravam as produções sobre o tema.

As buscas foram feitas em dois momentos, uma dia 19 de maio de 2022 e outra no 14 de junho de 2022, mantendo os mesmos critérios de inclusão e exclusão dos artigos,

mas acrescentando descritores na segunda rodada após identificação dessa necessidade pelos autores. Os descritores utilizados foram relatados no Quadro 1.

O processo de revisão com as seguintes etapas: busca sistematizada, filtragem por título e resumo, busca e exclusão de duplicatas, seleção de documentos após leitura completa, avaliação metodológica.

Equipe e funções dos pesquisadores

A equipe responsável foi composta por um pesquisador principal, discente do Programa de Pós Graduação em Educação Física da Universidade Federal do Maranhão (PPGEF-UFMA); dois coordenadores, um docente dos programas de Pós Graduação em Educação Física e Saúde Coletiva da UFMA, com experiência em epidemiologia da atividade física e fatores associados, e outro docente dos programas de Pós Graduação em Ciência da Computação da UFMA, com experiência na produção de revisões sistemáticas, desenvolvimento de softwares e computação aplicada à saúde; e dois pesquisadores auxiliares graduados em educação física com experiência em pesquisas epidemiológicas em atividade física e no uso de geoprocessamento de dados.

Os pesquisadores auxiliares foram responsáveis pela busca paralela e independente dos artigos, e auxiliarem na extração dos dados via formulário. O pesquisador principal ficou responsável por realizar a segunda camada de filtragem dos artigos com a leitura completa dos estudos pré selecionados, comandar a extração dos dados e a análise metodológica dos artigos. Os dois coordenadores revisaram todo o processo, incluindo as análises de qualidade metodológica e a escrita final dos documentos e foram responsáveis por solucionar discordâncias durante o processo de filtragem dos artigos pelos pesquisadores auxiliares.

Cr terios de inclus o e exclus o de artigos

Foram inclu dos estudos com dados originais e revisados por pares com as seguintes caracter sticas: desenhos observacionais; desfechos relacionados   atividade f sica ou comportamento sedent rio utilizando autorrelatos, an lise de sentimentos e geoprocessamento, com objetivos de vigil ncia epidemiol gica; aplica  o das m dias sociais digitais como fonte de dados.

Foram exclu dos da revis o documentos que n o utilizaram dados coletados em m dias sociais; ensaios cl nicos; artigos de opini o ou cartas ao editor; outras revis es sistem ticas ou de escopo, livros e cap tulos de livro; resumos expandidos para congressos; artigos *pr -print* e artigos publicados em revistas consideradas predat rias.

Extra  o de dados e an lise metodol gica

A extra  o das informa  es dos artigos selecionados foi feita com um formul rio inserido na plataforma Parsifal e utilizado pelo pesquisador principal, com participa  o dos dois pesquisadores auxiliares. A qualidade metodol gica dos artigos inclusos na revis o foi realizada em dois contextos. Um checklist com 10 cr terios, somando um total de 10 pontos, proposto por Kofod-petersen (2018). Identificamos a utiliza  o do mesmo instrumento na revis o sistem tica realizada por Barros et al. (2020), com caracter sticas metodol gicas similares ao nosso estudo. O instrumento foi aplicado pelo autor principal e revisado separadamente pelos dois pesquisadores de maior experi ncia.

Resultados

Vis o geral

Como mostra o fluxograma na figura 1, as buscas retornaram 6447 documentos, sendo a

maioria dos documentos obtidos na base de dados EMBASE (60,8%), 40% dos documentos foram obtidos na Medline/Pubmed e 8% na EI Compindex. Após a primeira filtragem desses documentos, 5000 foram rejeitados e 1382 artigos duplicados foram encontrados. Para segunda fase de filtragem, 65 artigos foram eleitos e 45 foram excluídos após a leitura aprofundada dos artigos inteiros. Ao final, 20 artigos foram incluídos na revisão final, tiveram seus dados extraídos e submetidos as avaliações de qualidade metodológica.

Dos 20 artigos incluídos na revisão final, apenas dois foram publicados antes de 2015 e 11 artigos publicados após 2018, sendo os dois mais recentes em 2022. Em relação à natureza dos estudos, a maioria é definida como transversal observacional e classificados nesta revisão como exploratórios-descritivos e analíticos (Quadro 3), ou seja, utilizaram um conjunto de análises estatísticas descritivas e inferenciais em suas metodologias, mas sem acompanhamento de grupos e uso de intervenções planejadas. Apenas um estudo foi classificado pelos próprios autores como longitudinal (SHAKERI HOSSEIN ABAD et al., 2022).

Origem e tipo de dados utilizados nos estudos

A maioria dos estudos utilizou exclusivamente o Twitter como mídia social para extração de dados (n=16), dois utilizaram exclusivamente o Instagram e outros dois utilizaram duas ou mais mídias sociais diferentes. Sete estudos utilizaram apenas autorrelatos (tweets ou legendas) em suas análises e outros dez estudos utilizaram autorrelatos e geolocalização juntos. Fotografias em conjunto com autorrelatos foram observadas em dois artigos e autorrelatos com dados de wearables em apenas um dos estudos.

Além da utilização de dados oriundos das mídias sociais, a maioria dos artigos

também utilizou outras fontes para caracterização da amostra ou compor os modelos de análise estatística. Quatro estudos apresentavam dados de saúde coletados por meio de inquéritos ou prontuários públicos e três utilizaram dados socioeconômicos, enquanto outros sete estudos mesclaram o uso de dados socioeconômicos e de saúde obtidos em bancos públicos de dados.

Em todos os estudos, os dados foram classificados como privados-gratuitos em relação a disponibilidade. Essa classificação foi dada por estes dados pertencerem a grupos empresariais privados, sendo disponibilizados gratuitamente para pesquisadores. Mas a gratuidade de acesso a esses dados pode ter sofrido alteração desde a publicação dos artigos revisados. Os artigos que fizeram uso de dados obtidos em bancos públicos de dados foram classificados por nós como públicos-gratuitos ou privados-gratuitos e somam um total de doze manuscritos.

Objetivos, características metodológicas e principais resultadas dos estudos revisados

Foram identificados dez estudos incluídos na fase final de revisão com a vigilância dos comportamentos associados à prática de atividade física como objetivo principal, ou seja, estudos que se enquadram no conceito de infovigilância (Quadro 5). A análise de conteúdo surge como objetivo de 13 artigos e 4 estudos investigaram a viabilidade do uso desses dados para fins de vigilância da atividade física.

Quanto as variáveis, grande parte dos artigos buscou mensurar de alguma forma os níveis de AF populacional (n=9). Em 8 estudos, os níveis de AF foram investigados conjuntamente com a presença de fatores associados como doenças crônicas não degenerativas, fatores de risco cardiovascular, qualidade do sono ou mesmo dor lombar crônica e percepções gerais sobre atividade física ou exercício físico. Outra parte dos

estudos investigou apenas as percepções sobre AF e exercício físico, associados ou não à outras variáveis (n=10).

Todos os estudos fizeram algum tipo de análise de conteúdo e 13 estudos utilizaram alguma forma de análise e classificação de sentimentos. O uso de machine learning foi identificado em nove artigos e algum tipo de análise manual para o tratamento e classificação dos dados, feita pelos próprios pesquisadores, esteve presente em 8 artigos. Ao cruzamos esses resultados, nota-se que em quatro dos nove estudos que utilizaram machine learning, a análise manual dos dados também esteve presente.

Três artigos descreveram a criação de bancos de dados (datasets) utilizando as mídias sociais para posterior análises de vigilâncias de comportamento. Nos 12 estudos em que foram realizadas análises estatísticas inferenciais e modelos de regressão, foram identificadas associações estatisticamente significantes ao cruzarem os dados derivados das mídias sociais com dados públicos oficiais sobre atividade física, indicadores de risco cardiovascular ou DCNTs. É importante considerar que em onze desses estudos, o Twitter foi a mídias sociais digital investigadas e apenas um estudo utilizou o Instagram, coletando dados de autorrelato das legendas das fotográficas publicadas por usuários.

Setes estudos que tiveram como objetivo identificar a viabilidade de uso dos dados oriundos de mídias sociais na investigação de desfechos relacionados à AF, sendo cinco desses estudos utilizando dados do Twitter e dois que utilizaram o Instagram. Nestes sete estudos os autores confirmaram a viabilidade de uso dessas duas mídias sociais como fonte de obtenção de dados sobre AF.

Análise de qualidade metodológica

Ao utilizarmos o instrumento padronizado por Kodof-Petersen (2018), os estudos

avaliados nesta revisão apresentaram uma pontuação média de 8 pontos, em uma escala que vai de 0 a 10. Mais da metade dos artigos avaliados ficaram com pontuação superior a sete, indicando boa qualidade metodológica. Destes, destacamos seis estudos que pontuaram acima da media (CESARE et al., 2019b; GORE; DIALLO; PADILLA, 2015; HUANG et al., 2020; LIU; CHEN; KUO, 2019; MEAH et al., 2020; MURALIDHARA; PAUL, 2018; NGUYEN et al., 2017a, 2017b; PAUL; DREDZE, 2014; SHAKERI HOSSEIN ABAD et al., 2022).

Por outro lado, seis estudos tiveram pontuação abaixo do valor médio, com destaque para dois artigos que ficaram com as pontuações mais baixas, inferiores a cinco pontos (HALLWARD; DUNCAN, 2021; MORROW et al., 2021b). Esses estudos possuem como característica comum um desenho de nuances qualitativas, que parecem não pontuar bem dentro do instrumento utilizado, sendo uma essa uma explicação plausível para o resultado apresentado.

Dois estudos foram classificados com nota máxima. Um deles avaliou a viabilidade de um algoritmo para identificação de tópicos relacionados a saúde no instagram, utilizando como dados para busca as legendas das fotos e hashtags, o podendo ser utilizado na identificação em tempo real e rotulagem de discussões sobre saúde, incluindo AF, nessa mídia social (Muralidharan & Paul, 2018). O outro estudo com nota máxima foi o de Cesare, Dwivedi, et al.(2019), que fez extração e tratamento de dados utilizando o algoritmo MALLET no Twitter, para identificar autorrelatos e categorizar sentimentos, combinou com análises de busca pelo Google Trends e dados de geolocalização para identificar prevalência de obesidade e fatores associados, entre eles a AF, nos Estados Unidos da América.

Discussão

O objetivo central deste estudo foi revisar a aplicação de dados obtidos em mídias sociais na vigilância de desfechos relacionados à prática de atividade física, especificamente os níveis de atividade física e percepções sobre sua prática autorrelatados por usuários dessas redes, avaliando a qualidade metodológica dos estudos atuais e descrevendo seus principais pontos. Assim, buscamos responder algumas questões norteadoras, discutidas a seguir.

Principais mídias sociais utilizadas e objetivos dos estudos revisados

Como demonstram os resultados, o Twitter é a principal fonte de dados utilizada nos estudos incluídos nesta revisão. Isso provavelmente se deve pelas características específicas dessa mídia social, descrita comumente como um micro blog onde o autorrelatado sobre a rotina pessoal é frequentemente feito pelos usuários, além de possuir uma API de acesso aberto para pesquisadores, ainda que atualmente possua algumas restrições (TORRES, 2020).

Em todos os estudos analisados, a análise de conteúdo era um objetivo em si. Isso decorre da natureza das próprias mídias sociais, já que há uma clara necessidade de interpretar os relatos dos usuários dessas mídias. A vigilância de comportamentos relacionados a prática de atividade física foi o outro objetivo de maior prevalência entre os estudos incluídos na revisão final. Esses resultados são muito próximos dos encontrados em revisão sistemática publicada recentemente no Journal of Medicine Internet Research, que teve por objetivo investigar e descrever as características metodológicas de estudos com dados obtidos na internet na saúde pública (BARROS; DUGGAN; REBHOLZ-SCHUHMANN, 2020).

Como a atividade física é abordada e mensurada nos estudos revisados

As mídias sociais digitais são espaços abertos para o usuário expor sentimentos, opiniões diversas, rotinas, fotografias, entre outras formas de conteúdos, mas ainda assim são produtos pertencentes a grupos empresariais, ou seja, os dados disponíveis nessas plataformas são privados, ainda que gratuitos e expostos publicamente. Nos estudos incluídos nesta revisão, os tipos de dados obtidos nessas plataformas foram majoritariamente o autorrelato e a geolocalização, combinados ou não com dados públicos, além de conteúdo secundário de inquéritos previamente realizados.

Essa combinação entre dados oriundos das mídias sociais e de outras fontes são evidentes em estudos cujo objetivos foram investigar associações entre desfechos diversos e naqueles em que a atividade física surge como variável secundária. Nesses estudos, os desfechos primários foram obesidade (CESARE et al., 2019b; GORE; DIALLO; PADILLA, 2015; NGUYEN et al., 2017a; PAUL; DREDZE, 2014; TURNER-MCGRIEVEY; BEETS, 2015), estilo de vida e felicidade (BUNNEY, P. E., ZINK, A. N., HOLM, A. A., BILLINGTON, C. J., & KOTZ, 2017; NGUYEN et al., 2016), doenças cardiovasculares (HUANG et al., 2020) e dores lombares crônicas (LYDECKER et al., 2021), sendo o relato de atividade física uma das variáveis secundárias associadas.

Os resultados encontrados demonstram que o tratamento dado aos dados sobre atividade física podem se enquadrar em dois grandes grupos: os estudos que avaliam o volume e frequência de publicações sobre atividade física e os estudos que utilizaram de técnicas de análise de sentimentos e rotulagem, seja manualmente ou por machine learning, para identificar percepções individuais sobre a prática de atividade física. Esses resultados são semelhantes aos identificados por Liu & Young (2019).

Outra metodologia explorada com frequência foi o geoprocessamento, incluindo os

estudos que tiveram como objetivo criar bancos de dados para vigilância e análise de vizinhanças específicas (BUNNEY, P. E., ZINK, A. N., HOLM, A. A., BILLINGTON, C. J., & KOTZ, 2017; HUANG et al., 2020; NGUYEN et al., 2016, 2017b). Essa metodologia representa uma evolução nas observações sistemáticas da atividade física e sua relação com os espaços urbanos, sendo possível graças aos avanços tecnológicos e disponibilidade dos dados do Sistema de Informação Geográfica (SIG), cada vez mais utilizado em pesquisas na área da saúde e também na epidemiologia da atividade física, com grande amplitude de objetivos, que vão desde o monitoramento da locomoção populacional até mudanças ambientais do espaço urbano e suas interações com a atividade física (LOPES et al., 2019).

A combinação entre diferentes métodos de pesquisa ajuda a caracterizar e pode ser explicada pela multidisciplinaridade dessas pesquisas. Nós consideramos essa característica um potencial evolutivo para as investigações sobre a atividade física e fatores associados em níveis populacionais. No entanto, várias são as limitações encontradas nos estudos avaliados, como abordado nos tópicos seguintes.

Viabilidade e qualidade metodológica dos estudos revisados.

A viabilidade de uso das mídias sociais foi pesquisada confirmada pelos estudos relacionados nesta revisão com variações metodológicas entre cada um deles. Contudo, podemos inferir que em todos eles, o que realmente foi testado foram técnicas específicas de coleta e tratamento dos dados, sejam algoritmos como o LDA e o ATAM (Lydecker et al., 2021; Paul & Dredze, 2014) ou modelagem de tópicos e análise de linguagem natural (Muralidharan & Paul, 2018) dentro do ambiente das mídias sociais.

Por este motivo, não é possível identificar uma padronização metodológica para investigações de vigilância da AF com dados obtidos em mídias sociais, mas também não seria correto traduzir este fato como uma advertência para o uso dessas plataformas como fonte de dados para as investigações, sendo que mesmo estudos com diferentes metodologias atestaram a viabilidade do uso desses dados. Além disto, a maioria dos estudos analisados apresentou boa qualidade metodológica, incluindo os seis estudos que investigaram especificamente a viabilidade.

Vários estudos combinaram diversas fontes de dados com os autorrelatos obtidos nas mídias sociais. Dos estudos que alcançaram pontuação igual ou superior a média na análise de qualidade, somente dois descreveram uso exclusivo de autorrelatos em suas análises (Meah et al., 2020; Muralidhara & Paul, 2018). Todos os outros combinaram autorrelatos com geolocalização. Esse fato é interessante, pois demonstra que o uso da geolocalização e do geoprocessamento parecem ser importantes na construção de uma vigilância da AF por meio das mídias sociais, com alguma associação com a qualidade desses estudos. Mas essa associação pode ser apenas o resultado de um viés do instrumento adotado para avaliação da qualidade metodológica, que parece priorizar estudos quantitativos.

Levando em consideração a qualidade metodológica média dos estudos avaliados e os resultados dos estudos que investigaram viabilidade indireta das mídias sociais como fonte de informação, ganha força a hipótese de que as mídias sociais podem ser uma importante ferramenta no monitoramento de desfechos relacionados à AF, permitindo a implementação de diversas metodologias. No entanto, algumas observações devem ser feitas sobre as limitações desses estudos, conforme será tratado no próximo tópico.

Limitações dos estudos revisados

A limitação abordada com maior frequência nos estudos incluídos nesta revisão chama a atenção por ser considerada um ponto crítico em pesquisas: a amostra coletada pode não ser representativa da população. Isso se deve ao fato de as pesquisas com dados captados em mídias sociais sofrerem uma forte influência das características sociodemográficas dos usuários dessas plataformas de comunicação. Essa limitação também foi encontrada na revisão sistemática feita por Barros et al. (2020), onde analisaram a aplicabilidade de dados da internet na vigilância de doenças contagiosas e não contagiosas.

Essa limitação é um ponto crítico na discussão da incorporação destes métodos de pesquisa e das mídias sociais como fonte confiáveis de dados sobre atividade física e saúde, pois confronta uma das principais vantagens apontadas nessas pesquisas: o volume de dados obtidos. O argumento sobre um volume de dados expressivo fica comprometido se não é possível confirmar uma representatividade dessa amostra. Além disto, outra limitação importante se refere ao geoprocessamento. Uma quantidade mínima dos usuários permite o acesso aos dados de geolocalização quando fazem uso das mídias sociais e isto também compromete a representatividade desses resultados (CESARE et al., 2019b; NGUYEN et al., 2017b).

É possível considerar que esses dados disponibilizados pelas plataformas de mídias sociais são produtos comercializáveis, mas produzidos sem nenhum tipo de remuneração pelos usuários. Portanto, o uso deles em pesquisas acadêmicas poderia ser alvo sim de questionamentos éticos. Outro debate que poderia ser posto é a responsabilidade dos pesquisadores pelos critérios de anonimato em relação aos dados, autorrelatos e localização desses usuários. Como exemplo, o Twitter, principal plataforma utilizada nesses estudos, informa aos usuários sobre a possibilidade de uso desses dados em pesquisas, mas ainda é

possível questionar se isso é suficiente para garantir o cumprimento dos parâmetros éticos em pesquisas científicas. Nenhum dos artigos incluídos nesta revisão discutiu esses temas.

Limitações desta revisão sistemática

Este estudo apresenta limitações em diversos aspectos, ainda que tenha cumprido seu papel investigativo e as etapas de sistematização de busca e seleção das publicações revisadas. A primeira de suas limitações é a quantidade e tipo de bases de dados utilizadas, que pode ter restringido o número de registros encontrados. A opção por não utilizar literatura cinza, pré-prints e revistas predatórias pode limitar também o número de documentos e de evidências durante a busca. Portanto, não é possível garantir que toda a evidência atual sobre o tema tenha sido sintetizada neste estudo.

Outra limitação corresponde ao viés de seleção, havendo a possibilidade de que percepções individuais e enviesadas possam interferir na seleção dos documentos. Contudo, entendemos que a presença dos tradicionais protocolos de busca pareada e cega dos artigos, além das etapas de revisão dos processos acompanhadas por dois pesquisadores experientes minimiza esta limitação.

A heterogeneidade metodológica dos artigos revisados impacta diretamente nos resultados desta revisão, dificultando uma comparação mais justa entre os estudos. No entanto, a análise da qualidade oferece direcionamentos sobre quais técnicas e desenhos de estudo podem ser mais interessantes para reproduzir por futuros pesquisadores.

Conclusões

As evidências coletadas e sintetizadas nesta revisão apresentaram qualidade média moderada e podem trazer informações quanti-qualitativas importantes principalmente na

identificação de sentimentos dos usuários dessas mídias sociais sobre a prática de AF.

Com base nos estudos revisados, sugerimos que a utilização dos autorrelatos, a análise de conteúdo e rotulagem de sentimentos sejam combinados com o geoprocessamento, quando possível. Além disso, visualizamos a possibilidade de combinação entre dados de mídias sociais e aqueles obtidos por meio de instrumentos tradicionais de pesquisas na área da epidemiologia da AF, como questionários e acelerômetros.

Até o momento, o Twitter segue como a principal mídia social digital utilizada com esse objetivo, mas a disponibilidade dos dados que essa plataforma oferece sofreu algumas modificações ao longo do tempo, não sendo mais completamente gratuita. Sugerimos que os pesquisadores interessados verifiquem sempre a política de uso dos dados de qualquer plataforma de AF em comunidades e faixas populacionais específicas: jovens adultos e localidades onde o acesso a internet seja democrático.

Referências

ALESSA, A.; FAEZIPOUR, M. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. **Journal of Medical Internet Research**, v. 21, n. 6, 2019.

BARROS, J. M.; DUGGAN, J.; REBHOLZ-SCHUHMANN, D. **The application of internet-based sources for public health surveillance (infoveillance): Systematic review** **Journal of Medical Internet Research**, 2020.

BAUMAN, A. et al. The international prevalence study on physical activity: Results from 20 countries. **International Journal of Behavioral Nutrition and Physical Activity**, v. 6,

p. 1–11, 2009.

BUNNEY, P. E., ZINK, A. N., HOLM, A. A., BILLINGTON, C. J., & KOTZ, C. M.
Leveraging geotagged Twitter data to examine neighborhood happiness. **Physiology & behavior**, v. 176, n. 801, p. 139–148, 2017.

CESARE, N. et al. Social media captures demographic and regional physical activity.
BMJ Open Sport and Exercise Medicine, v. 5, n. 1, 2019a.

CESARE, N. et al. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. **Palgrave Communications**, v. 5, n. 1, p. 106, 17 dez. 2019b.

CHEW, C.; EYSENBACH, G. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. **PLoS ONE**, v. 5, n. 11, p. 1–13, 2010.

EYSENBACH, G. Infodemiology: the epidemiology of (mis)information. **The American Journal of Medicine**, v. 113, n. 9, p. 763–765, dez. 2002.

EYSENBACH, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 244–248, 2006.

EYSENBACH, G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. **Journal of medical Internet research**, v. 11, n. 1, 2009.

GORE, R. J.; DIALLO, S.; PADILLA, J. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. **PLOS ONE**, v. 10, n. 9, p. e0133505, 2 set. 2015.

GUTHOLD, R. et al. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. **The Lancet Global Health**, v. 6, n. 10, p. e1077–e1086, 2018.

HALLWARD, L.; DUNCAN, L. R. “Compulsive exercise is a socially acceptable prison cell”: Exploring experiences with compulsive exercise across social media. **International Journal of Eating Disorders**, v. 54, n. 9, p. 1663–1671, 13 set. 2021.

HARRIS, J. K. et al. Messengers and Messages for Tweets That Used #thinspo and #fitspo Hashtags in 2016. **Preventing Chronic Disease**, v. 15, p. 170309, 4 jan. 2018.

HUANG, D. et al. Twitter-Derived Social Neighborhood Characteristics and Individual-Level Cardiometabolic Outcomes: Cross-Sectional Study in a Nationally Representative Sample. **Scientific reports**, v. 6, n. 1, p. e17969, 2020.

KOFOD-PETERSEN, A. How to do a structured literature review in computer science. **Researchgate**, n. May 2015, p. 1–7, 2018.

KRUK, J. Health and Economic Costs of Physical Inactivity. **Asian Pacific Journal of Cancer Prevention**, v. 15, n. 18, p. 7499–7503, 11 out. 2014.

LIU, S. et al. The Feasibility of Using Instagram Data to Predict Exercise Identity and Physical Activity Levels: Cross-sectional Observational Study. **Journal of Medical Internet Research**, v. 23, n. 4, p. e20954, 19 abr. 2021.

LIU, S.; CHEN, B.; KUO, A. Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study. **Journal of Medical Internet Research**, v. 21, n. 6, p. e12394, 3 jun. 2019.

LIU, S.; YOUNG, S. D. A survey of social media data analysis for physical activity surveillance. **J Forensic Leg Med.**, p. 33–36, 2019.

LOPES, A. A. DOS S. et al. O Sistema de Informação Geográfica em pesquisas sobre ambiente, atividade física e saúde. **Revista Brasileira de Atividade Física & Saúde**, v. 23, p. 1–11, 7 ago. 2019.

LYDECKER, J. A. et al. Investigating Individuals’ Perceptions Regarding the Context around the Low Back Pain Experience: Topic Modeling Analysis of Twitter Data. **Journal**

of **Medical Internet Research**, v. 5, n. 2, p. 564–568, 2021.

MAVRAGANI, A. Infodemiology and infoveillance: Scoping review. **Journal of Medical Internet Research**, v. 22, n. 4, 2020.

MEAH, V. L. et al. Knowledge translation and social media: Twitter data analysis of the 2019 Canadian Guideline for Physical Activity throughout Pregnancy. **Canadian Journal of Public Health**, v. 111, n. 6, p. 1049–1056, 9 dez. 2020.

MORROW, D. et al. Using social media to understand adults' experiences of physical activity as a child living with type 1 diabetes. **Practical Diabetes**, v. 38, n. 3, p. 8–13, 2021a.

MORROW, D. et al. Using social media to understand adults' experiences of physical activity as a child living with type 1 diabetes. **Practical Diabetes**, v. 38, n. 3, p. 8–13, 2 maio 2021b.

MOWERY, J. Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns. **Online Journal of Public Health Informatics**, v. 8, n. 3, p. 1–19, 28 dez. 2016.

MUNN, Z. et al. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. **BMC Medical Research Methodology**, v. 18, n. 1, p. 5, 10 dez. 2018.

MURALIDHARA, S.; PAUL, M. J. #Healthy Selfies: Exploration of Health Topics on Instagram. **JMIR Public Health and Surveillance**, v. 4, n. 2, p. e10150, 29 jun. 2018.

NGUYEN, Q. C. et al. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. **JMIR Public Health and Surveillance**, v. 2, n. 2, p. e158, 17 out. 2016.

NGUYEN, Q. C. et al. Geotagged US tweets as predictors of county-level health outcomes, 2015-2016. **American Journal of Public Health**, v. 107, n. 11, p. 1776–1782, 2017a.

NGUYEN, Q. C. et al. Twitter-derived neighborhood characteristics associated with obesity and diabetes. **Scientific Reports**, v. 7, n. 1, p. 16425, 27 dez. 2017b.

OSER, T. K. et al. Using Social Media to Broaden Understanding of the Barriers and Facilitators to Exercise in Adults With Type 1 Diabetes. **Journal of Diabetes Science and Technology**, v. 13, n. 3, p. 457–465, 12 maio 2019.

PAGE, M. J. et al. The PRISMA 2020 statement : an updated guideline for reporting systematic reviews. p. 1–11, 2021.

PAIXÃO, G. M. DE M. et al. Machine Learning na Medicina: Revisão e Aplicabilidade. **Arquivos Brasileiros de Cardiologia**, v. 118, n. 1, p. 95–102, 10 jan. 2022.

PAUL, M. J.; DREDZE, M. Discovering health topics in social media using topic models. **PLoS ONE**, v. 9, n. 8, 2014.

PRATT, M. et al. The cost of physical inactivity: moving into the 21st century: Table 1. **British Journal of Sports Medicine**, v. 48, n. 3, p. 171–173, fev. 2014.

SHAKERI HOSSEIN ABAD, Z. et al. **Physical activity, sedentary behavior, and sleep on Twitter_Multicountry and fully labeled public data set for digitalpublic health surveillance research.pdf**JMIR Public Health Surveill, , 2022. Disponível em:

<<https://publichealth.jmir.org/2022/2/e32355>>

SKENDER, S. et al. Accelerometry and physical activity questionnaires - a systematic review. **BMC Public Health**, v. 16, n. 1, p. 515, 16 dez. 2016.

SPRINGER, S.; ZIEGER, M.; STRZELECKI, A. The rise of infodemiology and infoveillance during COVID-19 crisis. **One Health**, v. 13, n. July, p. 100288, 2021.

TORRES, A. **Um novo passo para o futuro da pesquisa acadêmica com a API do Twitter**. Disponível em: <https://blog.twitter.com/pt_br/topics/product/2019/-um-novo-passo-para-o-futuro-da-pesquisa-academica-com-a-api-do->.

TURNER-MCGRIEVY, G. M.; BEETS, M. W. Tweet for health: using an online social

network to examine temporal trends in weight loss-related posts. **Translational**

Behavioral Medicine, v. 5, n. 2, p. 160–166, 29 jun. 2015.

VICKEY, T.; BRESLIN, J. Do As I Tweet, Not As I Do: Comparing Physical Activity Data Between Fitness Tweets and Healthy People 2020. **mHealth**, v. 1, n. 9, p. 1–7, 2015.

VOSEN, S.; SCHMIDT, T. Forecasting private consumption: survey-based indicators vs. Google trends. **Journal of Forecasting**, v. 30, n. 6, p. 565–578, set. 2011.

WHO. **Physical Inactivity: A Global Public Health Problem.**

WHO. **WHO Guidelines on physical activity and sedentary behaviour.** [s.l: s.n.].

WORLD HEALTH ORGANIZATION. **Managing the COVID-19 infodemic.** Geneva: [s.n.].

ZERAATKAR, K.; AHMADI, M. Trends of infodemiology studies: a scoping review.

Health Information and Libraries Journal, v. 35, n. 2, p. 91–120, 2018.

ZHANG, N. et al. Electronic word of mouth on twitter about physical activity in the United States: Exploratory infodemiology study. **Journal of Medical Internet Research**, v. 15, n. 11, p. 1–15, 2013.

ANEXOS

Anexo 1. Checklist de Kodof-Pedersen para avaliação da qualidade metodológica.

Quality Criteria	SIM (1 ponto)	Parcialmente (0,5 ponto)	Não ou não se aplica (0 ponto)
Existe uma clara afirmação do objetivo da pesquisa?			
O estudo é colocado no contexto de outros estudos e pesquisas?			
As decisões de design do sistema ou algorítmico são justificadas?			
O conjunto de dados de teste é reprodutível?			
O algoritmo de estudo é reprodutível?			
O procedimento experimental é completamente explicado e reprodutível?			
Está claramente indicado no estudo com quais outros algoritmos os algoritmos do estudo foram comparados?			
As métricas de desempenho utilizadas no estudo são explicadas e justificadas?			
Os resultados dos testes foram analisados minuciosamente?			