ÁREA DO CONHECIMENTO DE CIÊNCIAS DA VIDA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE

VINICIUS REMUS BALLOTIN

DESENVOLVIMENTO E VALIDAÇÃO DE UM FLUXO AUTOMATIZADO DE REVISÃO SISTEMÁTICA SOBRE SEMAGLUTIDA E DESFECHOS PERIOPERATÓRIOS EM COMPARAÇÃO COM UM PADRÃO DE REFERÊNCIA HUMANO

Caxias do Sul
2026

VINICIUS REMUS BALLOTIN

DESENVOLVIMENTO E VALIDAÇÃO DE UM FLUXO AUTOMATIZADO DE REVISÃO SISTEMÁTICA SOBRE SEMAGLUTIDA E DESFECHOS PERIOPERATÓRIOS EM COMPARAÇÃO COM UM PADRÃO DE REFERÊNCIA HUMANO

Tese apresentada à Universidade de Caxias do Sul, para obtenção do título de doutor em Ciências da Saúde.
Orientador: Prof. Dr. Luciano da Silva Selistre
Coorientador: Prof. Dr. Daniel Volquind

Caxias do Sul
2026

UNIVERSIDADE DE CAXIAS DO SUL PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE

COORDENADOR DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE PROF. DRA. VANDREA CARLA DE SOUZA

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
Sistema de Bibliotecas UCS - Processamento Técnico

B193d Ballotin, Vinicius Remus

Desenvolvimento e validação de um fluxo automatizado de revisão sistemática sobre semaglutida e desfechos perioperatórios em comparação com um padrão de referência humano [recurso eletrônico] / Vinicius Remus Ballotin. – 2026.

Dados eletrônicos.

Tese (Doutorado) - Universidade de Caxias do Sul, Programa de Pós-Graduação em Ciências da Saúde, 2026.

Orientação: Luciano da Silva Selistre.

Coorientação: Daniel Volquind.

Modo de acesso: World Wide Web

Disponível em: https://repositorio.ucs.br

1. Revisões sistemáticas (Pesquisa médica). 2. Semaglutida. 3. Automação. 4. Inteligência artificial. 5. Recuperação da informação. 6. Estudo de validação. I. Selistre, Luciano da Silva, orient. II. Volquind, Daniel, coorient. III. Título.

CDU 2. ed.: 61:001.8

Catalogação na fonte elaborada pela(o) bibliotecária(o)
Márcia Servi Gonçalves - CRB 10/1500

DESENVOLVIMENTO E VALIDAÇÃO DE UM FLUXO AUTOMATIZADO DE REVISÃO SISTEMÁTICA SOBRE SEMAGLUTIDA E DESFECHOS PERIOPERATÓRIOS EM COMPARAÇÃO COM UM PADRÃO DE REFERÊNCIA HUMANO

Vinicius Remus Ballotin

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciências da Saúde da Universidade de Caxias do Sul, como parte dos requisitos necessários para a obtenção do título de doutor em Ciências da Saúde. Linha de Pesquisa: Computação Aplicada às Ciências da Saúde.

Caxias do Sul, 21 de agosto de 2026.

Banca Examinadora

Prof. Dr. Leandro Corso
Universidade de Caxias do Sul

Profa. Dra. Valéria Weiss Angeli
Universidade de Caxias do Sul

Prof. Dr. Mário Wagner
Universidade Federal do Rio Grande do Sul

Prof. Dr. Luciano da Silva Selistre
Universidade de Caxias do Sul
Professor orientador

AGRADECIMENTOS

Agradeço à Universidade de Caxias do Sul e ao Programa de Pós-Graduação em Ciências da Saúde pela oportunidade de desenvolver esta pesquisa e pelo suporte oferecido ao longo da minha formação acadêmica.

Ao meu orientador, Prof. Dr. Luciano da Silva Selistre, agradeço pela confiança, disponibilidade, rigor científico e contribuições fundamentais para o desenvolvimento deste trabalho. Ao meu coorientador, Prof. Dr. Daniel Volquind, agradeço pelo apoio, pelas orientações e por compartilhar sua experiência acadêmica e profissional durante todas as etapas desta trajetória.

Aos professores integrantes da banca examinadora, agradeço pela leitura cuidadosa, pelas observações criteriosas e pelas recomendações que contribuíram para o aprimoramento desta tese.

Aos pesquisadores e colaboradores que participaram dos estudos que compõem este trabalho, agradeço pela dedicação, pelo diálogo científico e pela contribuição em cada etapa do projeto. Aos colegas do Programa de Pós-Graduação, professores e profissionais da instituição, agradeço pelas experiências compartilhadas e pelo apoio durante o doutorado.

À minha família e aos meus amigos, agradeço pela compreensão, pelo incentivo e pelo apoio constante durante os períodos de estudo, pesquisa e elaboração desta tese. A presença de vocês foi essencial para a conclusão deste percurso.

À minha namorada, agradeço pela alegria de compartilhar comigo este tempo, este caminho e esta etapa da vida.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho e para minha formação pessoal, profissional e científica.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

RESUMO

As revisões sistemáticas são ferramentas fundamentais para a síntese de evidências em saúde. No entanto, sua condução ainda exige grande investimento de tempo e trabalho manual, especialmente nas etapas de deduplicação bibliográfica, triagem de títulos e resumos e avaliação de textos completos.

Nesse contexto, este estudo teve como objetivo desenvolver e validar um fluxo automatizado para apoiar a condução de uma revisão sistemática sobre o uso de semaglutida no contexto perioperatório. As decisões de revisores treinados, orientados pelas diretrizes PRISMA, foram utilizadas como padrão de referência. Trata-se de um estudo metodológico, observacional e analítico, estruturado em três módulos sequenciais: deduplicação bibliográfica, triagem de títulos e resumos e avaliação dos artigos completos.

O desempenho do sistema foi avaliado por comparação direta com a seleção realizada pelos revisores humanos, utilizando métricas de sensibilidade, especificidade, precisão, valor preditivo negativo, escore F1, acurácia, coeficiente kappa de Cohen, índice Jaccard e redução da carga de trabalho manual. Além disso, os estudos finalmente incluídos na revisão sistemática foram acompanhados ao longo de todas as etapas do processo automatizado, a fim de verificar se sua identificação foi preservada.

O sistema apresentou desempenho elevado na deduplicação, especialmente em nível de agrupamento, e preservou os 12 estudos finalmente incluídos ao longo do fluxo automatizado. Na triagem de títulos e resumos, a concordância com a lista humana ampliada de recuperação de textos completos foi moderada, com sensibilidade de 0,703, embora nenhum estudo finalmente incluído tenha sido perdido. O módulo exploratório de texto completo apresentou sensibilidade de 1,000 na análise de segurança, com redução de 68,9% dos PDFs que exigiriam revisão humana prioritária.

Conclui-se que o SafeScreen apresenta potencial para reduzir a carga de trabalho em revisões sistemáticas, preservando os estudos finalmente incluídos e mantendo a rastreabilidade e a auditabilidade das decisões. Os resultados sustentam seu uso como ferramenta de apoio ao revisor, sob supervisão humana, sendo necessária validação externa para avaliar sua generalização.

Palavras-chave: revisão sistemática; automação; inteligência artificial; semaglutida; triagem bibliográfica; deduplicação; validação metodológica.

ABSTRACT

Systematic reviews are essential tools for synthesizing health evidence. However, their conduct still requires substantial investments of time and manual effort, particularly during bibliographic deduplication, title and abstract screening, and full-text assessment.

In this context, the present study aimed to develop and validate an automated workflow to support a systematic review of perioperative semaglutide use. Decisions made by trained reviewers, following PRISMA guidelines, were used as the reference standard. This was a methodological, observational, and analytical study structured into three sequential modules: bibliographic deduplication, title and abstract screening, and full-text assessment.

System performance was evaluated through direct comparison with the selections made by human reviewers using sensitivity, specificity, precision, negative predictive value, F1 score, accuracy, Cohen's kappa coefficient, Jaccard index, and manual workload reduction. In addition, the studies ultimately included in the systematic review were tracked throughout all stages of the automated process to determine whether their identification was preserved.

The system demonstrated strong performance in deduplication, particularly at the cluster level, and preserved all 12 studies ultimately included throughout the automated workflow. During title and abstract screening, agreement with the broader human full-text retrieval set was moderate, with a sensitivity of 0.703, although no study ultimately included was missed. The exploratory full-text module achieved a sensitivity of 1.000 in the safety analysis and reduced by 68.9% the number of PDFs requiring prioritized human review.

In conclusion, SafeScreen has the potential to reduce the workload of systematic reviews while preserving the studies ultimately included and maintaining decision traceability and auditability. The findings support its use as a reviewer-support tool under human supervision, although external validation is required to assess its generalizability.

Keywords: systematic review; automation; artificial intelligence; semaglutide; bibliographic screening; deduplication; methodological validation.

SUMÁRIO

LISTA DE ABREVIATURAS

AI – Artificial Intelligence

AMSTAR 2 – A Measurement Tool to Assess Systematic Reviews 2

API – Application Programming Interface

ASReview – Active Systematic Review

ASySD – Automated Systematic Search Deduplicator

CEP – Comitê de Ética em Pesquisa

CI – Confidence Interval

CNS – Conselho Nacional de Saúde

CSV – Comma-Separated Values

DOI – Digital Object Identifier

EC – Critério de Exclusão

EC1 – Manejo de Comorbidades Como Foco Principal

EC2 – Contraindicações à Semaglutida Como Foco Principal

EC3 – Ausência de Dados Específicos de Semaglutida

EC4 – Ausência de Dados Originais ou Desenho de Estudo Inelegível

EC5 – Publicação em Idioma Diferente do Inglês

EC6 – Dados Ausentes ou Insuficientes para Avaliação dos Desfechos

F1 – F1-score

GLP-1 – Glucagon-Like Peptide-1

GPT – Generative Pre-trained Transformer

IA – Inteligência Artificial

IC – Critério de Inclusão

IC1 – População Humana Adulta

IC2 – Semaglutida Explicitamente Estudada

IC3 – Desenho Elegível com Dados Originais e pelo Menos um Desfecho Perioperatório de Interesse

IC3A – Conteúdo Gástrico Residual ou Esvaziamento Gástrico

IC3B – Aspiração Pulmonar ou Brônquica

IC3C – Complicações Relacionadas à Via Aérea

IC3D – Suspensão ou Intervalo de Interrupção da Semaglutida Antes do Procedimento

IC3E – Complicações Perioperatórias Gerais

JSON - JavaScript Object Notation;

LLM – Large Language Model

ML – Machine Learning

PDF – Portable Document Format

PMCID – PubMed Central Identifier

PMID – PubMed Identifier

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PRISMA-S – Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search Extension

PROSPERO - International Prospective Register of Systematic Reviews

SafeScreen – Sistema Automatizado e Auditável para Deduplicação, Triagem e Priorização de Estudos em Revisões Sistemáticas

SysRev – Systematic Review Deduplicator - módulo de deduplicação do SafeScreen

TF-IDF – Term Frequency–Inverse Document Frequency

WSS – Work Saved over Sampling

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

As revisões sistemáticas permanecem o principal instrumento de síntese de evidências científicas, porém sua condução é dispendiosa em termos de tempo e recursos humanos, especialmente nas etapas de remoção de arquivos duplicados e na triagem de títulos e resumos (1-6). Em revisões com questões clínicas bem delimitadas, uma grande proporção dos registros é irrelevante ou duplicados, e o processamento manual desses registros pode consumir tempo considerável dos revisores sem acrescentar ganho proporcional à qualidade das conclusões (1, 7-9). Cresce, portanto, o interesse em estratégias de automação transparentes e reprodutíveis, capazes de reduzir a carga de trabalho sem comprometer a identificação dos estudos potencialmente elegíveis (3, 9, 10).

Grandes bases bibliográficas impõem dois desafios operacionais recorrentes. O primeiro diz respeito à heterogeneidade dos registros duplicados entre bases de dados: identificadores exatos podem estar ausentes, os dados bibliográficos podem ser incompletos, e o mesmo estudo pode aparecer simultaneamente como *preprint*, resumo de congresso e artigo publicado em periódico (11-14). O segundo desafio está na triagem de títulos e resumos, que exige a interpretação repetida dos critérios de elegibilidade, informações frequentemente incompletas nos metadados e que, em geral, só podem ser confirmadas após a leitura do texto completo (10, 15-17). As ferramentas de automação disponíveis podem acelerar ambos os processos, mas sua confiabilidade varia consideravelmente conforme a questão de pesquisa, a estrutura dos dados e a margem de erro considerada aceitável (10, 15, 16, 18-21).

O presente estudo abordou esses desafios no contexto de uma revisão focada no uso de semaglutida e nos desfechos perioperatórios. Esse tema é particularmente adequado para a validação metodológica, pois combina um domínio de intervenção específico com desfechos clínicos bem definidos, o que permite a construção de critérios de elegibilidade explícitos e regras de exclusão completamente auditáveis.

Além disso, o fluxo automatizado foi aplicado retrospectivamente ao corpus de uma revisão previamente concluída, com decisões humanas documentadas de acordo com o PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses), permitindo a comparação direta dos resultados automatizados com um padrão de referência humano, e não apenas com suposições heurísticas (4, 5, 22).

Desenvolveu-se, assim, um sistema automatizado estruturado com três componentes: deduplicação bibliográfica, triagem de títulos e resumos assistida por inteligência artificial, e um módulo exploratório de revisão de textos completos em PDF (Formato de Documento Portátil). A etapa de deduplicação foi realizada para identificar registros repetidos e preservar apenas a versão mais completa de cada referência. A etapa de triagem aplicou critérios de elegibilidade explícitos e, para os artigos que não puderam ser classificados com clareza apenas por esses critérios, utilizou um modelo estatístico conservador para decidir se deveriam ser encaminhados para avaliação do texto completo. O módulo de textos completos utilizou inteligência artificial para ler e avaliar os artigos selecionados com base em critérios predefinidos do protocolo de revisão, com o objetivo de verificar se a automação poderia ser aplicada de forma confiável também nessa etapa.

Diversas ferramentas computacionais têm sido desenvolvidas para auxiliar etapas específicas das revisões sistemáticas. O Rayyan oferece suporte à triagem colaborativa e à resolução de conflitos entre revisores, enquanto o ASReview utiliza aprendizado ativo com participação humana para priorizar registros potencialmente relevantes (10, 23, 24). Ferramentas como Abstrackr, EPPI-Reviewer e SWIFT-Active Screener auxiliam a triagem por meio da classificação e priorização automatizada dos registros, mantendo a participação do revisor humano na decisão final(15, 17). Na deduplicação, sistemas como Deduklick, Deduplicator e ASySD combinam normalização de metadados, identificadores bibliográficos e similaridade textual (11-13). Mais recentemente, sistemas integrados, como o ReviewGenie, passaram a reunir diferentes etapas do processo em fluxos modulares (25).

Este estudo teve como objetivo desenvolver e validar internamente o SafeScreen, um fluxo automatizado e auditável para deduplicação (módulo SysRev), triagem de títulos e resumos e priorização exploratória de textos completos, utilizando as decisões de revisores humanos como padrão de referência e avaliando o desempenho, os erros, a preservação dos estudos incluídos e a redução da carga de trabalho.

2. PERGUNTA NORTEADORA

Qual é o desempenho de um sistema automatizado da triagem bibliográfica, em comparação com um padrão de referência humano baseado nas diretrizes

PRISMA, quanto à identificação de estudos elegíveis, à preservação dos estudos finalmente incluídos e à redução da carga de trabalho manual em uma revisão sistemática focada no uso de semaglutida em contexto perioperatório?

3. OBJETIVOS

3.1 Objetivo Geral

Desenvolver e validar um sistema automatizado de triagem bibliográfica para revisões sistemáticas, avaliando seu desempenho em comparação com um padrão de referência humano baseado nas diretrizes PRISMA em uma revisão focada no uso de semaglutida em contexto perioperatório.

3.2 Objetivos Específicos

• Avaliar o desempenho do módulo de deduplicação bibliográfica na identificação e remoção de registros duplicados, em comparação com o padrão de referência humano.

• Avaliar o desempenho do módulo de triagem de títulos e resumos na identificação de estudos potencialmente elegíveis para a revisão sistemática.

• Quantificar a redução da carga de trabalho manual proporcionada pelo sistema automatizado em cada etapa do processo de revisão.

• Descrever o perfil de erros do sistema automatizado, identificando os tipos e as causas mais frequentes de classificações incorretas.

• Avaliar o desempenho do módulo exploratório de leitura automatizada de textos completos em PDF na identificação dos estudos elegíveis para inclusão final.

• Verificar se os estudos finalmente incluídos na revisão sistemática foram preservados ao longo de todas as etapas do processo automatizado.

4. HIPÓTESES

O sistema automatizado desenvolvido apresentará desempenho compatível com sua utilização como ferramenta de apoio ao processo de revisão conduzido por revisores humanos, preservando todos os estudos finalmente incluídos na revisão sistemática e proporcionando redução substancial da carga de trabalho manual.

4.1 Hipóteses Específicas

• O módulo de deduplicação bibliográfica apresentará elevada precisão e sensibilidade na identificação de registros duplicados, especialmente na análise em nível de agrupamento.

• O módulo de triagem de títulos e resumos priorizará sensibilidade, com preservação dos estudos finalmente incluídos no conjunto de referência e redução substancial do número de registros encaminhados à leitura manual de texto completo.

• A redução da carga de trabalho manual proporcionada pelo sistema será superior a 50% na etapa de triagem de títulos e resumos.

• O perfil de erros do sistema será composto predominantemente por falsos positivos, passíveis de correção em etapas subsequentes de revisão humana.

• O módulo exploratório de leitura automatizada de textos completos apresentará desempenho suficiente para atuar como ferramenta de apoio à decisão, com sensibilidade elevada e especificidade satisfatória.

• A preservação dos estudos finalmente incluídos ao longo de todas as etapas do fluxo automatizado constituirá evidência de segurança metodológica do sistema.

5. JUSTIFICATIVA

5.1 Relevância Científica

As revisões sistemáticas constituem o mais alto nível de evidência científica disponível para a tomada de decisão em saúde. No entanto, sua condução exige um volume considerável de tempo e recursos humanos, especialmente nas etapas de remoção de duplicatas e triagem de títulos e resumos. Com o crescimento exponencial da produção científica, esse problema tende a se agravar, tornando cada vez mais difícil manter revisões sistemáticas atualizadas e metodologicamente rigorosas.

Embora ferramentas de automação para revisões sistemáticas já existam, a literatura ainda carece de estudos que validem esses sistemas de forma rigorosa, comparando seus resultados diretamente com um padrão de referência humano baseado em diretrizes consagradas como o PRISMA. A maioria dos estudos disponíveis avalia ferramentas isoladas e em contextos metodológicos heterogêneos, dificultando a comparação entre abordagens e a adoção segura dessas tecnologias na prática. O presente estudo preenche essa lacuna ao desenvolver e validar internamente um sistema automatizado completo, da deduplicação bibliográfica à leitura de textos completos, em uma revisão sistemática focada, com comparação direta e auditável contra decisões humanas.

5.2 Relevância Profissional

Para pesquisadores e profissionais de saúde envolvidos na condução de revisões sistemáticas, a disponibilidade de um sistema automatizado confiável representa uma mudança significativa na forma de trabalho. A automação das etapas mais repetitivas e trabalhosas do processo permite redirecionar o esforço humano para as etapas que realmente exigem julgamento clínico e metodológico, como a avaliação crítica dos estudos incluídos e a síntese dos resultados. Além disso, a transparência e a auditabilidade do sistema desenvolvido neste estudo facilitam sua adoção em contextos acadêmicos e institucionais, uma vez que todas as decisões automatizadas podem ser inspecionadas e verificadas pelos revisores.

5.3 Relevância Social

Revisões sistemáticas de alta qualidade são a base para a elaboração de diretrizes clínicas, protocolos assistenciais e políticas públicas de saúde. Ao tornar o processo de revisão sistemática mais eficiente e acessível, a automação contribui

para que evidências científicas atualizadas cheguem mais rapidamente à prática clínica, beneficiando diretamente os pacientes. No contexto específico deste estudo, a geração de evidências mais ágeis sobre o uso de semaglutida em contexto perioperatório tem implicações diretas para a segurança dos pacientes submetidos a procedimentos cirúrgicos, um tema de crescente relevância clínica diante do aumento expressivo do uso desse medicamento nos últimos anos (26).

6. REFERENCIAL TEÓRICO

6.1 Revisões sistemáticas e o desafio contemporâneo da síntese de evidências

As revisões sistemáticas ocupam posição central na produção de conhecimento em saúde, pois permitem reunir, avaliar criticamente e sintetizar evidências disponíveis a partir de uma pergunta de pesquisa previamente definida. Diferentemente das revisões narrativas, nas quais a seleção dos estudos pode depender de julgamento subjetivo dos autores, as revisões sistemáticas exigem critérios explícitos de elegibilidade, estratégias de busca reproduzíveis, documentação do processo de seleção e avaliação crítica dos estudos incluídos. Essa estrutura metodológica torna as revisões sistemáticas instrumentos fundamentais para orientar decisões clínicas, diretrizes assistenciais e políticas públicas em saúde (2, 4, 5, 27).

No entanto, o mesmo rigor que confere confiabilidade às revisões sistemáticas também impõe elevada carga operacional. A formulação da pergunta, a elaboração da estratégia de busca, a recuperação de registros em múltiplas bases, a remoção de duplicatas, a triagem de títulos e resumos, a leitura de textos completos, a extração de dados e a avaliação do risco de viés constituem etapas sequenciais e interdependentes. Em revisões com grande volume de literatura, essas etapas podem exigir meses de trabalho e múltiplos revisores treinados. Estudos sobre tempo, custo e organização operacional das revisões demonstram que a condução de revisões sistemáticas pode demandar períodos prolongados, equipes numerosas e investimento financeiro relevante (1, 3).

Essa limitação operacional se torna mais relevante diante do crescimento contínuo da produção científica. Quanto maior o número de bases pesquisadas e quanto mais amplo o tema de interesse, maior a probabilidade de recuperar milhares de registros, muitos dos quais são irrelevantes ou duplicados. O custo humano de revisar manualmente esse volume de informação pode atrasar a atualização de evidências e comprometer sua utilidade prática. Nesse contexto, a automação passou a ser discutida não apenas como estratégia de conveniência, mas como possível resposta metodológica à sobrecarga informacional que afeta a síntese de evidências em saúde (3, 7, 25).

A literatura recente sugere que a automação deve ser compreendida como ferramenta de apoio, e não como substituição irrestrita do julgamento humano. O

interesse principal não é eliminar revisores do processo, mas reduzir tarefas repetitivas, priorizar registros relevantes, acelerar etapas mecânicas e permitir que especialistas concentrem esforço em decisões de maior complexidade metodológica. Essa distinção é essencial para o presente estudo, cujo objetivo não é substituir o processo humano de revisão, mas validar um sistema automatizado em comparação direta com decisões humanas baseadas nas diretrizes PRISMA.

6.2 Padronização metodológica, relato transparente e reprodutibilidade

A consolidação de diretrizes foi decisiva para aumentar a transparência das revisões sistemáticas. O PRISMA, publicado originalmente em 2009, estabeleceu um checklist e um fluxograma para documentar as etapas de identificação, triagem, elegibilidade e inclusão dos estudos (4). A versão PRISMA 2020 atualizou esses requisitos, incorporando avanços metodológicos e reforçando a necessidade de relatar de forma explícita estratégias de busca, critérios de seleção, uso de ferramentas automatizadas e decisões tomadas ao longo do processo (5, 22, 27).

A relevância do PRISMA para esta tese é dupla. Primeiro, ele fornece o padrão metodológico que orienta o processo humano usado como referência. Segundo, ele exige que qualquer ferramenta automatizada empregada em uma revisão seja descrita de forma transparente, permitindo avaliação externa de sua influência sobre a seleção dos estudos. Isso é particularmente importante quando decisões automatizadas podem excluir registros antes da leitura humana. Nesses casos, a reprodutibilidade do algoritmo, a clareza dos critérios e a possibilidade de auditoria tornam-se componentes centrais da validade metodológica.

A extensão PRISMA-S reforça essa preocupação ao enfatizar o relato completo das estratégias de busca em revisões sistemáticas (28). A busca bibliográfica é uma das etapas mais vulneráveis à baixa reprodutibilidade, pois pequenas variações em bases, filtros, termos, operadores booleanos e datas podem modificar substancialmente o conjunto de registros recuperados. Estudos metodológicos recentes também indicam que muitas estratégias de busca são relatadas de forma incompleta, limitando a capacidade de replicação e atualização das revisões (29).

Além do PRISMA, instrumentos como AMSTAR 2 contribuem para a avaliação crítica da qualidade metodológica de revisões sistemáticas. Shea et al. propuseram o

AMSTAR 2 como ferramenta para avaliar revisões que incluem estudos randomizados e não randomizados, incorporando domínios críticos relacionados a protocolo, busca, seleção, extração, risco de viés e interpretação dos achados(6). Para uma tese metodológica sobre automação, esse conjunto de diretrizes é relevante porque demonstra que a qualidade de uma revisão depende não apenas do resultado final, mas da rastreabilidade de cada decisão tomada durante o processo.

6.3 Carga de trabalho, custo e necessidade de automação

A automação em revisões sistemáticas emerge de uma necessidade concreta: reduzir tempo e esforço sem comprometer sensibilidade. Borah et al. destacaram que revisões sistemáticas exigem equipes e períodos prolongados de execução (1). Michelson e Reuter, por sua vez, estimaram que uma revisão sistemática custa, em média, US$ 141.194,80. O valor foi calculado considerando uma equipe de cinco pesquisadores, com dedicação média de 15 horas semanais durante 11 meses, equivalente a aproximadamente 1,72 anos-pessoa de trabalho, multiplicados por um custo anual de US$ 82.090 por cientista nos Estados Unidos, com base em dados de 2018. Os autores concluíram que o elevado consumo de tempo e recursos pode limitar a realização consistente de revisões sistemáticas, reforçando a necessidade de soluções automatizadas para reduzir a carga de trabalho e ampliar sua escalabilidade (3).

Clark et al. apresentaram um estudo de caso em que uma revisão sistemática completa foi conduzida em duas semanas com auxílio de ferramentas de automação (7). Embora esse tipo de experiência não elimine a necessidade de rigor, ela demonstra que fluxos estruturados podem reduzir substancialmente o tempo de execução quando aplicados a etapas específicas. De modo semelhante, Al-Marridi et al. descreveram o ReviewGenie como sistema automatizado para revisões sistemáticas, organizado em etapas que incluem busca, coleta, limpeza, deduplicação, filtragem e triagem de títulos e resumos (25).

Essa dependência do escopo é fundamental. Sistemas automatizados não resolvem perguntas mal formuladas. Ao contrário, tendem a amplificar ambiguidades quando critérios de inclusão e exclusão são vagos. Assim, a automação segura exige uma pergunta de pesquisa bem delimitada, critérios operacionais explícitos e mecanismos de verificação. O projeto desta tese se beneficia desse princípio porque

aplica o fluxo automatizado a uma revisão com intervenção específica — semaglutida — e desfechos perioperatórios definidos, o que permite transformar critérios clínicos em regras auditáveis.

6.4 Deduplicação bibliográfica: etapa inicial crítica e risco metodológico

A deduplicação é uma das primeiras etapas operacionais de uma revisão sistemática e influencia diretamente o conjunto de registros que seguirá para triagem. Quando revisores pesquisam múltiplas bases, o mesmo estudo pode aparecer repetidamente com variações de metadados, ausência de DOI, diferenças de título, nomes abreviados de autores ou registros provenientes de versões distintas do mesmo trabalho. A identificação incorreta dessas duplicatas pode gerar dois tipos de erro: manter registros repetidos, aumentando a carga de trabalho, ou remover registros únicos, introduzindo viés no processo de seleção.

Borissov et al. desenvolveram o Deduklick como algoritmo automatizado e explicável para deduplicação em revisões sistemáticas (11). Segundo os dados reportados, o sistema combina processamento de linguagem natural, normalização de metadados, distância de Levenshtein, agrupamento de registros semelhantes e regras criadas por especialistas em informação. O estudo relatou desempenho elevado, com médias de sensibilidade, precisão e F1 superiores ao processo manual, além de redução substancial no tempo de deduplicação (11).

A literatura sobre identificação de registros equivalentes e remoção de duplicatas também contribui para compreender os fundamentos técnicos dessa etapa. Embora nem todos os estudos de deduplicação estejam restritos ao contexto de revisões sistemáticas, abordagens baseadas em aprendizado, comparação textual e identificação de registros redundantes ajudam a explicar os desafios computacionais envolvidos na fusão de registros bibliográficos heterogêneos (30). Essa base conceitual é útil, desde que utilizada com cautela, pois a deduplicação em revisões sistemáticas tem implicações metodológicas específicas: a remoção indevida de um registro único pode comprometer a sensibilidade da revisão.

Forbes et al. avaliaram uma ferramenta denominada Deduplicator, comparando-a ao EndNote em revisões Cochrane. Conforme os dados reportados pelos autores, o Deduplicator reduziu o tempo médio de deduplicação e apresentou menor taxa de erros por mil registros. Entretanto, os autores destacaram limitações

importantes, incluindo a possibilidade de que avaliadores humanos cometessem erros semelhantes e a falta de uma definição universalmente aceita de duplicata (12). Essa observação é metodologicamente relevante, pois reforça que a avaliação da deduplicação deve considerar tanto os registros individuais quanto os agrupamentos de referências que representam o mesmo estudo.

Ferramentas de apoio à gestão de duplicatas e triagem, como sistemas web voltados para revisões sistemáticas, também foram propostas para reduzir a carga manual e aumentar a rastreabilidade do processo (31). Hair et al., ao desenvolverem o ASySD, abordaram a deduplicação como uma etapa passível de automação aberta, rápida e interoperável (13). A ferramenta combina correspondência exata e similaridade textual, demonstrando desempenho elevado em conjuntos biomédicos. McKeown e Mir, por sua vez, compararam métodos automatizados e destacaram que ferramentas específicas de deduplicação tendem a superar gerenciadores de referência tradicionais (14).

Em conjunto, esses estudos sustentam o desenho do módulo de deduplicação do presente projeto, que combina identificadores exatos, comparação por similaridade textual, critérios de segurança por autor e ano de publicação e agrupamento automatizado de registros relacionados. A contribuição específica da tese está em validar internamente uma abordagem de deduplicação em dois níveis: o nível do registro individual e o nível dos agrupamentos de referências que representam o mesmo estudo. Essa escolha é metodologicamente adequada porque duplicatas nem sempre aparecem como pares isolados; frequentemente, diferentes registros de uma mesma publicação formam conjuntos relacionados. Assim, a avaliação por agrupamentos tende a refletir melhor a tarefa real de deduplicação em revisões sistemáticas.

6.5 Triagem automatizada, aprendizado ativo e ferramentas clássicas

Após a deduplicação, a triagem de títulos e resumos é uma das etapas mais trabalhosas da revisão sistemática. Nessa fase, a maioria dos registros costuma ser excluída, mas a decisão exige leitura criteriosa para evitar perda de estudos elegíveis. O’Mara-Eves et al. identificaram o potencial da mineração de texto para apoiar a identificação de estudos em revisões sistemáticas, especialmente pela redução do número de registros que precisam ser avaliados manualmente (9). Cohen et al.

também demonstraram que classificadores automatizados poderiam reduzir a carga de trabalho na preparação de revisões (8).

Entre as ferramentas de aprendizado ativo, o ASReview ocupa posição de destaque. Van de Schoot et al. apresentaram uma estrutura aberta e transparente para revisões sistemáticas eficientes, baseada em aprendizado ativo com revisor no processo (10). O sistema aprende progressivamente com as decisões do revisor e prioriza os registros com maior probabilidade de relevância. A versão ASReview LAB v.2 introduziu múltiplos agentes e múltiplos revisores, mantendo a proposta de triagem eficiente e colaborativa (23).

O Rayyan, descrito por Ouzzani et al., representa outra ferramenta amplamente utilizada para triagem colaborativa de revisões sistemáticas(24). Embora não substitua a decisão dos revisores, oferece funcionalidades de organização, cegamento, resolução de conflitos e sugestões de classificação. Valizadeh et al. avaliaram o uso do Rayyan em revisões de acurácia diagnóstica, demonstrando potencial de efetividade para priorização e apoio à triagem (32).

Outros estudos, como os de Howard et al., Tsou et al. e van der Mierden et al., reforçam que ferramentas de triagem variam quanto ao tipo de algoritmo, necessidade de treinamento, forma de interação com revisores e desempenho em diferentes domínios (15, 17, 33). Essa heterogeneidade dificulta comparações diretas e indica que a validação deve ser contextualizada. Um desempenho adequado em uma revisão pode não se reproduzir em outra, especialmente quando a prevalência de estudos elegíveis, o vocabulário do tema e a clareza dos critérios de elegibilidade diferem substancialmente.

O fluxo automatizado proposto nesta tese se diferencia das ferramentas clássicas porque não depende de aprendizagem ativa com participação contínua do revisor. Em vez disso, utiliza critérios explícitos do protocolo para orientar a triagem dos registros e, em casos de dúvida, aplica uma etapa complementar para evitar a perda de estudos potencialmente relevantes. Dessa forma, a tese não propõe uma nova versão de ferramentas como ASReview ou Rayyan, mas valida um sistema automatizado, transparente e aplicado a uma revisão sistemática real.

6.6 Modelos de linguagem de grande porte e IA generativa em revisões sistemáticas

A emergência dos modelos de linguagem de grande porte ampliou o debate sobre automação em revisões sistemáticas. A arquitetura Transformer, descrita por Vaswani et al., permitiu o processamento paralelo de sequências textuais por mecanismos de atenção, tornando possível o desenvolvimento de modelos capazes de interpretar, classificar e gerar linguagem com alto grau de flexibilidade (34). Embora esse artigo não trate especificamente de revisões sistemáticas, ele fornece a base conceitual para compreender os modelos de linguagem atualmente utilizados em tarefas de triagem, extração e sumarização.

Estudos recentes avaliaram LLMs (Large Language Models) em triagem de títulos e resumos. Tran et al. investigaram o GPT-3.5 Turbo em cinco revisões sistemáticas, observando sensibilidades e especificidades variáveis conforme a regra de classificação utilizada (21). Esse achado é importante porque mostra que o desempenho do modelo não é fixo; ele depende do limiar, do prompt (instrução fornecida ao modelo) e da forma como a decisão é operacionalizada. Guo et al. também avaliaram modelos GPT (Generative Pre-trained Transformer) em triagem de artigos clínicos, relatando desempenho promissor, mas com limitações relacionadas à sensibilidade (19). Em revisões sistemáticas, essa limitação é crítica, pois falsos negativos podem excluir estudos elegíveis antes da avaliação humana.

Matsui et al. descreveram uma estratégia em três camadas para identificação de estudos elegíveis, utilizando modelos GPT em etapas sucessivas relacionadas a desenho do estudo, população e intervenção/controle (20). A lógica de dividir a decisão em camadas é semelhante ao raciocínio metodológico do sistema automatizado desta tese, que utiliza critérios sequenciais de elegibilidade para semaglutida, contexto perioperatório e desfechos de interesse. Essa convergência reforça a plausibilidade de estratégias estruturadas, nas quais a decisão automatizada não é uma classificação genérica, mas uma aplicação organizada de critérios do protocolo.

Dennstädt et al. avaliaram LLMs em conjuntos de revisões biomédicas e observaram variabilidade importante entre modelos e *datasets* (18). Khraisha et al. também demonstraram que o desempenho do GPT-4 (Generative Pre-trained Transformer 4) em revisões sistemáticas é influenciado pela diferença entre o número

de estudos incluídos e excluídos, pelo tipo de tarefa avaliada e pela forma como as instruções de classificação são apresentadas ao modelo(16). Kim et al., em revisão sistemática e metanálise sobre LLMs para triagem de títulos e resumos, identificaram desempenho global promissor, mas não suficiente para dispensar validação específica (35). Esses achados convergem para uma conclusão crítica: LLMs podem apoiar a triagem, mas sua adoção segura exige comparação contra padrão humano e avaliação explícita de falsos negativos.

Outros estudos recentes abordam diretamente oportunidades, desafios e desempenho comparativo de LLMs em diferentes tarefas de seleção de estudos. Galli et al. discutiram oportunidades e considerações metodológicas no uso de LLMs para triagem em revisões sistemáticas (36). Oami et al. avaliaram modelos de linguagem para triagem de citações em revisões sistemáticas e também investigaram o desempenho de LLMs na triagem de citações em contexto biomédico (37, 38). Estudos voltados ao desenvolvimento de prompts, como o de Homiar et al., reforçam que a formulação da instrução é parte central do desempenho em revisões de acurácia diagnóstica (39). De modo semelhante, trabalhos sobre uso de ChatGPT, Claude e outros modelos em diferentes domínios clínicos indicam que a comparação entre modelos e a validação por tarefa ainda são necessárias antes de sua incorporação plena em fluxos metodológicos sensíveis (40, 41).

A literatura mais recente também explora fluxos mais amplos de IA generativa em síntese de evidências. Lee propôs o A4SLR como um sistema automatizado baseado em agentes, no qual diferentes etapas da revisão sistemática são executadas por componentes específicos, incluindo a extração de dados com GPT-4 Turbo (42). Clark et al., em revisão sobre uso de IA generativa em síntese de evidências, identificaram interesse crescente, mas também limitações associadas à rápida evolução das ferramentas (43). Cao, em um artigo pré-publicado relatou a automação de revisões com LLMs e descreveu um fluxo de trabalho mais amplo para triagem e extração (44). Esses estudos indicam uma transição do uso de IA em tarefas isoladas para fluxos automatizados mais integrados, mas ainda com necessidade de validação rigorosa.

6.7 Segurança, alucinação e necessidade de supervisão humana

A principal barreira para o uso amplo de LLMs em revisões sistemáticas não é apenas o desempenho médio, mas a segurança metodológica. Em tarefas de triagem, um falso positivo aumenta o trabalho humano, mas um falso negativo pode excluir um estudo elegível e comprometer a validade da revisão. Por isso, a sensibilidade deve ser priorizada em relação à especificidade quando o sistema é utilizado para reduzir registros antes da avaliação humana.

Declarações de posição sobre o uso de IA em síntese de evidências reforçam que ferramentas automatizadas devem ser utilizadas de forma responsável, transparente e supervisionada (45, 46). Flemyng et al., no contexto Cochrane, destacaram a necessidade de cautela na incorporação de IA à síntese de evidências (45). Gartlehner et al. também discutiram a integração responsável da IA em revisões rápidas, ressaltando que o ganho de velocidade não deve ocorrer às custas de princípios metodológicos essenciais (46).

A literatura sobre alucinação em LLMs fornece base teórica para essa cautela. Rousta et al. discutiram alucinações em modelos de linguagem no contexto clínico, destacando que respostas plausíveis podem ser factualmente incorretas (47). Huang et al. analisaram princípios, taxonomia e desafios relacionados à alucinação (48), enquanto Tonmoy et al. revisaram técnicas de mitigação (49). Embora esses estudos não estejam todos focados exclusivamente em revisões sistemáticas, eles são relevantes porque a triagem automatizada depende da capacidade do modelo de fundamentar decisões no texto fornecido, sem extrapolar informações ausentes.

No projeto desta tese, essa preocupação aparece na decisão de tratar o módulo de leitura de textos completos como exploratório e de suporte, não como substituto da decisão humana final. Essa escolha é coerente com a literatura, pois reconhece o potencial dos LLMs para reduzir carga de trabalho, mas preserva a centralidade da validação humana em decisões críticas.

6.8 Métricas de validação: sensibilidade, especificidade, kappa e trabalho economizado

A validação de sistemas automatizados exige métricas alinhadas ao risco metodológico da tarefa. Em revisões sistemáticas, a sensibilidade representa a proporção de estudos elegíveis corretamente identificados e deve ser considerada

métrica prioritária. A especificidade mede a proporção de registros inelegíveis corretamente excluídos, enquanto a precisão indica a proporção de registros classificados como elegíveis que de fato são elegíveis. O escore F1 combina sensibilidade e precisão, sendo útil em cenários com classes desbalanceadas.

A métrica Work Saved over Sampling (WSS), discutida por Cohen et al. e O'Mara-Eves et al., é particularmente útil para avaliar o benefício operacional da automação (8, 9). Kusa et al. avançaram essa discussão ao demonstrar limitações da métrica tradicional de trabalho economizado e ao propor uma versão normalizada, que permite interpretar de forma mais direta a proporção de registros irrelevantes corretamente excluídos quando se mantém um nível fixo de sensibilidade(50). Essa contribuição é altamente relevante para esta tese, pois permite interpretar a economia de trabalho de forma condicionada à preservação da sensibilidade.

O kappa de Cohen é relevante para medir concordância entre o sistema automatizado e o padrão humano, ajustando a concordância esperada ao acaso. Landis e Koch propuseram categorias interpretativas amplamente utilizadas (51), enquanto McHugh discutiu a interpretação do kappa na confiabilidade interavaliador (52). No entanto, a aplicação do kappa em revisões sistemáticas deve ser cautelosa, pois a baixa prevalência de estudos elegíveis pode afetar o valor da métrica. Essa limitação reforça a necessidade de apresentar múltiplas métricas em conjunto.

Portanto, o desempenho do fluxo automatizado não deve ser interpretado apenas por acurácia global. Em bases altamente desbalanceadas, um sistema pode apresentar acurácia elevada simplesmente por excluir a maioria dos registros, mesmo que perca estudos relevantes. A avaliação deve priorizar: preservação dos estudos finalmente incluídos, sensibilidade, perfil de falsos negativos, especificidade em nível de sensibilidade aceitável, F1, kappa e redução da carga de trabalho manual.

6.9 Lacuna da literatura e justificativa da tese

A literatura analisada demonstra que há avanços importantes em deduplicação, triagem automatizada, aprendizado ativo e uso de LLMs em revisões sistemáticas. Entretanto, persistem lacunas metodológicas. Muitas ferramentas são avaliadas em tarefas isoladas, com base de dados heterogêneas, padrões de referência variáveis e baixa comparabilidade entre estudos. Além disso, parte das abordagens depende de aprendizado ativo com interação contínua do revisor,

enquanto outras utilizam LLMs de forma autônoma sem validação suficiente para garantir segurança metodológica.

Nesse cenário, a tese se justifica por propor e validar um fluxo automatizado aplicado a uma revisão sistemática real, com comparação direta contra decisões humanas baseadas no PRISMA. O diferencial não está apenas em usar automação, mas em documentar seu desempenho em módulos específicos, avaliar deduplicação em nível de registro e agrupamento, medir redução de carga de trabalho e acompanhar explicitamente a preservação dos estudos finalmente incluídos. Essa abordagem responde a uma necessidade identificada na literatura: desenvolver sistemas transparentes, auditáveis, reprodutíveis e avaliados contra padrão humano antes de sua adoção em processos de síntese de evidências.

Assim, a revisão da literatura sustenta que a automação em revisões sistemáticas é metodologicamente promissora, mas sua adoção segura depende de validação contextual, priorização de sensibilidade, análise de erros e manutenção do julgamento humano nas etapas críticas. A tese se insere precisamente nessa fronteira: entre a eficiência proporcionada por algoritmos e a segurança exigida pela metodologia científica.

7. METODOLOGIA

7.1 Delineamento

Trata-se de um estudo de validação metodológica, de caráter observacional e analítico, desenvolvido no contexto de uma revisão sistemática focada no uso de semaglutida em contexto perioperatório. O sistema automatizado foi projetado para apoiar, e não substituir, o processo convencional de revisão baseado nas diretrizes PRISMA, tendo suas decisões comparadas diretamente com as de revisores humanos treinados, que serviram como padrão de referência.

7.2 Local

O estudo foi conduzido na Universidade de Caxias do Sul, Caxias do Sul, Rio Grande do Sul, Brasil. O sistema automatizado foi implementado em ambiente de programação Python, utilizando a plataforma Google Colaboratory (Colab) para execução dos scripts de processamento bibliográfico e análise dos dados.

7.3 Amostra

A amostra deste estudo foi constituída retrospectivamente pelos registros bibliográficos recuperados na revisão sistemática complementar sobre o uso de semaglutida no contexto perioperatório, não tendo sido realizada nova busca bibliográfica especificamente para esta tese. Na revisão de origem, a busca foi conduzida desde o início de indexação até abril de 2026 nas bases Scopus, Web of Science, Embase, MEDLINE via PubMed, LILACS, Cochrane Library e OpenGrey, complementada pela busca manual nas listas de referências dos estudos recuperados. Após a exportação e consolidação, esses registros constituíram o corpus submetido ao fluxo automatizado de validação.

O padrão de referência humano foi igualmente derivado da revisão sistemática de origem. Nessa revisão, a seleção dos estudos foi realizada de forma sistemática e independente por dois revisores, com resolução das divergências por um terceiro revisor. Os revisores possuíam experiência prévia na seleção de estudos para revisões sistemáticas e foram treinados para conduzir o processo de acordo com os critérios estabelecidos no protocolo original. A seleção humana foi concluída antes da aplicação do fluxo automatizado. A revisão foi conduzida conforme as recomendações do PRISMA e do Cochrane Handbook, e seu protocolo foi registrado

no International Prospective Register of Systematic Reviews (PROSPERO), mantido pela University of York, sob o número CRD42024609601. Para a validação do sistema, foram utilizadas as decisões finais previamente registradas nas etapas de deduplicação, triagem de títulos e resumos e avaliação de textos completos. Os estudos finalmente incluídos na revisão constituíram, ainda, o conjunto de segurança utilizado para verificar sua preservação ao longo de todo o fluxo automatizado.

7.4 Critérios de inclusão

Foram incluídos na revisão sistemática estudos que atenderam simultaneamente aos seguintes critérios:

1. Pacientes adultos com idade igual ou superior a 18 anos;
2. Pacientes que receberam semaglutida no período pré-operatório, independentemente da duração do tratamento, do intervalo entre a última dose e o procedimento e do tipo de cirurgia ou procedimento anestésico realizado.
3. Ensaios clínicos randomizados e estudos observacionais — incluindo estudos de coorte, caso-controle e transversais — que relataram pelo menos um dos seguintes desfechos perioperatórios:
    - Conteúdo gástrico residual (desfecho primário);
    - Sintomas gastrointestinais;
    - Aspiração pulmonar;
    - Descontinuação do procedimento;
    - Gastroparesia.

Para fins de operacionalização dos critérios de elegibilidade no sistema, foram definidos três critérios de inclusão (IC): IC1, referente à população adulta; IC2, referente à exposição explícita à semaglutida; e IC3, referente à presença de desenho de estudo elegível, dados originais de pacientes e pelo menos um desfecho perioperatório de interesse.

7.5 Critérios de exclusão

Foram excluídos da revisão sistemática:

1. Pacientes com comorbidades significativas que, por si só, aumentam o risco de complicações perioperatórias, como doenças cardiovasculares ou respiratórias graves;
2. Pacientes com contraindicações ao uso de semaglutida;

3. Estudos que avaliaram exclusivamente outros agonistas do receptor de GLP-1 (*Glucagon-Like Peptide-1*) ou outros medicamentos para diabetes, sem incluir dados específicos sobre semaglutida;
4. Relatos de caso, séries de casos, revisões narrativas e opiniões de especialistas que não apresentaram dados originais ou comparações entre grupos;
5. Estudos publicados em idioma diferente do inglês, salvo quando uma versão traduzida estivesse prontamente disponível;
6. Estudos com dados ausentes de forma significativa, que impossibilitassem a análise e a interpretação adequadas dos desfechos de interesse.

Para fins de operacionalização dos critérios de elegibilidade no sistema, esses critérios foram codificados como critérios de exclusão (EC): EC1, manejo de comorbidades como foco principal; EC2, contraindicações à semaglutida como foco principal; EC3, ausência de dados específicos de semaglutida; EC4, ausência de dados originais ou desenho de estudo inelegível; EC5, publicação em idioma diferente do inglês; e EC6, dados ausentes ou insuficientes para avaliação dos desfechos.

7.6 Coleta de Dados

A coleta de dados deste estudo não envolveu a realização de uma nova busca bibliográfica. Foram utilizados os registros previamente recuperados na revisão sistemática complementar, cuja busca foi conduzida nas bases Scopus, Web of Science, Embase, MEDLINE via PubMed, LILACS, Cochrane Library e OpenGrey. As referências bibliográficas exportadas dessas bases foram consolidadas em um único arquivo no formato CSV (*Comma-Separated Values*), que constituiu o corpus submetido à validação. O processo automatizado foi estruturado em três módulos sequenciais, descritos a seguir.

7.7 Sistema automatizado

O sistema automatizado foi desenvolvido em linguagem Python para apoiar as etapas operacionais da revisão sistemática, especialmente a remoção de duplicatas, a triagem inicial de títulos e resumos e a avaliação dos textos completos em formato PDF. O fluxo foi estruturado em três módulos sequenciais: deduplicação bibliográfica, triagem de títulos e resumos e avaliação dos textos completos.

Inicialmente, os registros bibliográficos foram importados e submetidos a uma etapa de pré-processamento. Essa etapa incluiu a identificação automática das colunas disponíveis, a padronização dos campos bibliográficos e a normalização de informações como título, resumo, ano de publicação, primeiro autor e identificadores, quando disponíveis. Esse procedimento teve como objetivo reduzir variações de escrita e facilitar a comparação entre registros provenientes de diferentes bases de dados.

A etapa de deduplicação foi realizada para identificar registros repetidos e preservar a versão mais completa de cada referência. O sistema normalizou inicialmente os campos bibliográficos e realizou correspondência exata pelos identificadores DOI, PMID e PMCID, quando disponíveis. Na ausência de identificadores confiáveis, os registros foram comparados por título normalizado, ano de publicação, primeiro autor e resumo. A correspondência aproximada foi realizada com a biblioteca RapidFuzz, utilizando limiares de similaridade de 97 para títulos e 95 para resumos. As comparações entre resumos foram restritas àqueles com pelo menos 150 caracteres. Como regras de segurança, registros com anos de publicação incompatíveis ou primeiros autores discordantes não foram agrupados quando essas informações estavam disponíveis.

Neste estudo, definiu-se registro como cada entrada bibliográfica individual importada das bases de dados, ainda que diferentes registros pudessem representar a mesma publicação. O termo agrupamento (cluster) correspondeu ao conjunto de dois ou mais registros que, após a comparação dos identificadores e metadados bibliográficos, foram considerados representações da mesma publicação. Em cada agrupamento, um registro foi preservado como representante, enquanto os demais foram classificados como duplicatas. Os registros duplicados foram organizados em agrupamentos, mantendo-se a referência com maior escore de completude, definido pela presença de identificadores bibliográficos, dados de publicação — como autor, ano e periódico — e resumo mais completo.

Após a deduplicação, os registros únicos foram submetidos à triagem automatizada de títulos e resumos, implementada em Python 3.10 e executada no Google Colaboratory. A classificação principal foi baseada em regras derivadas dos critérios de elegibilidade da revisão sistemática, organizadas em três domínios: presença de termos relacionados à semaglutida ou aos agonistas do receptor de

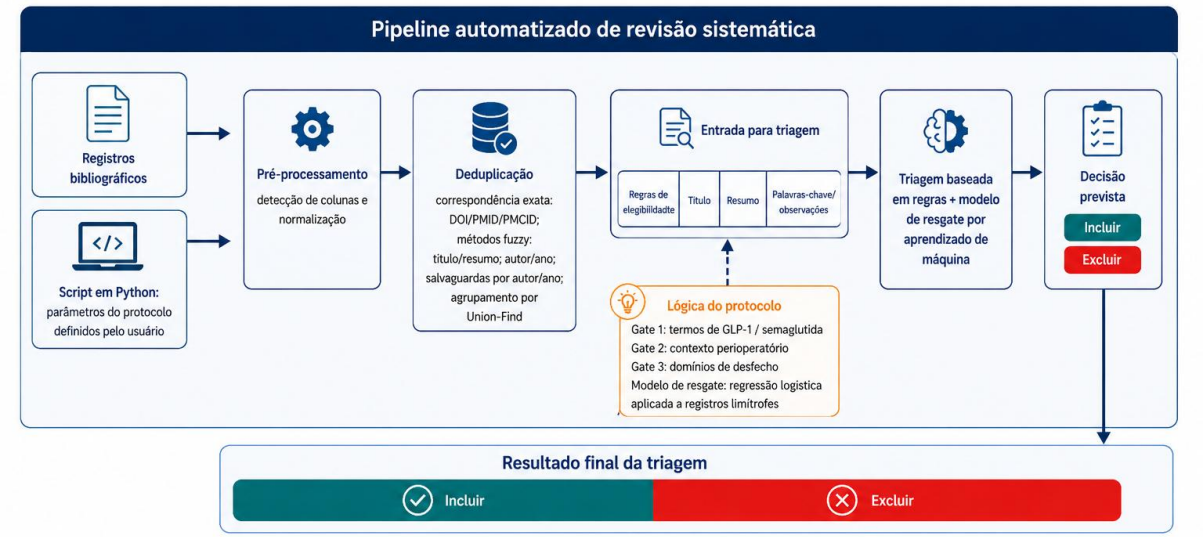
GLP-1; presença de contexto perioperatório, cirúrgico, anestésico, endoscópico ou procedural; e presença de pelo menos um desfecho de interesse. Também foram aplicadas exclusões operacionais para publicações não originais, estudos em animais, estudos pediátricos e registros sem informações específicas sobre semaglutida. Nos casos que atenderam aos critérios de intervenção e contexto, mas não ao domínio de desfecho, foi aplicada uma etapa conservadora de resgate por aprendizado de máquina, desenvolvida com *scikit-learn*, utilizando vetorização TF-IDF de palavras e caracteres, regressão logística com pesos balanceados e limiar de probabilidade de 0,90.

Nos casos em que o registro apresentava características duvidosas, foi aplicada uma etapa complementar de verificação com o objetivo de reduzir o risco de exclusão indevida de estudos potencialmente relevantes. Essa etapa foi utilizada apenas em situações limítrofes, especialmente quando o registro apresentava termos relacionados à intervenção e ao contexto perioperatório, mas não atendia claramente ao domínio de desfecho no título ou resumo. Dessa forma, o sistema priorizou a sensibilidade da triagem, aceitando maior número de falsos positivos para diminuir o risco de falsos negativos.

Por fim, os artigos selecionados para avaliação em texto completo foram processados em formato PDF. O conteúdo textual foi extraído com a biblioteca PyMuPDF e submetido, por meio da API da Anthropic, ao modelo de linguagem Claude Sonnet 4.6 (claude-sonnet-4-6). O modelo recebeu um comando padronizado baseado nos critérios de elegibilidade da revisão sistemática e retornou, em formato estruturado JSON, uma decisão de priorização ou não priorização, os critérios atendidos, a justificativa e o motivo de exclusão, quando aplicável. Foram utilizados temperatura igual a zero, limite de 1.200 tokens e até 35.000 caracteres extraídos por PDF, com uma solicitação independente para cada documento. Esse módulo foi considerado exploratório e utilizado apenas como apoio à priorização, mantendo-se a avaliação humana como referência final para a inclusão dos estudos.

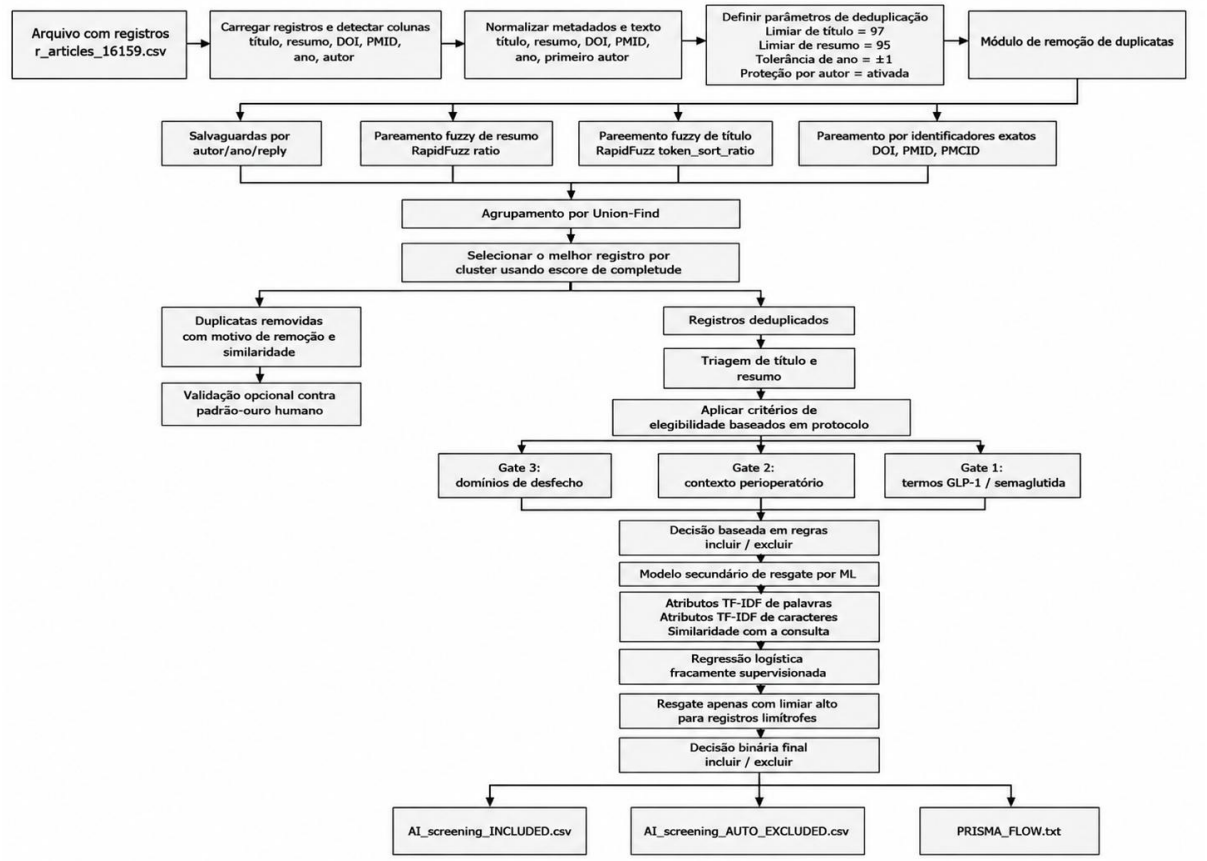
O fluxo geral do sistema automatizado está apresentado nas figuras 1 e 2, e os principais módulos, algoritmos e estratégias de decisão estão resumidos na Tabela 1 e a configuração de software, modelo e reprodutibilidade do fluxo automatizado disponível na tabela 2.

Figura 1. Fluxograma do sistema automatizado de deduplicação e triagem de títulos e resumos em revisão sistemática.



Fonte: Elaborada pelo autor.
Nota: DOI = Digital Object Identifier; PMID = PubMed Identifier; PMCID = PubMed Central Identifier; GLP-1 = glucagon-like peptide-1; ML = machine learning. Gate refere-se a critério operacional de elegibilidade aplicado na triagem automatizad

Figura 2. Fluxo metodológico do sistema automatizado de deduplicação e triagem bibliográfica com aplicação de critérios protocolares e modelo de resgate por aprendizado de máquina.



Nota: DOI = Digital Object Identifier; PMID = PubMed Identifier; PMCID = PubMed Central Identifier; GLP-1 = glucagon-like peptide-1; ML = aprendizado de máquina; TF-IDF = term frequency–inverse document frequency.
Nota: O fluxo de trabalho consiste em uma etapa de deduplicação determinística/comparação fuzzy, seguida por uma triagem de títulos e resumos orientada por protocolo. A deduplicação utiliza correspondência exata de identificadores, similaridade fuzzy de títulos e resumos, salvaguardas baseadas em autor/ano e agrupamento por Union-Find. A triagem é predominantemente baseada em regras, utilizando critérios de elegibilidade predefinidos, com aplicação de um modelo de regressão logística fracamente supervisionado apenas como mecanismo de resgate com limiar elevado para registros limítrofes.

Tabela 1 – Módulos implementados, algoritmos, técnicas de extração de atributos, estratégias de decisão e balanceamento no fluxo automatizado

Módulo	Algoritmos/classificadores implementados	Extração de atributos / entradas de pareamento	Estratégia de seleção/decisão	Estratégia de balanceamento
Módulo de remoção de duplicatas	Pareamento por identificadores exatos: DOI, PMID e PMCID; pareamento fuzzy por título; pareamento fuzzy por resumo; agrupamento por Union-Find	DOI, PMID, PMCID, título, resumo, ano e primeiro autor normalizados; similaridade de título por RapidFuzz; token-sort ratio; similaridade de resumo por RapidFuzz ratio	Detecção determinística/fuzzy de duplicatas com tolerância por ano, salvaguarda por autor e salvaguarda para reply/letter/erratum; o melhor registro foi retido por escore de completude	Não aplicável
Módulo de triagem por título/resumo	Triagem de elegibilidade baseada em regras, com modelo de regressão logística fracamente supervisionado como mecanismo de resgate	TF-IDF de palavras, TF-IDF de caracteres e similaridade com a pergunta; título, resumo, palavras-chave e notas utilizados como entradas textuais	Triagem protocolar em três critérios: 1) termos GLP-1/semaglutida; 2) contexto perioperatório; 3) domínios de desfecho. Registros limítrofes poderiam ser resgatados por modelo de regressão logística com limiar elevado	Pesos de classe inversamente proporcionais à frequência das classes (class_weight="balanced") na regressão logística de resgate

Fonte: Dados da pesquisa.
Nota: DOI = Digital Object Identifier; PMID = PubMed Identifier; PMCID = PubMed Central Identifier; GLP-1 = glucagon-like peptide-1; TF-IDF = term frequency–inverse document frequency. Diferentemente de sistemas clássicos de aprendizado ativo com revisor em ciclo iterativo, este fluxo automatizado não foi desenvolvido como uma estratégia iterativa de priorização. O módulo de remoção de duplicatas combinou pareamento por identificadores exatos, similaridade textual fuzzy e agrupamento por Union-Find. O módulo de triagem foi primariamente baseado em regras, com aprendizado de máquina utilizado apenas como mecanismo secundário de resgate para registros limítrofes. Nesse modelo de resgate, foram utilizados pesos balanceados entre as classes, atribuindo maior importância à classe menos frequente durante o treinamento, a fim de reduzir o risco de não recuperar registros potencialmente relevantes.

Tabela 2 – Configuração de software, modelo e reprodutibilidade do fluxo automatizado

Componente	Configuração
Ambiente de programação	Python (v. 3.10), Google Colaboratory
Bibliotecas de deduplicação	pandas (v. 2.2.2), RapidFuzz (v. 3.14.5), tqdm (v. 4.69.0)
Modelo de triagem	scikit-learn (v. 1.5.2); representações TF-IDF de palavras e caracteres com regressão logística
Modelo de IA para texto completo	claude-sonnet-4-6
Configurações da API	temperature = 0; max_tokens = 1200
Extração do texto completo	PyMuPDF (v. 1.28.0); get_text("blocks"); até 35.000 caracteres por PDF
Execução das consultas	5 de maio de 2026; Caxias do Sul, Brasil
Estratégia de prompt	Prompt fixo; uma solicitação independente à API por PDF
Materiais de auditoria	Arquivo S1; Figshare; Zenodo; GitHub

Nota: IA = inteligência artificial; API = interface de programação de aplicações; PDF = formato portátil de documento; TF-IDF = frequência do termo–inverso da frequência do documento. O módulo de IA para texto completo utilizou um modelo de prompt fixo e uma solicitação independente à API para cada PDF. Os caminhos dos arquivos e as chaves da API foram removidos do código compartilhado. Os PDFs de origem foram utilizados para extração textual no fluxo, mas não foram redistribuídos.
Fonte: Elaborado pelos autores (2026).

7.7.1 Módulo 1 — Deduplicação bibliográfica

Os registros bibliográficos foram submetidos a um processo automatizado de normalização textual, que converteu os campos de texto para letras minúsculas, removeu pontuação, eliminou espaços duplicados e padronizou os formatos de DOI e PMID. A identificação de duplicatas seguiu uma estratégia hierárquica composta pelas seguintes regras, aplicadas nesta ordem: correspondência exata por DOI; correspondência exata por PMID; correspondência exata por PMCID; correspondência exata por combinação de ano, primeiro autor e título; correspondência aproximada de resumos com limiar de similaridade de 95%; e correspondência aproximada de títulos com limiar de similaridade de 97%.

Para evitar a fusão incorreta de registros não relacionados, a correspondência aproximada foi protegida por dois critérios de segurança: registros com anos de publicação que diferissem em mais de um ano não eram agrupados, e a fusão por correspondência aproximada exigia concordância do primeiro autor quando essa informação estava disponível em ambos os registros. Artigos do tipo carta,

comentário, errata, editorial e resposta foram identificados e protegidos de fusão indevida com o artigo principal correspondente.

Os pares de duplicatas identificados foram agrupados por meio de uma estrutura de união e busca, permitindo a resolução de relações transitivas de duplicidade em agrupamentos. Dentro de cada agrupamento, o registro retido foi selecionado com base em um escore de completude, que privilegiou registros com maior número de campos preenchidos, identificadores exatos presentes e resumos mais extensos. O código-fonte do módulo de deduplicação bibliográfica está disponível no Apêndice A1.

7.7.2 Módulo 2 — Triagem de títulos e resumos

A triagem automatizada de títulos e resumos foi desenvolvida para classificar os registros únicos, após a deduplicação, em duas categorias: potencialmente elegíveis ou excluídos. Essa etapa foi orientada pelos critérios do protocolo da revisão sistemática e teve como prioridade maximizar a sensibilidade, reduzindo o risco de exclusão indevida de estudos potencialmente relevantes.

Inicialmente, o sistema identificou automaticamente as colunas bibliográficas disponíveis no arquivo de entrada, incluindo título, resumo, palavras-chave e notas, quando presentes. Esses campos foram combinados em um único campo textual para análise. Antes da aplicação dos critérios, o texto foi normalizado, com padronização para letras minúsculas e remoção de espaços duplicados, a fim de reduzir variações de escrita entre os registros provenientes de diferentes bases de dados.

A decisão primária foi baseada em três critérios sequenciais de triagem, denominados etapas operacionais. A primeira etapa avaliou a presença de termos relacionados à exposição de interesse, incluindo semaglutida, Ozempic, Wegovy, Rybelsus, GLP-1, GLP1-RA, agonistas do receptor de GLP-1, glucagon-like peptide-1 e termos relacionados a droga de interesse. Essa etapa correspondeu ao critério de exposição da revisão, isto é, uso de semaglutida ou de agonistas do receptor de GLP-1 com possibilidade de dados específicos para semaglutida.

A segunda etapa avaliou a presença de contexto perioperatório ou de procedimentos. Foram considerados termos relacionados ao período pré-operatório, intraoperatório, pós-operatório ou perioperatório, bem como expressões associadas a cirurgia, anestesia, sedação, endoscopia, colonoscopia, endoscopia digestiva alta,

gastroscopia, intubação, via aérea, jejum, NPO (nada por via oral), procedimento, laparoscopia e cirurgia bariátrica. Esse critério teve como objetivo excluir registros sobre semaglutida ou GLP-1 em contextos exclusivamente metabólicos, farmacológicos ou ambulatoriais, sem relação com procedimentos ou segurança perioperatória.

A terceira etapa avaliou a presença de pelo menos um domínio de desfecho perioperatório de interesse. Esse critério foi operacionalizado em cinco domínios não mutuamente exclusivos. O domínio IC3A correspondeu ao conteúdo gástrico residual, retenção gástrica, estômago cheio, esvaziamento gástrico, volume gástrico ou avaliação por ultrassonografia gástrica. O domínio IC3B correspondeu a aspiração pulmonar, aspiração bronquial, pneumonia aspirativa, pneumonite aspirativa ou regurgitação. O domínio IC3C incluiu complicações relacionadas à via aérea, como dificuldade de intubação, falha de intubação, intubação emergencial, sequência rápida, laringoespasmo, broncoespasmo ou complicações de manejo de via aérea. O domínio IC3D avaliou suspensão, retirada, interrupção, pausa, *washout* ou período de retenção da semaglutida ou de agonistas do receptor de GLP-1 antes de procedimentos. O domínio IC3E incluiu complicações perioperatórias, pós-operatórias, respiratórias, anestésicas ou cirúrgicas gerais.

Para que um registro fosse classificado como potencialmente elegível pela regra principal, era necessário que ele atendesse simultaneamente as três etapas: presença de termo relacionado a GLP-1/semaglutida, presença de contexto perioperatório ou procedural e presença de pelo menos um domínio de desfecho. Registros que falhavam em qualquer um desses critérios eram inicialmente classificados como excluídos.

Além dos critérios de inclusão, o sistema aplicou critérios operacionais de exclusão. Foram excluídos estudos não originais ou sem dados próprios, como revisões sistemáticas, metanálises, revisões narrativas, revisões de escopo, diretrizes, consensos, editoriais, cartas, comentários, opiniões de especialistas, relatos de caso e séries de casos. Também foram excluídos registros com marcadores de população pediátrica, estudos em animais e artigos relacionados exclusivamente a outros agonistas do receptor de GLP-1, como liraglutida, dulaglutida, exenatida, lixisenatida ou tirzepatida, quando não havia menção a semaglutida nem possibilidade de extração de dados específicos para semaglutida.

Idiomas diferentes do inglês também foram registrados como critério de exclusão operacional.

Para reduzir o risco de perda de estudos relevantes, foi incorporada uma etapa complementar de resgate por aprendizado de máquina. Essa etapa foi aplicada apenas aos registros classificados inicialmente como excluídos, mas que haviam atendido as primeiras duas etapas, ou seja, apresentavam termo relacionado a GLP-1/semaglutida e contexto perioperatório, sem identificação clara de domínio de desfecho no título ou resumo. O objetivo dessa etapa foi recuperar registros potencialmente elegíveis em que o desfecho de interesse estivesse descrito de forma indireta, incompleta ou pouco evidente nos dados.

O modelo complementar utilizou vetorização textual por TF-IDF (Term Frequency–Inverse Document Frequency) de palavras e de caracteres, além de medida de similaridade com uma pergunta estruturada da revisão. A classificação foi realizada por regressão logística com ponderação balanceada entre classes. Os registros incluídos pela regra principal foram usados como exemplos positivos, enquanto registros claramente sem termo GLP-1/semaglutida ou sem contexto perioperatório foram usados como exemplos negativos. O modelo foi utilizado apenas como mecanismo secundário de resgate, não como critério principal de triagem.

A inclusão por aprendizado de máquina exigiu probabilidade predita igual ou superior a 0,90. Esse limiar elevado foi adotado para manter a função conservadora do modelo, evitando a inclusão excessiva de registros irrelevantes. Assim, apenas registros com forte compatibilidade textual com a pergunta da revisão puderam ser recuperados por essa etapa complementar.

Ao final da triagem, cada registro recebeu uma decisão automatizada binária: incluído ou excluído. Para os registros incluídos, o sistema registrou as etapas atendidas, os domínios de desfecho identificados e a probabilidade estimada pelo modelo, quando aplicável. Para os registros excluídos, foram registrados os motivos operacionais de exclusão, como ausência de termo GLP-1/semaglutida, ausência de contexto perioperatório, ausência de domínio de desfecho, estudo não original, estudo animal, marcador pediátrico, outro agonista de GLP-1 sem dados específicos de semaglutida ou marcador de idioma não inglês. Essa estrutura permitiu rastrear as decisões automatizadas e compará-las posteriormente com o padrão de referência

humano. O código-fonte do módulo de triagem de títulos e resumos está disponível no Apêndice A2.

7.7.3 Módulo 3 — Leitura exploratória de textos completos

Os artigos selecionados para avaliação em texto completo foram organizados em formato PDF e processados em um módulo exploratório de leitura automatizada. Essa etapa teve como objetivo avaliar se um modelo de linguagem de grande porte poderia auxiliar na classificação preliminar dos textos completos, com base nos mesmos critérios de elegibilidade definidos no protocolo da revisão sistemática.

Inicialmente, os arquivos em PDF foram carregados a partir de uma pasta específica no ambiente Google Drive. O conteúdo textual de cada arquivo foi extraído por meio da biblioteca PyMuPDF. A extração foi realizada em blocos de texto, permitindo capturar não apenas parágrafos do corpo principal do artigo, mas também legendas, caixas de texto, fragmentos de tabelas, notas e outros elementos textuais presentes nas páginas. Essa estratégia foi adotada para aumentar a chance de identificar informações relevantes que, em alguns artigos, poderiam estar disponíveis apenas em tabelas ou seções específicas do texto completo.

Para padronizar o processamento e limitar o volume de entrada enviado ao modelo, foi extraído um máximo de 35.000 caracteres por artigo. Esse limite constituiu uma restrição técnica de entrada. Arquivos com extração textual insuficiente foram identificados como potencialmente problemáticos, especialmente quando o PDF apresentava características de imagem escaneada ou baixa capacidade de extração textual. Cada artigo processado recebeu um título derivado preferencialmente do conteúdo extraído ou, quando necessário, do nome do arquivo.

Os textos extraídos dos PDFs foram submetidos ao Claude Sonnet (identificador do modelo: claude-sonnet-4-6) por meio de API, permitindo a integração automatizada entre o código desenvolvido no Google Colab e a plataforma do modelo. Essa estratégia possibilitou o envio padronizado do conteúdo textual de cada artigo e o retorno estruturado das decisões de inclusão ou exclusão. O modelo foi configurado com temperatura igual a zero, com o objetivo de reduzir a variabilidade entre respostas e aumentar a reprodutibilidade das classificações automatizadas. O modelo recebeu uma instrução padronizada contendo os critérios de inclusão e exclusão da

revisão, além da orientação de responder exclusivamente em formato JSON (JavaScript Object Notation) estruturado.

A avaliação automatizada considerou três critérios principais de inclusão. Os critérios de elegibilidade estabelecidos na seção 7.4 foram operacionalizados no módulo de leitura de texto completo como critérios estruturados de inclusão (IC1, IC2 e IC3), utilizados para orientar a classificação automatizada dos artigos. O critério IC1 correspondeu à população adulta, sendo considerado atendido salvo quando o artigo indicava explicitamente população pediátrica ou estudo em animais. O critério IC2 exigiu que a semaglutida fosse explicitamente nomeada como medicamento estudado, incluindo os nomes semaglutida, Ozempic, Wegovy ou Rybelsus. Para atender a esse critério, a semaglutida deveria aparecer como parte da população, exposição, coorte, métodos, resultados ou tabela de caracterização dos pacientes, não sendo suficiente sua menção apenas na introdução ou contextualização geral do artigo.

O critério IC3 avaliou se o estudo apresentava desenho elegível e dados originais de pacientes, como ensaios clínicos ou estudos observacionais, incluindo coortes, estudos caso-controle ou transversais. Além disso, o artigo precisava relatar pelo menos um dos desfechos perioperatórios de interesse: conteúdo gástrico residual ou esvaziamento gástrico, aspiração pulmonar ou bronquial, complicações de via aérea, suspensão ou intervalo de interrupção da semaglutida antes do procedimento, ou complicações perioperatórias gerais.

Também foram aplicados critérios de exclusão. Os critérios de exclusão estabelecidos na seção 7.5 foram operacionalizados no módulo de leitura de texto completo como critérios estruturados de exclusão (EC1, EC2, EC3, EC4, EC5 e EC6), utilizados para orientar a classificação automatizada dos artigos. O modelo foi orientado a excluir artigos cujo foco principal fosse manejo de comorbidades sem relação direta com semaglutida, contraindicações à semaglutida como foco primário, estudos sem dados específicos de semaglutida, documentos sem dados originais de pacientes, artigos não escritos em inglês e estudos com dados quantitativos insuficientes para interpretação dos desfechos. Foram classificados como sem dados originais revisões narrativas, revisões sistemáticas, metanálises, editoriais, comentários, diretrizes, consensos, relatos de caso, séries de casos pequenas, cartas sem dados próprios e registros de ensaios clínicos sem resultados.

Para cada PDF, o modelo retornou uma decisão estruturada de inclusão ou exclusão. A saída incluiu a decisão final, o critério principal associado à decisão, o atendimento aos critérios IC1, IC2 e IC3, os domínios de desfecho identificados, o critério de exclusão quando aplicável, uma justificativa resumida, a evidência textual considerada mais relevante e uma explicação breve do raciocínio utilizado. As respostas foram salvas em arquivo CSV após o processamento de cada artigo, permitindo interromper e retomar a análise sem perda dos resultados já obtidos.

Esse módulo foi definido previamente como exploratório e de apoio à decisão. Portanto, suas classificações não substituíram a decisão humana final de inclusão dos estudos na revisão sistemática. O objetivo principal foi avaliar a utilidade do modelo como ferramenta auxiliar para priorização, rastreabilidade e redução da carga de trabalho na leitura de textos completos, mantendo o padrão humano como referência para validação das decisões automatizadas. O código-fonte do módulo de leitura exploratória de textos completos está disponível no Apêndice A3.

7.8 Análise de Dados

O desempenho do sistema automatizado foi avaliado retrospectivamente mediante a comparação de suas decisões com o padrão de referência humano previamente estabelecido na revisão sistemática original. O fluxo automatizado foi aplicado após a conclusão da revisão e não interferiu nas decisões dos revisores. Na deduplicação, as classificações automatizadas foram comparadas com o status de duplicidade registrado pelos revisores, definido a partir de título, DOI, PMID, PMCID, autoria, ano, periódico, resumo e identificadores de gerenciamento dos registros.

Na triagem de títulos e resumos, os registros retidos pelo sistema foram comparados com a lista de 118 registros selecionados pelos revisores humanos para recuperação do texto completo. No módulo exploratório de leitura de textos completos, a comparação foi restrita aos 90 PDFs efetivamente processados, utilizando-se a inclusão humana final como padrão de referência.

As comparações diretas com outras ferramentas foram restritas à deduplicação, etapa em que os resultados do Rayyan e do ASySD foram avaliados no mesmo corpus e contra o mesmo padrão humano. Ferramentas de aprendizado ativo, como ASReview e SWIFT-Active Screener, não foram comparadas

diretamente, pois dependem de interação iterativa com o revisor e de avaliação por ordenação, não produzindo uma classificação binária fixa diretamente equivalente.

Em todas as etapas, foi realizado o acompanhamento explícito dos 12 estudos finalmente incluídos na revisão, verificando se foram preservados após cada módulo do sistema automatizado.

7.9 Análise Estatística

As análises foram descritivas e realizadas por meio de rotinas desenvolvidas em Python, versão 3.10, executadas no ambiente Google Colaboratory. O desempenho do sistema automatizado foi quantificado a partir das matrizes de confusão específicas de cada etapa, utilizando-se as seguintes métricas:

- sensibilidade ou recall;
- especificidade;
- precisão ou valor preditivo positivo;
- valor preditivo negativo;
- escore F1;
- acurácia;
- coeficiente kappa de Cohen;
- índice de Jaccard;
- redução da carga de trabalho.

A redução da carga de trabalho foi calculada como a proporção de registros ou documentos removidos, excluídos ou não priorizados pelo sistema antes da avaliação humana correspondente, sendo interpretada como medida operacional, e não como evidência isolada de validade.

Os intervalos de confiança de 95% foram estimados de acordo com a natureza de cada medida. Para as proporções binomiais, foram utilizados intervalos pelo método de Wilson, com intervalos exatos de Clopper–Pearson apresentados secundariamente para proporções selecionadas. Para o escore F1, o kappa de Cohen, o índice de Jaccard e as diferenças pareadas entre os métodos de deduplicação, foram utilizados intervalos de confiança obtidos por bootstrap não paramétrico, com correção de viés e aceleração — BCa — e 10.000 reamostragens.

A reamostragem preservou a estrutura pareada dos registros ou agrupamentos, e os intervalos do kappa foram restringidos ao seu domínio teórico, de −1 a 1.

Para a deduplicação, foram realizadas análises no nível do registro individual e no nível dos agrupamentos definidos por título e ano, sendo a análise por agrupamento considerada o principal desfecho operacional. O sistema proposto foi comparado com o Rayyan e o ASySD, utilizando-se o mesmo corpus e o mesmo padrão de referência humano. Na triagem de títulos e resumos, o principal comparador foi a lista humana de registros selecionados para recuperação do texto completo. Para o módulo de leitura exploratória de textos completos, a análise foi restrita aos PDFs efetivamente processados, utilizando a inclusão humana final como padrão de referência. Não foi realizada análise de não inferioridade, pois se tratava de uma validação interna em um único tema, sem margem previamente estabelecida.

7.10 Aspectos Éticos

7.10.1 Riscos

O presente estudo não envolveu participantes humanos diretamente, não coletou dados pessoais ou clínicos individuais e não realizou intervenções de qualquer natureza. Os dados utilizados consistiram em registros bibliográficos e textos científicos empregados no contexto de uma revisão sistemática. Dessa forma, não houve riscos físicos, psicológicos, assistenciais ou de exposição de informações pessoais de pacientes.

Ainda assim, por se tratar de um estudo metodológico com uso de automação e inteligência artificial, foram reconhecidos riscos operacionais e científicos. Entre eles, destacam-se a possibilidade de classificação incorreta de registros, exclusão indevida de estudos potencialmente relevantes, inclusão excessiva de falsos positivos, falhas na extração de texto de arquivos em PDF, dependência da qualidade dos metadados bibliográficos e eventual inconsistência nas respostas dos modelos automatizados. Esses riscos poderiam afetar a validade metodológica da revisão caso não fossem monitorados.

Para minimizar esses riscos, o sistema foi desenvolvido com critérios explícitos de inclusão e exclusão, registro dos motivos de decisão automatizada, avaliação em múltiplas etapas e comparação direta com um padrão de referência humano. Além disso, os estudos finalmente incluídos na revisão sistemática foram rastreados ao

longo de todo o fluxo automatizado, permitindo verificar se algum estudo elegível havia sido perdido. O módulo de leitura de textos completos por IA foi tratado como exploratório e de apoio à decisão, não substituindo a avaliação humana final.

Assim, os riscos do estudo foram considerados mínimos, predominantemente metodológicos e operacionais, sem risco direto a participantes humanos. A principal medida de controle foi a manutenção da supervisão humana, da rastreabilidade das decisões e da validação do desempenho do sistema em comparação com o padrão humano.

7.10.2 Benefícios

Os resultados deste estudo poderão contribuir para o desenvolvimento de metodologias mais eficientes e confiáveis para a condução de revisões sistemáticas, reduzindo o tempo e os recursos necessários para a síntese de evidências científicas. A disponibilização de um sistema automatizado transparente e auditável pode beneficiar pesquisadores, instituições acadêmicas e, indiretamente, pacientes e profissionais de saúde que dependem de revisões sistemáticas atualizadas para embasar suas decisões clínicas.

Por se tratar de um estudo metodológico baseado exclusivamente em dados bibliográficos de domínio público, sem envolvimento de seres humanos, animais ou dados pessoais identificáveis, o estudo não requer submissão ao Comitê de Ética em Pesquisa, conforme a Resolução CNS nº 510/2016.

7.11 Disponibilidade dos dados, código e reprodutibilidade

Os dados utilizados na validação do fluxo automatizado foram disponibilizados publicamente no repositório Figshare, sob o DOI 10.6084/m9.figshare.32526549. Uma versão permanente e arquivada do código-fonte do SafeScreen, incluindo os scripts, parâmetros e materiais necessários à reprodução das análises, foi depositada no Zenodo, sob o DOI 10.5281/zenodo.21412862. A versão ativamente mantida do código está disponível no repositório público do projeto no GitHub (https://github.com/vrballotin/SafeScreen).

8. PRODUTO FINAL

Como produtos finais deste projeto, foram produzidos dois artigos científicos: o artigo principal, dedicado ao desenvolvimento e à validação do fluxo automatizado,

e o artigo complementar, referente à revisão sistemática sobre os desfechos perioperatórios associados ao uso de semaglutida. Adicionalmente, foi desenvolvido o SafeScreen, um protótipo de software integrado, modular e auditável para apoio à condução de revisões sistemáticas. O sistema reúne, em um único fluxo computacional, módulos sequenciais de deduplicação, triagem automatizada de títulos e resumos e priorização exploratória de textos completos por inteligência artificial. Implementado em Python, o protótipo gera arquivos de decisão e auditoria que permitem rastrear os critérios aplicados em cada etapa, mantendo a supervisão humana como componente central do processo. O código-fonte, os parâmetros e a documentação foram disponibilizados publicamente para favorecer sua transparência, reprodutibilidade e desenvolvimento futuro.

8.1 Artigo principal

O artigo principal consiste em um estudo original de validação metodológica, descrevendo o desenvolvimento e a validação interna do sistema automatizado de triagem bibliográfica para revisões sistemáticas. O manuscrito incluiu os três módulos do sistema — deduplicação bibliográfica, triagem de títulos e resumos e leitura exploratória de textos completos — e reportou as métricas de desempenho obtidas em cada etapa.

8.1.1 Resultado geral do fluxo

O corpus bibliográfico operacional incluiu 16.159 registros. Após a etapa automatizada de deduplicação, 2.788 registros foram removidos como duplicatas, resultando em 13.371 registros únicos encaminhados à triagem de títulos e resumos. Nessa etapa, o sistema automatizado classificou 123 registros como potencialmente elegíveis e excluiu 13.248 registros. Entre os 123 registros incluídos pela IA na triagem inicial, 90 PDFs foram localizados e avaliados no módulo exploratório de leitura de texto completo, enquanto 33 registros não foram encontrados em formato PDF para essa etapa. A revisão sistemática realizada pelos revisores humanos selecionou 118 artigos para avaliação em texto completo, dos quais 12 compuseram a amostra final incluída na revisão sistemática. O fluxo operacional completo está

apresentado na Tabela 3 e no fluxograma da Figura 3 e as métricas de desempenhos específicas por etapas na Tabela 4.

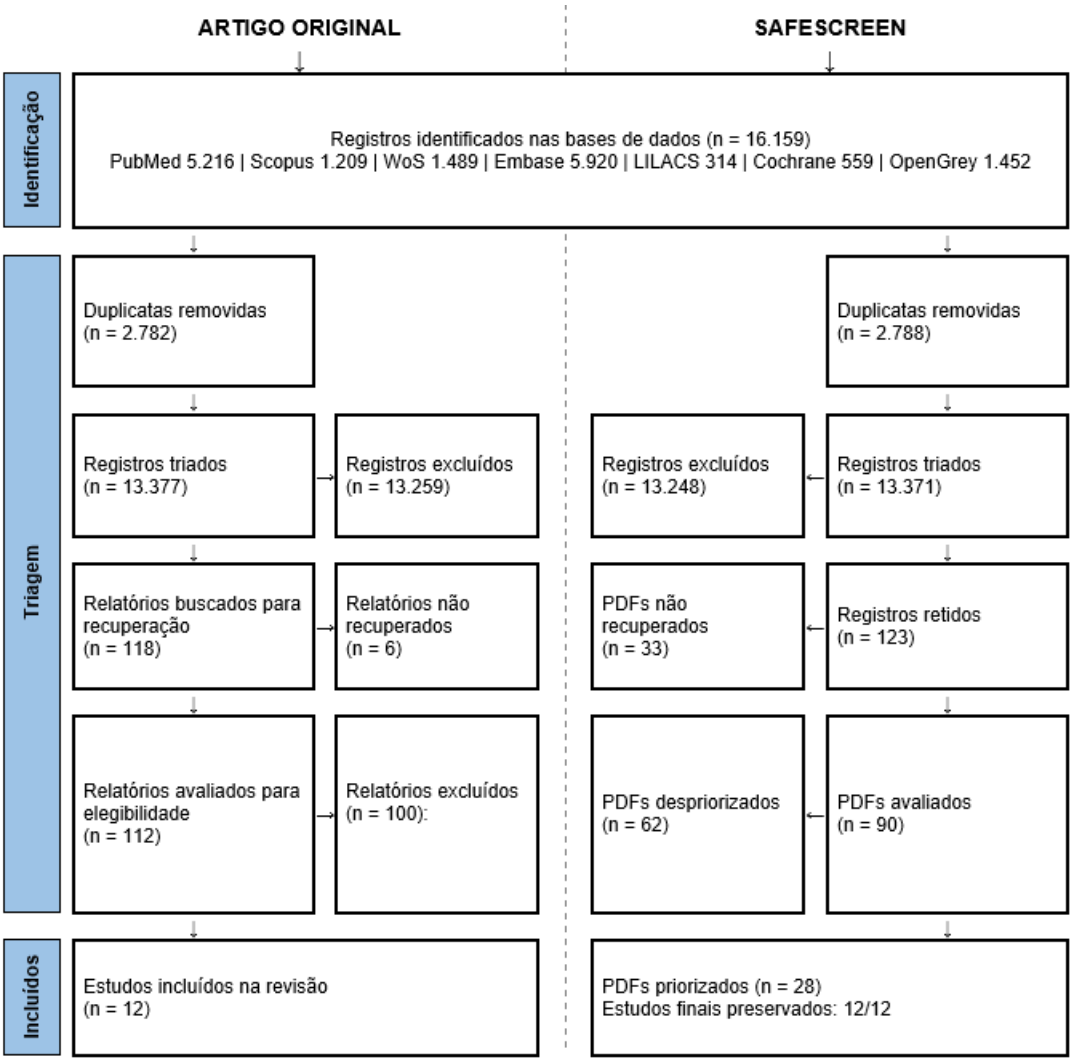
Tabela 3 – Fluxo operacional

Etapa	Resultado
Registros iniciais considerados no fluxo PRISMA operacional	16.159
Duplicatas removidas pelo fluxo automatizado SysRev	2.788
Registros após deduplicação e enviados à triagem título/resumo	13.371
Registros incluídos pela IA na triagem título/resumo	123
Registros excluídos pela IA na triagem título/resumo	13.248
Registros incluídos pela IA sem PDF localizado para leitura automatizada	33
PDFs avaliados pela IA no módulo de texto completo	90
PDFs incluídos pela IA no módulo de texto completo	28
PDFs excluídos pela IA no módulo de texto completo	62
Artigos selecionados pelo humano para avaliação em texto completo	118
Estudos finalmente incluídos pelo humano	12

Fonte: Dados da pesquisa.

Nota: PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses; SysRev = systematic review – módulo de deduplicação; IA = inteligência artificial; PDF = Portable Document Format. – Fluxogram prisma

Figura 3. Fluxograma PRISMA comparativo da seleção de estudos: artigo original versus SafeScreen.



Fonte: adaptado de Page MJ et al. BMJ. 2021;372:n71. doi: 10.1136/bmj.n71.

Tabela 4 – Métricas de desempenho específicas por etapa para deduplicação e triagem automatizadas.

(continua)

Etapa/comparação	Nível	Sensibilidade/recall	Especificidade	Precisão/VPP	VPN	Escore F1	Acurácia	Kappa de Cohen	Outros resultados
Fluxo de deduplicação proposto	Registro	0,800 (0,785–0,815)	0,958 (0,954–0,961)	0,798 (0,783–0,813)	0,958 (0,955–0,962)	0,799 (0,787–0,810)	0,931 (0,927–0,935)	0,757 (0,743–0,770)	Jaccard 0,666 (0,649–0,681)
Fluxo de deduplicação proposto	Agrupamento título–ano	0,933 (0,921–0,943)	0,990 (0,988–0,992)	0,941 (0,930–0,951)	0,989 (0,987–0,990)	0,937 (0,929–0,944)	0,982 (0,980–0,984)	0,927 (0,917–0,935)	Jaccard 0,882 (0,867–0,895)
Deduplicação pelo Rayyan	Registro	0,651 (0,633–0,668)	0,928 (0,924–0,933)	0,654 (0,636–0,671)	0,927 (0,923–0,932)	0,652 (0,637–0,667)	0,881 (0,875–0,885)	0,580 (0,563–0,597)	Jaccard 0,484 (0,468–0,500)
Deduplicação pelo Rayyan	Agrupamento título–ano	0,899 (0,885–0,911)	0,978 (0,975–0,980)	0,871 (0,856–0,885)	0,983 (0,980–0,985)	0,885 (0,874–0,895)	0,966 (0,963–0,969)	0,865 (0,853–0,877)	Jaccard 0,793 (0,776–0,809)
Deduplicação pelo ASySD	Registro	0,250 (0,234–0,266)	0,970 (0,967–0,973)	0,637 (0,608–0,665)	0,861 (0,856–0,867)	0,359 (0,340–0,378)	0,846 (0,841–0,852)	0,290 (0,270–0,310)	Jaccard 0,219 (0,205–0,233)
Deduplicação pelo ASySD	Agrupamento título–ano	0,411 (0,389–0,432)	0,983 (0,980–0,985)	0,800 (0,775–0,824)	0,908 (0,903–0,913)	0,543 (0,521–0,564)	0,900 (0,895–0,905)	0,493 (0,471–0,516)	Jaccard 0,373 (0,353–0,393)
Triagem de títulos e resumos versus lista humana de recuperação de textos completos	Agrupamento título–ano	0,703 (0,616–0,778)	0,997 (0,996–0,998)	0,675 (0,588–0,751)	0,997 (0,996–0,998)	0,689 (0,618–0,753)	0,994 (0,993–0,996)	0,686 (0,615–0,749)	Jaccard 0,525 (0,448–0,602); redução da carga de trabalho de 99,1%; estudos finalmente incluídos preservados: 12/12

Tabela 4 – Métricas de desempenho específicas por etapa para deduplicação e triagem automatizadas.

(conclusão)

Etapa/comparação	Nível	Sensibilidade/recall	Especificidade	Precisão/ VPP	VPN	Escore F1	Acurácia	Kappa de Cohen	Outros resultados
IA exploratória para textos completos versus análise de segurança da inclusão humana final	90 PDFs processados	1,000 (0,758–1,000)	0,795 (0,692–0,870)	0,429 (0,265–0,609)	1,000 (0,942–1,000)	0,600 (0,387–0,762)	0,822 (0,731–0,888)	0,508 (0,308–0,684)	Jaccard 0,429 (0,265–0,609); 62/90 PDFs não priorizados (68,9%); FN = 0; estudos finalmente incluídos preservados: 12/12

Nota: VPP = valor preditivo positivo; VPN = valor preditivo negativo; FN = falso negativo; IA = inteligência artificial; PDF = Portable Document Format. Os valores correspondem às estimativas pontuais com intervalos de confiança de 95%. Foram utilizados intervalos de Wilson para as métricas de proporção binomial. Para o escore F1 e o kappa de Cohen, foram utilizados intervalos bootstrap com correção de viés e aceleração (BCa), com 10.000 reamostragens. O denominador do módulo de IA para textos completos foi restrito aos 90 PDFs efetivamente processados pelo módulo.
Fonte: Dados da pesquisa (2026)

8.1.2 Deduplicação bibliográfica

Na avaliação da deduplicação em nível de registro, o fluxo automatizado SysRev (módulo de deduplicação do SafeScreen) apresentou sensibilidade de 0,800, especificidade de 0,958, precisão de 0,798, escore F1 de 0,799, acurácia de 0,931 e kappa de Cohen de 0,757. Na análise em nível de agrupamento título–ano, considerada mais representativa da tarefa operacional de deduplicação bibliográfica, o desempenho foi superior, com sensibilidade de 0,933, especificidade de 0,990, precisão de 0,941, escore F1 de 0,937, acurácia de 0,982 e kappa de Cohen de 0,927. Esses achados indicam concordância quase perfeita entre o fluxo automatizado e o padrão de referência humano em nível de agrupamento. Detalhes nas tabelas 4 a 7.

Tabela 5 – Desempenho em nível de registro

Métrica	Resultado
TP	2.226
FP	562
TN	12.815
FN	556
Precisão	0,798
Sensibilidade	0,800
Especificidade	0,958
F1-score	0,799
Acurácia	0,931
Kappa	0,757
IC 95% da sensibilidade	0,785–0,815
IC 95% da especificidade	0,954–0,961

Fonte: Dados da pesquisa.

Nota: TP = verdadeiros positivos, correspondentes aos registros duplicados corretamente identificados; FP = falsos positivos, registros únicos classificados incorretamente como duplicados; TN = verdadeiros negativos, registros únicos corretamente preservados; FN = falsos negativos, registros duplicados não identificados pelo sistema; F1-score = média harmônica entre precisão e sensibilidade; Kappa = coeficiente kappa de Cohen, que expressa a concordância ajustada pelo acaso; IC 95% = intervalo de confiança de 95%, calculado pelo método de Wilson para sensibilidade e especificidade.

Tabela 6 – Desempenho em nível de agrupamento

Métrica	Resultado
TP	1.876
FP	117
TN	11.836
FN	135
Precisão	0,941
Sensibilidade	0,933
Especificidade	0,990
F1-score	0,937
Acurácia	0,982
Kappa	0,927
IC 95% da sensibilidade	0,921–0,943
IC 95% da especificidade	0,988–0,992

Fonte: Dados da pesquisa.
Nota: TP = verdadeiros positivos, correspondentes aos registros duplicados corretamente identificados; FP = falsos positivos, registros únicos classificados incorretamente como duplicados; TN = verdadeiros negativos, registros únicos corretamente preservados; FN = falsos negativos, registros duplicados não identificados pelo sistema; F1-score = média harmônica entre precisão e sensibilidade; Kappa = coeficiente kappa de Cohen, que expressa a concordância ajustada pelo acaso; IC 95% = intervalo de confiança de 95%, calculado pelo método de Wilson para sensibilidade e especificidade.

Tabela 7 – Sensibilidade e especificidade do fluxo automatizado de remoção de duplicatas em comparação com o padrão humano

Nível de validação	Verdadeiros negativos, n	Falsos negativos, n	Falsos positivos, n	Verdadeiros positivos, n	Sensibilidade, % (IC 95%)	Especificidade, % (IC 95%)
Análise em nível de registro	12.815	556	562	2.226	80,0 (78,5–81,5)	95,8 (95,4–96,1)
Análise em nível de agrupamento	11.836	135	117	1.876	93,3 (92,1–94,3)	99,0 (98,8–99,2)

Fonte: Dados da pesquisa.
Nota: IC = intervalo de confiança;

8.1.2.1 Mapeamento da deduplicação

Para permitir a comparação entre a deduplicação automatizada e o padrão humano, os registros removidos como duplicatas foram pareados por chaves bibliográficas disponíveis. No conjunto de duplicatas identificado pelo padrão humano, 1.688 registros foram pareados por DOI, 7 por PMID e 1.087 por combinação de título e ano, sem registros não pareados. No conjunto removido pelo sistema automatizado, os 2.788 registros apresentaram chave de pareamento disponível, também sem registros não pareados. Esse resultado indica que a comparação entre

os conjuntos pôde ser realizada de forma completa e rastreável. O mapeamento está resumido na tabela 8.

Tabela 8 – Mapeamento da remoção de duplicatas

Fonte	Método de pareamento	n
Padrão humano	DOI	1.688
Padrão humano	PMID	7
Padrão humano	título + ano	1.087
Padrão humano	não pareado	0
SysRev	chave direta	2.788
SysRev	não pareado	0

Fonte: Dados da pesquisa.
Nota: DOI = Digital Object Identifier; PMID = PubMed Identifier; SysRev = fluxo automatizado de revisão sistemática – módulo de deduplicação.

8.1.3 Triagem automatizada de títulos e resumos

8.1.3.1 Resultado global da triagem IA

Na triagem de títulos e resumos, o sistema automatizado avaliou 13.371 registros após deduplicação. Destes, 123 foram classificados como potencialmente elegíveis e encaminhados para avaliação em texto completo, enquanto 13.248 foram excluídos. Assim, o fluxo automatizado reduziu em 99,08% o número de registros que exigiriam avaliação humana inicial, mantendo todos os estudos que compuseram a amostra final da revisão. Detalhes nas tabelas 9 a 13.

Tabela 9 – Resultado global da triagem IA

Resultado	n	%
Registros triados após deduplicação	13.371	100%
Incluídos pela IA	123	0,92%
Excluídos pela IA	13.248	99,08%

Fonte: Dados da pesquisa.
Nota: IA = inteligência artificial.

Tabela 10 – Funcionamento dos critérios

Critério / flag	Registros positivos
Gate 1 — termo GLP-1/semaglutida	1.386
Gate 2 — contexto perioperatório	8.318
Gate 3 — domínio de desfecho	2.097
Semaglutida no título	858
Resumo ausente/vazio	6.142
Estudo não original	1.409
Marcador pediátrico	571
Estudo animal	550
Outro GLP-1 sem semaglutida	60
Marcador de não inglês	0

Fonte: Dados da pesquisa.
Nota: Gate = critério operacional de triagem automatizada; GLP-1 = glucagon-like peptide-1.

Tabela 11 – Domínios de desfecho detectados

Domínio	Registros com domínio positivo
Conteúdo gástrico residual/esvaziamento gástrico	697
Aspiração/regurgitação	296
Complicações de via aérea	90
Suspensão/retirada de GLP-1	24
Complicações perioperatórias/respiratórias	1.283

Fonte: Dados da pesquisa.
Nota: GLP-1 = glucagon-like peptide-1.

Tabela 12 – Domínios entre os 123 incluídos pela IA

Domínio	n entre incluídos pela IA
Conteúdo gástrico residual	86
Aspiração/regurgitação	34
Complicações de via aérea	7
Suspensão/retirada de GLP-1	11
Complicações perioperatórias/respiratórias	24

Fonte: Dados da pesquisa.
Nota: Os domínios não são mutuamente exclusivos; um mesmo registro pode apresentar mais de um domínio. IA = inteligência artificial; GLP-1 = glucagon-like peptide-1.

Tabela 13 – Motivos de exclusão na triagem automatizada

Motivo	n
Ausência de termo GLP-1/semaglutida	11.985
Ausência de domínio de desfecho	11.271
Ausência de contexto perioperatório	5.053
Tipo de estudo não original	1.409
Marcador pediátrico	571
Estudo animal	550
Outro GLP-1 sem semaglutida	60

Fonte: Dados da pesquisa.
Nota: Os motivos de exclusão não são mutuamente exclusivos; um mesmo registro pode ter sido classificado em mais de uma categoria. GLP-1 = glucagon-like peptide-1.

Entre os registros excluídos pela IA, os motivos mais frequentes foram ausência de termo relacionado a GLP-1/semaglutida, ausência de domínio de desfecho relevante e ausência de contexto perioperatório. Também foram aplicadas regras duras de exclusão para estudos não originais, estudos animais, marcadores pediátricos e registros relacionados exclusivamente a outros agonistas GLP-1 sem dados específicos de semaglutida. Como múltiplos motivos de exclusão podiam ser atribuídos ao mesmo registro, essas categorias não foram mutuamente exclusivas.

8.1.3.2 Modelo de resgate por aprendizado de máquina

A classificação primária por regras identificou 120 registros como potencialmente elegíveis. O modelo secundário de aprendizado de máquina, aplicado apenas a registros limítrofes que preencheram os dois primeiros critérios, mas falharam no domínio de desfecho, resgatou 3 registros adicionais. Assim, a decisão final da IA incluiu 123 registros para avaliação em texto completo. Detalhes na tabela 14.

Tabela 14 – Modelo de resgate por aprendizado de máquina.

Resultado	n
Incluídos por regra direta	120
Resgatados pelo modelo de ML	3
Total incluído pela IA	123

Fonte: Dados da pesquisa.
Nota: IA = inteligência artificial. ML = machine learning.

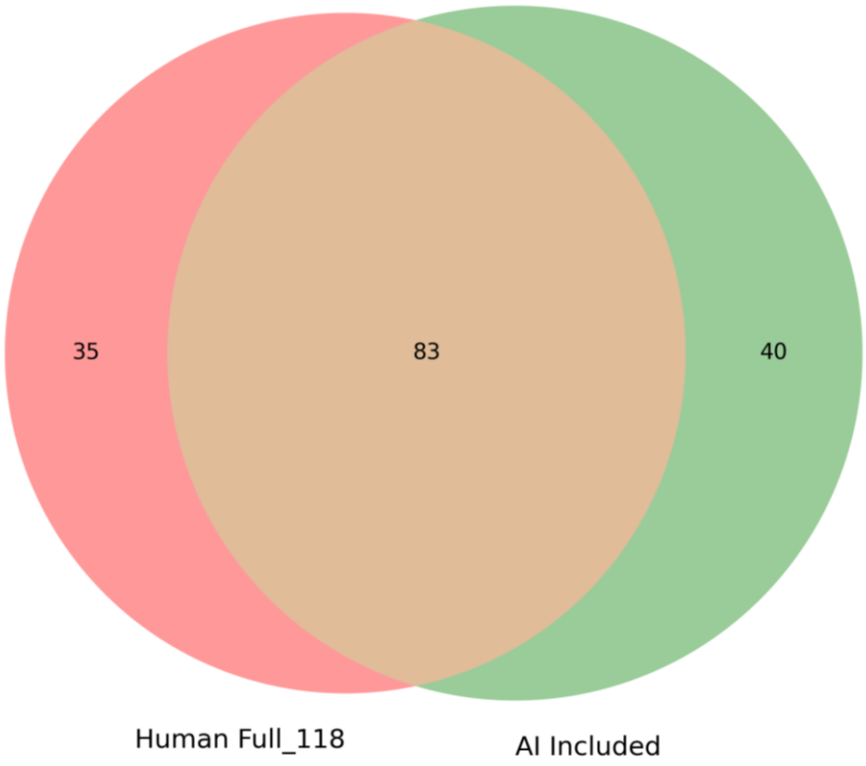
8.1.3.3 Concordância com a lista humana de 118 registros selecionados para recuperação do texto completo

Na comparação entre os 123 registros retidos pela triagem automatizada e os 118 registros selecionados pelos revisores humanos para recuperação do texto completo, 83 registros foram selecionados por ambos os métodos. Quarenta registros foram retidos exclusivamente pelo sistema automatizado, enquanto 35 foram selecionados apenas pelos revisores humanos (Figura 4). Considerando a lista humana de 118 registros como padrão de referência, essa comparação resultou em sensibilidade de 0,703, precisão de 0,675, escore F1 de 0,689 e índice de Jaccard de 0,525, conforme apresentado nas Tabelas 15 e 16.

Entre os 83 registros retidos por ambos os métodos, 12 corresponderam aos estudos finalmente incluídos e 71 foram excluídos em etapas posteriores da avaliação

humana. Embora os 35 registros selecionados apenas pelos revisores tenham sido classificados como falsos negativos nessa comparação, nenhum deles pertencia ao conjunto final de 12 estudos. Portanto, apesar das divergências em relação à lista humana ampliada, todos os estudos finalmente incluídos foram preservados pelo fluxo automatizado. Os 71 registros retidos simultaneamente pelo sistema e pelos revisores humanos, mas não incluídos no conjunto final, estão apresentados no Apêndice C.

Figura 4. Sobreposição entre a seleção humana para texto completo e os registros retidos pela triagem automatizada de títulos e resumos.



Fonte: Dados da pesquisa.

Nota: A figura apresenta a sobreposição entre os 118 registros selecionados pelos revisores humanos para recuperação do texto completo e os 123 registros classificados como potencialmente elegíveis pela IA na triagem de títulos e resumos. A área central representa os 83 registros presentes em ambas as listas. A área exclusiva da seleção humana representa 35 registros encaminhados pelos revisores humanos, mas não selecionados pela IA. A área exclusiva da IA representa 40 registros selecionados pela IA, mas não encaminhados pelos revisores humanos para texto completo. Esta análise avalia a concordância entre as listas amplas de triagem, não a inclusão final dos estudos na revisão sistemática.

Tabela 15 – Concordância entre a triagem automatizada e a lista humana de 118 registros selecionados para recuperação do texto completo

Métrica	Resultado
Registros na lista humana ampliada	118
Registros incluídos pela IA, total bruto	123
Sobreposição entre IA e humano	83
Selecionados pelo humano, mas não pela IA	35
Selecionados pela IA, mas não pelo humano	40
Sensibilidade / recall	0,703
Precisão / PPV	0,675
F1-score	0,689
Índice de Jaccard	0,525
Estudos finais capturados pela IA	12/12

Fonte: Dados da pesquisa.
Nota: IA = inteligência artificial; PPV = positive predictive value. A análise utilizou pareamento conservador um-para-um entre registros humanos e registros incluídos pela IA.

Tabela 16 – Checagem dos 35 registros selecionados pelos revisores humanos, mas não retidos pela triagem automatizada

(continua)

Nº	Artigos	Critério/Domínios	Motivo provável para não inclusão pela IA
1	Effect of bariatric surgery on cardio-psycho-metabolic outcomes in severe obesity: a randomized controlled trial	Sem específico IC3	Não continha termos de semaglutida/GLP-1RA no título/resumo; estudo sobre cirurgia bariátrica e desfechos metabólicos.
2	Adjustable Gastric Band Surgery or Medical Management in Patients With Type 2 Diabetes: a Randomized Clinical Trial	Sem específico IC3	Não continha termos de semaglutida/GLP-1RA no título/resumo; estudo comparou cirurgia bariátrica e manejo clínico do diabetes.
3	Increased volume of gastric contents in diabetic patients undergoing renal transplantation: lack of effect with cisapride	IC3A, IC3B	Estudo relevante para volume gástrico/risco de aspiração em diabéticos, mas sem exposição a semaglutida/GLP-1RA; provável falha do Gate 1.
4	The impact of Hafnia alvei HA4597™ on weight loss and glycaemic control after bariatric surgery	Sem específico IC3	Protocolo de probiótico após cirurgia bariátrica; sem semaglutida/GLP-1RA e sem desfechos de RGC/aspiração.
5	Ultrasound-guided stellate ganglion block accelerates postoperative gastrointestinal function recovery following laparoscopic radical gastrectomy for gastric cancer	IC3A, IC3E	Desfecho gastrointestinal pós-operatório, mas intervenção anestésica/bloqueio do gânglio estrelado, sem semaglutida/GLP-1RA.

Tabela 16 – Checagem dos 35 registros selecionados pelos revisores humanos, mas não retidos pela triagem automatizada

(continuação)

Nº	Artigos	Critério/Domínios	Motivo provável para não inclusão pela IA
6	Gastric Volume in Diabetic Patients after Overnight Fasting vs Clear Liquid Ingestion Two Hours before Surgery Using Ultrasonography	IC3A, IC3B	Avalia volume gástrico perioperatório em diabéticos, mas não exposição a semaglutida/GLP-1RA.
7	Ultrasound-guided assessment of gastric residual volume in patients receiving three types of clear fluids	IC3A	Avalia volume gástrico residual após líquidos claros; sem semaglutida/GLP-1RA.
8	Perioperative Versus Postoperative Glycemia Control in Cardiac Surgery Patients	Sem IC3 específico	Controle glicêmico em cirurgia cardíaca, sem semaglutida/GLP-1RA e sem domínio de RGC/aspiração.
9	Perioperative Considerations for Patients on GLP1 Agonists	IC3A, IC3B, IC3C, IC3D	Texto de considerações perioperatórias, provavelmente não original/narrativo; resultados em GLP-1RA como classe, sem extração específica de semaglutida.
10	Glucagon-Like Peptide-1 agonists in perioperative medicine: to suspend or not to suspend, that is the question	IC3C, IC3D	Sem resumo/metadados suficientes e provável comentário/opinião; GLP-1RA como classe, sem dados originais extraíveis.
11	Aspiration under anesthesia: what happens after we sound the GLP-1 receptor agonist alarm?	IC3B, IC3C	Comentário/editorial sem resumo e sem dados originais extraíveis; GLP-1RA como classe.
12	Perioperative Management of Patients with Diabetes and Cancer: Challenges and Opportunities	Sem IC3 específico	Revisão/visão geral sobre diabetes e câncer no perioperatório; sem foco em semaglutida/GLP-1RA.
13	Excluded stomach and biliopancreatic limb distention in patients treated with GLP-1 receptor agonists after Roux-en-Y gastric bypass: a case series	IC3A, IC3E	Série de casos após bypass com GLP-1RA como classe; sem separação específica de semaglutida e contexto não diretamente perioperatório de anestesia.

Tabela 16 – Checagem dos 35 registros selecionados pelos revisores humanos, mas não retidos pela triagem automatizada

(continuação)

Nº	Artigos	Critério/Domínios	Motivo provável para não inclusão pela IA
14	Moving forward: long-term outcomes of gastric emptying in diabetic gastroparesis patients managed medically at a GI motility center	IC3A	Gastroparesia diabética, mas sem exposição a semaglutida/GLP-1RA e sem contexto perioperatório.
15	GLP-1 agonists increase surgical risk	IC3B, IC3C	Provável comentário/opinião; sem resumo/dados originais e GLP-1RA como classe.
16	GLP-1 Receptor Agonists Do Not Increase Aspiration During Upper Endoscopy in Patients With Diabetes	IC3B, IC3E	Estudo de GLP-1RA como classe; provável ausência de resultados separados para semaglutida.
17	GLP-1 Receptor Agonists Use before Endoscopy Is Associated with Low Retained Gastric Contents	IC3A, IC3D	Estudo de GLP-1RA como classe em endoscopia; provável ausência de resultados separados para semaglutida.
18	Anesthesia and GLP-1 receptor agonists: proceed with caution!	IC3A, IC3B, IC3C	Editorial/comentário; sem dados originais extraíveis e sem separação de semaglutida.
19	Semaglutide as an adjunct for weight loss after bariatric surgery	Sem IC3 específico	Semaglutida presente, mas objetivo foi perda ponderal após bariátrica; sem desfechos perioperatórios de RGC/aspiração/anestesia.
20	Real fasting times and incidence of pulmonary aspiration in children	IC3B	Estudo pediátrico sobre jejum/aspiração; sem semaglutida/GLP-1RA, e população pediátrica era critério de exclusão da IA.
21	GLP-1 Receptor Agonists Increase Solid Gastric Residue Rates on Upper Endoscopy Especially in Patients With Complicated Diabetes	IC3A	GLP-1RA como classe e ausência de resumo completo no arquivo; sem confirmação de resultados específicos para semaglutida.
22	Preoperative continuation of GLP-1 receptor agonists. Response to Br J Anaesth 2024; 133: 437–8	IC3C, IC3D	Resposta/carta ao editor; não original e GLP-1RA como classe, sem dados extraíveis.
23	Practice guidelines for preoperative fasting and pharmacologic agents to reduce pulmonary aspiration risk	IC3B, IC3C	Diretriz de jejum pré-operatório; não original e sem semaglutida/GLP-1RA.

Tabela 16 – Checagem dos 35 registros selecionados pelos revisores humanos, mas não retidos pela triagem automatizada

(continuação)

Nº	Artigos	Critério/Domínios	Motivo provável para não inclusão pela IA
24	The Use of Semaglutide in Patients With Renal Failure—A Retrospective Cohort Study	Sem IC3 específico	Semaglutida presente, mas estudo renal/metabólico; sem contexto perioperatório ou domínio RGC/aspiração.
25	Effects of GLP-1 and Other Gut Hormone Receptors on the Gastrointestinal Tract and Implications in Clinical Practice	IC3A	Revisão sobre efeitos gastrointestinais de incretinas; não original e sem contexto perioperatório específico.
26	Obesity, heart failure with preserved ejection fraction, and the role of GLP-1 receptor agonists	Sem IC3 específico	Revisão cardiometabólica/HFpEF; sem contexto perioperatório ou desfechos de RGC/aspiração.
27	Safe Continuation of GLP-1 Receptor Agonists at Endoscopy: A Case Series of 57 Adults Undergoing Endoscopic Sleeve Gastroplasty	IC3A, IC3D, IC3E	Série de casos de endoscopia com GLP-1RA como classe; provável ausência de separação para semaglutida.
28	Real-world clinical evidence of oral semaglutide on metabolic abnormalities in type 2 diabetes	Sem IC3 específico	Semaglutida oral em diabetes tipo 2; sem contexto perioperatório/anestésico e sem RGC/aspiração.
29	Perioperative Considerations for Patients on Semaglutide	IC3A, IC3B, IC3C, IC3D	Revisão/considerações perioperatórias sobre semaglutida; excluída pela IA por desenho não original, apesar de ser relevante para discussão.
30	Rates of GLP-1 receptor agonist use and aspiration events associated with anesthesia at a Canadian academic teaching centre	IC3B, IC3E	GLP-1RA como classe; sem resumo completo no arquivo e provável ausência de resultados específicos para semaglutida.
31	Preoperative continuation of GLP-1 receptor agonists. Response to Br J Anaesth 2024; 133: 437–8	IC3C, IC3D	Resposta/carta ao editor; não original e GLP-1RA como classe, sem dados extraíveis.
32	Point-of-Care Ultrasound Aids in the Management of Patient Taking Semaglutide Before Surgery: A Case Report	IC3A, IC3B, IC3C	Relato de caso de semaglutida antes da cirurgia; provavelmente excluído pela regra da IA para relato de caso.

Tabela 16 – Checagem dos 35 registros selecionados pelos revisores humanos, mas não retidos pela triagem automatizada

(conclusão)

Nº	Artigos	Critério/Domínios	Motivo provável para não inclusão pela IA
33	Risk of pulmonary aspiration during semaglutide use and anesthesia in a fasting patient: a case report with tomographic evidence	IC3A, IC3B, IC3C	Relato de caso de aspiração em uso de semaglutida; provavelmente excluído pela regra da IA para relato de caso.
34	Aspiration risk and strategic approach for patients receiving GLP-1 receptor agonists undergoing elective surgery	IC3B, IC3C, IC3D	Revisão/opinião estratégica; GLP-1RA como classe e sem dados originais extraíveis.
35	Perioperative GLP-1 receptor agonist use and retained gastric contents: A retrospective analysis of patients undergoing elective upper endoscopy	IC3A, IC3E	Análise retrospectiva de GLP-1RA como classe; sem resumo completo no arquivo e provável ausência de resultados específicos para semaglutida.

Nota: IC3A = conteúdo gástrico residual/esvaziamento gástrico/gastroparesia; IC3B = aspiração; IC3C = via aérea/conduta anestésica; IC3D = suspensão/continuação perioperatória de GLP-1RA; IC3E = complicações perioperatórias/endoscópicas ou eventos adversos.

8.1.3.4 Rastreamento dos 12 estudos finais

A análise de rastreamento dos estudos finais demonstrou que todos os 12 estudos incluídos na revisão sistemática foram preservados pelo fluxo automatizado. Os 12/12 estudos estavam presentes na lista de registros incluídos pela IA na triagem de títulos e resumos e também no conjunto humano de 118 artigos encaminhados para leitura em texto completo. O pareamento ocorreu por título exato em 7 estudos e por DOI em 5 estudos. Não houve perda de nenhum estudo finalmente incluído ao longo da triagem automatizada.

Tabela 17 – Rastreamento dos 12 estudos finais

Ponto de verificação	Resultado
Estudos finais no Full_final_12.xlsx	12
Presentes na lista incluída pela IA	12/12
Presentes no conjunto humano Full_118	12/12
Pareados por título exato	7/12
Pareados por DOI	5/12

Fonte: Dados da pesquisa.

Nota: Full_final_12.xlsx corresponde ao arquivo de referência com os 12 estudos finalmente incluídos pela revisão humana. Full_118 corresponde ao conjunto humano ampliado de artigos encaminhados para avaliação em texto completo. DOI = Digital Object Identifier. O

pareamento por título exato e por DOI foi utilizado para confirmar a presença dos estudos finais nos conjuntos avaliados.

8.1.4 Módulo exploratório de leitura de texto completo por IA

Entre os 123 registros classificados como potencialmente elegíveis pela IA na triagem de títulos e resumos, 90 artigos foram localizados em formato PDF e submetidos ao módulo exploratório de leitura de texto completo. Os 33 registros restantes não estavam disponíveis em formato PDF no momento da execução do código. Embora tenha sido realizada tentativa de obtenção dos textos completos por contato com os autores correspondentes, não houve retorno em tempo hábil para inclusão desses artigos na etapa automatizada de leitura integral. A avaliação automatizada dos 90 textos completos resultou na classificação de 28 artigos como incluídos e 62 como excluídos, correspondendo, respectivamente, a 31,1% e 68,9% dos PDFs avaliados. Dos 90 PDFs processados, 38 (42,2%) apresentaram texto extraído superior ao limite operacional de 35.000 caracteres e, portanto, tiveram apenas os primeiros 35.000 caracteres submetidos ao modelo. Os outros 52 documentos (57,8%) foram processados integralmente.

Após a classificação dos textos completos, os motivos de exclusão foram analisados conforme os critérios operacionais previamente definidos. Entre os 62 artigos excluídos pela IA, o motivo mais frequente foi EC4, correspondente a estudos não originais, revisões, editoriais ou documentos sem dados próprios, identificado em 37 artigos. O segundo motivo mais frequente foi EC3, relacionado à ausência de semaglutida ou à impossibilidade de identificar dados específicos para semaglutida, observado em 24 artigos. Apenas um artigo foi excluído por não atender ao critério IC1.

Entre os 28 artigos classificados como incluídos pela IA, 15 foram enquadrados pelo critério IC3, relacionado à presença de desfechos perioperatórios de interesse, enquanto 13 foram classificados como atendendo aos critérios gerais de inclusão. Esses resultados indicam que o módulo de leitura de texto completo foi capaz de aplicar os critérios do protocolo de forma estruturada, separando documentos com dados originais potencialmente relevantes daqueles sem elegibilidade para a revisão. Detalhes nas tabelas 18 a 20.

Para fins de transparência e rastreabilidade, o Apêndice B apresenta a classificação individual dos artigos avaliados pelo módulo exploratório de leitura de

texto completo, incluindo a decisão da IA, os critérios aplicados, as justificativas resumidas para inclusão ou exclusão e a lista dos artigos não encontrados em formato PDF para essa etapa.

Tabela 18 – Decisões da IA em texto completo.

Decisão	n
Incluído	28
Excluído	62
Total	90

Fonte: Dados da pesquisa.
Nota: Dos 123 registros incluídos pela IA na triagem de títulos e resumos, 90 foram localizados em formato PDF e avaliados no módulo de texto completo; 33 não foram encontrados para leitura integral.

Tabela 19 – Critérios de decisão usados pela IA

Critério de exclusão	n
EC4 — estudo não original ou sem dados próprios	37
EC3 — semaglutida ausente ou não específica	24
IC1 não atendido	1

Fonte: Dados da pesquisa.
Nota: EC = critério de exclusão; IC = critério de inclusão; EC4 = estudo não original ou sem dados próprios, incluindo revisão, editorial, comentário, diretriz, relato de caso, série de casos ou documento sem dados próprios de pacientes; EC3 = ausência de semaglutida como medicamento estudado ou ausência de dados específicos para semaglutida; IC1 = critério de população adulta.

Tabela 20 – Critérios de decisão usados pela IA entre incluídos

Critério de inclusão	n
IC3	15
IC	13

Fonte: Dados da pesquisa.
Nota: IC = critério de inclusão; IC3 = presença de pelo menos um desfecho perioperatório de interesse. Nesta tabela, IC3 não indica exclusão ou ausência de critério, mas sim que a IA registrou o desfecho perioperatório como principal critério de inclusão. IC indica que a IA classificou o artigo como incluído por atendimento global aos critérios do protocolo.

8.1.4.1 Comparação texto completo IA com o humano

No módulo exploratório de leitura de textos completos, a IA avaliou 90 PDFs, classificando 28 como incluídos e 62 como excluídos. A comparação com o padrão humano final demonstrou que os 12 estudos finalmente incluídos pela revisão sistemática estavam presentes entre os PDFs avaliados e foram corretamente classificados como incluídos pela IA. A matriz de confusão demonstrou 12 verdadeiros positivos, 16 falsos positivos, nenhum falso negativo e 62 verdadeiros negativos. A sensibilidade foi de 1,000, a especificidade de 0,795, a precisão de

0,429, o valor preditivo negativo de 1,000, o F1-score de 0,600, a acurácia de 0,822 e o kappa de Cohen de 0,508. A IA reduziu em 68,9% o número de textos completos que exigiriam revisão humana prioritária, sem excluir nenhum estudo finalmente incluído. Detalhes nas tabelas 21 e 22.

Tabela 21 – Comparação texto completo IA vs humano

	Humano incluiu	Humano excluiu	Total
IA incluiu	12	16	28
IA excluiu	0	62	62
Total	12	78	90

Fonte: Dados da pesquisa.
Nota: A tabela compara a decisão da IA no módulo exploratório de leitura de texto completo com o padrão humano final de inclusão da revisão sistemática. Os 12 estudos incluídos pelo humano foram corretamente classificados como incluídos pela IA. Os 16 casos em que a IA incluiu e o humano excluiu correspondem a falsos positivos. Não houve falsos negativos, pois nenhum estudo finalmente incluído pelo humano foi excluído pela IA.

Tabela 22 – Métricas IA x humano

Métrica	Resultado
Sensibilidade	1,000
Especificidade	0,795
Precisão/VPP	0,429
VPN	1,000
F1-score	0,600
Acurácia	0,822
Kappa de Cohen	0,508
Jaccard	0,429
Redução de carga de trabalho	0,689
Falsos positivos entre incluídos pela IA	16/28 = 57,1%
Falsos negativos	0

Fonte: Dados da pesquisa.
Nota: IA = inteligência artificial; VPP = valor preditivo positivo VPN = valor preditivo negativo. As métricas foram calculadas a partir da matriz de confusão do módulo de leitura de texto completo, considerando 90 PDFs avaliados. A redução da carga de trabalho corresponde à proporção de textos completos excluídos pela IA antes da revisão humana prioritária. Falsos positivos correspondem aos artigos classificados como incluídos pela IA, mas não pertencentes ao conjunto final de estudos incluídos pelo padrão humano. Falsos negativos correspondem aos estudos incluídos pelo humano, mas excluídos pela IA.

A comparação individual dos 90 PDFs processados, incluindo as decisões automatizadas, as decisões humanas, os critérios aplicados e a classificação de cada registro como verdadeiro positivo, falso positivo, verdadeiro negativo ou discordância identificada na auditoria, encontra-se apresentada no Apêndice D.

8.1.4.2 Perfil dos erros

Não foram observados falsos negativos entre os estudos finalmente incluídos. Todos os 12 estudos finais foram recuperados na triagem de títulos e resumos e mantidos como incluídos no módulo de leitura de textos completos após correção do pareamento bibliográfico. Esse achado é metodologicamente central, pois a principal ameaça de uma ferramenta automatizada em revisões sistemáticas é a exclusão precoce de estudos elegíveis. Detalhes na tabela 23.

Tabela 23 – Perfil dos erros

Etapa	Falsos negativos
Triagem título/resumo para os 12 estudos finais	0
Leitura de texto completo corrigida	0

Fonte: Dados da pesquisa.

8.1.4.3 Análise dos registros retidos ou priorizados que não integraram o conjunto final

O perfil de erro do sistema foi caracterizado predominantemente por falsos positivos, e não por falsos negativos. No módulo de leitura de texto completo, 16 artigos foram inicialmente classificados como incluídos pela IA, mas não pertenciam ao conjunto final de 12 estudos selecionados pelo padrão humano.

A checagem manual individual desses 16 artigos demonstrou que a principal razão para não inclusão no padrão humano final foi a ausência de dados específicos extraíveis para semaglutida. Em 9 artigos, os resultados foram apresentados para agonistas do receptor de GLP-1 como classe, sem separação específica da semaglutida; em outros 4, a extração quantitativa específica dos usuários de semaglutida não foi possível. Além disso, 1 artigo avaliou o desfecho fora do contexto perioperatório e 1 não reportou desfechos de interesse compatíveis com a revisão. Após verificação humana, um artigo inicialmente contabilizado como falso positivo foi considerado potencialmente elegível em auditoria posterior. Essa discordância sugeriu possível inconsistência na decisão humana original, mas não modificou o padrão de referência previamente estabelecido, composto por 12 estudos, nem as estimativas primárias de desempenho do sistema.

Esse comportamento é coerente com uma estratégia de triagem sensível e conservadora, na qual registros com potencial relevância são retidos para avaliação humana subsequente. A lista individual dos artigos inicialmente classificados como

falsos positivos, bem como o motivo da não inclusão ou a reclassificação após checagem humana, está apresentada nas Tabelas 24 e 25.

Tabela 24 – Registros retidos ou priorizados pela IA que não integraram o conjunto humano final

Etapa	Resultado
Registros retidos tanto pela IA quanto pelos revisores humanos, mas não incluídos no conjunto final — triagem de títulos e resumos	71
Artigos priorizados pela IA no módulo de texto completo, mas ausentes do conjunto humano final original	16
Resultado após checagem manual individual	15 falsos positivos confirmados e 1 discordância de auditoria considerada potencialmente elegível

Fonte: Dados da pesquisa.

Nota: Os 71 registros da etapa de títulos e resumos foram selecionados tanto pelo módulo automatizado quanto pelos revisores humanos para recuperação do texto completo, mas não integraram o conjunto final de 12 estudos. No módulo de texto completo, os 16 casos corresponderam a artigos inicialmente contabilizados como falsos positivos por não constarem no conjunto humano final original. Após checagem individual, 15 permaneceram classificados como falsos positivos e um foi considerado potencialmente elegível em auditoria posterior. O padrão de referência original e as métricas primárias foram mantidos.

Tabela 25 – Checagem manual dos 16 artigos classificados como incluídos pela IA e não pertencentes ao conjunto humano final

Nº	Artigos	Critério/ Domínios	Motivo para não inclusão
1	Fasting Duration, Not Timing of Last GLP-1 Dose, Predicts Aspiration Risk	IC3A	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
2	Ultrasound assessment of preoperative gastric volume in fasted diabetic patients	IC3A, IC3B, IC3D, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
3	Incidence of aspiration pneumonitis and emergent intubation in patients	IC3B, IC3C, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem

			separação específica da semaglutida
4	The impact of semaglutide on retained gastric contents	IC3A, IC3B	Não foi possível realizar extração quantitativa específica dos usuários de semaglutida
5	Risk of Aspiration Pneumonitis After Elective Esophagogastroduodenoscopy	IC3B, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
6	Association of glucagon-like peptide receptor 1 agonist therapy with perioperative outcomes	IC3A, IC3B, IC3C, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
7	Real-World Impact of GLP-1 Receptor Agonists on Endoscopic Patient Outcomes	IC3A, IC3B, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
8	Influence of semaglutide use on the presence of residual gastric solids	IC3A	Desfecho avaliado fora do contexto perioperatório
9	Glucagon-like peptide-1 receptor agonist use and perioperative outcomes	IC3B, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
10	Risk of Postoperative Aspiration Pneumonia in Patients Undergoing ACDF	IC3B, IC3E	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
11	Effect of discontinuing semaglutide prior to surgery on perioperative outcomes	IC3A, IC3B, IC3D	Não foi possível realizar extração quantitativa específica dos usuários de semaglutida

12	The impact of pre-procedural diet on the risk of residual gastric content	IC3A	Não foi possível realizar extração quantitativa específica dos usuários de semaglutida
13	Semaglutide and Postoperative Outcomes in Nondiabetic Patients	IC3E	Desfechos de interesse não reportados
14	Postoperative Aspiration Pneumonia Among Adults Using GLP-1 Receptor Agonists	IC3B, IC3E	Não foi possível realizar extração quantitativa específica dos usuários de semaglutida
15	Glucagon-like-peptide-1 receptor agonist use and the risk of pulmonary aspiration	IC3B	Os resultados foram apresentados para GLP-1RA como classe, sem separação específica da semaglutida
16	Impact of GLP-1 Receptor Agonists on Increased Residual Gastric Content	IC3A, IC3D	Reclassificado como potencialmente elegível após checagem humana.

Fonte: Dados da pesquisa.

Nota: Esta tabela apresenta os artigos inicialmente classificados como falsos positivos no módulo exploratório de leitura de texto completo. Foram 15 falsos positivos confirmados e 1 discordância de auditoria considerada potencialmente elegível; o padrão de referência original e as métricas primárias foram mantidos. GLP-1RAs = agonistas do receptor de GLP-1; IC3A = conteúdo gástrico residual/esvaziamento gástrico; IC3B = aspiração pulmonar ou bronquial; IC3C = complicações de via aérea; IC3D = suspensão ou intervalo de interrupção da semaglutida/GLP-1; IC3E = complicações perioperatórias gerais.

8.1.5 Estimativa de economia de tempo

Além da redução numérica da carga de trabalho, foi realizada uma estimativa exploratória do tempo potencialmente poupado pelo fluxo automatizado baseado na experiência profissional dos pesquisadores ao realizar a revisão sistemática. Como o tempo real da revisão humana original e da execução computacional não foi cronometrado prospectivamente, essa análise foi apresentada como estimativa operacional, baseada no volume de registros processados em cada etapa e em tempos médios assumidos para avaliação manual.

Na revisão humana original, estimou-se que a deduplicação bibliográfica demandaria aproximadamente 15,5 horas, considerando 20 segundos para avaliação de cada registro duplicado removido. A triagem manual de 13.371 registros após deduplicação, assumindo 1 minuto por registro, demandaria aproximadamente 222,9 horas. A leitura dos 118 artigos selecionados para avaliação em texto completo, considerando 20 minutos por artigo, demandaria aproximadamente 39,3 horas. Assim, o tempo humano total estimado para essas etapas seria de 277,7 horas.

Com o uso do fluxo automatizado, o tempo humano residual estimado seria de 11,4 horas, correspondente à conferência dos 123 registros priorizados na triagem de títulos e resumos e dos 28 PDFs priorizados no módulo de texto completo. O tempo computacional estimado para execução das etapas automatizadas variou entre 8 e 17 minutos. Dessa forma, no cenário adotado, o fluxo automatizado teria potencial para reduzir aproximadamente 266,3 horas de trabalho humano.

Como análise de sensibilidade, considerando um cenário mais conservador no qual o tempo necessário para a execução manual de cada etapa fosse 50% inferior ao estimado, a carga de trabalho humana total seria reduzida de 277,7 para aproximadamente 138,9 horas. Mesmo nesse cenário, substancialmente mais eficiente que o adotado na análise principal, a diferença em relação ao tempo requerido pelo fluxo automatizado permaneceria expressiva. Esse resultado sugere que a estimativa de economia de tempo não depende exclusivamente das premissas adotadas para a duração das tarefas manuais, mantendo-se relevante mesmo quando se assume uma velocidade de processamento humano duas vezes maior

Esses valores devem ser interpretados como estimativas exploratórias, e não como medidas cronometradas. A economia real pode variar conforme experiência dos revisores, complexidade dos registros, qualidade dos resumos, disponibilidade dos textos completos, desempenho computacional, latência da API e necessidade de consenso entre avaliadores. Os valores estimados encontram-se na Tabela 26.

Tabela 26 – Estimativa comparativa de tempo entre um cenário de processamento manual aplicado ao corpus automatizado e o fluxo assistido

Etapa	Volume utilizado no cenário manual	Tempo humano estimado no cenário manual	Saída após IA	Tempo Computacional estimado da IA	Tempo humano residual após IA	Economia potencial de trabalho humano
Deduplicação bibliográfica	2.788 duplicatas removidas	15,5 h	2.788 duplicatas removidas automaticamente	1-2 min	0 h	15,5 h
Triagem de títulos/resumos	13.371 registros	222,9 h	123 registros	2-5 min	2,1 h	220,8 h
Leitura de texto completo	118 artigos	39,3 h	28 PDFs priorizados	5-10min	9,3 h	30,0 h
Total	—	**277,7 h**	—	**8–17 min**	**11,4 h**	**266,3 h**

Fonte: Dados da pesquisa.
Nota: IA = inteligência artificial; PDF = Portable Document Format. A estimativa considerou 20 segundos por duplicata removida, 1 minuto por registro na triagem de títulos e resumos e 20 minutos por artigo/PDF na leitura de texto completo. Os tempos humanos e computacionais são estimativas operacionais e não representam cronometragem prospectiva. O tempo computacional pode variar conforme desempenho do ambiente Google Colab, latência da API, limites de requisição e características dos arquivos processados.

8.1.6 Custos

O estudo não envolveu custos assistenciais, coleta de dados clínicos, contratação de equipe externa ou aquisição de equipamentos laboratoriais. Os custos diretos estiveram relacionados principalmente à infraestrutura computacional e ao uso de ferramentas digitais necessárias para desenvolvimento, execução e validação do fluxo automatizado.

No módulo exploratório de leitura de textos completos, foi utilizada a API do Claude para avaliação automatizada dos PDFs. O código foi configurado para extrair até 35.000 caracteres por artigo e retornar uma decisão estruturada de inclusão ou exclusão. Considerando os 90 PDFs avaliados e a precificação por tokens do modelo utilizado, o custo direto estimado da API para essa etapa foi baixo, situando-se aproximadamente entre US$ 3,78 e US$ 4,32. Esse valor pode variar conforme o número real de tokens processados, o tamanho dos textos extraídos, o tamanho das respostas geradas e eventuais mudanças na precificação da plataforma.

Além da API, houve custos operacionais associados ao uso de ferramentas digitais, como a assinatura do Google Colab para ampliação da capacidade

computacional durante o processamento dos dados. Esses custos foram considerados despesas de infraestrutura digital do projeto.

Do ponto de vista econômico, o baixo custo direto da execução computacional contrasta com a economia potencial de tempo humano estimada nas etapas de deduplicação, triagem e leitura de textos completos. Assim, o sistema automatizado apresentou custo operacional reduzido em relação à carga de trabalho humano potencialmente poupada, especialmente nas etapas repetitivas e de alto volume da revisão sistemática. Detalhes na Tabela 27.

Tabela 27 – Estimativa de custo da API Claude Sonnet 4.6

Componente	Estimativa de tokens	Preço usado	Custo estimado
Entrada dos 90 PDFs	720.000–900.000 tokens	US$ 3 / 1 milhão	US$ 2,16–2,70
Saída dos 90 PDFs	até 108.000 tokens	US$ 15 / 1 milhão	até US$ 1,62
Total estimado da API Claude	—	—	US$ 3,78–4,32

Fonte: Dados da pesquisa.
Nota: API = Application Programming Interface, ou interface de programação de aplicações. A estimativa considerou 90 PDFs avaliados, até 35.000 caracteres por artigo, aproximadamente 8.000–10.000 tokens de entrada por PDF e até 1.200 tokens de saída por resposta. A precificação considerada foi de US$ 3,00 por milhão de tokens de entrada e US$ 15,00 por milhão de tokens de saída para Claude Sonnet 4.6.

8.1.7 Resultado final das hipóteses

As hipóteses principais foram sustentadas pelos resultados obtidos. O módulo de deduplicação apresentou desempenho elevado, especialmente na análise em nível de agrupamento, indicando boa capacidade do sistema em identificar registros duplicados e preservar a versão bibliográfica mais completa de cada referência.

Na etapa de triagem de títulos e resumos, o sistema priorizou sensibilidade e preservou todos os estudos finalmente incluídos pelo padrão humano. Embora tenha mantido maior número de falsos positivos, não houve falsos negativos entre os estudos finais, reforçando o caráter conservador da triagem automatizada.

No módulo exploratório de leitura de textos completos, a IA também apresentou desempenho favorável como ferramenta de apoio. Todos os estudos finalmente incluídos pelo padrão humano foram classificados como incluídos pela IA, sem ocorrência de falsos negativos. Na auditoria pós-hoc dos artigos priorizados pela IA e não pertencentes ao conjunto humano final, um artigo inicialmente classificado como falso positivo foi considerado potencialmente elegível após reavaliação

humana. Esse achado não modificou o padrão de referência humano original, composto por 12 estudos, nem as estimativas primárias de desempenho do sistema.

Esse resultado reforça que a análise dos falsos positivos não deve ser interpretada apenas como erro do sistema, mas também como uma etapa de auditoria capaz de identificar discordâncias ou possíveis falhas da seleção humana inicial. O perfil geral de erro permaneceu predominantemente conservador, com retenção de artigos potencialmente relevantes para avaliação posterior, em vez de exclusão indevida de estudos elegíveis.

Em conjunto, esses resultados indicam que o sistema automatizado apresentou desempenho compatível com sua finalidade principal: reduzir a carga operacional da revisão sistemática sem comprometer a preservação dos estudos relevantes. Dessa forma, o fluxo desenvolvido mostra-se promissor como ferramenta auditável, conservadora e reprodutível de apoio à condução de revisões sistemáticas, mantendo a decisão humana como referência final. Detalhes na tabela 28.

Tabela 28 – Resultado final das hipóteses.

Hipótese	Resultado
Deduplicação teria alta sensibilidade e precisão	Sustentada, especialmente em nível de agrupamento
Triagem priorizaria sensibilidade	Parcialmente sustentada: todos os 12 estudos finais foram preservados, embora a sensibilidade em relação à lista humana ampliada de 118 registros tenha sido de 0,703.
Redução de trabalho >50% na triagem título/resumo	Sustentada operacionalmente
Erros seriam predominantemente falsos positivos	Sustentada
Texto completo teria sensibilidade elevada	Sustentada: sensibilidade de 1,000 nos 90 PDFs processados, com preservação dos 12 estudos finalmente incluídos.
Estudos finais seriam preservados	Sustentada: 12/12

Fonte: Dados da pesquisa.

8.1.8 Síntese dos resultados

Em síntese, o fluxo automatizado apresentou desempenho favorável para apoio à condução de revisão sistemática. A etapa de deduplicação demonstrou alta concordância com o padrão humano, especialmente na análise em nível de agrupamento, indicando boa capacidade de identificar registros duplicados e preservar a versão bibliográfica mais completa de cada referência.

Na triagem de títulos e resumos, o sistema reduziu de forma expressiva o volume de registros a serem avaliados, encaminhando 123 dos 13.371 registros pós-deduplicação para avaliação posterior. Apesar dessa redução, todos os 12 estudos finalmente incluídos pelo padrão humano foram preservados. Embora a especificidade e a precisão tenham sido limitadas quando comparadas à inclusão final humana, não houve falsos negativos entre os estudos finais, o que indica comportamento conservador da ferramenta.

O módulo exploratório de leitura de textos completos também apresentou desempenho favorável como estratégia de apoio e priorização. A IA classificou como incluídos todos os estudos pertencentes ao conjunto final de 12 estudos do padrão humano, sem ocorrência de falsos negativos, e reduziu em 68,9% a carga de leitura prioritária. A análise dos falsos positivos mostrou que a maior parte das discordâncias ocorreu por ausência de dados específicos extraíveis para semaglutida ou por apresentação dos resultados para agonistas do receptor de GLP-1 como classe. Além disso, um artigo inicialmente classificado como falso positivo mostrou-se potencialmente elegível após checagem humana, sugerindo que a auditoria dos falsos positivos também pode auxiliar na identificação de discordâncias do processo humano inicial.

Portanto, o fluxo automatizado não deve ser interpretado como substituto da decisão humana final, mas como ferramenta de triagem, priorização, auditoria e redução de carga operacional. Em conjunto, os achados indicam que o sistema apresenta rastreabilidade, comportamento conservador e segurança metodológica adequados para uso supervisionado em revisões sistemáticas.

8.1.9 Limitações e riscos metodológicos

Apesar dos resultados favoráveis, este estudo apresenta limitações inerentes ao seu delineamento metodológico. A validação foi conduzida em uma revisão sistemática específica, focada em semaglutida e desfechos perioperatórios, de modo que a generalização do fluxo automatizado para outras perguntas clínicas, outras intervenções ou áreas do conhecimento exigirá validações adicionais. Além disso, o padrão de referência humano, embora baseado em decisões registradas por revisores treinados e orientados pelas diretrizes PRISMA, também está sujeito a

inconsistências, variações de julgamento e limitações próprias do processo manual de revisão.

Outro ponto metodológico relevante diz respeito à dependência da qualidade dos metadados bibliográficos e dos arquivos em PDF. Registros sem DOI, títulos incompletos, resumos ausentes, duplicatas com metadados divergentes e textos completos com extração textual inadequada podem afetar o desempenho do sistema automatizado. Da mesma forma, o módulo de leitura de textos completos por inteligência artificial deve ser interpretado como ferramenta exploratória e de apoio à decisão, não como substituto da avaliação humana final. Esse cuidado é especialmente importante porque modelos de linguagem podem apresentar erros de interpretação, inconsistência na classificação ou inferências não sustentadas pelo texto avaliado.

Vale destacar uma limitação que se refere ao ponto de corte de 35.000 caracteres aplicado ao texto extraído dos PDFs. Esse limite afetou 38 dos 90 documentos processados (42,2%) devido a uma restrição técnica dos pacotes utilizados. A truncagem pode ter omitido informações relevantes localizadas nas porções finais dos documentos, produzindo representação desigual entre artigos de diferentes extensões e influenciando sua classificação em qualquer direção. Embora nenhum dos 12 estudos incluídos no conjunto final de referência tenha sido incorretamente despriorizado, esse resultado não exclui a possibilidade de impacto sobre outras classificações nem garante desempenho semelhante em novos conjuntos documentais.

Em relação ao tempo de execução, embora o fluxo automatizado tenha demonstrado redução expressiva da carga de trabalho manual nas etapas de triagem e avaliação exploratória de textos completos, seu desenvolvimento exigiu investimento inicial relevante em programação, testes, depuração, normalização de dados, ajuste de critérios, pareamento com o padrão humano e verificação dos resultados. Assim, o ganho operacional tende a ser maior após a implementação inicial do sistema, quando o fluxo pode ser reutilizado, adaptado e aplicado a novas revisões com menor necessidade de reestruturação completa.

Quanto aos custos, o estudo não envolveu despesas assistenciais, coleta de dados clínicos, contratação de equipe externa ou aquisição de equipamentos laboratoriais. Os principais custos estiveram relacionados à infraestrutura

computacional e ao uso de ferramentas digitais, incluindo utilização de API de modelo de linguagem de grande porte, assinatura de ferramenta de IA generativa para apoio à redação, revisão e organização metodológica, e assinatura do Google Colab para ampliar a capacidade computacional disponível durante o processamento dos dados. Esses custos foram operacionais, vinculados à execução computacional do projeto, e não alteram a natureza metodológica, bibliográfica e de baixo risco da pesquisa.

8.1.10 Perspectivas futuras

Como perspectivas futuras, propõe-se a validação externa do SafeScreen em revisões sistemáticas de diferentes áreas do conhecimento, com distintos volumes de registros, prevalências de elegibilidade e critérios de inclusão e exclusão. Pretende-se também avançar no desenvolvimento do SafeScreen como software, com interface acessível ao usuário e integração dos diferentes módulos em uma única plataforma, facilitando sua utilização por pesquisadores sem conhecimento de programação. Estudos futuros poderão ainda avaliar diferentes modelos de inteligência artificial e o impacto do sistema sobre tempo, custos e carga de trabalho em cenários prospectivos. Os resultados desta validação deram origem a manuscrito científico submetido ao periódico PLOS ONE, atualmente em processo editorial, cuja versão submetida é apresentada no Apêndice E. A eventual publicação do estudo poderá ampliar a divulgação, a avaliação por pares e a disseminação do SafeScreen na comunidade científica. Essas etapas contribuirão para avaliar e ampliar a generalização, a robustez e a aplicabilidade do SafeScreen na condução de revisões sistemáticas.

8.2 Artigo complementar

Tratou-se da revisão sistemática e meta-análise sobre os desfechos perioperatórios associados ao uso de semaglutida seguindo o protocolo convencional. O manuscrito, intitulado “Residual gastric content and aspiration-related outcomes in perioperative patients treated with semaglutide: A systematic review and meta-analysis”, encontra-se em fase final de publicação no World Journal of Gastrointestinal Surgery (26). Página inicial disponível na Figura 5 e o artigo na íntegra disponível no Apêndice F.

Figura 5. Página inicial do artigo complementar aceito para publicação no World Journal of Gastrointestinal Surgery.

World Journal of Gastrointestinal Surgery

Submit a Manuscript: https://www.f6publishing.com

World J Gastrointest Surg ; 0(0): 121449

DOI: 10.4240/wjgs.121449

ISSN 1948-9366 (online)

META-ANALYSIS

Residual gastric content and aspiration-related outcomes in perioperative patients treated with semaglutide: A systematic review and meta-analysis

Vinicius Remus Ballotin, Pedro Lucas Carneiro Ferreira, Valentina Tonin de Almeida, Carolina da Silva Borges, Henrique Tonin de Almeida, Rodrigo Giovanardi Pandolfo da Silva, Daniel Volquind, Luciano da Silva Selistre

Peer review: Externally peer reviewed.
Peer-review model: Single blind
Peer-review report's classification
Scientific quality: Grade B, Grade C, Grade C
Novelty: Grade C, Grade C, Grade C
Creativity or innovation: Grade C, Grade C, Grade C
Scientific significance: Grade B, Grade C, Grade C
P-Reviewer: Gökdere OG, Assistant Professor, MD, Türkiye; Pappachan JM, Editor, FRCP, MD, MRCP, Professor, Senior Researcher, United Kingdom
Received: March 25, 2026
Revised: April 23, 2026
Accepted: June 3, 2026
Processing time: 142 Days and 18.8 Hours

Vinicius Remus Ballotin, Department of Health Sciences, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

Pedro Lucas Carneiro Ferreira, Valentina Tonin de Almeida, Carolina da Silva Borges, Henrique Tonin de Almeida, Rodrigo Giovanardi Pandolfo da Silva, School of Medicine, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

Daniel Volquind, Department of Anesthesia and Pain Medicine, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

Luciano da Silva Selistre, Department of Nephrology and Biostatistics, University of Caxias do Sul, Caxias do Sul, Brazil, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

ORCID number: Vinicius Remus Ballotin 0000-0002-2659-2249; Pedro Lucas Carneiro Ferreira 0000-0002-3872-830X; Valentina Tonin de Almeida 0000-0002-7748-6374; Carolina da Silva Borges 0009-0006-6605-7677; Henrique Tonin de Almeida 0009-0008-0242-740X; Rodrigo Giovanardi Pandolfo da Silva 0009-0000-7434-7734; Daniel Volquind 0000-0002-8298-823X; Luciano da Silva Selistre 0000-0002-0152-0636.

Co-first authors: Vinicius Remus Ballotin and Pedro Lucas Carneiro Ferreira.

Corresponding author: Vinicius Remus Ballotin, MD, Principal Investigator, Researcher, Department of Health Sciences, University of Caxias do Sul, Campus-Sede Rua Francisco Getúlio Vargas, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil. vrballotin@ucs.br

Abstract

BACKGROUND

Semaglutide, a long-acting glucagon-like peptide-1 receptor agonist, is increasingly prescribed for type 2 diabetes and obesity. Its perioperative safety remains uncertain due to concerns regarding delayed gastric emptying, increased residual gastric content (RGC), and possible aspiration-related complications.

AIM

To determine whether perioperative semaglutide use is associated with RGC, pulmonary aspiration, digestive symptoms, pro-cedure discontinuation, and

Fonte: Elaborada pelo autor.

Os agonistas do receptor do peptídeo semelhante ao glucagon-1 (GLP-1RAs) assumiram papel central no tratamento da obesidade e do diabetes mellitus tipo 2,

condições de elevada e crescente prevalência mundial (53, 54). Estima-se que centenas de milhões de adultos vivam com diabetes, sendo mais de 90% dos casos classificados como tipo 2, enquanto a obesidade afeta uma parcela expressiva da população global e mantém tendência de crescimento (55, 56). Nesse contexto, a semaglutida, um GLP-1RA de longa ação, passou a ser amplamente utilizada tanto para o controle glicêmico quanto para o tratamento da obesidade, aumentando significativamente a probabilidade de pacientes em uso desse medicamento serem submetidos a cirurgias, exames endoscópicos e procedimentos sob sedação.

Além de estimular a secreção de insulina dependente da glicose e reduzir a liberação de glucagon, a semaglutida retarda o esvaziamento gástrico por mecanismos que incluem relaxamento do fundo gástrico, aumento da complacência gástrica, redução da contratilidade antral e aumento do tônus pilórico (57-59). Embora esse efeito possa diminuir ao longo do tratamento em razão da taquifilaxia, o atraso do esvaziamento gástrico pode persistir em alguns pacientes, aumentando a preocupação com conteúdo gástrico residual, regurgitação e aspiração pulmonar durante a anestesia ou sedação (60-62).

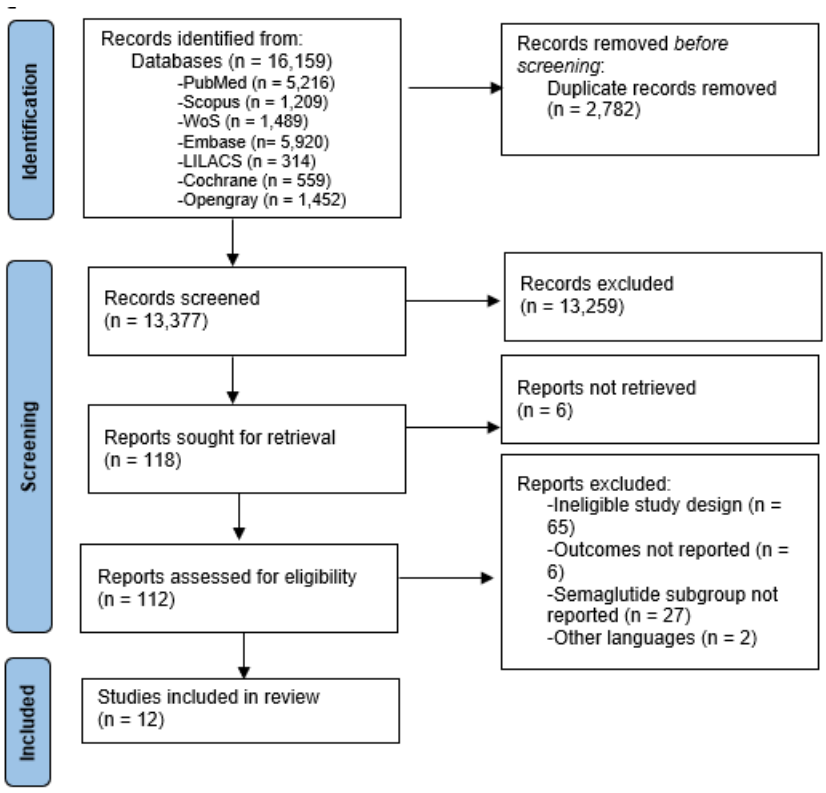
Diante do número crescente de usuários, mesmo complicações pouco frequentes podem apresentar impacto clínico relevante, levando a mudanças na estratégia anestésica, adiamentos, cancelamentos de procedimentos e maior utilização de recursos hospitalares. Entretanto, as recomendações atuais para o manejo perioperatório dos GLP-1RAs ainda se baseiam em evidências limitadas e heterogêneas, provenientes principalmente de relatos de caso e estudos observacionais (59, 63, 64).

Assim, este artigo complementar teve como objetivo avaliar se o uso perioperatório de semaglutida está associado ao aumento de conteúdo gástrico residual, aspiração pulmonar, sintomas digestivos, suspensão ou descontinuação de procedimentos e gastroparesia, em comparação com pacientes controles. A revisão foi conduzida conforme as recomendações PRISMA e o Manual Cochrane, com protocolo registrado no PROSPERO sob o número CRD42024609601 (Apêndice G). Foram pesquisadas sete bases de dados eletrônicas: Scopus, Web of Science, Embase, MEDLINE/PubMed, LILACS, Cochrane Library e OpenGrey, desde a origem das bases até abril de 2026.

A seleção dos estudos, extração dos dados e avaliação da qualidade metodológica foram realizadas de forma independente por revisores treinados. A qualidade dos estudos observacionais foi avaliada pela escala Newcastle-Ottawa Scale, e a certeza da evidência foi classificada por meio da abordagem GRADE. Os desfechos foram sintetizados por meio de metanálise, utilizando risco relativo como medida de efeito. Considerando a heterogeneidade clínica e metodológica esperada entre os estudos, o modelo de efeitos aleatórios foi utilizado como análise principal.

A busca identificou 16.159 registros. Após a remoção de 2.782 duplicatas, 13.377 títulos e resumos foram avaliados, dos quais 13.259 foram excluídos. Foram procurados 118 relatórios para recuperação em texto completo; seis não foram recuperados e 112 foram avaliados quanto à elegibilidade. Após a exclusão de 100 relatórios, 12 estudos observacionais foram incluídos na análise final, totalizando 49.651 pacientes, dos quais 11.022 eram usuários de semaglutida. A Figura 6 apresenta o fluxograma PRISMA do estudo complementar.

Figura 6. Fluxograma PRISMA do processo de identificação, triagem, elegibilidade e inclusão dos estudos na revisão sistemática complementar.



Fonte: Elaborado pelos autores (2026), conforme o PRISMA 2020.

O principal achado da metanálise foi a associação entre o uso perioperatório de semaglutida e maior risco de conteúdo gástrico residual. O conteúdo gástrico residual foi observado em 239 de 1.690 usuários de semaglutida, correspondendo a 14,1%, e em 322 de 6.106 controles, correspondendo a 5,2%. Isso representou uma diferença absoluta de risco de 8,9 pontos percentuais e um risco relativo agrupado de 3,49 (IC95%: 1,78–6,83). A heterogeneidade foi substancial, com I^2 = 87,8%, indicando importante variabilidade entre os estudos.

Os sintomas digestivos também foram mais frequentes entre os usuários de semaglutida. Esse desfecho ocorreu em 27 de 156 pacientes expostos à semaglutida, correspondendo a 17,3%, e em 43 de 1.342 controles, correspondendo a 3,2%. O risco relativo agrupado foi de 5,65 (IC95%: 3,61–8,84), com diferença absoluta de risco de 14,1 pontos percentuais.

Por outro lado, os desfechos clínicos mais graves e menos frequentes permaneceram incertos. A aspiração pulmonar foi rara nos dois grupos, ocorrendo em 15 de 6.222 usuários de semaglutida e em 80 de 21.519 controles, com risco relativo de 1,46 (IC95%: 0,29–7,24). Esse intervalo de confiança amplo demonstra elevada imprecisão e impossibilita concluir de forma definitiva se há aumento, ausência ou redução do risco de aspiração pulmonar associado à semaglutida.

A descontinuação de procedimentos ocorreu em 84 de 5.979 usuários de semaglutida e em 100 de 18.776 controles, com risco relativo de 1,74 (IC95%: 0,59–5,11). A gastroparesia foi relatada em 20 de 663 usuários de semaglutida e em 81 de 2.824 controles, com risco relativo de 1,43 (IC95%: 0,71–2,87). Esses achados também foram considerados incertos, devido ao pequeno número de eventos e aos intervalos de confiança amplos.

A certeza da evidência foi classificada como moderada para conteúdo gástrico residual e sintomas digestivos, baixa para gastroparesia e descontinuação de procedimentos, e muito baixa para aspiração pulmonar. Esses resultados indicam que a semaglutida parece estar associada a marcadores intermediários de esvaziamento gástrico retardado, especialmente conteúdo gástrico residual e sintomas digestivos, mas ainda não há evidência suficiente para estabelecer associação causal direta com aspiração pulmonar ou outros eventos clínicos graves.

A análise de sensibilidade demonstrou que a associação entre semaglutida e conteúdo gástrico residual permaneceu consistente após a exclusão sequencial de

cada estudo individual, sugerindo que nenhum estudo isolado foi responsável pelo efeito observado. No entanto, a heterogeneidade permaneceu elevada, reforçando que diferenças entre populações, tipos de procedimento, definições de conteúdo gástrico residual, tempo de suspensão da medicação e estratégias perioperatórias podem ter influenciado os resultados.

Esses achados devem ser interpretados com cautela. Todos os estudos incluídos foram observacionais, não foram identificados ensaios clínicos randomizados, e variáveis importantes, como tempo de suspensão da semaglutida, fase de escalonamento da dose, tempo de jejum, uso de ultrassonografia gástrica e conduta anestésica, foram reportadas de forma incompleta ou heterogênea. Além disso, o conteúdo gástrico residual é um marcador substituto e não deve ser interpretado como evidência direta de aumento do risco de aspiração pulmonar.

Apesar dessas limitações, o artigo complementar demonstra a aplicabilidade prática do sistema automatizado desenvolvido nesta tese. A ferramenta permitiu apoiar uma revisão sistemática real, com grande volume inicial de registros, múltiplas etapas de triagem, extração estruturada de dados e síntese quantitativa dos resultados. Dessa forma, o artigo complementar reforça a utilidade do sistema automatizado não apenas como método experimental, mas como ferramenta aplicada à produção de evidência científica em uma questão clínica atual e relevante para a anestesiologia.

Portanto, este artigo complementar contribui para a tese em dois aspectos principais. Primeiro, a utilização dos dados bibliográficos para a aplicação prática do fluxo automatizado desenvolvido no artigo principal. Segundo, contribui para a literatura perioperatória ao sintetizar evidências atuais sobre semaglutida, conteúdo gástrico residual e desfechos relacionados à aspiração, destacando a necessidade de estudos prospectivos com definições padronizadas, protocolos uniformes de jejum e suspensão medicamentosa, e avaliação sistemática de desfechos clínicos.

9. CRONOGRAMA

Tabela 29 – Cronograma

Etapa	Início	Término
Elaboração do projeto	01/11/2024	31/05/2026
Revisão de literatura	01/11/2024	31/03/2026
Elaboração dos materiais e métodos	01/11/2024	28/02/2026
Busca e coleta dos registros bibliográficos	01/11/2024	30/11/2024
Desenvolvimento do sistema automatizado	01/12/2024	28/02/2025
Triagem manual pelos revisores humanos/padrão de referência	01/12/2024	31/03/2025
Validação do módulo de deduplicação	01/03/2025	30/04/2026
Validação do módulo de triagem de títulos e resumos	01/04/2025	30/04/2026
Validação do módulo de leitura de textos completos	01/05/2025	30/05/2026
Coleta e organização dos dados de desempenho	30/05/2025	01/07/2025
Análise estatística dos resultados	01/07/2025	30/08/2025
Elaboração do artigo científico principal	01/09/2025	10/05/2026
Elaboração da revisão sistemática/artigo complementar	01/11/2025	10/05/2026
Envio do material à banca examinadora	30/06/2026	30/06/2026
Apresentação da qualificação do projeto	04/07/2026	04/07/2026
Correções solicitadas pela banca examinadora	05/07/2026	30/07/2026
Submissão dos artigos a periódicos científicos	30/05/2026	30/07/2026

Fonte: Material produzido pelos autores**.**

10. ORÇAMENTO

Tabela 30 – Materiais de Consumo

Material	Finalidade	Status	Custo estimado	Fonte de financiamento
Assinatura Google Colab Pro	Execução dos scripts de processamento bibliográfico e desenvolvimento do sistema automatizado	Adquirido	US$ 20,00/mês	Recursos próprios
Assinatura Claude (Anthropic)	Auxílio na redação científica, desenvolvimento e revisão do sistema automatizado	Adquirido	US$ 20,00/mês	Recursos próprios
API Claude (Anthropic)	Integração do modelo de linguagem de grande porte no módulo de leitura de textos completos	Adquirido	Variável (por uso)	Recursos próprios
Assinatura Google Gemini	Auxílio na redação científica e revisão do projeto	Adquirido	Variável/mês	Recursos próprios
Revisão gramatical e de estilo em inglês	Revisão profissional dos manuscritos antes da submissão aos periódicos	Adquirido	US$ 140,00/artigo	Recursos próprios
Revisão Gramatical Inglês artigo complementar	Revisão profissional do manuscrito durante processo de publicação	Adquirido	R$ 2053,58	Recursos próprios
Taxa de publicação World Journal of Gastrointestinal Surgery	Taxa de publicação do artigo complementar	Adquirido	US$ 3023,00	Recursos próprios

Fonte: Material produzido pelos autores**.**

Tabela 31 – Materiais Permanentes

Material	Finalidade	Status	Custo estimado	Fonte de financiamento
Computador pessoal	Desenvolvimento do sistema automatizado, redação dos manuscritos e análise dos dados	Adquirido	—	—
Acesso à internet	Acesso às bases de dados bibliográficas, plataformas de programação e ferramentas de inteligência artificial	Adquirido	—	—

Fonte: Material produzido pelos autores**.**

11. CONCLUSÃO

O SafeScreen demonstrou potencial para reduzir a carga de trabalho em revisões sistemáticas, mantendo a rastreabilidade e a auditabilidade das decisões. O fluxo preservou todos os 12 estudos finalmente incluídos, com desempenho favorável na deduplicação e na priorização de textos completos. Os resultados sustentam seu uso como ferramenta de apoio ao revisor, e não como substituto da decisão humana, sendo necessária validação externa para avaliar sua generalização.

REFERÊNCIAS

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ open. 2017;7(2):e012545. doi:10.1136/bmjopen-2016-012545

2. Cumpston MS, McKenzie JE, Welch VA, Brennan SE. Strengthening systematic reviews in public health: guidance in the Cochrane Handbook for Systematic Reviews of Interventions. Journal of Public Health. 2022;44(4):e588-e92. doi:10.1093/pubmed/fdac036

3. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemporary clinical trials communications. 2019;16:100443. doi:10.1016/j.conctc.2019.100443

4. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Bmj. 2009;339. doi:10.1136/bmj.b2535

5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj. 2021;372. doi:10.1136/bmj.n71

6. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. bmj. 2017;358. doi:10.1136/bmj.j4008

7. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. Journal of clinical epidemiology. 2020;121:81-90. doi:https://doi.org/10.1016/j.jclinepi.2020.01.008

8. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association. 2006;13(2):206-19. doi:10.1197/jamia.M1929

9. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current

approaches. Systematic reviews. 2015;4(1):5. doi:https://doi.org/10.1186/2046-4053-4-5

10. Van De Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nature machine intelligence. 2021;3(2):125-33. doi:10.1038/s42256-020-00287-7

11. Borissov N, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, et al. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. Systematic Reviews. 2022;11(1):172. doi:10.1186/s13643-022-02045-9

12. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: Deduplicator. Systematic reviews. 2024;13(1):206. doi:10.1186/s13643-024-02619-9

13. Hair K, Bahor Z, Macleod M, Liao J, Sena ES. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. BMC biology. 2023;21(1):189. doi:10.1186/s12915-023-01686-z

14. McKeown S, Mir ZM. Considerations for conducting systematic reviews: A follow-up study to evaluate the performance of various automated methods for reference de-duplication. Research Synthesis Methods. 2024;15(6):896-904. doi:10.1002/jrsm.1736

15. Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, et al. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environment International. 2020;138:105623. doi:https://doi.org/10.1016/j.envint.2020.105623

16. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods. 2024;15(4):616-26. doi:10.1002/jrsm.1715

17. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-

Reviewer. Systematic reviews. 2020;9(1):73. doi:https://doi.org/10.1186/s13643-020-01324-7

18. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Systematic reviews. 2024;13(1):158. doi:10.1186/s13643-024-02575-4

19. Guo E, Gupta M, Deng J, Park Y-J, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. Journal of Medical Internet Research. 2024;26:e48996. doi:10.2196/48996

20. Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. Journal of Medical Internet Research. 2024;26:e52758. doi:10.2196/52758

21. Tran V-T, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. Annals of internal medicine. 2024;177(6):791-9. doi:https://doi.org/10.7326/M23-3389

22. Sohrabi C, Franchi T, Mathew G, Kerwan A, Nicola M, Griffin M, et al. PRISMA 2020 statement: What's new and the importance of reporting guidelines. Elsevier; 2021. p. 105918.

23. de Bruin J, Lombaers P, Kaandorp C, Teijema J, van der Kuil T, Yazan B, et al. ASReview LAB v. 2: Open-source text screening with multiple agents and a crowd of experts. Patterns. 2025. doi:10.1016/j.patter.2025.101318

24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic reviews. 2016;5(1):210. doi:https://doi.org/10.1186/s13643-016-0384-4

25. Al-Marridi AZ, Bensaid A, Ulde SM, Khwaileh T. ReviewGenie: a novel automated system for systematic reviews—an exploratory study in speech and language disorders. Systematic Reviews. 2025;14(1):167. doi:10.1186/s13643-025-02895-z

26. Remus Ballotin V, Carneiro Ferreira PL, Tonin de Almeida V, da Silva Borges C, Tonin de Almeida H, Giovanardi Pandolfo da Silva R, et al. Residual gastric content and aspiration-related outcomes in perioperative patients treated with semaglutide: A

systematic review and meta-analysis. World J Gastrointest Surg. 2026. doi:https://doi.org/10.4240/wjgs.121449

27. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. bmj. 2021;372. doi:10.1136/bmj.n160

28. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic reviews. 2021;10(1):39. doi:https://doi.org/10.1186/s13643-020-01542-z

29. Rethlefsen ML, Brigham TJ, Price C, Moher D, Bouter LM, Kirkham JJ, et al. Systematic review search strategies are poorly reported and not reproducible: a cross-sectional metaresearch study. Journal of Clinical Epidemiology. 2024;166:111229. doi:10.1016/j.jclinepi.2023.111229

30. Ravikanth M, Korra S, Mamidisetti G, Goutham M, Bhaskar T. An efficient learning based approach for automatic record deduplication with benchmark datasets. Scientific Reports. 2024;14(1):16254. doi:10.1038/s41598-024-63242-1

31. Escaldelai FMD, Escaldelai L, Bergamaschi DP. Sistema “Apoio à Revisão Sistemática”: solução web para gerenciamento de duplicatas e seleção de artigos elegíveis. Revista Brasileira de Epidemiologia. 2022;25:e220030. doi:10.1590/1980-549720220030

32. Valizadeh A, Moassefi M, Nakhostin-Ansari A, Hosseini Asl SH, Saghab Torbati M, Aghajani R, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. BMC Medical Research Methodology. 2022;22(1):160. doi:https://doi.org/10.1186/s12874-022-01631-8

33. Van der Mierden S, Tsaioun K, Bleich A, Leenaars CH. Software tools for literature screening in systematic reviews in biomedical research. Altex. 2019;36(3):508-17. doi:https://doi.org/10.14573/altex.1902131

34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30. doi:10.48550/arXiv.1706.03762

35. Kim JK, Rickard M, Dangle P, Batra N, Chua ME, Khondker A, et al. Evaluating large language models for title/abstract screening: a systematic review and meta-

analysis & development of new tool. Journal of Medical Artificial Intelligence. 2025;8:34. doi:10.21037/jmai-24-408

36. Galli C, Gavrilova AV, Calciolari E. Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations. Information. 2025;16(5):378. doi:10.3390/info16050378

37. Oami T, Okada Y, Nakada T-a. Optimal large language models to screen citations for systematic reviews. Research Synthesis Methods. 2025:1-17. doi:10.1017/rsm.2025.10014

38. Oami T, Okada Y, Nakada T-a. Performance of a large language model in screening citations. JAMA network open. 2024;7(7):e2420496. doi:10.1001/jamanetworkopen.2024.20496

39. Homiar A, Thomas J, Ostinelli EG, Kennett J, Friedrich C, Cuijpers P, et al. Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ mental health. 2025;28(1). doi:https://doi.org/10.1136/bmjment-2025-301762

40. Insuk S, Boonpattharatthiti K, Booncharoen C, Chaipitak P, Rashid M, Veettil SK, et al. How well do ChatGPT and Claude perform in study selection for systematic review in obstetrics. Journal of Medical Systems. 2025;49(1):110. doi:https://doi.org/10.1007/s10916-025-02246-4

41. Trad F, Yammine R, Charafeddine J, Chakhtoura M, Rahme M, El-Hajj Fuleihan G, et al. Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. BMC medical research methodology. 2025;25(1):130. doi:https://doi.org/10.1186/s12874-025-02583-5

42. Lee K, Paek H, Ofoegbu N, Rube S, Higashi MK, Dawoud D, et al. A4SLR: an agentic AI-assisted systematic literature review framework to augment evidence synthesis for HEOR and HTA. Value in Health. 2025. doi:10.1016/j.jval.2025.08.002

43. Clark J, Barton B, Albarqouni L, Byambasuren O, Jowsey T, Keogh J, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods. 2025:1-19. doi:10.1017/rsm.2025.16

44. Cao C, Arora R, Cento P, Budak A, Manta K, Farahani E, et al. Automation of systematic reviews with large language models. medRxiv. 2025:2025.06. 13.25329541. doi:10.1101/2025.06.13.2532954

45. Flemyng E, Noel-Storr A, Macura B, Gartlehner G, Thomas J, Meerpohl JJ, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. SAGE Publications Sage CA: Los Angeles, CA; 2025. p. cl2. 70074.

46. Gartlehner G, Nussbaumer-Streit B, Hamel C, Garritty C, Griebler U, King VJ, et al. Responsible Integration of Artificial Intelligence in Rapid Reviews: A Position Statement From the Cochrane Rapid Reviews Methods Group. Cochrane Evidence Synthesis and Methods. 2025;3(6):e70063. doi:10.1002/cesm.70063

47. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interactive journal of medical research. 2025;14(1):e59823. doi:i:10.2196/59823

48. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. 2025;43(2):1-55. doi:10.1145/3703155

49. Tonmoy S, Zaman S, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:240101313. 2024. doi:10.48550/arXiv.2401.01313

50. Kusa W, Lipani A, Knoth P, Hanbury A. An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. Intelligent Systems with Applications. 2023;18:200193. doi:10.1016/j.iswa.2023.200193

51. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74. doi:10.2307/2529310

52. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82. doi:10.11613/BM.2012.031

53. Firkins SA, Yates J, Shukla N, Garg R, Vargo JJ, Lembo A, et al. Clinical outcomes and safety of upper endoscopy while on glucagon-like peptide-1 receptor agonists. Clinical Gastroenterology and Hepatology. 2025;23(5):872-3. e3. doi:10.1016/j.cgh.2024.03.013

54. Nersessian RS, da Silva LM, Carvalho MAS, Silveira SQ, Abib AC, Bellicieri FN, et al. Relationship between residual gastric content and peri-operative

semaglutide use assessed by gastric ultrasound: a prospective observational study. Anaesthesia. 2024;79(12):1317-24. doi:10.1111/anae.16454

55. International Diabetes F. IDF Diabetes Atlas. 11th ed. ed. Brussels, Belgium: International Diabetes Federation; 2025.

56. World Obesity F. Prevalence of Obesity London: World Obesity Federation; 2026 [updated 2026/04/13. Available from: https://www.worldobesity.org/about/about-obesity/prevalence-of-obesity.

57. Jalleh RJ, Plummer MP, Marathe CS, Umapathysivam MM, Quast DR, Rayner CK, et al. Clinical consequences of delayed gastric emptying with GLP-1 receptor agonists and tirzepatide. The Journal of Clinical Endocrinology & Metabolism. 2025;110(1):1-15.

58. Kindel TL, Wang AY, Wadhwa A, Schulman AR, Sharaiha RZ, Kroh M, et al. Multi-society clinical practice guidance for the safe use of glucagon-like peptide-1 receptor agonists in the perioperative period. Surgical endoscopy. 2025;39(1):180-3. doi:10.1007/s00464-024-11263-2

59. Klonoff DC, Kim SH, Galindo RJ, Joseph JI, Garrett V, Gombar S, et al. Risks of peri-and postoperative complications with glucagon-like peptide-1 receptor agonists. Diabetes, Obesity and Metabolism. 2024;26(8):3128-36. doi:10.1111/dom.15636

60. Mendes FF, Carvalho LIM, Lopes MB. Glucagon-Like Peptide-1 agonists in perioperative medicine: to suspend or not to suspend, that is the question. SciELO Brasil; 2024. p. 844538.

61. Silveira SQ, da Silva LM, Abib AdCV, de Moura DTH, de Moura EGH, Santos LB, et al. Relationship between perioperative semaglutide use and residual gastric content: a retrospective analysis of patients undergoing elective upper endoscopy. Journal of clinical anesthesia. 2023;87:111091. doi:10.1016/j.jclinane.2023.111091

62. Zaffar D, Hamza M, Borissov MB, Dy KC, Hwang SY, Hsieh P, et al. Su1100 ASSOCIATION BETWEEN PERIOPERATIVE GLP-1 AGONIST USE AND POST OPERATIVE COMPLICATIONS IN PATIENTS UNDERGOING UPPER ENDOSCOPY. Gastroenterology. 2024;166(5):S-653-S-4. doi:10.1016/S0016-5085(24)01955-3

63. Sen S, Potnuru PP, Hernandez N, Goehl C, Praestholm C, Sridhar S, et al. Glucagon-like peptide-1 receptor agonist use and residual gastric content before anesthesia. JAMA surgery. 2024;159(6):660-7. doi:10.1001/jamasurg.2024.0111

64. Welk B, McClure JA, Carter B, Clarke C, Dubois L, Clemens KK. No association between semaglutide and postoperative pneumonia in people with diabetes undergoing elective surgery. Diabetes, Obesity and Metabolism. 2024;26(9):4105-10. doi:doi: 10.1111/dom.15711

65. Marino EC, Negretto LAF, Ribeiro RS, Momesso D, Feitosa ACR, Toyoshima MTK, et al. Rastreamento e Controle da Hiperglicemia no Perioperatório – Posicionamento Conjunto da Sociedade Brasileira de Diabetes (SBD), Sociedade Brasileira de Anestesiologia (SBA) e Associação Brasileira para o Estudo da Obesidade e Síndrome Metabólica (ABESO). Sociedade Brasileira de Diabetes; 2025.

APÊNDICES

APÊNDICE A – CÓDIGO-FONTE DO SISTEMA AUTOMATIZADO DE REVISÃO SISTEMÁTICA

APÊNDICE A1 – Código do módulo de deduplicação bibliográfica

```
# ============================================================
#  SysRev Deduplicator — v3
# ============================================================

# ── CELL 1 · Install ──────────────────────────────────────
!pip -q install pandas rapidfuzz tqdm

# ── CELL 2 · Mount Drive and set paths ────────────────────
import os
from google.colab import drive

DRIVE_ROOT = "/content/drive"
if not os.path.ismount(DRIVE_ROOT):
    drive.mount(DRIVE_ROOT)
else:
    print("Drive already mounted")

# ▼▼▼  EDITABLE ▼▼▼
OUT_DIR   = [CAMINHO_REMOVIDO]
INPUT_CSV = os.path.join(OUT_DIR, "r_articles_16159.csv")
# ▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲▲

OUT_DEDUP = os.path.join(OUT_DIR, "deduplicated.csv")
OUT_DUPS  = os.path.join(OUT_DIR, "duplicates_removed.csv")
OUT_PAIRS = os.path.join(OUT_DIR, "duplicate_pairs.csv")

os.makedirs(OUT_DIR, exist_ok=True)
print("INPUT exists:", os.path.exists(INPUT_CSV))

# ── CELL 3 · Thresholds EDITABLE (edit here to tune) ──────
TITLE_THR      = 97    # token-sort-ratio threshold for title fuzzy match (0-100)
ABS_THR        = 95    # character ratio threshold for abstract fuzzy match (0-100)
ABS_MIN_LEN    = 150   # ignore abstracts shorter than this (chars)
ABS_PREFIX_LEN = 80    # blocking prefix length for abstract grouping

CHECK_DOI      = True
CHECK_PMID     = True
CHECK_ABSTRACT = True
CHECK_TITLE    = True

# ── Year / author guards on fuzzy matching ────────────────
# When both records have a year, skip the match if they differ
# by more than this value. Set to None to disable.
YEAR_TOLERANCE = 1     # allow ±1 year (handles preprint → journal)

# When both records have a first author, require them to match.
# Set to False to disable.
AUTHOR_GUARD   = True

# ── CELL 4 · Load & inspect ───────────────────────────────
import pandas as pd
import numpy as np

encodings_to_try = [
    "utf-8-sig",
    "utf-8",
    "cp1252",
    "latin1",
```

```
    "ISO-8859-1"
]

df = None
used_encoding = None

for enc in encodings_to_try:
    try:
        df = pd.read_csv(INPUT_CSV, low_memory=False, encoding=enc)
        used_encoding = enc
        print(f"CSV loaded successfully with encoding: (65)")
        break
    except UnicodeDecodeError as e:
        print(f"Failed with encoding (65): {e}")

if df is None:
    raise ValueError("Could not read CSV with the tested encodings.")

print(f"Shape: {df.shape}")
print("Columns:", list(df.columns))
df.head(3)

# ── CELL 5 · Deduplication ──────────────────────────────
import re
from rapidfuzz import fuzz
from tqdm.auto import tqdm

# ── helpers ─────────────────────────────────────────────

def norm_text(s):
    if s is None or (isinstance(s, float) and np.isnan(s)):
        return ""
    return re.sub(r"\s+", " ", str(s).lower().strip())

def norm_title(s):
    s = norm_text(s)
    return re.sub(r"\s+", " ", re.sub(r"[^a-z0-9\s]", " ", s)).strip()

def norm_abstract(s):
    s = norm_text(s)
    return re.sub(r"\s+", " ", re.sub(r"[^a-z0-9\s]", " ", s)).strip()

def norm_doi(s):
    s = norm_text(s)
    for prefix in ("https://doi.org/", "http://doi.org/", "doi:"):
        s = s.replace(prefix, "")
    m = re.search(r"(10\.\d{4,9}/\S+)", s)
    return m.group(1).rstrip(" .);,") if m else ""

def norm_pmid(s):
    s = norm_text(s)
    m = re.search(r"\b(\d{6,10})\b", s)
    return m.group(1) if m else ""

def norm_year(s):
    s = norm_text(s)
    m = re.search(r"\b(19|20)\d{2}\b", s)
    return m.group(0) if m else ""

def first_author_key(s):
    s = norm_text(s)
    if s.strip() in ("", "nan", "none"):
        return ""
    s = re.sub(r"[^a-z\s,;]", "", s)
```

```
    parts = re.split(r"[,;]", s)
    tokens = parts[0].strip().split()
    return tokens[-1] if tokens else ""

def find_col(df, keys):
    lc = {c.lower(): c for c in df.columns}
    for k in keys:
        if k in lc:
            return lc[k]
    for c in df.columns:
        if any(k in c.lower() for k in keys):
            return c
    return None

def years_compatible(ya, yb):
    """Return True if years are compatible (same, missing, or within tolerance)."""
    if YEAR_TOLERANCE is None:
        return True
    if not ya or not yb:
        return True   # one missing → don't penalise
    try:
        return abs(int(ya) - int(yb)) <= YEAR_TOLERANCE
    except ValueError:
        return True

def authors_compatible(aa, ab):
    """Return True if first-author keys are compatible."""
    if not AUTHOR_GUARD:
        return True
    if not aa or not ab:
        return True   # one missing → don't penalise
    return aa == ab

# ── detect columns ────────────────────────────────────────

title_col   = find_col(df, ["title"])
doi_col     = find_col(df, ["doi"])
pmid_col    = find_col(df, ["pmid", "pubmed_id", "pubmedid", "pubmed id"])
pmc_col     = find_col(df, ["pmc_id", "pmcid", "pmc id"])
year_col    = find_col(df, ["year", "publication_year", "pubyear"])
auth_col    = find_col(df, ["authors", "author"])
abs_col     = find_col(df, ["abstract", "abstract note", "abstract_note", "summary"])
journal_col = find_col(df, ["journal", "source", "publication", "container-title"])
pages_col   = find_col(df, ["pages", "page"])
vol_col     = find_col(df, ["volume"])
iss_col     = find_col(df, ["issue", "number"])

print("\nDetected columns:")
for name, col in [("title", title_col), ("doi", doi_col), ("pmid", pmid_col),
                  ("pmcid", pmc_col), ("year", year_col), ("authors", auth_col),
                  ("abstract", abs_col), ("journal", journal_col)]:
    print(f"  {name:10s}: {col}")

if title_col is None:
    raise ValueError("Title column not found — check your CSV headers above.")

# ── normalise ─────────────────────────────────────────────

d = df.copy()
d["_idx"]        = range(len(d))
d["_title_norm"] = d[title_col].astype(str).apply(norm_title)
d["_doi"]        = d[doi_col].astype(str).apply(norm_doi)      if doi_col  else ""
d["_pmid"]       = d[pmid_col].astype(str).apply(norm_pmid)    if pmid_col else ""
d["_pmc"]        = d[pmc_col].astype(str).apply(norm_pmid)     if pmc_col  else ""
```

```
d["_abs_norm"]   = d[abs_col].astype(str).apply(norm_abstract) if abs_col  else ""
d["_year"]       = d[year_col].astype(str).apply(norm_year)    if year_col else ""
d["_auth_key"]   = d[auth_col].astype(str).apply(first_author_key) if auth_col else ""

# ── reply / letter / erratum guard ──────────────────────────
# Titles ending with these words are distinct article types and
# must never be merged with the original article.
REPLY_SUFFIXES = r"[\-
\s]*(reply|comment|letter|erratum|correction|editorial|response|rejoinder)\s*$"
d["_is_reply"] = d["_title_norm"].str.contains(REPLY_SUFFIXES, regex=True, na=False)
print(f"  Reply/letter/erratum records flagged: {d['_is_reply'].sum()}")

# ── completeness score (prefer the most complete record) ─────
# Higher weights for fields most useful for downstream screening

score_cols = [c for c in [title_col, doi_col, pmid_col, pmc_col, year_col,
                          auth_col, abs_col, journal_col, pages_col, vol_col, iss_col] if
c]

def completeness(row):
    s = sum(1 for c in score_cols
            if pd.notna(row.get(c)) and str(row.get(c, "")).strip()
            and str(row.get(c, "")).lower() not in ("nan", "none", ""))
    # Bonus for high-value identifiers
    s += 4 * bool(row.get("_doi"))
    s += 3 * bool(row.get("_pmid"))
    s += 1 * bool(row.get("_pmc"))
    # Bonus for long abstract (more useful for screening)
    abs_len = len(str(row.get("_abs_norm", "")))
    if abs_len >= 200: s += 3
    elif abs_len >= 50: s += 1
    # Bonus for having journal + volume + pages (full bibliographic record)
    has_journal = journal_col and pd.notna(row.get(journal_col)) and
str(row.get(journal_col, "")).strip() not in ("", "nan")
    has_vol     = vol_col and pd.notna(row.get(vol_col)) and str(row.get(vol_col,
"")).strip() not in ("", "nan")
    has_pages   = pages_col and pd.notna(row.get(pages_col)) and str(row.get(pages_col,
"")).strip() not in ("", "nan")
    s += 2 * (has_journal and has_vol and has_pages)
    return s

d["_score"] = d.apply(completeness, axis=1)

# ── Union-Find ──────────────────────────────────────────────

class UnionFind:
    def __init__(self, n):
        self.p = list(range(n)); self.r = [0] * n
    def find(self, x):
        while self.p[x] != x:
            self.p[x] = self.p[self.p[x]]; x = self.p[x]
        return x
    def union(self, a, b):
        ra, rb = self.find(a), self.find(b)
        if ra == rb: return
        if   self.r[ra] < self.r[rb]: self.p[ra] = rb
        elif self.r[ra] > self.r[rb]: self.p[rb] = ra
        else: self.p[rb] = ra; self.r[ra] += 1

uf    = UnionFind(len(d))
pairs = []   # list of (idx_a, idx_b, match_type, similarity)

def link_exact(col, label):
    groups = d[d[col] != ""].groupby(col)["_idx"].apply(list)
```

```
    for _, idxs in groups.items():
        if len(idxs) < 2: continue
        for j in idxs[1:]:
            uf.union(idxs[0], j)
            pairs.append((idxs[0], j, label, 100))

# ── exact ID matches ──────────────────────────────────────

if CHECK_DOI:  link_exact("_doi",  "DOI_EXACT")
if CHECK_PMID: link_exact("_pmid", "PMID_EXACT")
link_exact("_pmc", "PMCID_EXACT")

# ── year + first-author + title exact ─────────────────────
# Catches same paper from different databases when DOI/PMID missing.
# Uses full title norm (not just first 40 chars) to avoid false merges
# on numbered series (WP1 vs WP2, Part I vs Part II, etc.)
d["_yat"] = d["_year"] + "|" + d["_auth_key"] + "|" + d["_title_norm"]
_yat_valid = (d["_year"].ne("") & d["_auth_key"].ne("") &
              d["_title_norm"].str.len().gt(20))
link_exact_mask = d[_yat_valid].groupby("_yat")["_idx"].apply(list)
for _, idxs in link_exact_mask.items():
    if len(idxs) < 2: continue
    for j in idxs[1:]:
        uf.union(idxs[0], j)
        pairs.append((idxs[0], j, "YEAR_AUTHOR_TITLE_EXACT", 100))

# ── abstract fuzzy (all-pairs within prefix block) ────────
# Guards: year must be compatible, author must match if both present
if CHECK_ABSTRACT and abs_col:
    d["_abs_pref"] = d["_abs_norm"].str[:ABS_PREFIX_LEN]
    blocks = d[d["_abs_norm"].str.len() >= ABS_MIN_LEN].groupby("_abs_pref", dropna=False)
    for _, g in tqdm(blocks, desc="Abstract blocks"):
        idxs  = g["_idx"].tolist()
        abss  = g["_abs_norm"].tolist()
        years = g["_year"].tolist()
        auths = g["_auth_key"].tolist()
        for i in range(len(idxs)):
            for j in range(i + 1, len(idxs)):
                if not years_compatible(years[i], years[j]):
                    continue
                if not authors_compatible(auths[i], auths[j]):
                    continue
                sim = fuzz.ratio(abss[i], abss[j])
                if sim >= ABS_THR:
                    uf.union(idxs[i], idxs[j])
                    pairs.append((idxs[i], idxs[j], f"ABSTRACT_FUZZY_{ABS_THR}", sim))

# ── title fuzzy (all-pairs within prefix block) ───────────
# Uses token_sort_ratio only — preserves extra words like "Reply".
# Guards: reply type must match, year must be compatible,
#         author must match if both present.
if CHECK_TITLE:
    d["_title_pref"] = d["_title_norm"].str[:20]
    title_blocks = d[d["_title_norm"].str.len() > 10].groupby("_title_pref", dropna=False)
    for _, g in tqdm(title_blocks, desc="Title blocks"):
        idxs     = g["_idx"].tolist()
        titles   = g["_title_norm"].tolist()
        is_reply = g["_is_reply"].tolist()
        years    = g["_year"].tolist()
        auths    = g["_auth_key"].tolist()
        for i in range(len(idxs)):
            for j in range(i + 1, len(idxs)):
                # Never merge a reply/letter with a non-reply record
                if is_reply[i] != is_reply[j]:
```

```
                    continue
                if not years_compatible(years[i], years[j]):
                    continue
                if not authors_compatible(auths[i], auths[j]):
                    continue
                sim = fuzz.token_sort_ratio(titles[i], titles[j])
                if sim >= TITLE_THR:
                    uf.union(idxs[i], idxs[j])
                    pairs.append((idxs[i], idxs[j], f"TITLE_FUZZY_{TITLE_THR}", sim))

# ── select best record per cluster ──────────────────────────

d["_cluster"] = d["_idx"].apply(uf.find)
d_sorted = d.sort_values(["_cluster", "_score"], ascending=[True, False])
keep_idx = set(d_sorted.groupby("_cluster")["_idx"].first().values)

# ── label each duplicate with the strongest removal reason ──

reason_rank = {
    "DOI_EXACT": 7, "PMID_EXACT": 6, "PMCID_EXACT": 5,
    "YEAR_AUTHOR_TITLE_EXACT": 4,
    f"ABSTRACT_FUZZY_{ABS_THR}": 3,
    f"TITLE_FUZZY_{TITLE_THR}": 2,
    "CLUSTER_DUPLICATE": 1,
}

pairs_df = pd.DataFrame(pairs, columns=["idx_a", "idx_b", "match_type", "similarity"])
best_reason = {}
for _, row in pairs_df.iterrows():
    removed = row.idx_b if row.idx_a in keep_idx else (row.idx_a if row.idx_b in keep_idx
else None)
    if removed is None: continue
    cur = best_reason.get(removed)
    if cur is None or reason_rank.get(row.match_type, 0) > reason_rank.get(cur["type"], 0):
        best_reason[removed] = {"type": row.match_type, "sim": row.similarity}

# ── build output dataframes ─────────────────────────────────

internal = [c for c in d.columns if c.startswith("_")]
keep_mask  = d["_idx"].isin(keep_idx)
dedup      = d[keep_mask].drop(columns=internal)
duplicates = d[~keep_mask].drop(columns=internal).copy()
duplicates["dup_reason"]     = duplicates.index.map(
    lambda i: best_reason.get(d.loc[i, "_idx"], {}).get("type", "CLUSTER_DUPLICATE"))
duplicates["dup_similarity"] = duplicates.index.map(
    lambda i: best_reason.get(d.loc[i, "_idx"], {}).get("sim", ""))

# ── report ──────────────────────────────────────────────────

print("\n" + "=" * 50)
print(f"  Total input records  : {len(d):>7,}")
print(f"  Unique (kept)        : {len(dedup):>7,}")
print(f"  Duplicates removed   :
{len(duplicates):>7,}  ({len(duplicates)/len(d)*100:.1f}%)")
print("=" * 50)
print("\nRemoval reasons:")
print(duplicates["dup_reason"].value_counts().to_string())

# ── save ────────────────────────────────────────────────────

dedup.to_csv(OUT_DEDUP,  index=False, encoding="utf-8-sig")
duplicates.to_csv(OUT_DUPS, index=False, encoding="utf-8-sig")
pairs_df.to_csv(OUT_PAIRS, index=False, encoding="utf-8-sig")
```

```
print(f"\nSaved to Drive:")
print(f"  {OUT_DEDUP}")
print(f"  {OUT_DUPS}")
print(f"  {OUT_PAIRS}")

# ── CELL 6 · Spot-check: review sample pairs ──────────────────
import textwrap

title_col_local = title_col

print(f"Showing 20 sample duplicate pairs\n{'─' * 70}")
sample = pairs_df.sample(min(20, len(pairs_df)), random_state=42)
for _, row in sample.iterrows():
    ia, ib = int(row.idx_a), int(row.idx_b)
    ta = df.iloc[ia][title_col_local] if title_col_local else "(no title)"
    tb = df.iloc[ib][title_col_local] if title_col_local else "(no title)"
    print(f"[{row.match_type}  sim={row.similarity:.0f}]")
    print(f"  A: {textwrap.shorten(str(ta), 90)}")
    print(f"  B: {textwrap.shorten(str(tb), 90)}")
    print()
```

APÊNDICE A2 – Código do módulo de triagem automatizada de títulos e resumos

```
# ============================================================
# PRISMA TITLE/ABSTRACT SCREENING — v4
# Semaglutide + perioperative outcomes | Colab
#
# APPROACH: 3-gate rules (primary) + ML query similarity (secondary)
#   Gate 1 — GLP-1 / semaglutide mention
#   Gate 2 — Perioperative context
#   Gate 3 — ≥1 of 5 IC3 outcome domains
#
#   ML model rescues borderline Gate-3 misses at very high threshold.
#   Final decision is always binary: include / exclude.
#
# Outputs (Google Drive):
#   1) AI_screening_ALL.csv
#   2) AI_screening_INCLUDED.csv
#   3) AI_screening_AUTO_EXCLUDED.csv
#   4) PRISMA_FLOW.txt
# ============================================================

import os, re, sys, subprocess

# -----------------------
# 0) FIX ENV (Colab-safe pins)
# -----------------------
PINNED = {
    "numpy": "1.26.4",
    "pandas": "2.2.2",
    "scipy": "1.13.1",
    "scikit-learn": "1.5.2",
}

def get_ver(pkg):
    try:
        import importlib.metadata as md
        return md.version(pkg)
    except Exception:
        return None
```

```
need_fix = any(get_ver(p) != v for p, v in PINNED.items())
if need_fix:
    pkgs = [f"{p}=={v}" for p, v in PINNED.items()]
    print("Fixing package versions...")
    subprocess.check_call([sys.executable, "-m", "pip", "install", "-q",
                           "--upgrade", "--force-reinstall"] + pkgs)
    print("Versions fixed. Restarting runtime...")
    os.kill(os.getpid(), 9)

# -----------------------
# 1) Imports
# -----------------------
import numpy as np
import pandas as pd
from scipy.sparse import hstack, csr_matrix
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics.pairwise import cosine_similarity

# -----------------------
# 2) Mount Google Drive
# -----------------------
from google.colab import drive
DRIVE_ROOT = "/content/drive"
if not os.path.ismount(DRIVE_ROOT):
    drive.mount(DRIVE_ROOT)
else:
    print("Drive already mounted")

# -----------------------
# 3) PATHS - EDITABLE
# -----------------------
OUT_DIR   = [CAMINHO_REMOVIDO]
os.makedirs(OUT_DIR, exist_ok=True)

INPUT_CSV = os.path.join(OUT_DIR, "deduplicated.csv")
OUT_ALL   = os.path.join(OUT_DIR, "AI_screening_ALL.csv")
OUT_INCL  = os.path.join(OUT_DIR, "AI_screening_INCLUDED.csv")
OUT_EXCL  = os.path.join(OUT_DIR, "AI_screening_AUTO_EXCLUDED.csv")
OUT_PRIS  = os.path.join(OUT_DIR, "PRISMA_FLOW.txt")

if not os.path.exists(INPUT_CSV):
    raise FileNotFoundError(f"INPUT_CSV not found: {INPUT_CSV}")

# ============================================================
# TUNING KNOBS
#   MIN_OUTCOME_DOMAINS : raise to 2 if still too many results
#   ML_RESCUE_THR       : probability threshold to rescue a record
#                         that passed Gates 1+2 but failed Gate 3
#                         Set 0.0 to disable ML rescue entirely
# ============================================================
MIN_OUTCOME_DOMAINS = 1
ML_RESCUE_THR       = 0.90

# -----------------------
# 4) HELPERS
# -----------------------
def norm_text(s):
    s = "" if s is None or (isinstance(s, float) and np.isnan(s)) else str(s)
    return re.sub(r"\s+", " ", s.lower().strip())

def find_col(df, keys):
    cols = {c.lower(): c for c in df.columns}
```

```
    for k in keys:
        if k.lower() in cols:
            return cols[k.lower()]
    for c in df.columns:
        if any(k.lower() in c.lower() for k in keys):
            return c
    return None

def any_match(text, patterns):
    return any(re.search(p, text) for p in patterns)

# -----------------------
# 5) LOAD DATA
# -----------------------
df = pd.read_csv(INPUT_CSV)
n_after_dedup = len(df)
print(f"\nCSV loaded | {df.shape[0]} rows, {df.shape[1]} cols")

title_col = find_col(df, ["title"])
abs_col   = find_col(df, ["abstract", "summary", "abstract_note", "abstract note"])
kw_col    = find_col(df, ["keywords", "keyword"])
notes_col = find_col(df, ["notes", "note"])

print(f"Columns → title: {title_col} | abstract: {abs_col} | "
      f"keywords: {kw_col} | notes: {notes_col}")

if title_col is None:
    raise ValueError(f"No Title column found. Available: {list(df.columns)}")
if abs_col is None:
    df["_abstract_tmp"] = ""
    abs_col = "_abstract_tmp"
    print("WARNING: No abstract column — using empty strings.")

def safe_col(c):
    return df[c].astype(str) if c and (c in df.columns) else pd.Series([""] * len(df))

n_missing_abstract = (
    df[abs_col].isna().sum() +
    (df[abs_col].astype(str).str.strip().isin(["", "nan"])).sum()
)
print(f"Records with missing/empty abstract: {n_missing_abstract}")

df["_title_norm"]    = safe_col(title_col).apply(norm_text)
df["_abstract_norm"] = safe_col(abs_col).apply(norm_text)
df["_has_abstract"]  = (~df["_abstract_norm"].str.strip().isin(["", "nan"])).astype(int)

df["_text"] = (
    safe_col(title_col) + " " +
    safe_col(abs_col)   + " " +
    safe_col(kw_col)    + " " +
    safe_col(notes_col)
).apply(norm_text)

# -----------------------
# 6) TERM LISTS — protocol-mapped (EDITABLE - another researchers)
# -----------------------

# ── GATE 1: GLP-1 / Semaglutide (IC2) ──────────────────────────────
# Includes semaglutide by name AND GLP-1 class terms.
# Papers studying the GLP-1 class without naming semaglutide explicitly
# are still relevant (semaglutide is a GLP-1 RA).
# Specificity is enforced by Gates 2 and 3.
SEMAGLUTIDE_TERMS = [
    r"\bsemaglutide\b", r"\bozempic\b", r"\bwegovy\b", r"\brybelsus\b",
```

```
    r"\bglp[- ]?1\b",
    r"\bglp[- ]?1 receptor agonist\w*",
    r"\bglucagon[- ]like peptide[- ]?1\b",
    r"\bglucagon[- ]like peptide 1 receptor agonist\w*",
    r"\bglp1[- ]?ra\b",
    r"\bglp[- ]?1[- ]?ra\b",
    r"\bincretin\w*",
]

# ── GATE 2: Perioperative context (IC3 header) ──────────────────────────
PERIOP_TERMS = [
    r"\bperioperat\w*", r"\bpreoperat\w*", r"\bintraoperat\w*", r"\bpostoperat\w*",
    r"\bsurger\w*", r"\bsurgical\b", r"\boperative\b",
    r"\banesth\w*", r"\banaesth\w*", r"\bsedat\w*",
    r"\bendoscop\w*", r"\bcolonoscop\w*", r"\begd\b", r"\bgastroscop\w*",
    r"\bintubat\w*", r"\bairway\b",
    r"\bfasting\b", r"\bnpo\b", r"\bnil by mouth\b",
    r"\blaparoscop\w*", r"\bbariatric\b",
    r"\bprocedur\w*",
]

# ── GATE 3: Outcome domains IC3a–e ──────────────────────────────────────

# IC3a – Residual gastric content
# CORRECTED: "gastric emptying", "gastric volume", and "gastroparesis"
# now require proximity to a perioperative term (within 80 chars).
# Rationale: these terms appear heavily in diabetes pharmacology papers
# (e.g. "gastric emptying rate", "fasting gastric volume") where they
# describe metabolic endpoints, NOT perioperative safety.
# "fasting" deliberately excluded from proximity trigger — it is
# extremely common in diabetes abstracts as "fasting glucose/insulin".
_PERIOP_PROX =
r"(periop\w*|preoperat\w*|anesth\w*|surger\w*|intubat\w*|endoscop\w*|sedation|aspiration|np
o|nil by mouth)"
GASTRIC_TERMS = [
    r"\bresidual gastric\b",        # always specific to periop safety
    r"\bgastric content\b",         # always specific
    r"\bstomach content\b",         # always specific
    r"\bfull stomach\b",            # always specific
    r"\bretained gastric\b",        # always specific
    r"\bgastric ultrasound\b",      # always anesthesia context
    r"\bgastric residual\b",        # always specific
    # gastroparesis only when near periop terms
    r"\bgastropar\w*.{0,80}" + _PERIOP_PROX,
    _PERIOP_PROX + r".{0,80}\bgastropar\w*",
    # gastric emptying only when near periop terms (not "fasting glucose")
    r"\bgastric emptying\b.{0,80}" + _PERIOP_PROX,
    _PERIOP_PROX + r".{0,80}\bgastric emptying\b",
    # gastric volume only when near periop/ultrasound terms
    r"\bgastric volume\b.{0,80}" + _PERIOP_PROX,
    _PERIOP_PROX + r".{0,80}\bgastric volume\b",
    r"\bgastric volume\b.{0,80}\bultrasound\b",
    r"\bultrasound\b.{0,80}\bgastric volume\b",
]

# IC3b – Aspiration / regurgitation
# "aspiration" alone excluded — too generic (needle aspiration, joint aspiration)
ASPIRATION_TERMS = [
    r"\bpulmonary aspiration\b",
    r"\bbronchial aspiration\b",
    r"\baspiration pneumonitis\b",
    r"\baspiration pneumonia\b",
    r"\bregurgitat\w*",
    r"\bsilent aspiration\b",
```

```
]

# IC3c – Airway management complications
AIRWAY_TERMS = [
    r"\bdifficult intubat\w*",
    r"\bdifficult airway\b",
    r"\bfailed intubat\w*",
    r"\bemergent intubat\w*",
    r"\brapid sequence intubat\w*",
    r"\brsi\b",
    r"\bairway complication\w*",
    r"\bairway management\b",
    r"\blaryngospasm\b",
    r"\bbronchospasm\b",
]

# IC3d – Semaglutide/GLP-1 withdrawal / hold period
WITHDRAWAL_TERMS = [
    r"\bpreoperative (hold|discontinu|withhold|cessation|suspension)\w*",
    r"\bsemaglutide\b.{0,80}(withdr\w*|held|hold\w*|stop\w*|discontinu\w*|pause\w*|cessatio
n|washout|suspend\w*)",
    r"(withdr\w*|held|hold\w*|stop\w*|discontinu\w*|pause\w*|cessation|washout|suspend\w*).
{0,80}\bsemaglutide\b",
    r"\bglp[- ]?1\b.{0,80}(withdr\w*|discontinu\w*|hold|washout)",
]

# IC3e – Overall perioperative complications
# CORRECTED: added "respiratory complication\w*" and "postoperative respiratory\b"
# These patterns are needed to catch titles like:
# "...Risk of Postoperative Respiratory Complications" (JAMA 2024, Dixit et al.)
COMPLICATION_TERMS = [
    r"\bperioperative complication\w*",
    r"\bpostoperative complication\w*",
    r"\bsurgical complication\w*",
    r"\banesthetic complication\w*",
    r"\bperioperative (outcome|risk|safety|adverse)\w*",
    r"\bpostoperative (outcome|morbidity|mortality|adverse)\w*",
    r"\badverse perioperative\b",
    r"\bintraoperative (event|complication)\w*",
    r"\brespiratory complication\w*",    # NEW: catches "postoperative respiratory
complications"
    r"\bpostoperative respiratory\b",    # NEW: catches "risk of postoperative
respiratory..."
]

ALL_OUTCOME_DOMAINS = [
    GASTRIC_TERMS,
    ASPIRATION_TERMS,
    AIRWAY_TERMS,
    WITHDRAWAL_TERMS,
    COMPLICATION_TERMS,
]
OUTCOME_DOMAIN_NAMES = [
    "IC3a_gastric", "IC3b_aspiration", "IC3c_airway",
    "IC3d_withdrawal", "IC3e_complication",
]

# ── HARD EXCLUSION RULES ──────────────────────────────────────────

# EC4: Non-original study types
NON_ORIGINAL_TERMS = [
    r"\bsystematic review\b", r"\bmeta[- ]?analysis\b", r"\bmetaanalysis\b",
    r"\bscoping review\b",    r"\bnarrative review\b",  r"\bliterature review\b",
    r"\bguideline\b",         r"\bpractice guideline\b",
```

```
    r"\bconsensus statement\b", r"\bconsensus\b",
    r"\bcase report\b",       r"\bcase series\b",
    r"\beditorial\b",         r"\bletter to the editor\b",
    r"\bcommentary\b",        r"\bexpert opinion\b",    r"\bposition statement\b",
]

# EC3: Other specific GLP-1 agents — excluded ONLY if no semaglutide AND
# no GLP-1 class term present (pure competitor-drug paper)
OTHER_GLP1_TERMS = [
    r"\bliraglutide\b", r"\bsaxenda\b",   r"\bvictoza\b",
    r"\bdulaglutide\b", r"\btrulicity\b",
    r"\bexenatide\b",   r"\bbyetta\b",    r"\bbydureon\b",
    r"\blixisenatide\b", r"\badlyxin\b",
    r"\btirzepatide\b", r"\bmounjaro\b",  r"\bzepbound\b",
]

# EC1 proxy: Pediatric / non-adult populations
PEDIATRIC_TERMS = [
    r"\bpediatric\b", r"\bpaediatric\b", r"\bchildren\b",
    r"\badolescent\w*", r"\bneonate\w*", r"\binfant\w*", r"\bnewborn\b",
]

# Animal studies
ANIMAL_TERMS = [
    r"\bmouse\b", r"\bmice\b", r"\brat\b", r"\brats\b",
    r"\bporcine\b", r"\bcanine\b", r"\bovine\b", r"\bmurine\b",
    r"\banimal model\w*", r"\brodent\w*", r"\bin[- ]vivo\b",
]

# EC5: Non-English markers (conservative)
NON_ENGLISH_TERMS = [
    r"\bresumen\b", r"\bresumo\b", r"\bzusammenfassung\b", r"\brésumé\b",
    r"\babstract.{0,20}(en español|em português|auf deutsch|en français)\b",
]

# -----------------------
# 7) CLASSIFY EACH RECORD
# -----------------------
def classify(row):
    text  = row["_text"]
    title = row["_title_norm"]

    # ── Gates ──
    gate1 = any_match(text, SEMAGLUTIDE_TERMS)
    gate2 = any_match(text, PERIOP_TERMS)
    domain_hits = [any_match(text, d) for d in ALL_OUTCOME_DOMAINS]
    n_domains   = sum(domain_hits)
    gate3       = n_domains >= MIN_OUTCOME_DOMAINS

    # ── Hard exclusion ──
    ex_non_orig    = any_match(text, NON_ORIGINAL_TERMS)
    ex_pediatric   = any_match(text, PEDIATRIC_TERMS)
    ex_animal      = any_match(text, ANIMAL_TERMS)
    ex_non_english = any_match(text, NON_ENGLISH_TERMS)

    has_other_glp1    = any_match(text, OTHER_GLP1_TERMS)
    has_sema_explicit = any_match(text, [r"\bsemaglutide\b", r"\bozempic\b",
                                         r"\bwegovy\b",      r"\brybelsus\b"])
    has_glp1_class    = any_match(text, [r"\bglp[- ]?1\b",
                                         r"\bglucagon[- ]like peptide[- ]?1\b",
                                         r"\bglp1[- ]?ra\b", r"\bglp[- ]?1[- ]?ra\b"])
    # EC3: only exclude if competitor drug paper with NO semaglutide and NO GLP-1 class
    ex_other_glp1_only = has_other_glp1 and (not has_sema_explicit) and (not
has_glp1_class)
```

```
    hard_exclude = ex_non_orig or ex_animal or ex_pediatric or ex_other_glp1_only

    # Exclusion reasons — mapped to protocol criteria for audit trail
    reasons = []
    if ex_non_orig:        reasons.append("EC4:non_original")
    if ex_animal:          reasons.append("animal_study")
    if ex_pediatric:       reasons.append("EC1_proxy:pediatric")
    if ex_other_glp1_only: reasons.append("EC3:other_GLP1_only")
    if ex_non_english:     reasons.append("EC5:non_english_marker")
    if not gate1:          reasons.append("no_GLP1_semaglutide_term")
    if not gate2:          reasons.append("no_periop_context")
    if not gate3:          reasons.append("no_outcome_domain")

    # Primary decision — strict binary, no uncertain category
    if gate1 and gate2 and gate3 and not hard_exclude:
        decision = "include"
    else:
        decision = "exclude"

    return {
        "gate1_glp1":            int(gate1),
        "gate2_periop":          int(gate2),
        "gate3_outcome":         int(gate3),
        "n_outcome_domains":     n_domains,
        **{OUTCOME_DOMAIN_NAMES[i]: int(domain_hits[i]) for i in range(5)},
        "sema_in_title":         int(any_match(title, SEMAGLUTIDE_TERMS)),
        "ex_non_original":       int(ex_non_orig),
        "ex_pediatric":          int(ex_pediatric),
        "ex_animal":             int(ex_animal),
        "ex_other_glp1_only":    int(ex_other_glp1_only),
        "ex_non_english":        int(ex_non_english),
        "flag_missing_abstract": int(not row["_has_abstract"]),
        "exclusion_reasons":     ";".join(reasons),
        "rule_decision":         decision,
    }

print("\nApplying 3-gate rules...")
results = df.apply(classify, axis=1).apply(pd.Series)
df = pd.concat([df, results], axis=1)

print("Rule-based decisions:")
print(df["rule_decision"].value_counts())

# -----------------------
# 8) ML MODEL (secondary — rescues borderline Gate-3 misses only)
#
# Uses word TF-IDF + character TF-IDF (catches morphological variants)
# + cosine similarity to the structured query.
# Trained on rule-includes (positive) vs clear non-GLP1/non-periop (negative).
# Only applied to records that passed Gates 1+2 but failed Gate 3,
# and only at a very high confidence threshold (ML_RESCUE_THR).
# -----------------------
QUERY_TEXT = """
(glucagon-like peptide-1 OR GLP-1 OR GLP1 OR GLP1-RA OR semaglutide OR ozempic OR wegovy OR
rybelsus)
(endoscopy OR upper endoscopy OR EGD OR gastroscopy OR colonoscopy OR sedation OR
anesthesia OR surgery OR perioperative)
(residual gastric content OR retained gastric OR gastric emptying OR gastric volume OR
aspiration OR regurgitation OR pneumonia OR respiratory complications)
(withholding OR discontinuation OR hold OR suspension OR cessation OR washout)
""".strip()

query = norm_text(QUERY_TEXT)
```

```
# Word TF-IDF
vec_word = TfidfVectorizer(ngram_range=(1, 2), min_df=2, max_df=0.90,
                           sublinear_tf=True, stop_words="english")
Xw = vec_word.fit_transform(df["_text"].values)
qw = vec_word.transform([query])
sim_word = cosine_similarity(Xw, qw).ravel()

# Character TF-IDF (catches morphological variants, abbreviations)
vec_char = TfidfVectorizer(analyzer="char_wb", ngram_range=(3, 5),
                           min_df=2, max_df=0.95)
Xc = vec_char.fit_transform(df["_text"].values)
qc = vec_char.transform([query])
sim_char = cosine_similarity(Xc, qc).ravel()

df["sim_query"] = np.maximum(sim_word, sim_char)

# Weak labels: rule-includes = positive, clear non-GLP1 or non-periop = negative
train_pos = df[df["rule_decision"] == "include"]
train_neg = df[(df["gate1_glp1"] == 0) | (df["gate2_periop"] == 0)]
train_neg = train_neg.sample(
    n=min(len(train_neg), len(train_pos) * 3),
    random_state=42
)
train_df = pd.concat([train_pos.assign(_wl=1), train_neg.assign(_wl=0)])

USE_MODEL = len(train_df) >= 50
if USE_MODEL:
    Xw_tr = vec_word.transform(train_df["_text"].values)
    Xc_tr = vec_char.transform(train_df["_text"].values)
    X_tr  = hstack([Xw_tr, Xc_tr,
                    csr_matrix(train_df["sim_query"].values.reshape(-1, 1))],
                   format="csr")
    clf = LogisticRegression(max_iter=4000, class_weight="balanced",
                             C=0.5, random_state=42, n_jobs=-1)
    clf.fit(X_tr, train_df["_wl"].astype(int).values)

    X_all = hstack([Xw, Xc,
                    csr_matrix(df["sim_query"].values.reshape(-1, 1))],
                   format="csr")
    df["p_include_model"] = clf.predict_proba(X_all)[:, 1]
    print(f"\nModel trained | pos={len(train_pos)} neg={len(train_neg)}")
    print(df["p_include_model"].describe(percentiles=[0.50, 0.75, 0.90, 0.95, 0.99]))
else:
    df["p_include_model"] = np.nan
    print("Skipping ML — insufficient training data.")

# -----------------------
# 9) FINAL DECISION
#    Rules are primary. ML only rescues records that:
#      - Passed Gate 1 (GLP-1 term) ✓
#      - Passed Gate 2 (periop context) ✓
#      - Failed Gate 3 (no outcome domain matched) ✗
#      - Pass no hard-exclusion rule
#      - Have model probability ≥ ML_RESCUE_THR
# -----------------------
df["ai_decision"] = df["rule_decision"]

if USE_MODEL and ML_RESCUE_THR > 0:
    rescue_mask = (
        (df["rule_decision"]      == "exclude") &
        (df["gate1_glp1"]         == 1) &
        (df["gate2_periop"]       == 1) &
```

```
        (df["ex_non_original"]    == 0) &
        (df["ex_animal"]          == 0) &
        (df["ex_pediatric"]       == 0) &
        (df["ex_other_glp1_only"] == 0) &
        (df["p_include_model"]    >= ML_RESCUE_THR)
    )
    df.loc[rescue_mask, "ai_decision"] = "include"
    df.loc[rescue_mask, "exclusion_reasons"] = (
        df.loc[rescue_mask, "exclusion_reasons"] + ";rescued_by_ML"
    )
    print(f"\nML rescued {int(rescue_mask.sum())} records (threshold p≥{ML_RESCUE_THR})")

print("\nFinal decisions:")
print(df["ai_decision"].value_counts())

# -----------------------
# 10) PRISMA FLOW
# -----------------------
n_include = int((df["ai_decision"] == "include").sum())
n_exclude = int((df["ai_decision"] == "exclude").sum())
excl_tmp  = df[df["ai_decision"] == "exclude"]

def ce(col):
    return int((excl_tmp[col] == 1).sum()) if col in excl_tmp.columns else 0

n_no_sema    = int((excl_tmp["gate1_glp1"] == 0).sum())
n_no_periop  = int(((excl_tmp["gate1_glp1"] == 1) &
                    (excl_tmp["gate2_periop"] == 0)).sum())
n_no_outcome = int(((excl_tmp["gate1_glp1"] == 1) &
                    (excl_tmp["gate2_periop"] == 1) &
                    (excl_tmp["gate3_outcome"] == 0)).sum())

prisma_text = f"""
============================================================
PRISMA FLOW — Title/Abstract Screening (AI-assisted v4)
============================================================

Records after deduplication:               {n_after_dedup}
  of which missing/empty abstract:         {n_missing_abstract}

── Screening logic ──────────────────────────────────────
  Primary:   3-gate rule (ALL gates required)
  Secondary: ML rescue at p≥{ML_RESCUE_THR} (Gate 1+2 pass, Gate 3 miss only)

  Gate 1 — GLP-1 / semaglutide mention (IC2)
  Gate 2 — Perioperative context (IC3 header)
  Gate 3 — ≥{MIN_OUTCOME_DOMAINS} of 5 IC3 outcome domain(s):
           IC3a: residual gastric content
           IC3b: aspiration / regurgitation
           IC3c: airway management complications
           IC3d: GLP-1/semaglutide withdrawal period
           IC3e: perioperative / respiratory complications

  Any gate failure OR hard exclusion rule → EXCLUDED
  No uncertain category — binary include/exclude only.

Title/Abstract Screening:
  ├── INCLUDED : {n_include}
  └── EXCLUDED : {n_exclude}

Records forwarded to full-text review: {n_include}

────────────────────────────────────────────────────────────
Exclusion breakdown:
```

```
  No GLP-1/semaglutide term  [Gate 1 fail]: {n_no_sema}
  No perioperative context   [Gate 2 fail]: {n_no_periop}
  No outcome domain          [Gate 3 fail]: {n_no_outcome}

  EC4 – Non-original study type:            {ce("ex_non_original")}
       (reviews, meta-analyses, guidelines,
        editorials, case reports/series)
  EC3 – Other GLP-1 only (no semaglutide):  {ce("ex_other_glp1_only")}
  EC1 proxy – Pediatric population:         {ce("ex_pediatric")}
  Animal studies:                           {ce("ex_animal")}
  EC5 – Non-English markers:                {ce("ex_non_english")}

  Note: one record can match multiple reasons.

============================================================
FOR YOUR PRISMA FLOWCHART:
  "Records screened"                  = {n_after_dedup}
  "Records excluded (title/abstract)" = {n_exclude}
  "Full-text articles assessed"       = {n_include}
============================================================
"""
print(prisma_text)
with open(OUT_PRIS, "w", encoding="utf-8") as f:
    f.write(prisma_text)
print(f"PRISMA flow saved: {OUT_PRIS}")

# -----------------------
# 11) EXPORT CSVs
# -----------------------
drop_cols = ["_text", "_title_norm", "_abstract_norm", "_abstract_tmp",
             "_has_abstract", "rule_decision", "_wl"]
out_df  = df.drop(columns=drop_cols, errors="ignore").copy()
incl_df = out_df[out_df["ai_decision"] == "include"].copy()
excl_df = out_df[out_df["ai_decision"] == "exclude"].copy()

out_df.to_csv(OUT_ALL,  index=False, encoding="utf-8-sig")
incl_df.to_csv(OUT_INCL, index=False, encoding="utf-8-sig")
excl_df.to_csv(OUT_EXCL, index=False, encoding="utf-8-sig")

print(f"\nFiles saved:")
print(f"  ALL      : {len(out_df):>5} rows → {OUT_ALL}")
print(f"  INCLUDED : {len(incl_df):>5} rows → {OUT_INCL}")
print(f"  EXCLUDED : {len(excl_df):>5} rows → {OUT_EXCL}")

# -----------------------
# 12) SANITY PREVIEW
# -----------------------
tcol = title_col if title_col in incl_df.columns else "title"

print(f"\n── INCLUDED sample (first 15) ──")
for i, (_, row) in enumerate(incl_df.head(15).iterrows()):
    domains  = [OUTCOME_DOMAIN_NAMES[j] for j in range(5) if
row.get(OUTCOME_DOMAIN_NAMES[j], 0) == 1]
    rescued  = " 🔁ML" if "rescued_by_ML" in str(row.get("exclusion_reasons","")) else ""
    p        = row.get("p_include_model", float("nan"))
    print(f"  [{i+1}] {str(row.get(tcol,''))[:80]}{rescued}")
    print(f"       domains: {', '.join(domains) or 'none'} | p={p:.2f}")

print(f"\n── EXCLUDED sample (first 10) ──")
for i, (_, row) in enumerate(excl_df.head(10).iterrows()):
    print(f"  [{i+1}] {str(row.get(tcol,''))[:80]}")
    print(f"       reasons: {row.get('exclusion_reasons','')}")
```

```
print(f"\nDone. {n_include} included | {n_exclude} excluded")
print(f"Forwarded to full-text review: {n_include} records")
```

APÊNDICE A3 – Código do módulo de leitura exploratória de textos completos em PDF

```
# Full-text review - Claude AI screener v7
# Features:
#   - temperature=0, max_tokens=1200 (reproducible, no truncation)
#   - IC2: semaglutide must be in methods/results (not just intro)
#   - EC4: clear accept/reject, label-agnostic
#   - EC6: relaxed - any document with numbers (%, OR, p, CI) passes
#   - PDF: get_text("blocks") extracts tables+prose on ALL PyMuPDF versions
#   - Titles from filename, no manual overrides
# Semaglutide + perioperative outcomes
#
# SETUP:
#   1) Save Anthropic API key in: BASE_DIR/anthropic_key.txt
#   2) Upload PDFs to: BASE_DIR/full_text_review/pdfs/
#   3) Run - saves after every PDF, safe to stop/resume

import os, re, sys, subprocess, time, json

subprocess.check_call([sys.executable, "-m", "pip", "install", "-q",
                       "PyMuPDF", "anthropic", "matplotlib"])

# EDITABLE #

from google.colab import drive
if not os.path.ismount("/content/drive"):
    drive.mount("/content/drive")
else:
    print("Drive already mounted")

BASE_DIR      = [CAMINHO_REMOVIDO]
PDF_DIR       = os.path.join(BASE_DIR, "full_text_review", "pdfs")
OUT_DIR       = os.path.join(BASE_DIR, "full_text_review")
DECISIONS_CSV = os.path.join(OUT_DIR,  "full_text_decisions_claude.csv")
OUT_INCL      = os.path.join(OUT_DIR,  "full_text_INCLUDED.csv")
OUT_EXCL      = os.path.join(OUT_DIR,  "full_text_EXCLUDED.csv")
OUT_SUMMARY   = os.path.join(OUT_DIR,  "full_text_summary.txt")
os.makedirs(PDF_DIR, exist_ok=True)
os.makedirs(OUT_DIR, exist_ok=True)

# ------------------------------------------------------------------
# API KEY - EDITABLE
# ------------------------------------------------------------------
KEY_FILE = os.path.join(BASE_DIR, "anthropic_key.txt")
if os.path.exists(KEY_FILE):
    with open(KEY_FILE) as f:
        api_key = f.read().strip()
    print("API key loaded from file")
elif os.environ.get("ANTHROPIC_API_KEY"):
    api_key = os.environ["ANTHROPIC_API_KEY"]
else:
    api_key = input("Paste your Anthropic API key: ").strip()
    with open(KEY_FILE, "w") as f:
        f.write(api_key)
    print("Key saved to " + KEY_FILE)

import fitz
```

```
import anthropic
import pandas as pd
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
from collections import Counter

client = anthropic.Anthropic(api_key=api_key)

# ------------------------------------------------------------------
# RESUME CONTROL
# FORCE_FRESH = True  -> delete old CSV and rescreen all PDFs
# FORCE_FRESH = False -> resume from last stopped point
# ------------------------------------------------------------------
FORCE_FRESH = True

if FORCE_FRESH and os.path.exists(DECISIONS_CSV):
    os.remove(DECISIONS_CSV)
    print("Old decisions file deleted - starting fresh")

# ------------------------------------------------------------------
# PROTOCOL PROMPT - sent to Claude for every article -
# EDITABLE for another reasearchers
# ------------------------------------------------------------------
SYSTEM_PROMPT = """You are a systematic review screener for a study on
SEMAGLUTIDE and PERIOPERATIVE OUTCOMES.

INCLUSION CRITERIA - ALL must be met:

IC1  Adults aged 18 years and older.
     Assume TRUE unless paper explicitly studies children or animals.

IC2  SEMAGLUTIDE must be EXPLICITLY NAMED as a studied drug.
     Accepted names: semaglutide, Ozempic, Wegovy, Rybelsus.

     INCLUDE if ANY of these are true:
       - Semaglutide is named in the METHODS or RESULTS as a
         studied drug (sole, primary, or one of several)
       - A drug breakdown table in methods/results lists semaglutide
         with patient counts or percentages
       - The exposure/cohort definition explicitly includes sema-
         glutide patients (e.g. "patients on semaglutide were...")

     EXCLUDE as EC3 if:
       - Semaglutide appears ONLY in the introduction or background
         as a general example, but is NOT part of the study cohort
       - Semaglutide does not appear anywhere in the full text
       - Only unnamed GLP-1 receptor agonists are studied

     CRITICAL: Semaglutide mentioned ONLY in the introduction or
     background section does NOT meet IC2. It must be part of the
     actual study population described in methods or results.

IC3  RCT or observational study (cohort, case-control, cross-sectional)
     with ORIGINAL PATIENT DATA reporting at least ONE of:
       IC3a  Residual gastric content
       IC3b  Bronchial or pulmonary aspiration incidence
       IC3c  Airway management complications (perioperative)
       IC3d  Semaglutide withdrawal or hold period before procedure
       IC3e  Overall perioperative complications

     EXCLUDE as EC4 if the document is any of these:
       - Narrative review, systematic review, or meta-analysis
```

```
       - Editorial, commentary, or opinion piece
       - Guideline or consensus statement
       - Case report or case series (1-5 patients)
       - Clinical trial registration without results
       - Letter or correspondence that ONLY comments on another
         paper without presenting its own patient data

     ACCEPT (NOT EC4) if the document has ALL of these:
       - Its own patient cohort (own n=)
       - Its own study methods described
       - Its own quantitative results reported
     This applies regardless of label: a "research letter" or
     "conference abstract" with own data IS original research.

EXCLUSION CRITERIA:

EC1  Comorbidity management is the PRIMARY focus, not semaglutide.
EC2  Contraindications to semaglutide as primary focus.
EC3  Semaglutide NOT explicitly named as a studied drug anywhere.
EC4  No original patient data (see IC3 rules above).
EC5  Non-English language.
EC6  Significant missing data impeding analysis.
     ALSO apply EC6 to any document (abstract or full paper) that
     reports ONLY qualitative statements with NO numbers at all.
     A document PASSES EC6 (do NOT exclude) if it contains ANY of:
       - Percentages (e.g. "9.4% had retained content", "17.6%")
       - Counts per group (e.g. "8/33 patients", "n=232")
       - Odds or risk ratios (OR=, aOR=, RR=, HR=)
       - P-values (p=0.001, p<0.05)
       - Confidence intervals (95% CI)
     A document FAILS EC6 only if it contains NO quantitative
     outcome data whatsoever - e.g. only says "GLP-1 users had
     more retained gastric content" with no supporting numbers.
     NOTE: Section headers (Methods/Results) are NOT required.
     A conference abstract with numbers PASSES EC6.

Respond ONLY with a valid JSON object. No text before or after. No markdown.

{
  "decision": "include or exclude",
  "criterion": "IC or EC1 or EC2 or EC3 or EC4 or EC5 or EC6 or IC1 or IC2 or IC3",
  "ic1_met": true or false,
  "ic2_semaglutide": true or false,
  "ic2_note": "exact quote from paper where semaglutide is named as studied drug, or state
it is absent",
  "ic3_design_ok": true or false,
  "ic3a": true or false,
  "ic3b": true or false,
  "ic3c": true or false,
  "ic3d": true or false,
  "ic3e": true or false,
  "exclusion_criterion": "EC1/EC2/EC3/EC4/EC5/EC6 or null",
  "exclusion_reason": "one sentence why excluded, or null",
  "key_evidence": "exact text from paper showing semaglutide is or is not studied",
  "reasoning": "2-3 sentences explaining your decision"
}"""

# ----------------------------------------------------------------
# PDF EXTRACTION
# ----------------------------------------------------------------
# Uses get_text("blocks") which works on ALL PyMuPDF versions and
# extracts every text region: prose, table cells, text boxes,
# captions, footnotes. This ensures drug names that appear only
# in tables (not in prose) are always captured.
```

```
# ---------------------------------------------------------------
def extract_pdf(path, max_chars=35000):
    try:
        doc = fitz.open(path)
        all_text = ""

        for page in doc:
            # get_text("blocks") returns all text blocks including
            # table cells as separate rectangular regions.
            # Each block: (x0, y0, x1, y1, text, block_no, block_type)
            # Sort by vertical then horizontal position for reading order.
            blocks = page.get_text("blocks")
            blocks = sorted(blocks, key=lambda b: (round(b[1]/10)*10, b[0]))
            page_text = "\n".join(
                b[4].strip() for b in blocks
                if len(b) > 4 and b[4].strip()
            )
            all_text += page_text + "\n"
            if len(all_text) >= max_chars:
                break

        doc.close()
        cleaned = all_text[:max_chars].strip()
        return cleaned if len(cleaned) > 150 else None

    except Exception:
        return None

def get_title(pdf_text, fallback):
    skip = [
        r"^https?://", r"^\d+$", r"^doi\s*:",
        r"^(citation|submitted|received|accepted|published)",
        r"^(abstract|introduction|background|methods?|results?|discussion)",
        r"^(copyright|all rights|open access|this is an open)",
        r"^(figure|table|fig\.|supplementary)",
        r"r\s+e\s+s\s+e\s+a\s+r\s+c\s+h",
        r"^(review article|original article|case report|editorial|letter|commentary)",
        r"^\d+\s*(division|department|university|hospital)",
        r"^(as a library|nlm provides|pubmed)",
        r"^(special issue|citation:|willson|al sakka)",
        r"^schedules", r"^measuring continuous", r"^submit a",
    ]
    for line in pdf_text.split("\n"):
        line = line.strip()
        if not (35 <= len(line) <= 220):
            continue
        if any(re.search(p, line, re.IGNORECASE) for p in skip):
            continue
        if sum(c.isalpha() for c in line) / max(len(line), 1) < 0.55:
            continue
        return line[:130]
    return fallback[:130]

def title_from_filename(fname):
    name = os.path.splitext(fname)[0]
    name = re.sub(r"^\d+_", "", name)
    name = name.replace("_", " ").replace("-", " ")
    if name == name.upper():
        name = name.title()
    return name[:120]

# ---------------------------------------------------------------
# CLAUDE API CALL
# ---------------------------------------------------------------
```

```
def screen_with_claude(pdf_text, filename, retries=3):
    user_msg = (
        "Filename: " + filename + "\n\n"
        "Article full text (first ~35000 chars):\n"
        + "-" * 60 + "\n"
        + pdf_text + "\n"
        + "-" * 60 + "\n\n"
        "Apply the protocol and respond with JSON only."
    )
    for attempt in range(retries):
        try:
            response = client.messages.create(
                model="claude-sonnet-4-6",
                max_tokens=1200,
                temperature=0,
                system=SYSTEM_PROMPT,
                messages=[{"role": "user", "content": user_msg}]
            )
            raw = response.content[0].text.strip()
            raw = re.sub(r"^```(?:json)?\s*", "", raw)
            raw = re.sub(r"\s*```$", "", raw)
            result = json.loads(raw)
            if "decision" not in result:
                raise ValueError("Missing decision field")
            return result
        except json.JSONDecodeError:
            if attempt < retries - 1:
                time.sleep(3)
            else:
                return error_result("JSON parse error")
        except anthropic.RateLimitError:
            wait = 30 * (attempt + 1)
            print("    Rate limit - waiting " + str(wait) + "s...")
            time.sleep(wait)
        except Exception as e:
            if attempt < retries - 1:
                time.sleep(5)
            else:
                return error_result("API error: " + str(e))
    return error_result("Max retries exceeded")

def error_result(msg):
    return {
        "decision": "error", "criterion": "ERR",
        "ic1_met": None, "ic2_semaglutide": None, "ic2_note": "",
        "ic3_design_ok": None,
        "ic3a": False, "ic3b": False, "ic3c": False,
        "ic3d": False, "ic3e": False,
        "exclusion_criterion": None,
        "exclusion_reason": msg,
        "key_evidence": "",
        "reasoning": msg,
    }

# ------------------------------------------------------------------
# LOAD PREVIOUS PROGRESS
# ------------------------------------------------------------------
if os.path.exists(DECISIONS_CSV):
    done_df      = pd.read_csv(DECISIONS_CSV)
    done_df["decision"] = done_df["decision"].fillna("").astype(str)
    already_done = set(done_df["filename"].astype(str).tolist())
    print("Resuming - " + str(len(already_done)) + " articles already screened")
else:
    done_df      = pd.DataFrame()
```

```
    already_done = set()

# ------------------------------------------------------------------
# SCREENING LOOP
# ------------------------------------------------------------------
all_pdfs = sorted([f for f in os.listdir(PDF_DIR) if f.lower().endswith(".pdf")])
pending  = [f for f in all_pdfs if f not in already_done]

if not all_pdfs:
    raise FileNotFoundError("No PDFs found in " + PDF_DIR)

print("\nPDFs found : " + str(len(all_pdfs)))
print("Done       : " + str(len(already_done)))
print("To screen  : " + str(len(pending)) + "\n")

new_rows = []

for i, fname in enumerate(pending, 1):
    n_done = len(already_done) + i
    print("[" + str(n_done) + "/" + str(len(all_pdfs)) + "] " + fname[:65])

    title    = title_from_filename(fname)
    pdf_text = extract_pdf(os.path.join(PDF_DIR, fname))

    if pdf_text is None:
        print("  WARNING: could not extract text - possibly scanned PDF")
        result = error_result("PDF text extraction failed - scanned image")
    else:
        print("  " + str(len(pdf_text)) + " chars - sending to Claude...", end=" ",
flush=True)
        result = screen_with_claude(pdf_text, fname)
        dec    = result.get("decision", "error")
        crit   = result.get("criterion", "?")
        icon   = "[INCLUDE]" if dec == "include" else "[EXCLUDE]"
        print(icon + " " + crit)

        if dec == "include":
            domains = [d.upper() for d in ["ic3a","ic3b","ic3c","ic3d","ic3e"]
                       if result.get(d) is True]
            print("  Domains : " + (", ".join(domains) if domains else "none flagged"))
            print("  Sema    : " + str(result.get("ic2_note",""))[:70])
        else:
            print("  Reason  : " + str(result.get("exclusion_reason",""))[:80])

    new_rows.append({
        "filename":           fname,
        "title":              title,
        "decision":           result.get("decision", "error"),
        "criterion":          result.get("criterion", "ERR"),
        "exclusion_reason":   result.get("exclusion_reason", ""),
        "reasoning":          result.get("reasoning", ""),
        "key_evidence":       result.get("key_evidence", ""),
        "ic1_met":            result.get("ic1_met", ""),
        "ic2_semaglutide":    result.get("ic2_semaglutide", ""),
        "ic2_note":           result.get("ic2_note", ""),
        "ic3_design_ok":      result.get("ic3_design_ok", ""),
        "ic3a":               result.get("ic3a", False),
        "ic3b":               result.get("ic3b", False),
        "ic3c":               result.get("ic3c", False),
        "ic3d":               result.get("ic3d", False),
        "ic3e":               result.get("ic3e", False),
        "exclusion_criterion":result.get("exclusion_criterion", ""),
    })
```

```
    all_rows = pd.concat([done_df, pd.DataFrame(new_rows)], ignore_index=True)
    all_rows.to_csv(DECISIONS_CSV, index=False, encoding="utf-8-sig")
    time.sleep(1)

# ---------------------------------------------------------------
# FINAL RESULTS
# ---------------------------------------------------------------
final = pd.concat([done_df, pd.DataFrame(new_rows)], ignore_index=True) \
        if new_rows else done_df.copy()

for col in ["decision","criterion","exclusion_reason","reasoning",
            "key_evidence","ic2_note","title"]:
    if col not in final.columns:
        final[col] = ""
    final[col] = final[col].fillna("").astype(str)

# Normalize boolean columns
for col in
["ic3a","ic3b","ic3c","ic3d","ic3e","ic2_semaglutide","ic1_met","ic3_design_ok"]:
    if col in final.columns:
        final[col] = final[col].astype(str).str.lower().isin(["true","1","yes"])

inc = final[final["decision"] == "include"].copy()
exc = final[final["decision"] == "exclude"].copy()
err = final[~final["decision"].isin(["include","exclude"])].copy()

n_tot = len(final)
n_inc = len(inc)
n_exc = len(exc)
n_err = len(err)

CRIT_LABELS = {
    "IC":  "All criteria met - INCLUDED",
    "IC1": "IC1 not met: not adult / animal population",
    "IC2": "IC2 not met: semaglutide not studied",
    "IC3": "IC3 not met: no perioperative outcome reported",
    "EC1": "EC1: comorbidity management as primary focus",
    "EC2": "EC2: contraindication to semaglutide focus",
    "EC3": "EC3: other GLP-1 only, no semaglutide data",
    "EC4": "EC4: not original research (review/editorial/case report)",
    "EC5": "EC5: non-English language",
    "EC6": "EC6: missing or insufficient outcome data",
    "ERR": "Error: PDF unreadable",
}

print("\n" + "="*65)
print("SCREENING COMPLETE")
print("="*65)
print("  Total    : " + str(n_tot))
print("  Include  : " + str(n_inc))
print("  Exclude  : " + str(n_exc))
if n_err > 0:
    print("  Errors   : " + str(n_err))
    for _, r in err.iterrows():
        print("    " + r["filename"] + ": " + str(r.get("exclusion_reason",""))[:60])

crit_counts = Counter(exc["criterion"].tolist())
print("\n  Exclusions by criterion:")
for c in ["EC4","EC3","IC2","IC3","EC1","EC5","EC6","IC1"]:
    n = crit_counts.get(c, 0)
    if n:
        print("    " + c + " (" + str(n) + "): " + CRIT_LABELS.get(c,""))

print("\n  Included articles:")
```

```
for i, (_, r) in enumerate(inc.iterrows(), 1):
    domains = [d.upper() for d in ["ic3a","ic3b","ic3c","ic3d","ic3e"] if r.get(d) is True]
    print("  " + str(i).rjust(2) + ". " + r["title"][:72])
    print("      Domains : " + (", ".join(domains) if domains else "see reasoning"))
    print("      Sema    : " + str(r.get("ic2_note",""))[:70])
    print("      Reason  : " + str(r.get("reasoning",""))[:80])

inc.to_csv(OUT_INCL, index=False, encoding="utf-8-sig")
exc.to_csv(OUT_EXCL, index=False, encoding="utf-8-sig")
print("\n  Saved: " + OUT_INCL)
print("  Saved: " + OUT_EXCL)

# ---------------------------------------------------------------
# EXCLUSION CHART
# ---------------------------------------------------------------
if n_exc > 0:
    chart = {CRIT_LABELS.get(c,c): n for c,n in crit_counts.items() if n > 0}
    rs    = sorted(chart.items(), key=lambda x: -x[1])

    fig, ax = plt.subplots(figsize=(11, max(3.5, len(rs)*0.7+1.5)))
    fig.patch.set_facecolor("#FAFAFA")
    ax.set_facecolor("#FAFAFA")
    cmap = matplotlib.colormaps.get_cmap("tab10")
    bars = ax.barh([x[0] for x in rs][::-1],
                   [x[1] for x in rs][::-1],
                   color=[cmap(j/max(len(rs),1)) for j in range(len(rs))],
                   height=0.55, edgecolor="white")
    for bar, (_, val) in zip(bars, rs[::-1]):
        ax.text(bar.get_width()+0.05,
                bar.get_y()+bar.get_height()/2,
                str(val), va="center", fontsize=11, fontweight="bold")
    ax.set_xlabel("Articles excluded", fontsize=11)
    ax.set_title(
        "Full-Text Exclusion Reasons - Claude AI Screening\n"
        "(n = " + str(n_exc) + " excluded / " + str(n_tot) + " assessed)",
        fontsize=12, fontweight="bold")
    ax.set_xlim(0, max(x[1] for x in rs)+1.8)
    for sp in ["top","right"]:
        ax.spines[sp].set_visible(False)
    ax.grid(axis="x", alpha=0.3)
    plt.tight_layout()
    p = os.path.join(OUT_DIR, "exclusion_reasons.png")
    plt.savefig(p, dpi=300, bbox_inches="tight")
    plt.close()
    print("  Chart: " + p)

# ---------------------------------------------------------------
# PRISMA FLOW
# ---------------------------------------------------------------
excl_detail = "\n".join(
    str(n) + "x  " + CRIT_LABELS.get(c,c)[:42]
    for c,n in sorted(crit_counts.items(), key=lambda x:-x[1])[:6]
)

fig, ax = plt.subplots(figsize=(8, 10))
fig.patch.set_facecolor("#FAFAFA")
ax.set_facecolor("#FAFAFA")
ax.axis("off")
ax.set_xlim(0,10)
ax.set_ylim(0,13)

def box(ax, x, y, w, h, txt, col="#1565C0", fs=9.5, bg="#E3F2FD"):
    ax.add_patch(mpatches.FancyBboxPatch(
        (x-w/2, y-h/2), w, h,
```

```
        boxstyle="round,pad=0.15",
        facecolor=bg, edgecolor=col, linewidth=1.8, zorder=2))
    ax.text(x, y, txt, ha="center", va="center", fontsize=fs,
            color="#1A1A1A", zorder=3, multialignment="center", linespacing=1.5)

def arr(ax, x, y1, y2):
    ax.annotate("", xy=(x, y2+0.02), xytext=(x, y1-0.02),
                arrowprops=dict(arrowstyle="-|>", color="#555", lw=1.5))

box(ax, 5, 12.0, 5.5, 0.8,
    "AI title/abstract screening\nn = " + str(n_tot) + " articles",
    col="#1565C0", bg="#E3F2FD")
arr(ax, 5, 11.6, 10.85)
box(ax, 5, 10.45, 5.5, 0.75,
    "Full-text articles assessed\nn = " + str(n_tot),
    col="#6A1B9A", bg="#F3E5F5")

ax.add_patch(mpatches.FancyBboxPatch(
    (6.85, 7.8), 3.1, 2.8,
    boxstyle="round,pad=0.1",
    facecolor="#FDECEA", edgecolor="#C62828", linewidth=1.4, zorder=2))
ax.text(8.4, 9.2,
        "Excluded: n = " + str(n_exc) + "\n\n" + excl_detail,
        ha="center", va="center", fontsize=7.5,
        color="#5B0000", zorder=3, multialignment="left", linespacing=1.45)
ax.annotate("", xy=(6.85, 10.45), xytext=(7.75, 10.45),
            arrowprops=dict(arrowstyle="-|>", color="#C62828", lw=1.2))

arr(ax, 5, 10.07, 9.2)
box(ax, 5, 8.8, 5.5, 0.75,
    "Studies included in final review\nn = " + str(n_inc),
    col="#1B5E20", bg="#E8F5E9", fs=11)

pct = (n_inc+n_exc)/n_tot if n_tot > 0 else 0
ax.add_patch(mpatches.FancyBboxPatch(
    (1.5, 7.5), 7*pct, 0.4,
    boxstyle="round,pad=0.05",
    facecolor="#1B5E20" if pct==1 else "#1565C0", alpha=0.75, zorder=2))
ax.add_patch(mpatches.FancyBboxPatch(
    (1.5, 7.5), 7, 0.4,
    boxstyle="round,pad=0.05",
    facecolor="none", edgecolor="#AAA", linewidth=1, zorder=3))
ax.text(5, 7.72,
        str(n_inc+n_exc) + "/" + str(n_tot) + " screened  ("
        + str(round(pct*100)) + "%)",
        ha="center", va="center", fontsize=9,
        color="white" if pct > 0.3 else "#333",
        fontweight="bold", zorder=4)

ax.set_title(
    "Full-Text Review - PRISMA Flow\nSemaglutide + Perioperative Outcomes",
    fontsize=12, fontweight="bold", pad=10)
plt.tight_layout()
pf = os.path.join(OUT_DIR, "fulltext_prisma_flow.png")
plt.savefig(pf, dpi=300, bbox_inches="tight")
plt.close()
print("  PRISMA flow: " + pf)

# ---------------------------------------------------------------
# SUMMARY TEXT
# ---------------------------------------------------------------
top3 = "; ".join(
    str(n) + " (" + CRIT_LABELS.get(c,c).split(":")[0] + ")"
    for c,n in sorted(crit_counts.items(), key=lambda x:-x[1])[:3]
```

```
) if crit_counts else "not available"

summary = (
    "\n"
    "================================================================\n"
    "FULL-TEXT REVIEW SUMMARY\n"
    "Semaglutide + perioperative outcomes | Claude AI screening v3\n"
    "================================================================\n\n"
    "INCLUSION CRITERIA\n"
    "  IC1  Adults >= 18 years\n"
    "  IC2  Preoperative semaglutide treatment (any duration)\n"
    "  IC3  RCT or observational study reporting >=1 of:\n"
    "         IC3a  Residual gastric content\n"
    "         IC3b  Bronchial / pulmonary aspiration\n"
    "         IC3c  Airway management complications\n"
    "         IC3d  Semaglutide withdrawal / hold period\n"
    "         IC3e  Overall perioperative complications\n\n"
    "EXCLUSION CRITERIA\n"
    "  EC1  Comorbidity management as primary focus\n"
    "  EC2  Contraindication focus\n"
    "  EC3  Other GLP-1 only - no semaglutide data\n"
    "  EC4  Case report / series / review / editorial\n"
    "  EC5  Non-English language\n"
    "  EC6  Missing / insufficient outcome data\n\n"
    "PRISMA NUMBERS\n"
    "  Full-text assessed : " + str(n_tot) + "\n"
    "  Excluded           : " + str(n_exc) + "\n"
    "  Included           : " + str(n_inc) + "\n"
    "  Errors             : " + str(n_err) + "\n\n"
    "EXCLUSION BREAKDOWN\n"
    + "\n".join("  " + str(n).rjust(3) + "  " + CRIT_LABELS.get(c,c)
                for c,n in sorted(crit_counts.items(), key=lambda x:-x[1]))
    + "\n\n"
    "SUGGESTED RESULTS TEXT\n"
    "  Full-text review was conducted on " + str(n_tot) + " articles.\n"
    "  After applying predefined inclusion and exclusion criteria,\n"
    "  " + str(n_exc) + " articles were excluded. Main exclusion reasons:\n"
    "  " + top3 + ".\n"
    "  A total of " + str(n_inc) + " studies met all inclusion criteria\n"
    "  and were included in the final systematic review.\n\n"
    "================================================================\n"
    "Files saved to: " + OUT_DIR + "\n"
    "  full_text_decisions_claude.csv\n"
    "  full_text_INCLUDED.csv\n"
    "  full_text_EXCLUDED.csv\n"
    "  exclusion_reasons.png\n"
    "  fulltext_prisma_flow.png\n"
    "  full_text_summary.txt\n"
    "================================================================\n"
)
print(summary)
with open(OUT_SUMMARY, "w", encoding="utf-8") as f:
    f.write(summary)

print("=" * 65)
print("DONE  |  " + str(n_inc) + " included  |  " + str(n_exc) + " excluded  |  " +
str(n_err) + " errors")
print("=" * 65)
```

APÊNDICE B – CLASSIFICAÇÃO INDIVIDUAL DOS ARTIGOS AVALIADOS NO MÓDULO EXPLORATÓRIO DE LEITURA DE TEXTO COMPLETO POR IA

Quadro B1 – Artigos incluídos pela IA no módulo de texto completo

Nº	Artigos	Decisão da IA	Critério/Domínios
1	05_No association between semaglutide and postoperative pneumonia	Incluído	IC3B, IC3E
2	09_CLINICAL OUTCOMES AND SAFETY OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS	Incluído	IC3A, IC3B, IC3E
3	103_GLP-1 receptor agonist increase retained gastric contents on EGD	Incluído	IC3A, IC3E
4	10_SEMAGLUTIDE IS AN INDEPENDENT RISK FACTOR FOR RETAINED GASTRIC CONTENTS	Incluído	IC3A, IC3E
5	113_Gastric ultrasound in patients receiving semaglutide: a prospective study	Incluído	IC3A, IC3D
6	114_Fasting Duration, Not Timing of Last GLP-1 Dose, Predicts Aspiration Risk	Incluído	IC3A
7	115_Ultrasound assessment of preoperative gastric volume in fasted diabetic patients	Incluído	IC3A, IC3B, IC3D, IC3E
8	15_INCIDENCE OF ASPIRATION PNEUMONITIS AND EMERGENT INTUBATION IN PATIENTS	Incluído	IC3B, IC3C, IC3E
9	18_THE IMPACT OF SEMAGLUTIDE ON RETAINED GASTRIC CONTENTS	Incluído	IC3A, IC3B
10	19_GLUCAGON-LIKE PEPTIDE 1 RECEPTOR AGONISTS PRIOR TO ENDOSCOPY	Incluído	IC3A, IC3D
11	22_ASSOCIATION BETWEEN PERIOPERATIVE GLP-1 AGONIST USE AND POSTOPERATIVE COMPLICATIONS	Incluído	IC3A, IC3B, IC3E
12	38_Glucagon-like peptide-1 receptor agonists before upper gastrointestinal endoscopy	Incluído	IC3B, IC3E
13	42_Effect of various perioperative semaglutide interruption intervals on residual gastric content	Incluído	IC3A, IC3B, IC3D
14	49_Relationship between perioperative semaglutide use and residual gastric content	Incluído	IC3A, IC3B, IC3D
15	57_Glucagon-Like Peptide-1 Receptor Agonist Use and Residual Gastric Content	Incluído	IC3A, IC3D
16	58_Risk of Aspiration Pneumonitis After Elective Esophagogastroduodenoscopy	Incluído	IC3B, IC3E
17	59_Association of glucagon-like peptide receptor 1 agonist therapy with perioperative outcomes	Incluído	IC3A, IC3B, IC3C, IC3E
18	65_Real-World Impact of GLP-1 Receptor Agonists on Endoscopic Patient Outcomes	Incluído	IC3A, IC3B, IC3E
19	67_Relationship between residual gastric content and peri-operative semaglutide use	Incluído	IC3A, IC3B, IC3D
20	68_Influence of semaglutide use on the presence of residual gastric solids	Incluído	IC3A
21	71_Glucagon-like peptide-1 receptor agonist use and perioperative outcomes	Incluído	IC3B, IC3E

22	77_Risk of Postoperative Aspiration Pneumonia in Patients Undergoing ACDF	Incluído	IC3B, IC3E
23	82_Effect of discontinuing semaglutide prior to surgery on perioperative outcomes	Incluído	IC3A, IC3B, IC3D
24	86_THE IMPACT OF PRE-PROCEDURAL DIET ON THE RISK OF RESIDUAL GASTRIC CONTENT	Incluído	IC3A
25	90_Semaglutide and Postoperative Outcomes in Nondiabetic Patients	Incluído	IC3E
26	94_Postoperative Aspiration Pneumonia Among Adults Using GLP-1 Receptor Agonists	Incluído	IC3B, IC3E
27	98_Glucagon-like-peptide-1 receptor agonist use and the risk of pulmonary aspiration	Incluído	IC3B
28	99_Impact of GLP-1 Receptor Agonists on Increased Residual Gastric Content	Incluído	IC3A, IC3D

Fonte: Dados da pesquisa.
Nota: IC3A = conteúdo gástrico residual/esvaziamento gástrico; IC3B = aspiração pulmonar ou bronquial; IC3C = complicações de via aérea; IC3D = suspensão ou intervalo de interrupção da semaglutida; IC3E = complicações perioperatórias gerais. EC3 = estudos sobre outros agonistas de GLP-1 sem dados específicos de semaglutida; EC4 = relato de caso, série de casos, revisão, editorial, comentário ou artigo sem dados originais; IC1 = critério de população adulta não atendido.

Quadro B2 – Artigos excluídos pela IA no módulo de texto completo

Nº	Artigos	Decisão da IA	Critério de exclusão
1	01_a-pilot-study-examining-the-effects-of-glp-1-receptor-blockade	Excluído	EC3
2	02_Semaglutide Treatment for Hyperglycaemia After Renal Transplant	Excluído	EC4
3	03_Gastric Emptying and Metabolic Response After a Laparoscopic Procedure	Excluído	EC3
4	04_Efficacy and Safety of GLP-1 Medicines for Type 2 Diabetes and Obesity	Excluído	EC4
5	06_Implications of GLP-1 agonist use on airway management	Excluído	EC4
6	07_Preoperative GLP-1 Receptor Agonist Use and Risk of Postoperative Outcomes	Excluído	EC3
7	08_Long term metabolic effects and perioperative safety assessment	Excluído	EC3
8	104_GLP-1 receptor agonists and delayed gastric emptying implications	Excluído	EC4
9	105_Multisociety Clinical Practice Guidance for the Safe Use of GLP-1 RAs	Excluído	EC4
10	109_Use of gastric ultrasound to identify GLP-1RA users at high risk	Excluído	EC4
11	111_Point-of-care gastric ultrasound: redefining aspiration risk	Excluído	EC4
12	112_Tirzepatide for Obstructive Sleep Apnea	Excluído	EC3
13	116_Can the incidence of voluntarily reported side effects support clinical decisions	Excluído	EC4

14	119_Gastroparesis, a diabetic complication causing further complications	Excluído	EC4
15	11_THE UNSEEN CONSEQUENCES OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS	Excluído	EC3
16	121_Glucagon-Like Peptide-1 Receptor Agonists in Gynecologic Surgery	Excluído	EC4
17	122_Clinical Consequences of Delayed Gastric Emptying With GLP-1	Excluído	EC4
18	12_THE IMPACT OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS USE	Excluído	EC3
19	13_ASSOCIATION OF GLP-1RA USE AND ASPIRATION PNEUMONIA AFTER THE PROCEDURE	Excluído	EC3
20	14_GLP-1 RA USE PREDICTS PERIOPERATIVE OUTCOMES	Excluído	EC3
21	16_Erratum Diabetes Care in the Hospital Standards of Care in Diabetes	Excluído	EC4
22	17_ARE GLP-1 RECEPTOR AGONISTS ASSOCIATED WITH INCREASED INCIDENCE	Excluído	EC3
23	20_CLINICAL IMPACT OF GLP-1RA ON ENDOSCOPY: A RETROSPECTIVE COHORT	Excluído	EC3
24	21_OUTCOMES ASSOCIATED WITH GLP-1 RECEPTOR AGONISTS	Excluído	EC3
25	23_INCREASED RISK OF RESIDUAL GASTRIC CONTENT IN PATIENTS ON GLP-1	Excluído	EC3
26	24_Pulmonary aspiration of gastric contents in two patients taking semaglutide	Excluído	EC4
27	27_BARIATRIC ENDOSCOPIC ANTRAL MYOTOMY TECHNICAL FEASIBILITY	Excluído	EC3
28	28_Residual Gastric Content and Perioperative Semaglutide Use	Excluído	EC4
29	30_Frequency of GLP-1 analog use in diabetic patients with delayed gastric emptying	Excluído	EC3
30	32_A pre-clinical animal study of combined intragastric balloon	Excluído	IC1
31	33_One-anastomosis jejunal interposition with gastric resection	Excluído	EC3
32	34_One-year follow-up and physiological alterations following endoscopic procedure	Excluído	EC3
33	37_Considerations of delayed gastric emptying with peri-operative GLP-1 use	Excluído	EC4
34	40_A randomized crossover study of the effects of glutamine	Excluído	EC3
35	41_A novel weight-reducing operation: lateral subtotal gastrectomy	Excluído	EC3
36	43_Perioperative management of patients on GLP-1 receptor agonists	Excluído	EC4
37	44_Effects of GLP-1 receptor agonists on upper gastrointestinal outcomes	Excluído	EC3
38	46_The role of gastric function in control of food intake and body weight	Excluído	EC4

39	47_GLP-1RA Essentials in Gastroenterology Side Effect Management	Excluído	EC4
40	48_Comparison of societal guidance on perioperative management of GLP-1 RAs	Excluído	EC4
41	50_Perioperative management of long-acting GLP-1 receptor agonists	Excluído	EC4
42	51_Is it necessary to stop GLP-1 receptor agonists before procedures?	Excluído	EC3
43	52_Risks of peri- and postoperative complications with GLP-1 RAs	Excluído	EC3
44	54_Impact of GLP-1 Receptor Agonists in Gastrointestinal Endoscopy	Excluído	EC4
45	55_Should We Stop GLP-1 Receptor Agonists Before Procedures?	Excluído	EC4
46	56_Anesthesia Considerations for a Patient on Semaglutide and Delayed Gastric Emptying	Excluído	EC4
47	61_GLP-1 Receptor Agonists Used for Medically Supervised Weight Loss	Excluído	EC4
48	62_GLP-1 Agonists and General Anesthesia Perioperative Management	Excluído	EC4
49	64_Perioperative GLP-1 receptor agonists-induced delayed gastric emptying	Excluído	EC4
50	66_Implications of Ozempic and Other Semaglutide Medications for Surgery	Excluído	EC4
51	69_Emerging therapies in the treatment of diabesity beyond GLP-1	Excluído	EC3
52	70_Laparoscopic sleeve gastrectomy: more than a restrictive bariatric procedure	Excluído	EC4
53	72_The Use of GLP-1 Agonists in the Perioperative Period	Excluído	EC4
54	76_Letter to Editor on “As Few as Three Months of Preoperative Semaglutide”	Excluído	EC4
55	78_Peri-Operative Use of GLP-1 Agonists	Excluído	EC3
56	80_The Effect of GLP-1 Receptor Agonists on Perioperative Outcomes	Excluído	EC4
57	81_Preoperative Metoclopramide to Enhance Gastric Emptying in Patients Using GLP-1 RAs	Excluído	EC4
58	83_The fasting and the furious	Excluído	EC4
59	84_Risk and prevention of perioperative pulmonary aspiration caused by GLP-1 RAs	Excluído	EC4
60	92_The pandemic increase of GLP-1 receptor agonists use and the perioperative risk	Excluído	EC4
61	93_GLP-1 Receptor Agonists and Risk of Adverse Events	Excluído	EC4
62	95_GLP-1 Receptor Agonists in the Peri-Operative Period	Excluído	EC4

Fonte: Dados da pesquisa.

Nota: IC3A = conteúdo gástrico residual/esvaziamento gástrico; IC3B = aspiração pulmonar ou bronquial; IC3C = complicações de via aérea; IC3D = suspensão ou intervalo de interrupção da semaglutida; IC3E = complicações perioperatórias gerais. EC3 = estudos sobre outros agonistas de GLP-1 sem dados específicos de semaglutida; EC4 = relato de caso, série de casos, revisão, editorial, comentário ou artigo sem dados originais; IC1 = critério de população adulta não atendido.

Quadro B3 – Artigos dos 123 incluídos pela IA que não foram encontrados

Nº nos 123	Artigo	Ano	DOI	Status
25	Retained Gastric Contents after Adequate Fasting Associated with GLP-1 Receptor Agonist Use: A Report of 3 Cases	—	—	Não encontrado
26	Particulate Gastric Contents in Patients Prescribed Glucagon-Like Peptide 1 Receptor Agonists After Appropriate Perioperative Fasting: A Report of 2 Cases	—	—	Não encontrado
29	Frequency of GLP-1 analog use in diabetic patients with delayed GES and their demographic profile	—	—	Não encontrado
31	Are early dumping and postprandial hypoglycemia after Roux-en-Y gastric bypass related events?	—	—	Não encontrado
35	Liraglutide therapy in elective surgical patients with type 2 diabetes	—	—	Não encontrado
36	Preoperative GLP-1 Receptor Agonists and Risk of Postoperative Respiratory Complications-Reply	—	10.1001/jama.2024.10032	Não encontrado
39	The use of laparoscopic gastric corrugation (placation) for treatment of diabetes mellitus type I. Our experience	—	—	Não encontrado
45	Glucagon-like Peptide-1 Agonists: What the Orthopaedic Surgeon Needs to Know	2024	10.2106/JBJS.RVW.23.00167	Não encontrado
53	Point-of-Care Gastric Ultrasound to Identify a Full Stomach on a Diabetic Patient Taking a Glucagon-Like Peptide 1 Receptor Agonist	—	10.1213/XAA.0000000000001751	Não encontrado
60	Obesity and anesthesia	—	10.1097/ACO.0000000000001377	Não encontrado
63	Update on Perioperative Medication Management for the Hand Surgeon: A Focus on Diabetes, Weight Loss, Rheumatologic, and Antithrombotic Medications	—	10.1016/j.jhsa.2024.05.018	Não encontrado
73	Gastric Emptying With Metoclopramide in GLP-1	2025	—	Não encontrado

	Agonist Patients Undergoing Elective Surgery			
74	Gastric Ultrasound, as an Indicator for Intubation in Short non-Laparoscopic Surgery in Patients Receiving GLP-1 Receptor Agonist. A Comparative Randomized Study	2025	—	Não encontrado
75	Airway Management in the Age of GLP-1 Receptor Agonists	—	10.1016/j.anclin.2025.10.010	Não encontrado
79	Impact of a One-Week Discontinuation of Semaglutide on Gastric Retention and Endoscopic Mucosal Visibility Scores: A Single-Center, Prospective, Observational Study	—	—	Não encontrado
85	Effect of Prolonged Fasting Time on Gastric Residual Volume in Patients Taking GLP1 Receptor Agonists	—	—	Não encontrado
87	Perioperative Risk of Aspiration and Postoperative Complications With GLP-1 Therapy (PROTECT) Study	—	10.1016/S0016-5085(25)03352-9	Não encontrado
88	Reduced Risk of Retained Gastric Contents After Institutional Implementation of Holding GLP-1 Receptor Agonists Prior to Endoscopic Evaluation	—	10.1016/S0016-5085(25)03351-7	Não encontrado
89	Evaluate the Detection of Retained Gastric Contents and Assess Safety Using the Flower Capsule Endoscopy in Healthy Individuals and GLP-1 Receptor Agonist Users	—	—	Não encontrado
91	Effect of Fasting Recommendations on Residual Gastric Contents Among Patients Using Glucagon-like Peptide-1 Receptor Agonists: A Randomised Controlled Trial	—	—	Não encontrado
96	Coming full circle: could glucagon-like peptide-1 receptor agonists confer a net perioperative benefit?	2026	10.1016/j.bja.2025.10.003	Não encontrado
97	Managing patients taking glucagon-like peptide-1 receptor agonists: proceed with caution	2026	10.1016/j.bja.2025.10.004	Não encontrado

100	Prevalence and predictors of residual gastric content in patients with type 2 diabetes on GLP-1 receptor agonists: A prospective observational study	2026	10.1111/dom.70537	Não encontrado
101	Assessment of Gastric Content Using Gastric Ultrasound in Patients on Glucagon-Like Peptide-1 Receptor Agonists Before Anesthesia	2026	10.1213/ANE.0000000000007764	Não encontrado
102	Glucagon Like Peptide-1 Receptor Agonist Use Not Related With Gastric Volume After Preoperative Fasting: A Prospective Observational Study	2025	10.1093/milmed/usaf510	Não encontrado
106	Risk of periprocedural regurgitation of gastric contents in patients on glucagon-like peptide-1 receptor agonists: a two-centre nested case-control study	2026	10.1016/j.bja.2025.12	Não encontrado
107	Perioperative management of patients taking GLP-1 receptor agonists: implementation of complex fasting recommendations. Response to Br J Anaesth 2025; 136: 375-6	2026	10.1016/j.bja.2025.12.022	Não encontrado
108	A restrictive fasting regimen is not necessary in patients taking glucagon-like peptide-1 receptor agonists. Response to Br J Anaesth 2025; 135: 1816-8	2026	10.1016/j.bja.2025.11.022	Não encontrado
110	Impact of GLP-1 receptor agonists on gastrointestinal function and symptoms	2025	10.1080/17474124.2025.2579117	Não encontrado
117	Impact of GLP-1 Receptor Agonists on Retained Gastric Contents During Esophagogastroduodenoscopy: A Propensity Score-Matched Study	2025	10.1111/den.70016	Não encontrado
118	Risk of perioperative cardiorespiratory complications and mortality associated with preoperative GLP-1 receptor agonist use in type 2 diabetes mellitus: a nationwide propensity-score matched study	2026	10.1016/j.bja.2025.08.003	Não encontrado

120	Incretin impact on gastric function in obesity: physiology, and pharmacological, surgical and endoscopic treatments	2025	10.1113/JP287535	Não encontrado
123	Preoperative GLP-1 Receptor Agonist Use and Postoperative Outcomes Following Operative Ankle Fracture Repair in Patients With Type 2 Diabetes: A National Database Study	2025	10.1177/10711007251364187	Não encontrado

Fonte: Dados da pesquisa.
Nota: Esta tabela apresenta os registros classificados como potencialmente elegíveis pela IA na etapa de triagem de títulos e resumos, mas que não foram localizados em formato PDF para avaliação no módulo exploratório de leitura de texto completo. Assim, dos 123 registros incluídos pela IA na triagem inicial, 90 foram avaliados em texto completo e 33 não foram encontrados em PDF.

APÊNDICE C – REGISTROS RETIDOS NA TRIAGEM AUTOMATIZADA E SELECIONADOS PELOS REVISORES HUMANOS, MAS NÃO INCLUÍDOS NO CONJUNTO FINAL

Quadro C1 – Registros retidos pela triagem automatizada e selecionados pelos revisores humanos para recuperação do texto completo, mas não incluídos no conjunto final

Nº	Artigos	Critério da IA
1	Preoperative GLP-1 Receptor Agonists and Risk of Postoperative Respiratory Complications	IC3E
2	Association of glucagon-like peptide receptor 1 agonist therapy with the presence of gastric contents in fasting patients undergoing endoscopy under anesthesia care: a historical cohort study	IC3A, IC3B
3	THE IMPACT OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS USE PRIOR TO ESOPHAGOGASTRODUODENOSCOPY	IC3A, IC3B
4	GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONIST (GLP-1 RA) USE PREDICTS RETAINED GASTRIC CONTENTS BUT NOT INCREASED RISK OF MAJOR ADVERSE ANESTHESIA-RELATED EVENTS DURING UPPER ENDOSCOPY	IC3A, IC3B
5	INCIDENCE OF ASPIRATION PNEUMONITIS AND EMERGENT INTUBATION IN PATIENTS ON GLUCAGON-LIKE PEPTIDE-1 INHIBITORS UNDERGOING UPPER ENDOSCOPY AND COLONOSCOPY: A COMPARATIVE STUDY	IC3A, IC3B, IC3C
6	THE UNSEEN CONSEQUENCES OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS (GLP-1RA) IN PATIENTS UNDERGOING UPPER ENDOSCOPY	IC3A, IC3C
7	ASSOCIATION OF GLP-1RA USE AND ASPIRATION PNEUMONIA AFTER THE ENDOSCOPIC PROCEDURE: REAL-WORLD DATA FROM A LARGE NATIONAL DATABASE	IC3B

8	1247 ARE GLP-1 RECEPTOR AGONISTS ASSOCIATED WITH INCREASED INCIDENCE OF PNEUMONIA AFTER ENDOSCOPY?	IC3A, IC3B
9	THE IMPACT OF SEMAGLUTIDE ON RETAINED GASTRIC CONTENTS DURING ENDOSCOPY	IC3A, IC3D
10	CLINICAL IMPACT OF GLP-1RA ON ENDOSCOPY: A RETROSPECTIVE COHORT STUDY IN ENDOSCOPY PATIENTS	IC3A
11	883 OUTCOMES ASSOCIATED WITH GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONIST USE IN PATIENTS UNDERGOING OUTPATIENT UPPER ENDOSCOPY	IC3A, IC3B
12	INCREASED RISK OF RESIDUAL GASTRIC CONTENT IN PATIENTS ON GLP-1 AGONISTS UNDERGOING ESOPHAGOGASTRODUODENOSCOPY	IC3A, IC3C
13	Pulmonary aspiration of gastric contents in two patients taking semaglutide for weight loss	IC3A, IC3B
14	Considerations of delayed gastric emptying with peri-operative use of glucagon-like peptide-1 receptor agonists	ML rescue; no specific IC3 domain
15	Risk of Aspiration Pneumonitis After Elective Esophagogastroduodenoscopy in Patients on Glucagon-Like Peptide-1 Receptor Agonists	IC3A, IC3B
16	Effects of glucagon-like peptide-1 receptor agonists on upper endoscopy in diabetic and nondiabetic patients	IC3A
17	Perioperative management of long-acting glucagon-like peptide-1 (GLP-1) receptor agonists: concerns for delayed gastric emptying and pulmonary aspiration	IC3A, IC3B
18	Is it necessary to stop glucagon-like peptide-1 receptor agonists prior to endoscopic procedure? A retrospective study	IC3A, IC3B
19	Risks of peri- and postoperative complications with glucagon-like peptide-1 receptor agonists	IC3A, IC3E
20	Point-of-Care Gastric Ultrasound to Identify a Full Stomach on a Diabetic Patient Taking a Glucagon-Like Peptide 1 Receptor Agonist	IC3A
21	Should We Stop Glucagon-Like Peptide-1 Receptor Agonists Before Surgical or Endoscopic Procedures? Balancing Limited Evidence With Clinical Judgment	IC3B
22	Anesthesia Considerations for a Patient on Semaglutide and Delayed Gastric Emptying	IC3A, IC3B
23	Perioperative glucagon-like peptide-1 receptor agonists-induced gastroparesis - Is gastric ultrasound the answer?	IC3A
24	Glucagon-like peptide-1 receptor agonist use and perioperative outcomes after anterior cervical discectomy and fusion: a propensity-matched cohort study	IC3E
25	Gastric Emptying With Metoclopramide in GLP-1 Agonist Patients Undergoing Elective Surgery	IC3A
26	Gastric Ultrasound, as an Indicator for Intubation in Short non-Laparoscopic Surgery in Patients Receiving Glucagon-Like Peptide-1 (GLP1) Receptor Agonist. A comparative Randomized Study	IC3A
27	Airway Management in the Age of GLP-1 Receptor Agonists	IC3C
28	Letter to Editor on “As Few as Three Months of Preoperative Semaglutide Exposure Prior to Total Knee Arthroplasty Is	IC3D, IC3E

	Associated With Reduced Postoperative Adverse Events in Patients Who Have Type II Diabetes”	
29	Risk of Postoperative Aspiration Pneumonia in Patients Undergoing ACDF With Preoperative Use of GLP-1 Receptor Agonizts	IC3B
30	Peri-Operative Use of Glucagon-Like Peptide 1 (GLP-1) Agonists Is Associated with Retained Gastric Contents During Upper Endoscopy and Procedure Abortion, But Not With Procedural Complications	IC3A
31	Impact of a One-Week Discontinuation of Semaglutide on Gastric Retention and Endoscopic Mucosal Visibility Scores: A Single-Center, Prospective, Observational Study	IC3D
32	The Effect of Glucagon-like Peptide-1 Receptor Agonists (GLP-1 RA) on Gastric Emptying Before Elective Surgery	IC3A
33	Preoperative Metoclopramide to Enhance Gastric Emptying in Patients on Glucagon-Like Peptide-1 Receptor Agonists for Weight Loss: A Randomised Controlled Trial With Ultrasound Assessment	IC3A
34	Effect of discontinuing semaglutide prior to surgery on perioperative glycemic control: A pilot study	IC3D
35	The fasting and the furious: reconciling fasting guidelines with glucagon-like peptide-1 receptor agonists, ‘Sip-til-Send’ policies and gastric ultrasound	IC3A
36	Risk and prevention of perioperative pulmonary aspiration caused by delayed gastric emptying associated with semaglutide	IC3A, IC3B
37	Effect of Prolonged Fasting Time on Gastric Residual Volume in Patients Taking GLP1 Receptor Agonists	IC3A
38	THE IMPACT OF PRE-PROCEDURAL DIET ON THE RISK OF RESIDUAL GASTRIC CONTENTS IN EGD WITH GLP-1 RECEPTOR AGONISTS	IC3A
39	PERIOPERATIVE RISK OF ASPIRATION AND POSTOPERATIVE COMPLICATIONS WITH GLP-1 THERAPY (PROTECT) STUDY	IC3E
40	REDUCED RISK OF RETAINED GASTRIC CONTENTS AFTER INSTITUTIONAL IMPLEMENTATION OF HOLDING GLUCAGON-LIKE PEPTIDE 1 RECEPTOR AGONISTS PRIOR TO ENDOSCOPIC EVALUATION	IC3A
41	Evaluate the Detection of Retained Gastric Contents and Assess Safety Using the Flower Capsule Endoscopy in Healthy Individuals and GLP-1 Receptor Agonist Users	IC3A
42	Semaglutide and Postoperative Outcomes in Nondiabetic Patients Following Body Contouring Surgery	IC3D, IC3E
43	Effect of Fasting Recommendations on Residual Gastric Contents Among Patients Using Glucagon-like Peptide-1 Receptor Agonists: A Randomised Controlled Trial	IC3A
44	The “pandemic” increase of GLP-1 receptor agonists use and the time of discontinuation before anesthesia: Something new?	IC3D
45	Glucagon-like Peptide 1 Receptor Agonists and Risk of Adverse Anesthesia Events: Caution Warranted When Interpreting Associations With Pulmonary Aspiration	IC3B
46	Postoperative Aspiration Pneumonia Among Adults Using GLP-1 Receptor Agonists	IC3B

47	Coming full circle: could glucagon-like peptide-1 receptor agonists confer a net perioperative benefit?	IC3A, IC3B, IC3E
48	Managing patients taking glucagon-like peptide-1 receptor agonists: proceed with caution	IC3A, IC3B
49	Glucagon-like-peptide-1 (GLP-1) receptor agonist use and the risk of pulmonary aspiration in patients undergoing surgery	IC3B
50	Glucagon-Like Peptide-1 Receptor Agonists in Gynecologic Surgery	IC3A
51	Risk of perioperative cardiorespiratory complications and mortality associated with preoperative glucagon-like peptide-1 receptor agonist use in type 2 diabetes mellitus: a nationwide-score matched	IC3A, IC3B, IC3E
52	Impact of Glucagon-like Peptide-1 Receptor Agonists (GLP-1 RAs) on Increased Residual Gastric Content in Patients With and Without Concurrent Colonoscopy: A Retrospective Case-Control Study	IC3A
53	Prevalence and predictors of residual gastric content in patients with type 2 diabetes on GLP-1 receptor agonists: A prospective observational study	IC3A
54	Assessment of Gastric Content Using Gastric Ultrasound in Patients on Glucagon-Like Peptide-1 Receptor Agonists Before Anesthesia	IC3A
55	Glucagon Like Peptide-1 Receptor Agonist Use Not Related With Gastric Volume After Preoperative Fasting: A Prospective Observational Study	IC3A
56	GLP-1 receptor agonists and delayed gastric emptying: implications for invasive cardiac interventions and surgery	IC3A
57	Multisociety Clinical Practice Guidance for the Safe Use of Glucagon-like Peptide-1 Receptor Agonists in the Perioperative Period	IC3A
58	Risk of periprocedural regurgitation of gastric contents in patients on glucagon-like peptide-1 receptor agonists: a two-centre nested case-control study	IC3B, IC3E
59	Perioperative management of patients taking glucagon-like peptide-1 receptor agonists: implementation of complex fasting recommendations. Response to Br J Anaesth 2025; 136: 375-6	IC3A
60	A restrictive fasting regimen is not necessary in patients taking glucagon-like peptide-1 receptor agonists. Response to Br J Anaesth 2025; 135: 1816-8	IC3A, IC3B
61	Use of gastric ultrasound to identify GLP-1RA users at high risk of aspiration during surgery	IC3A
62	Impact of GLP-1 receptor agonists on gastrointestinal function and symptoms	IC3B
63	Point-of-care gastric ultrasound: Redefining aspiration risk assessment in anesthesia	IC3A, IC3B
64	Fasting Duration, Not Timing of Last GLP-1 Dose, Predicts Aspiration Risk Indicators: A Prospective Point-of-Care Gastric Ultrasound Study	IC3A
65	Ultrasound assessment of preoperative gastric volume in fasted diabetic surgical patients: A prospective observational cohort study on the effects of glucagon-like peptide-1 agonists on gastric emptying	IC3A

66	Can the incidence of voluntarily reported side effects support the argument for a clinically relevant GLP-1RA tachyphylaxis? Comment on Br J Anaesth 2025; 134:1486-1496	IC3A, IC3B, IC3E
67	Impact of Glucagon-Like Peptide-1 Receptor Agonists on Retained Gastric Contents During Esophagogastroduodenoscopy: A Propensity Score-Matched Study	IC3A, IC3B
68	Gastroparesis, a diabetic complication causing further, even serious, complications: How to prevent its worsening?	IC3A, IC3B
69	Incretin impact on gastric function in obesity: physiology, and pharmacological, surgical and endoscopic treatments	IC3A
70	Clinical Consequences of Delayed Gastric Emptying With GLP-1 Receptor Agonists and Tirzepatide	IC3A, IC3B
71	Preoperative GLP-1 Receptor Agonist Use and Postoperative Outcomes Following Operative Ankle Fracture Repair in Patients With Type 2 Diabetes: A National Database Study	IC3E

Fonte: Dados da pesquisa.

Nota: Este quadro apresenta os 71 registros presentes simultaneamente entre os 123 registros retidos pelo módulo automatizado de títulos e resumos e os 118 registros selecionados pelos revisores humanos para recuperação do texto completo, mas que não integraram o conjunto final de 12 estudos. Esses registros não correspondem aos 16 falsos positivos inicialmente identificados no módulo exploratório de leitura de textos completos; IC3A = conteúdo gástrico residual/esvaziamento gástrico; IC3B = aspiração pulmonar ou brônquica; IC3C = complicações de via aérea; IC3D = suspensão ou intervalo de interrupção de semaglutida/GLP-1RA; IC3E = complicações perioperatórias gerais.

APÊNDICE D – COMPARAÇÃO INDIVIDUAL ENTRE A CLASSIFICAÇÃO DA INTELIGÊNCIA ARTIFICIAL EM TEXTO COMPLETO E AS DECISÕES DO PADRÃO DE REFERÊNCIA HUMANO

Quadro D1 – Comparação individual entre a classificação da inteligência artificial no módulo de texto completo e as decisões do padrão de referência humano

Nº	Artigo	Decisão da IA	Critério/domínio da IA	Decisão humana no texto completo	Motivo humano/categoria de exclusão	Classificação na validação	Comentário
1	05_No association between semaglutide and postoperative pneumonia	Incluído	IC3B, IC3E	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
2	09_CLINICAL OUTCOMES AND SAFETY OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS	Incluído	IC3A, IC3B, IC3E	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
3	103_GLP-1 receptor agonist increase retained gastric contents on EGD	Incluído	IC3A, IC3E	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
4	10_SEMAGLUTIDE IS AN INDEPENDENT RISK FACTOR FOR RETAINED GASTRIC CONTENTS	Incluído	IC3A, IC3E	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
5	113_Gastric ultrasound in patients receiving semaglutide: a prospective study	Incluído	IC3A, IC3D	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
6	114_Fasting Duration, Not Timing of Last GLP-1 Dose, Predicts Aspiration Risk	Incluído	IC3A	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.
7	115_Ultrasound assessment of preoperative gastric volume in fasted diabetic patients	Incluído	IC3A, IC3B, IC3D, IC3E	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.

8	15_INCIDENCE OF ASPIRATION PNEUMONITIS AND EMERGENT INTUBATION IN PATIENTS	Incluído	IC3B, IC3C, IC3E	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.
9	18_THE IMPACT OF SEMAGLUTIDE ON RETAINED GASTRIC CONTENTS	Incluído	IC3A, IC3B	Excluído	Não foi possível realizar extração quantitativa específica para semaglutida.	FP	Inclusão excessiva pela IA em relação ao conjunto humano final original.
10	19_GLUCAGON-LIKE PEPTIDE 1 RECEPTOR AGONISTS PRIOR TO ENDOSCOPY	Incluído	IC3A, IC3D	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
11	22_ASSOCIATION BETWEEN PERIOPERATIVE GLP-1 AGONIST USE AND POSTOPERATIVE COMPLICATIONS	Incluído	IC3A, IC3B, IC3E	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
12	38_Glucagon-like peptide-1 receptor agonists before upper gastrointestinal endoscopy	Incluído	IC3B, IC3E	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
13	42_Effect of various perioperative semaglutide interruption intervals on residual gastric content	Incluído	IC3A, IC3B, IC3D	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
14	49_Relationship between perioperative semaglutide use and residual gastric content	Incluído	IC3A, IC3B, IC3D	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
15	57_Glucagon-Like Peptide-1 Receptor Agonist Use and Residual Gastric Content	Incluído	IC3A, IC3D	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
16	58_Risk of Aspiration Pneumonitis After Elective Esophagogastroduodenoscopy	Incluído	IC3B, IC3E	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.
17	59_Association of glucagon-like peptide receptor 1 agonist	Incluído	IC3A, IC3B,	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem	FP	Inclusão excessiva pela IA.

	therapy with perioperative outcomes		IC3C, IC3E		separação específica para semaglutida.		
18	65_Real-World Impact of GLP-1 Receptor Agonists on Endoscopic Patient Outcomes	Incluído	IC3A, IC3B, IC3E	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.
19	67_Relationship between residual gastric content and peri-operative semaglutide use	Incluído	IC3A, IC3B, IC3D	Incluído	—	VP	Estudo incluído ao final e preservado pelo módulo de IA.
20	68_Influence of semaglutide use on the presence of residual gastric solids	Incluído	IC3A	Excluído	Desfecho avaliado fora do contexto perioperatório.	FP	Inclusão excessiva pela IA.
21	71_Glucagon-like peptide-1 receptor agonist use and perioperative outcomes	Incluído	IC3B, IC3E	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.
22	77_Risk of Postoperative Aspiration Pneumonia in Patients Undergoing ACDF	Incluído	IC3B, IC3E	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem separação específica para semaglutida.	FP	Inclusão excessiva pela IA.
23	82_Effect of discontinuing semaglutide prior to surgery on perioperative outcomes	Incluído	IC3A, IC3B, IC3D	Excluído	Não foi possível realizar extração quantitativa específica para semaglutida.	FP	Inclusão excessiva pela IA.
24	86_THE IMPACT OF PRE-PROCEDURAL DIET ON THE RISK OF RESIDUAL GASTRIC CONTENT	Incluído	IC3A	Excluído	Não foi possível realizar extração quantitativa específica para semaglutida.	FP	Inclusão excessiva pela IA.
25	90_Semaglutide and Postoperative Outcomes in Nondiabetic Patients	Incluído	IC3E	Excluído	Os desfechos de interesse não foram relatados.	FP	Inclusão excessiva pela IA.
26	94_Postoperative Aspiration Pneumonia Among Adults Using GLP-1 Receptor Agonists	Incluído	IC3B, IC3E	Excluído	Não foi possível realizar extração quantitativa específica para semaglutida.	FP	Inclusão excessiva pela IA.
27	98_Glucagon-like-peptide-1 receptor agonist use and the risk of pulmonary aspiration	Incluído	IC3B	Excluído	Resultados apresentados para a classe dos GLP-1RA, sem	FP	Inclusão excessiva pela IA.

					separação específica para semaglutida.		
28	99_Impact of GLP-1 Receptor Agonists on Increased Residual Gastric Content	Incluído	IC3A, IC3D	Excluído no conjunto final original; reclassificado como elegível após auditoria humana	Discordância identificada na auditoria; a verificação humana posterior sustentou a elegibilidade.	FP / discordância de auditoria	Manter sinalizado; não interpretar como inclusão confirmada autonomamente pela IA.
29	01_a-pilot-study-examining-the-effects-of-glp-1-receptor-blockade	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
30	02_Semaglutide Treatment for Hyperglycaemia After Renal Transplant	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
31	03_Gastric Emptying and Metabolic Response After a Laparoscopic Procedure	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
32	04_Efficacy and Safety of GLP-1 Medicines for Type 2 Diabetes and Obesity	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
33	06_Implications of GLP-1 agonist use on airway management	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

34	07_Preoperative GLP-1 Receptor Agonist Use and Risk of Postoperative Outcomes	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
35	08_Long term metabolic effects and perioperative safety assessment	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
36	104_GLP-1 receptor agonists and delayed gastric emptying implications	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
37	105_Multisociety Clinical Practice Guidance for the Safe Use of GLP-1 RAs	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
38	109_Use of gastric ultrasound to identify GLP-1RA users at high risk	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
39	111_Point-of-care gastric ultrasound: redefining aspiration risk	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
40	112_Tirzepatide for Obstructive Sleep Apnea	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
41	116_Can the incidence of voluntarily reported side effects support clinical decisions	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

42	119_Gastroparesis, a diabetic complication causing further complications	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
43	11_THE UNSEEN CONSEQUENCES OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
44	121_Glucagon-Like Peptide-1 Receptor Agonists in Gynecologic Surgery	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
45	122_Clinical Consequences of Delayed Gastric Emptying With GLP-1	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
46	12_THE IMPACT OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS USE	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
47	13_ASSOCIATION OF GLP-1RA USE AND ASPIRATION PNEUMONIA AFTER THE PROCEDURE	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
48	14_GLP-1 RA USE PREDICTS PERIOPERATIVE OUTCOMES	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
49	16_Erratum Diabetes Care in the Hospital Standards of Care in Diabetes	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

50	17_ARE GLP-1 RECEPTOR AGONISTS ASSOCIATED WITH INCREASED INCIDENCE	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
51	20_CLINICAL IMPACT OF GLP-1RA ON ENDOSCOPY: A RETROSPECTIVE COHORT	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
52	21_OUTCOMES ASSOCIATED WITH GLP-1 RECEPTOR AGONISTS	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
53	23_INCREASED RISK OF RESIDUAL GASTRIC CONTENT IN PATIENTS ON GLP-1	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
54	24_Pulmonary aspiration of gastric contents in two patients taking semaglutide	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
55	27_BARIATRIC ENDOSCOPIC ANTRAL MYOTOMY TECHNICAL FEASIBILITY	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
56	28_Residual Gastric Content and Perioperative Semaglutide Use	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
57	30_Frequency of GLP-1 analog use in diabetic patients with delayed gastric emptying	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
58	32_A pre-clinical animal study of combined intragastric balloon	Excluído	IC1	Excluído	Critério de população humana adulta não atendido ou registro não humano/pré-clínico.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

59	33_One-anastomosis jejunal interposition with gastric resection	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
60	34_One-year follow-up and physiological alterations following endoscopic procedure	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
61	37_Considerations of delayed gastric emptying with peri-operative GLP-1 use	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
62	40_A randomized crossover study of the effects of glutamine	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
63	41_A novel weight-reducing operation: lateral subtotal gastrectomy	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
64	43_Perioperative management of patients on GLP-1 receptor agonists	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
65	44_Effects of GLP-1 receptor agonists on upper gastrointestinal outcomes	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
66	46_The role of gastric function in control of food intake and body weight	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

67	47_GLP-1RA Essentials in Gastroenterology Side Effect Management	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
68	48_Comparison of societal guidance on perioperative management of GLP-1 RAs	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
69	50_Perioperative management of long-acting GLP-1 receptor agonists	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
70	51_Is it necessary to stop GLP-1 receptor agonists before procedures?	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
71	52_Risks of peri- and postoperative complications with GLP-1 RAs	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
72	54_Impact of GLP-1 Receptor Agonists in Gastrointestinal Endoscopy	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
73	55_Should We Stop GLP-1 Receptor Agonists Before Procedures?	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
74	56_Anesthesia Considerations for a Patient on Semaglutide and Delayed Gastric Emptying	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

					casos, revisão, editorial, comentário, diretriz ou similar).		
75	61_GLP-1 Receptor Agonists Used for Medically Supervised Weight Loss	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
76	62_GLP-1 Agonists and General Anesthesia Perioperative Management	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
77	64_Perioperative GLP-1 receptor agonists-induced delayed gastric emptying	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
78	66_Implications of Ozempic and Other Semaglutide Medications for Surgery	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
79	69_Emerging therapies in the treatment of diabesity beyond GLP-1	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
80	70_Laparoscopic sleeve gastrectomy: more than a restrictive bariatric procedure	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
81	72_The Use of GLP-1 Agonists in the Perioperative Period	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

82	76_Letter to Editor on 'As Few as Three Months of Preoperative Semaglutide'	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
83	78_Peri-Operative Use of GLP-1 Agonists	Excluído	EC3	Excluído	Exposição a outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida.	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
84	80_The Effect of GLP-1 Receptor Agonists on Perioperative Outcomes	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
85	81_Preoperative Metoclopramide to Enhance Gastric Emptying in Patients Using GLP-1 RAs	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
86	83_The fasting and the furious	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
87	84_Risk and prevention of perioperative pulmonary aspiration caused by GLP-1 RAs	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
88	92_The pandemic increase of GLP-1 receptor agonists use and the perioperative risk	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.
89	93_GLP-1 Receptor Agonists and Risk of Adverse Events	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

					casos, revisão, editorial, comentário, diretriz ou similar).		
90	95_GLP-1 Receptor Agonists in the Peri-Operative Period	Excluído	EC4	Excluído	Tipo de publicação inelegível ou ausência de dados originais extraíveis (relato ou série de casos, revisão, editorial, comentário, diretriz ou similar).	VN	A exclusão pela IA concordou com o conjunto final de referência humano.

Fonte: Dados da pesquisa.

Nota: IC3A = conteúdo gástrico residual/esvaziamento gástrico; IC3B = aspiração pulmonar ou brônquica; IC3C = complicações de via aérea; IC3D = suspensão ou intervalo de interrupção de semaglutida/GLP-1RA; IC3E = complicações perioperatórias gerais; EC3 = outro GLP-1RA ou dados em nível de classe sem resultados extraíveis específicos para semaglutida; EC4 = relato de caso, série de casos, revisão, editorial, comentário, diretriz ou outro desenho não original/inelegível; IC1 = critério de população humana adulta não atendido. GLP-1RA = agonista do receptor do peptídeo semelhante ao glucagon-1; IA = inteligência artificial; VP = a IA incluiu/priorizou e o humano incluiu; FP = a IA incluiu/priorizou, mas o humano excluiu ou o artigo não integrou o conjunto final original; VN = a IA excluiu e o humano excluiu. Não ocorreram falsos negativos entre os 90 PDFs processados. Para um registro positivo pela IA, a auditoria humana subsequente sugeriu elegibilidade; o caso foi sinalizado como discordância de auditoria. Os títulos dos artigos foram mantidos no idioma original para preservar sua identificação bibliográfica

APÊNDICE E – ARTIGO CIENTÍFICO SOBRE DESENVOLVIMENTO E VALIDAÇÃO DO SAFESCREEN SUBMETIDO AO PLOS ONE EM PROCESSO DE REVISÃO

PLOS One

Can artificial intelligence reduce systematic review workload? Internal validation of an auditable screening workflow

--Manuscript Draft--

Manuscript Number:	PONE-D-26-36175
Article Type:	Research Article
Full Title:	Can artificial intelligence reduce systematic review workload? Internal validation of an auditable screening workflow
Short Title:	Auditable AI workflow for systematic reviews
Corresponding Author:	Vinicius Remus Ballotin, M.D University of Caxias do Sul: Universidade de Caxias do Sul Caxias do Sul, Rio Grande do Sul BRAZIL
Keywords:	systematic reviews; evidence synthesis; deduplication; citation screening; Machine learning; large language models; validation
Abstract:	Objective: To develop and internally validate an automated, auditable workflow for systematic review deduplication, title/abstract screening, and exploratory full-text prioritization, using protocol-based human reviewer decisions as the reference standard. Methods: We used bibliographic records from a systematic review of semaglutide and perioperative outcomes. The workflow included three sequential modules: deterministic and fuzzy deduplication; protocol-derived title/abstract screening with a conservative machine-learning rescue step; and an exploratory generative artificial intelligence (AI) module for full-text prioritization. Automated decisions were compared with protocol-based human decisions documented within a PRISMA-guided review process using stage-specific confusion matrices, sensitivity, specificity, precision, F1 score, accuracy, Cohen's kappa, overlap measures, and workload reduction estimates. Results: The corpus contained 16,159 records. Automated deduplication removed 2,788 duplicates, leaving 13,371 records for title and abstract screening. At title-year cluster level, deduplication achieved sensitivity 0.933, precision 0.941, F1 score 0.937, Jaccard index 0.882, and Cohen's kappa 0.927, yielding higher point estimates than Rayyan and ASySD. The title and abstract module retained 123 records. Compared with the 118 records selected by human reviewers for full-text retrieval, recall was 0.703, precision 0.675, F1 score 0.689, and Jaccard overlap 0.525. The workflow failed to retain 35 human-selected records, although none belonged to the final included set; all 12 studies finally included by human reviewers were retained. Among 90 PDFs processed by the full-text AI module, 28 were prioritized and 62 deprioritized. Using final human inclusion as the reference standard, sensitivity was 1.000, precision 0.429, F1 score 0.600, and Cohen's kappa 0.508. Conclusion: This study demonstrates the feasibility of an auditable workflow for systematic review deduplication, title and abstract screening,

	and AI-assisted full-text prioritization. Although externally unvalidated and topic-specific, the framework provided transparent decision tracking and preserved all studies ultimately included in the reference review, supporting further evaluation as a human-supervised evidence synthesis tool.
Order of Authors:	Vinicius Remus Ballotin, M.D
	Pedro Lucas Carneiro Ferreira
	Daniel Volquind, PhD
	Luciano da Silva Selistre, PhD
	Lucas Goldmann Bigarella, M.D
Opposed Reviewers:	
Additional Information:	
Question	**Response**

Can artificial intelligence reduce systematic review workload? Internal validation of an auditable screening workflow

Short title: Auditable AI workflow for systematic reviews

Article type: Research Article

Authors: Vinicius Remus Ballotin[1*], Pedro Lucas Carneiro Ferreira[2], Daniel Volquind[3], Luciano da Silva Selistre[4], Lucas Goldmann Bigarella[5].

[1]Department of Health Sciences, University of Caxias do Sul, Caxias do Sul, 95070-560, Rio Grande do Sul, Brazil, vrballotin@ucs.br

[2]School of Medicine, University of Caxias do Sul, Caxias do Sul, 95070-560, Rio Grande do Sul, Brazil, plcferreira@ucs.br

[3]Department of Anesthesia and Pain Medicine, University of Caxias do Sul, Caxias do Sul, 95070-560, Rio Grande do Sul, Brazil, danielvolquind@gmail.com

[4]Department of Nephrology and Biostatistics, University of Caxias do Sul, Caxias do Sul, 95070-560, Rio Grande do Sul, Brazil, lsselist@ucs.br

[5]Institute of Cardiology, University Foundation of Cardiology of Rio Grande do Sul, Porto Alegre, 90620-001, Rio Grande do Sul, Brazil, lgbigarella@ucs.br

***Corresponding author: Vinicius Remus Ballotin,** Department of Health Sciences, University of Caxias do Sul, Brazil. vrballotin@ucs.br

Abstract

Objective: To develop and internally validate an automated, auditable workflow for systematic review deduplication, title/abstract screening, and exploratory full-text prioritization, using protocol-based human reviewer decisions as the reference standard.

Methods: We used bibliographic records from a systematic review of semaglutide and perioperative outcomes. The workflow included three sequential modules: deterministic and fuzzy deduplication; protocol-derived title/abstract screening with a conservative machine-learning rescue step; and an exploratory generative artificial intelligence (AI) module for full-text prioritization. Automated decisions were compared with protocol-based human decisions documented within a PRISMA-guided review process using stage-specific confusion matrices, sensitivity, specificity, precision, F1 score, accuracy, Cohen's kappa, overlap measures, and workload reduction estimates.

Results: The corpus contained 16,159 records. Automated deduplication removed 2,788 duplicates, leaving 13,371 records for title and abstract screening. At title-year cluster level, deduplication achieved sensitivity 0.933, precision 0.941, F1 score 0.937, Jaccard index 0.882, and Cohen's kappa 0.927, yielding higher point estimates than Rayyan and ASySD. The title and abstract module retained 123 records. Compared with the 118 records selected by human reviewers for full-text retrieval, recall was 0.703, precision 0.675, F1 score 0.689, and Jaccard overlap 0.525. The workflow failed to retain 35 human-selected records, although none belonged to the final included set; all 12 studies finally included by human reviewers were retained. Among 90 PDFs processed by the full-text AI module, 28 were prioritized and 62 deprioritized. Using final human

inclusion as the reference standard, sensitivity was 1.000, precision 0.429, F1 score 0.600, and Cohen's kappa 0.508.

Conclusion: This study demonstrates the feasibility of an auditable workflow for systematic review deduplication, title and abstract screening, and AI-assisted full-text prioritization. Although externally unvalidated and topic-specific, the framework provided transparent decision tracking and preserved all studies ultimately included in the reference review, supporting further evaluation as a human-supervised evidence synthesis tool.

Keywords: systematic reviews; evidence synthesis; deduplication; citation screening; machine learning; large language models; validation

Introduction

Systematic reviews are central to biomedical evidence synthesis, but remain slow, labor-intensive, and vulnerable to operational error, particularly during record management, deduplication, screening, full-text retrieval, and documentation[1-6]. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 and PRISMA-S emphasize transparent reporting of selection decisions, search methods, and reproducibility safeguards[7-10].

Deduplication and screening carry different risks. Over-aggressive deduplication may wrongly merge distinct studies, while false-negative screening may exclude eligible evidence before full assessment[11-14]. Automation tools, including text mining, active learning, rule-based classifiers, deduplication systems, and large language models, may reduce workload, but their performance varies by topic, metadata quality, eligibility criteria, and workflow stage[15-25]. Accuracy alone is therefore insufficient; preservation of included studies, error location, and auditability are more important.

This secondary methodological validation used the corpus and reviewer decisions from a completed perioperative semaglutide review to evaluate an auditable workflow for deduplication, title and abstract screening, and exploratory full-text prioritization. Automated decisions were compared with human decisions to assess agreement, errors, preservation of final inclusions, reproducibility, and workload reduction.

Automation bottlenecks in systematic reviews

Automation may reduce the time and cost of systematic reviews, but its value depends on the task being automated and the tolerance for error at that stage. A tool suitable for prioritizing records may be unsafe for exclusion, and performance in one topic may not generalize when

terminology is sparse, abstracts are incomplete, or eligibility depends on full-text assessment. Integrated systems such as ReviewGenie reflect a shift from single-task tools toward modular workflows for search support, data collection, cleaning, deduplication, filtering, and title and abstract screening[26]. These systems better reflect real review practice, but their value remains context dependent.

Bibliographic deduplication and record linkage

Bibliographic deduplication has moved beyond simple reference-manager matching. Tools such as Deduklick, Deduplicator, ASySD, web-based systems, and record-linkage approaches combine exact identifiers, metadata normalization, text similarity, clustering, and explainable rules[11-14, 27, 28]. These methods address a key systematic review risk: reducing workload without removing unique studies. Exact DOI or PMID matching is specific but incomplete, whereas fuzzy title or abstract matching improves recall but may incorrectly merge distinct reports. Deduplication is therefore better viewed as a record-linkage problem, supporting evaluation at both record and cluster levels, because multiple records may represent the same study and the retained citation must remain adequate for downstream screening.

Title and abstract screening and active learning

Title and abstract screening are a major target for automation because most retrieved records are excluded and eligible studies are uncommon. Early citation-classification and text-mining studies showed that workload can be reduced by ranking or classifying records according to probable relevance[16, 17]. ASReview extended this approach through open-source active learning with human-in-the-loop screening[18], later adding multi-agent and multi-reviewer functionality[29]. Rayyan is widely used for collaborative screening and conflict resolution,

although it primarily supports workflow management rather than autonomous eligibility assessment[30, 31]. Evaluations of Abstrackr, EPPI-Reviewer, and SWIFT-Active Screener show that performance varies by dataset, prevalence of included studies, metadata quality, and separation between eligible and ineligible records[19, 21, 32, 33]. Thus, screening performance is not an intrinsic property of a tool, but depends on the review question, eligibility criteria, and acceptable trade-off between workload reduction and false-negative risk.

Large language models and full-text eligibility assessment

Large language models have increased interest in automating systematic review tasks, including title and abstract screening, full-text eligibility assessment, and data extraction[34]. GPT-based and other large language model (LLM) approaches have shown variable sensitivity and specificity, with performance influenced by prompt design, decision thresholds, review topic, input text, and comparator standard[20, 22-25, 35-44]. This variability matters because screening is not a generic classification task: high average accuracy is insufficient if eligible studies are excluded before human assessment. Position statements on AI in evidence synthesis therefore emphasize transparency, validation, human oversight, and caution against replacing established methods without adequate evidence[45, 46]. Concerns about hallucination, unstable outputs, and limited reproducibility further support conservative use, especially for full-text decisions[47-49]. In this context, LLM outputs are better treated as prioritization or audit signals than as final eligibility decisions[50].

Validation metrics and safety priorities

Performance metrics should match the decision being evaluated. In citation screening, accuracy is a weak primary measure because eligible studies are uncommon; a system may exclude

many irrelevant records yet fail if it misses studies requiring full-text review. Sensitivity, negative predictive value, false-negative review, and preservation of the final included set are therefore more informative than accuracy alone[16, 17, 51].

Work saved is useful but not a validity measure. It estimates avoided manual screening and depends on eligible-study prevalence and the recall threshold considered acceptable[51]. Cohen's kappa also requires caution in highly imbalanced datasets and should be interpreted alongside the confusion matrix[52, 53].

These concerns are greater in staged workflows. False positives mainly increase workload, whereas false exclusions during deduplication or title and abstract screening may remove studies before human assessment. We therefore report performance separately for deduplication, title and abstract screening, and exploratory full-text prioritization, while tracking all studies finally included by human reviewers across the workflow.

Methods

Ethics and reporting considerations

Formal ethics committee review was not required because the study used bibliographic metadata, reviewer decisions, and published articles only. It involved no human participants, clinical interventions, animals, patient-level data, or identifiable information. PDFs were used only for text extraction and automated classification; CHART-relevant generative-AI details were documented where applicable.

Study design

This secondary methodological validation study used the bibliographic corpus and protocol-based reviewer decisions from a completed systematic review and meta-analysis of

perioperative semaglutide use[54]. The clinical review provided the validation dataset; the present study evaluated a supervised automated workflow applied to that dataset. Automated decisions were compared with human decisions recorded during a PRISMA-guided review process. The analysis focused on agreement with reviewers, preservation of the final included studies, stage-specific errors, auditability, and workload reduction. Fig 1 summarizes the workflow and its validation against the human reference standard.

Fig 1. Automated workflow and validation structure. The workflow normalized bibliographic data, applied exact and fuzzy deduplication with predefined safeguards, screened titles and abstracts using protocol-derived eligibility gates, and used a conservative machine-learning rescue step for borderline records. An exploratory AI module was used for full-text prioritization. Automated decisions and audit outputs were compared with protocol-based reviewer decisions from the completed systematic review.

Human reference standard

PRISMA was used to document record flow through identification, deduplication, title and abstract screening, full-text assessment, and final inclusion. Eligibility followed the prespecified review protocol. Records were screened independently by two reviewers (VRB and PLCF) at title and abstract and full-text stages, with disagreements resolved by a third reviewer (CSB). Reference lists were searched manually. The deduplication reference standard was based on duplicate status recorded during the original review, using title, DOI, PMID, PMCID, authorship, year, journal, abstract, and record-management identifiers. The automated workflow was applied retrospectively after the human review and did not inform the reference standard.

Phase 1: Dataset source and bibliographic data

The validation dataset was derived from the completed semaglutide perioperative systematic review and meta-analysis. All bibliographic information, search records, database sources, record counts, and human selection decisions were taken from the original review files and supplementary material. No new bibliographic search was conducted for the present methodological validation.

In the original review, seven electronic sources were searched from inception through April 2026: Scopus, Web of Science, EMBASE, MEDLINE/PubMed, Latin American and Caribbean Health Sciences Literature, Cochrane Library, and OpenGrey. The search yielded 16,159 records. After duplicate removal, 13,377 records were screened by title and abstract. Human reviewers selected 118 records for full-text retrieval; six reports could not be retrieved, leaving 112 full-text articles assessed for eligibility. Twelve studies were included in the final review. The full search strategy, database-specific syntax, and study-selection flow were reported in the original review and its supplementary material.

For the present study, these bibliographic records and reviewer decisions were used as the reference dataset for internal validation of the automated workflow. Records were consolidated as comma-separated values files and processed in Python (v. 3.10) using Google Colaboratory. The workflow normalized titles, abstracts, publication years, first authors, and bibliographic identifiers, including DOI, PMID, and PMCID.

Phase 2: Automated workflow modules

Deduplication module

The deduplication module used a hierarchical record-linkage strategy. Bibliographic fields were first normalized, including title, abstract, publication year, first author, DOI, PMID, and

PMCID. Exact matching was then performed using DOI, PMID, and PMCID when available. Records without reliable identifiers were compared using normalized title, year, first author, and abstract fields.

Fuzzy matching was performed with RapidFuzz (v. 3.14.5). The workflow used a title similarity threshold of 97 and an abstract similarity threshold of 95. Abstract comparisons were restricted to records with at least 150 characters, with blocking based on the first 80 characters of the normalized abstract. To reduce inappropriate merging, fuzzy matches were constrained by publication year and first-author safeguards. Records with discordant years beyond the prespecified tolerance or incompatible first authors were not merged when these fields were available.

Candidate duplicate links were resolved into clusters using a Union-Find structure. Within each cluster, one record was retained according to a completeness score that prioritized the presence of bibliographic identifiers, publication metadata, and a more complete abstract.

Title and abstract screening module

After deduplication, unique records were screened using rules derived from the eligibility criteria of the original review. The primary decision used three sequential gates: presence of glucagon-like peptide-1 (GLP-1) receptor agonist or semaglutide-related terms; presence of a perioperative, procedural, endoscopic, surgical, anesthetic, aspiration, or airway-management context; and presence of at least one outcome domain relevant to the review.

Operational exclusions were applied for non-original publications, animal studies, pediatric records, records concerning other GLP-1 receptor agonists without semaglutide-specific information, and non-English markers. Records that met the intervention and context gates but failed the outcome gate were eligible for a machine-learning rescue step. This step used term

frequency–inverse document frequency (TF-IDF) word and character features with logistic regression and balanced class weights.

The rescue threshold was set at 0.90. This high threshold limited rescue to records with strong textual compatibility with the review question. The title and abstract module was therefore designed to favor sensitivity at the workflow level while keeping the machine-learning rescue step conservative. Each record received a binary automated decision and an audit label describing the gates met or the reason for exclusion.

Exploratory full-text AI module

Records retained after title and abstract screening were eligible for AI-assisted full-text processing when a PDF was available. Text was extracted from PDFs using PyMuPDF (v. 1.28.0) with block-level extraction get_text("blocks"). Extracted text was submitted through the Anthropic Claude application programming interface (API) for structured classification against the inclusion and exclusion criteria of the original review.

The model used was Claude Sonnet (model identifier: claude-sonnet-4-6). For each PDF, a maximum of 35,000 extracted characters was submitted. API calls used temperature 0 and max_tokens = 1200. The model was instructed to return a structured JavaScript Object Notation (JSON) object containing the eligibility decision, criteria met, exclusion reason if applicable, key supporting evidence, and brief reasoning. Ninety independent API calls were performed, one per PDF, using the same prompt template and settings for all articles. Calls were executed from Caxias do Sul, Brazil, on May 5, 2026.

This module was exploratory and was used for full-text prioritization only. Its outputs did not replace human full-text assessment and did not modify the human reference standard. Reporting of the generative-AI component included CHART-relevant details where applicable,

including model identification, access route, prompt structure, query date and location, number of independent API calls, input limits, output format, reference standard, and reproducibility materials[55]. Further details are provided in S1 File, Tables S1–S9.

Phase 3: Validation against human decisions

Automated decisions were compared with protocol-based reviewer decisions from the completed systematic review. Where a binary comparator was available, performance was summarized using confusion matrices, sensitivity/recall, specificity, precision, negative predictive value, F1 score, accuracy, and Cohen's kappa. Jaccard overlap was used when agreement between automated and human-selected sets was the more relevant comparison.

Direct comparator analyses were limited to deduplication, for which Rayyan and ASySD outputs could be evaluated against the same human reference standard. Screening platforms such as ASReview and SWIFT-Active Screener were not directly compared because they require active-learning, reviewer-in-the-loop, ranking-based evaluation rather than fixed binary classification.

Safety was assessed by tracking whether any study finally included by human reviewers was removed during deduplication, excluded at title and abstract screening, or not prioritized by the full-text AI module. False negatives among final included studies were treated as the key error, because early exclusion may prevent later human assessment.

Workload reduction was reported as the proportion of records or PDFs removed, excluded, or deprioritized before manual assessment at each stage. It was interpreted as an operational outcome, not evidence of validity, because work-saved estimates depend on eligible-study prevalence, assumptions about manual review, and the recall threshold considered acceptable[51].

Error analysis and workload estimation

Error analysis was performed separately for each workflow stage. False positives and false negatives were reviewed to determine whether discrepancies arose from duplicate adjudication, incomplete bibliographic metadata, absence of relevant information in the title or abstract, overly broad automated retention, PDF unavailability, AI over-inclusion, or disagreement with the human reference decision.

Particular attention was given to studies finally included by human reviewers. Any automated exclusion of a finally included study was treated as a critical safety error. For the full-text AI module, false positives were reviewed manually and categorized as AI over-inclusion, GLP-1 receptor agosnsit class-level reporting without semaglutide-specific extractable data, absence of eligible perioperative outcomes, non-original or ineligible study design, or possible inconsistency in the human reference standard.

Workload reduction was estimated as the proportion of records or PDFs removed, excluded, or deprioritized by the automated workflow before manual review at the corresponding stage. These estimates were interpreted as operational measures only and were not used as evidence of classification validity.

Workload and cost estimation

Workload reduction was estimated as a scenario-based operational outcome. Manual workload for the original review was estimated from the number of duplicate records removed, records screened by title and abstract, and full-text reports assessed by human reviewers. The assisted-workflow workload was estimated from the number of records or PDFs remaining for

human review after automated deduplication, title and abstract screening, and full-text AI prioritization.

The time assumptions were prespecified as 20 seconds per duplicate record adjudicated, 1 minute per title and abstract record screened, and 20 minutes per full-text article assessed. Computational time was recorded separately and was not added to human workload. API cost was reported as direct API usage cost only. Development time, debugging, reviewer training, institutional subscriptions, article retrieval, and manual verification of false-positive automated inclusions were not included.

Statistical analysis

Analyses were descriptive. Automated classifications were compared with the human reference standard defined for each workflow stage. For binary classifications, we calculated sensitivity/recall, specificity, precision, negative predictive value, F1 score, accuracy, workload reduction, and Cohen's kappa. Jaccard index was used when agreement was based on overlap between automated and human-selected record sets rather than a complete classification table.

Ninety-five percent confidence intervals were estimated according to the measure. Wilson score intervals were used for binomial proportions, with exact Clopper–Pearson intervals reported as secondary estimates for selected key proportions. For F1 score, Cohen's kappa, Jaccard overlap, and pairwise differences between deduplication methods, 95% confidence intervals were estimated by non-parametric bootstrap with bias-corrected and accelerated intervals using 10,000 resamples. Bootstrap resampling preserved the paired structure of records or clusters, and kappa intervals were constrained to the theoretical range from −1 to 1.

Deduplication was assessed at record and title-year cluster levels, with cluster-level analysis treated as the main operational assessment. The proposed workflow was compared with

Rayyan and ASySD using the same corpus and human duplicate-removal reference standard. For title and abstract screening, the main comparator was the human full-text retrieval set; retention of the final included studies was reported separately as a safety outcome. For the exploratory full-text AI module, analyses were limited to the PDFs processed by the module, using human final inclusion status as the reference standard within that set.

No non-inferiority analysis was performed because this was a single-topic internal validation without a prespecified margin. Cohen's kappa was interpreted alongside the underlying confusion matrices because of its sensitivity to outcome prevalence and class imbalance[52, 53].

Software, model configuration, and reproducibility

The automated workflow was implemented in Python (v. 3.10) and executed in Google Colaboratory. The deduplication module used pandas (v. 2.2.2), RapidFuzz (v. 3.14.5), and tqdm (v. 4.69.0); the title and abstract screening module used scikit-learn (v. 1.5.2) with TF-IDF word and character features and logistic regression; and the exploratory full-text module used PyMuPDF (v. 1.28.0) for PDF text extraction and the Anthropic Claude API with Claude Sonnet (model identifier: claude-sonnet-4-6) for structured full-text prioritization.

Prespecified parameters were kept constant across the workflow. These included a title fuzzy-matching threshold of 97, an abstract fuzzy-matching threshold of 95, and a machine-learning rescue threshold of 0.90 for title and abstract screening. For the full-text module, the model identifier, prompt template, input limit, temperature, token limit, execution date, and number of API calls are reported in the full-text AI methods section and summarized in Fig 2.

Fig 2. CHART-style methodological diagram for the exploratory generative-AI full-text module. The diagram summarizes model identification, API access, prompt development, query strategy, performance evaluation, and reproducibility details for the AI-assisted full-text

prioritization module. The module was evaluated against protocol-based human full-text decisions and was not used to replace human assessment or modify the reference standard.

The workflow generated CSV outputs and audit files for deduplication decisions, title and abstract classifications, exclusion reasons, full-text AI decisions, and validation comparisons. File paths and API keys were removed from the shared code. To support reproducibility, the source code, fixed prompt template, parameter settings, and audit-output structure are provided in the GitHub repository. Table 1 summarizes the software, model, and reproducibility configuration of the workflow.

Table 1. Software, model, and reproducibility configuration of the automated workflow.

Component	Configuration
Programming environment	Python (v. 3.10), Google Colaboratory
Deduplication libraries	pandas (v. 2.2.2), RapidFuzz (v. 3.14.5), tqdm (v. 4.69.0)
Screening model	scikit-learn (v. 1.5.2); TF-IDF word and character features with logistic regression
Full-text AI model	claude-sonnet-4-6
API settings	temperature = 0; max_tokens = 1200
Full-text extraction	PyMuPDF (v. 1.28.0); get_text("blocks"); up to 35,000 characters per PDF
Query execution	May 5, 2026; Caxias do Sul, Brazil
Prompt strategy	Fixed prompt; one independent API request per PDF

Component	Configuration
Audit materials	S1 File; Figshare; Zenodo; GitHub

Note. AI = artificial intelligence; API = application programming interface; PDF = Portable Document Format; TF-IDF = term frequency–inverse document frequency. The full-text AI module used a fixed prompt template and one independent API request per PDF. File paths and API keys were removed from the shared code. Source PDFs were used for text extraction within the workflow but were not redistributed.

Data availability

The data underlying the findings of this study are openly available in Figshare at doi:10.6084/m9.figshare.32526549. The archived version of the SafeScreen source code is permanently available in Zenodo at doi:10.5281/zenodo.21412862. The actively maintained source-code repository is available on GitHub at https://github.com/vrballotin/SafeScreen.

Results

Dataset flow

The operational bibliographic corpus contained 16,159 records. Automated deduplication removed 2,788 duplicates, leaving 13,371 unique records for title/abstract screening. The title/abstract module classified 123 records as potentially eligible and excluded 13,248 records. Of the 123 records retained by the automated screening module, 90 full-text PDFs were located and processed by the exploratory full-text AI module; 33 were not available as PDFs for automated full-text evaluation. Human reviewers selected 118 records for full-text retrieval after title and abstract screening. Six reports could not be retrieved, leaving 112 full-text articles assessed for

eligibility; 12 studies were finally included. The dataset flow is shown in Fig 3 and summarized in Table 2.

Fig 3. PRISMA-style flow of the automated workflow and human reference standard. The figure shows the operational corpus, automated duplicate removal, title and abstract screening, PDF availability for the full-text AI module, AI-based full-text prioritization, and comparison with the human reference standard. The human reference standard comprised 118 records selected for full-text retrieval, six reports not retrieved, 112 full-text reports assessed for eligibility, and 12 studies finally included. The dashed arrow indicates comparison with the human reference standard, not automated final inclusion.

Table 2. Study dataset and workflow modules.

Stage/module	Input	Automated logic	Output	Human comparator
Operational corpus	16,159 bibliographic records	Import and field normalization	Corpus prepared for deduplication	Search exports from the underlying systematic review
Deduplication	16,159 records	Exact DOI/PMID/PMCID matching; fuzzy title and abstract matching; year and author safeguards; Union-Find clustering;	2,788 duplicate records removed; 13,371 unique records retained	Human duplicate decisions from the completed review

Stage/module	Input	Automated logic	Output	Human comparator
		completeness-based record retention		
Title and abstract screening	13,371 unique records	Protocol-derived eligibility gates; hard exclusions; TF-IDF logistic-regression rescue at high threshold	123 records retained; 13,248 records excluded	Human records selected for full-text retrieval (n = 118) and final included studies (n = 12)
Exploratory full-text AI module	90 PDFs among the 123 retained records	PDF text extraction; Claude API; structured inclusion/exclusion decision and rationale	28 records prioritized; 62 records excluded	Human full-text decisions among processed PDFs; final included studies tracked

Note. PDF availability limited the automated full-text module to 90 of the 123 records retained after title and abstract screening. DOI = Digital Object Identifier; PMID = PubMed Identifier; PMCID = PubMed Central Identifier; TF-IDF = term frequency–inverse document frequency; AI = artificial intelligence; PDF = Portable Document Format.

Deduplication performance

The proposed workflow removed 2,788 duplicate records and was compared with Rayyan and ASySD using the same corpus and the same human duplicate-removal reference standard. The

human reference set contained 2,782 duplicate records and 2,011 duplicate title-year clusters. Mapping coverage was complete for all three automated outputs. The proposed workflow was mapped directly by record key, whereas Rayyan and ASySD required metadata-based mapping using DOI, PMID, and title-year fields.

At record level, the proposed workflow yielded 2,226 true positives, 562 false positives, 12,815 true negatives, and 556 false negatives. Precision was 0.798 (95% CI, 0.783–0.813), sensitivity/recall 0.800 (95% CI, 0.785–0.815), specificity 0.958 (95% CI, 0.954–0.961), F1 score 0.799 (95% CI, 0.788–0.810), accuracy 0.931 (95% CI, 0.927–0.935), Jaccard index 0.666 (95% CI, 0.649–0.681), and Cohen's kappa 0.757 (95% CI, 0.744–0.770). Rayyan had lower record-level sensitivity and agreement, with precision 0.654, sensitivity 0.651, F1 score 0.652, Jaccard index 0.484, and kappa 0.580. ASySD had high specificity but lower sensitivity, with precision 0.637, sensitivity 0.250, F1 score 0.359, Jaccard index 0.219, and kappa 0.290.

At title-year cluster level, performance improved for all methods. The proposed workflow yielded 1,876 true-positive clusters, 117 false-positive clusters, 11,836 true-negative clusters, and 135 false-negative clusters. Precision was 0.941 (95% CI, 0.930–0.951), sensitivity/recall 0.933 (95% CI, 0.921–0.943), specificity 0.990 (95% CI, 0.988–0.992), F1 score 0.937 (95% CI, 0.929–0.944), accuracy 0.982 (95% CI, 0.980–0.984), Jaccard index 0.882 (95% CI, 0.867–0.895), and Cohen's kappa 0.927 (95% CI, 0.917–0.935). By comparison, Rayyan achieved cluster-level precision 0.871, sensitivity 0.899, F1 score 0.885, Jaccard index 0.793, and kappa 0.865. ASySD achieved precision 0.800, sensitivity 0.411, F1 score 0.543, Jaccard index 0.373, and kappa 0.493.

Pairwise point estimates favored the proposed workflow over both comparators. At record level, the proposed workflow exceeded Rayyan by 0.150 in sensitivity, 0.147 in F1 score, 0.182 in Jaccard index, and 0.177 in kappa. It exceeded ASySD by 0.550 in sensitivity, 0.440 in F1 score,

0.447 in Jaccard index, and 0.467 in kappa. At title-year cluster level, the proposed workflow exceeded Rayyan by 0.034 in sensitivity, 0.052 in F1 score, 0.088 in Jaccard index, and 0.062 in kappa, and exceeded ASySD by 0.522 in sensitivity, 0.394 in F1 score, 0.509 in Jaccard index, and 0.433 in kappa. These findings support the cluster-level analysis as the more informative operational assessment, because duplicate citations usually occur as groups of related records rather than isolated record pairs. The results are summarized in Table 3.

Table 3. Stage-specific performance metrics for automated deduplication and screening.

Stage/comparison	Level	Sensitivity/recall	Specificity	Precision/PPV	NPV	F1 score	Accuracy	Cohen's kappa	Other result
Proposed deduplication workflow	Record	0.800 (0.785–0.815)	0.958 (0.954–0.961)	0.798 (0.783–0.813)	0.958 (0.955–0.962)	0.799 (0.787–0.810)	0.931 (0.927–0.935)	0.757 (0.743–0.770)	Jaccard 0.666 (0.649–0.681)
Proposed deduplication workflow	Title-year cluster	0.933 (0.921–0.943)	0.990 (0.988–0.992)	0.941 (0.930–0.951)	0.989 (0.987–0.990)	0.937 (0.929–0.944)	0.982 (0.980–0.984)	0.927 (0.917–0.935)	Jaccard 0.882 (0.867–0.895)
Rayyan deduplication	Record	0.651 (0.633–0.668)	0.928 (0.924–0.933)	0.654 (0.636–0.671)	0.927 (0.923–0.932)	0.652 (0.637–0.667)	0.881 (0.875–0.885)	0.580 (0.563–0.597)	Jaccard 0.484 (0.468–0.500)
Rayyan deduplication	Title-year cluster	0.899 (0.885–0.911)	0.978 (0.975–0.980)	0.871 (0.856–0.885)	0.983 (0.980–0.985)	0.885 (0.874–0.895)	0.966 (0.963–0.969)	0.865 (0.853–0.877)	Jaccard 0.793 (0.776–0.809)

Stage/comparison	Level	Sensitivity/recall	Specificity	Precision/PPV	NPV	F1 score	Accuracy	Cohen's kappa	Other result
ASySD deduplication	Record	0.250 (0.234–0.266)	0.970 (0.967–0.973)	0.637 (0.608–0.665)	0.861 (0.856–0.867)	0.359 (0.340–0.378)	0.846 (0.841–0.852)	0.290 (0.270–0.310)	Jaccard 0.219 (0.205–0.233)
ASySD deduplication	Title-year cluster	0.411 (0.389–0.432)	0.983 (0.980–0.985)	0.800 (0.775–0.824)	0.908 (0.903–0.913)	0.543 (0.521–0.564)	0.900 (0.895–0.905)	0.493 (0.471–0.516)	Jaccard 0.373 (0.353–0.393)
Title/abstract screening vs human full-text retrieval list	Record	0.703 (0.616–0.778)	0.997 (0.996–0.998)	0.675 (0.588–0.751)	0.997 (0.996–0.998)	0.689 (0.618–0.753)	0.994 (0.993–0.996)	0.686 (0.615–0.749)	Jaccard 0.525 (0.448–0.602); workload reduction 99.1%; final included studies retained 12/12
Exploratory full-text AI vs final human inclusion safety analysis	90 processed PDFs	1.000 (0.758–1.000)	0.795 (0.692–0.870)	0.429 (0.265–0.609)	1.000 (0.942–1.000)	0.600 (0.387–0.762)	0.822 (0.731–0.888)	0.508 (0.308–0.684)	Jaccard 0.429 (0.265–0.609); deprioritized 62/90 PDFs (68.9%); FN=0; final

Stage/comparison	Level	Sensitivity/recall	Specificity	Precision/PPV	NPV	F1 score	Accuracy	Cohen's kappa	Other result
									included studies preserved 12/12

Note. PPV = positive predictive value; NPV = negative predictive value; FN = false negative. Values are point estimates with 95% confidence intervals. Wilson score intervals were used for binomial proportion metrics. Bias-corrected and accelerated bootstrap intervals with 10,000 resamples were used for F1 score and Cohen's kappa. The full-text AI denominator was restricted to the 90 PDFs actually processed by the AI module.

Title/abstract screening performance

The title and abstract module screened 13,371 post-deduplication records. It classified 123 records as potentially eligible and excluded 13,248, corresponding to 0.92% retained and 99.08% excluded.

Compared with the 118 records selected by human reviewers for full-text retrieval, 83 records were retained by both the automated workflow and human reviewers. Forty records were retained by the automated workflow but were not selected by human reviewers, and 35 human-selected records were not retained by the automated workflow.

The resulting confusion matrix contained 83 true positives, 40 false positives, 13,213 true negatives, and 35 false negatives. Sensitivity/recall was 0.703 (95% CI, 0.616–0.778), specificity 0.997 (95% CI, 0.996–0.998), precision 0.675 (95% CI, 0.588–0.751), negative predictive value 0.997 (95% CI, 0.996–0.998), F1 score 0.689 (95% CI, 0.618–0.753), accuracy 0.994 (95% CI, 0.993–0.996), Jaccard overlap 0.525 (95% CI, 0.448–0.602), and Cohen's kappa 0.686 (95% CI, 0.615–0.749). The results are summarized in Table 3.

The human full-text retrieval list was broader than the final included set and was therefore a conservative comparator for title and abstract screening. The main limitation of this stage was incomplete reproduction of that broader human retrieval set, rather than loss of finally included studies. All 12 studies finally included by human reviewers were retained by the automated title and abstract module. Post hoc review of the 35 human-selected records not retained by the automated module showed that none belonged to the final included set. Reasons for non-retention included absence of semaglutide or GLP-1 receptor agonist exposure terms in the title or abstract, GLP-1 receptor agonist class-level reporting without semaglutide-specific extractable data, non-original publication type, case-report design, insufficient title or abstract information, or absence of a perioperative residual gastric content or aspiration-related

outcome. The 35 human-selected records not retained by the automated title and abstract module are listed in S1 File, Tables S1 and S2.

Full-text AI module

The exploratory full-text AI module processed 90 PDFs, which was the denominator for all full-text performance metrics. The module prioritized 28 PDFs for human assessment and deprioritized 62, corresponding to a workload reduction of 68.9% (95% CI, 58.7–77.5) among processed full texts.

Using final human inclusion as the reference standard within the processed-PDF set, all 12 final included studies were prioritized by the AI module. The confusion matrix contained 12 true positives, 16 false positives, 62 true negatives, and 0 false negatives.

Sensitivity/recall was 1.000 (95% CI, 0.758–1.000), specificity 0.795 (95% CI, 0.692–0.870), precision 0.429 (95% CI, 0.265–0.609), negative predictive value 1.000 (95% CI, 0.942–1.000), F1 score 0.600 (95% CI, 0.387–0.762), accuracy 0.822 (95% CI, 0.731–0.888), Jaccard overlap 0.429 (95% CI, 0.265–0.609), and Cohen's kappa 0.508 (95% CI, 0.308–0.684). The results are summarized in Table 3.

These results should be interpreted as a safety analysis of full-text prioritization, not as a validation of autonomous full-text eligibility assessment. The comparator was the final included set, which is narrower than the full human eligibility-assessment process. Precision remained limited: 16 AI-prioritized PDFs were not part of the final human inclusion set and required human review. Therefore, all AI-prioritized and AI-deprioritized decisions require human confirmation before use in study selection. Article-level full-text AI classifications are provided in S1 File, Tables S3–S6.

Error analysis

Errors were mainly attributable to over-retention, not to loss of finally included studies. No study finally included by human reviewers was excluded by the automated title and abstract module or by the full-text AI module.

At the title and abstract stage, 123 records were retained by the automated module. Of these, 12 were finally included, 40 were not selected by human reviewers for full-text retrieval, and 71 were selected for further human assessment but later excluded during the human review process. Thus, 111 retained records were false positives relative to the final included set, although they did not represent missed eligible studies. Title and abstract false positives relative to final inclusion are summarized in S1 File, Table S2.

In the full-text module, 16 articles were initially prioritized by the AI module but were not part of the original final human inclusion set. Manual review showed that most discrepancies were due to GLP-1 receptor agonist class-level reporting without semaglutide-specific extractable data, absence of semaglutide-specific quantitative results, or outcomes outside the eligible perioperative domains. One article initially counted as a false positive was judged eligible after post hoc human checking. This audit finding did not modify the original human reference standard or the primary performance estimates. The audit of AI-prioritized articles not included in the original final human set is provided in S1 File, Table S7.

Workload and cost estimates

Using the assumptions specified in S1 File, Table S8, the estimated manual workload for the original review process was 277.7 hours. In the assisted-workflow scenario, the estimated residual human workload was 11.4 hours, in addition to 8–17 minutes of computational processing. The estimated human work saved was 266.3 hours.

These estimates are scenario-based operational approximations. They were not derived from prospective time-motion measurement and should not be interpreted as externally generalizable productivity effects. The estimates exclude software development, debugging,

reviewer training, institutional subscriptions, article retrieval, and manual verification of false-positive automated inclusions.

The direct API cost for processing 90 full-text PDFs was estimated at US$3.78–4.32. This figure reflects API usage only and excludes broader infrastructure and implementation costs. The assumptions used for workload and cost estimation are detailed in S1 File, Table S8.

Discussion

Principal findings

In this single-topic internal validation, a staged and auditable workflow reduced the number of records requiring manual review while retaining all studies finally included by human reviewers. Performance was strongest for deduplication, particularly at cluster level, where agreement with the human reference standard was high. Title and abstract screening was operationally efficient, retaining 123 of 13,371 unique records, but its overlap with the 118 records selected by human reviewers for full-text retrieval was only moderate. Thus, the automated module did not miss any finally included study, but it did not reproduce the broader and more cautious human retrieval set (Tables 3 and 4; Fig 3).

Table 4. Underlying counts for stage-specific validation analyses.

Stage/comparison	TP or overlap	FP or AI-only	TN	FN or human-only	Comment
Deduplication, record level	2,226	562	12,815	556	Record-level agreement was lower than cluster-level agreement.
Deduplication, cluster level	1,876	117	11,836	135	Cluster-level analysis better reflected duplicate grouping.
Title/abstract vs human retrieval	83	40	13,213	35	One human-selected record was not found in AI_screening_ALL and was

Stage/comparison	TP or overlap	FP or AI-only	TN	FN or human-only	Comment
					conservatively counted as a false negative. All 12 final included studies were retained.
Full-text AI vs final inclusion	12	16	62	0	The AI module processed 90 PDFs, prioritized 28, and deprioritized 62. All 12 final included studies were prioritized by the AI module.

Note. TP = true positive; FP = false positive; TN = true negative; FN = false negative; AI = artificial intelligence.

The exploratory full-text AI module showed a similar trade-off. Among the 90 PDFs processed, it prioritized all 12 studies included in the original human reference set and had no false negatives in the prespecified safety analysis. A post hoc audit identified one AI-prioritized article initially classified as a false positive that was later judged potentially eligible after human checking[56]. This did not alter the original human reference standard or primary estimates, but suggests that an auditable prioritization layer may help detect occasional discrepancies in study selection. Precision remained low, and agreement beyond chance was only moderate. These findings support use for prioritization, audit, and workload management under human supervision, but not autonomous exclusion or final eligibility decisions (Tables 3 and 4; S1 File, Table S7).

For deduplication, the proposed workflow was directly benchmarked against Rayyan and ASySD outputs applied to the same corpus and evaluated against the same human duplicate-removal reference standard. These comparisons supported higher cluster-level recall, F1 score, Jaccard overlap, and Cohen's kappa for the proposed workflow. However, this was

not a full comparative effectiveness study of review automation platforms, because ASReview, SWIFT-Active Screener, and other screening-oriented systems were not tested in the title and abstract stage[11-14, 18, 30, 31]. The findings therefore support internal feasibility and risk profiling, not superiority across systems or generalizability across review topics.

Methodological implications for systematic review automation

Automation in evidence synthesis should be evaluated according to the risk attached to each decision, rather than by overall accuracy alone. A system may be useful for ranking, flagging, or prioritizing records, but still be unsuitable for autonomous exclusion. In title and abstract screening, and especially in full-text assessment, the key safety issue is whether potentially eligible studies remain available for human review. False positives mainly increase workload and can be corrected later, whereas false negatives may remove evidence before eligibility assessment. Sensitivity, false-negative assessment, preservation of the final included set, and auditability are therefore more informative than accuracy alone[16, 17, 51]. In practice, this workflow should remain reviewer-led: its outputs may support prioritization, audit, and discrepancy detection, but should not determine exclusion without human assessment. Autonomous exclusion would require broader evidence, including external validation, prespecified safety thresholds, comparisons with established tools, and procedures for sampling or adjudicating excluded records, particularly when eligibility depends on full-text nuance, extractable outcomes, or intervention-specific reporting[45, 46].

Deduplication as a cluster-level problem

The deduplication results were broadly consistent with evaluations of dedicated deduplication systems, although comparisons are limited by differences in datasets, duplicate definitions, metadata completeness, and unit of analysis. In this validation, cluster-level performance was high, with sensitivity/recall 0.933, specificity 0.990, precision 0.941, F1 score 0.937, accuracy 0.982, and Cohen's κ 0.927. These estimates fall within the range reported for

high-performing tools, although below the most favorable results from dedicated systems such as Deduklick, ASySD, and the Systematic Review Accelerator Deduplicator[11-14]. This supports the deduplication module as a strong component of the workflow, while avoiding claims of superiority across settings.

Screening performance and false-negative risk

The title and abstract screening results were less robust than the deduplication findings and require a more cautious interpretation. The module reduced 13,371 unique records to 123 retained records and preserved all 12 studies in the original final included set, yielding no false negatives in the prespecified safety analysis. However, agreement with the broader human full-text retrieval set was only moderate: Among 118 records selected by human reviewers, 83 overlapped with the automated set, 35 were not retained by the automated module, and 40 additional records were retained, corresponding to sensitivity/recall 0.703, precision 0.675, F1 score 0.689, and Jaccard overlap 0.525. This distinction is important. Missing a record selected for full-text retrieval is not equivalent to missing a study that ultimately contributed to the review, but it remains methodologically relevant. The missed retrieval records indicate that title and abstract rules may fail when abstracts are sparse, intervention terms are absent, outcomes are implicit, or eligibility depends on full-text information. These findings support use for prioritization and workload reduction under human-in-the-loop conditions, rather than autonomous exclusion[16-19, 21].

Exploratory full-text AI module and error structure

The main limitation of the full-text AI module was not recognition of topical relevance, but discrimination between relevance and protocol-level eligibility. This distinction is consistent with previous LLM screening studies. Tran et al. reported high sensitivity after prompt optimization, but limited specificity, supporting use for prioritization rather than exclusion[25]. Guo et al. found high overall accuracy and specificity, but lower sensitivity for

included studies, illustrating why accuracy alone is unsafe in screening tasks[23]. Matsui et al. improved sensitivity with a layered GPT-based strategy, although performance remained dependent on prompt structure and screening rules[24]. Similarly, Khraisha et al. and Oami et al. suggest that LLMs can support screening and extraction, but not replace reviewers, because performance remains context-, prompt-, and model-dependent[20, 41, 42].

In the present study, most discrepancies involved articles close to the review question but not eligible under the protocol. The commonest error was class-level GLP-1 receptor agonist reporting without extractable semaglutide-specific perioperative data. Thus, AI-prioritized records often appeared clinically relevant, but failed eligibility because of intervention specificity, outcome definition, study design, or lack of extractable data.

The post hoc audit also identified one AI-prioritized article that was later judged potentially eligible after human review. This did not alter the original reference standard or primary estimates, but suggests that structured AI outputs may occasionally help identify reviewer discrepancies. This potential benefit should be interpreted cautiously, as the module also produced multiple false positives and relied on a commercial generative-AI API. Prompt dependence, model updates, output instability, and hallucination remain important limitations for full-text eligibility decisions[45-49].

Auditability, reproducibility, and practical use

The practical value of this workflow lies in traceability rather than automation alone. Each stage produced explicit outputs—duplicate clusters, screening labels, full-text classifications, validation metrics, and error audits—that could be compared with reviewer decisions and inspected after the fact. This is important because reproducible study selection requires more than reporting exclusion counts; reviewers should be able to reconstruct why consequential decisions were made and where errors occurred[5, 7, 9]. For AI-assisted full-text prioritization, documenting fixed prompts, model settings, input limits, execution dates, and

article-level outputs improves transparency, but does not replace human judgment. Consistent with current guidance, these systems should support documentation, prioritization, and quality control, not reviewer-led eligibility assessment[45, 46, 55].

Workload and cost implications

The workload estimates should be interpreted against the substantial resource burden of systematic reviews. Previous reports have described median timelines of 41 weeks to journal submission, accelerated reviews requiring 61–71 person-hours, and costs that may exceed tens of thousands of dollars[1, 3, 15, 26]. In this study, the assisted-workflow scenario reduced estimated human workload from 277.7 to 11.4 hours, with 8–17 minutes of computational processing and a direct API cost of US$3.78–4.32 for 90 PDFs. These figures suggest a meaningful reduction in operational burden, particularly for screening and full-text prioritization. However, they are not direct productivity estimates, as they were based on prespecified assumptions rather than prospective time-motion measurement and excluded development time, debugging, reviewer training, article retrieval, subscriptions, and verification of false positives. Workload reduction is therefore a practical advantage only if eligible studies are preserved and excluded records remain auditable[51].

Strengths

This study has several strengths. It used protocol-based reviewer decisions from a completed PRISMA-guided review as the reference standard, evaluated the workflow as a staged process rather than as an isolated tool, and reported performance separately for deduplication, title and abstract screening, and full-text prioritization. The analysis also included explicit thresholds, audit labels, reproducibility materials, and multiple performance metrics rather than relying on accuracy alone. Most importantly, all studies in the original final human inclusion set were tracked across the automated workflow, allowing assessment of the

error most relevant to systematic review validity: loss of an eligible study before human assessment.

Limitations

This study has several limitations. It was a single-topic internal validation, limiting generalizability to reviews with different terminology, eligibility criteria, or metadata quality. Rule-based screening may misclassify records with sparse abstracts, absent intervention terms, or implicit outcomes. The full-text AI module was limited by PDF availability, text extraction quality, and reliance on a commercial generative-AI API. The human reference standard was also imperfect[57]; one AI-prioritized article was judged potentially eligible post hoc, without altering the primary estimates. Apart from deduplication benchmarks, the workflow was not directly compared with established screening tools. Finally, workload and cost estimates were scenario-based and excluded development, retrieval, subscription, verification, and opportunity costs, and should not be interpreted as generalizable productivity gains.

Future work

Future work should move from internal validation to external testing across review topics and study designs. Protocol-derived gates should be adapted to each question and piloted against human screening to assess false-negative risk. Open-source code and audit templates may support replication, but do not establish validity. Further studies should compare the workflow with established deduplication and screening tools, test alternative thresholds, and use prespecified safety rules, sampling of excluded records, and adjudication before automated exclusion is considered[11-14, 18, 30, 31].

Conclusion

This study shows the feasibility of an auditable workflow for deduplication, title and abstract screening, and AI-assisted full-text prioritization. It preserved all studies included in

the reference review, but did not fully reproduce the broader human retrieval set and had limited full-text precision. The workflow should therefore be used for supervised prioritization and audit, not autonomous exclusion. External validation across topics, teams, and comparator tools is needed before broader adoption.

Acknowledgments

The authors thank the University of Caxias do Sul for institutional support during the development of this doctoral project.

References

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ open. 2017;7(2):e012545. doi: 10.1136/bmjopen-2016-012545.

2. Cumpston MS, McKenzie JE, Welch VA, Brennan SE. Strengthening systematic reviews in public health: guidance in the Cochrane Handbook for Systematic Reviews of Interventions. Journal of Public Health. 2022;44(4):e588-e92. doi: 10.1093/pubmed/fdac036.

3. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemporary clinical trials communications. 2019;16:100443. doi: 10.1016/j.conctc.2019.100443.

4. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Bmj. 2009;339. doi: 10.1136/bmj.b2535.

5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj. 2021;372. doi: 10.1136/bmj.n71.

6. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. bmj. 2017;358. doi: 10.1136/bmj.j4008.

7. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. bmj. 2021;372. doi: 10.1136/bmj.n160.

8. Rethlefsen ML, Brigham TJ, Price C, Moher D, Bouter LM, Kirkham JJ, et al. Systematic review search strategies are poorly reported and not reproducible: a cross-sectional metaresearch study. Journal of Clinical Epidemiology. 2024;166:111229. doi: 10.1016/j.jclinepi.2023.111229.

9. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic reviews. 2021;10(1):39. doi: https://doi.org/10.1186/s13643-020-01542-z.

10. Sohrabi C, Franchi T, Mathew G, Kerwan A, Nicola M, Griffin M, et al. PRISMA 2020 statement: What's new and the importance of reporting guidelines. Elsevier; 2021. p. 105918.

11. Borissov N, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, et al. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. Systematic Reviews. 2022;11(1):172. doi: 10.1186/s13643-022-02045-9.

12. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: Deduplicator. Systematic reviews. 2024;13(1):206. doi: 10.1186/s13643-024-02619-9.

13. Hair K, Bahor Z, Macleod M, Liao J, Sena ES. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. BMC biology. 2023;21(1):189. doi: 10.1186/s12915-023-01686-z.

14. McKeown S, Mir ZM. Considerations for conducting systematic reviews: A follow-up study to evaluate the performance of various automated methods for reference de-duplication. Research Synthesis Methods. 2024;15(6):896-904. doi: 10.1002/jrsm.1736.

15. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. Journal of clinical epidemiology. 2020;121:81-90. doi: https://doi.org/10.1016/j.jclinepi.2020.01.008.

16. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association. 2006;13(2):206-19. doi: 10.1197/jamia.M1929.

17. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews. 2015;4(1):5. doi: https://doi.org/10.1186/2046-4053-4-5.

18. Van De Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nature machine intelligence. 2021;3(2):125-33. doi: 10.1038/s42256-020-00287-7.

19. Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, et al. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environment International. 2020;138:105623. doi: https://doi.org/10.1016/j.envint.2020.105623.

20. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and

extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods. 2024;15(4):616-26. doi: 10.1002/jrsm.1715.

21. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Systematic reviews. 2020;9(1):73. doi: https://doi.org/10.1186/s13643-020-01324-7.

22. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Systematic reviews. 2024;13(1):158. doi: 10.1186/s13643-024-02575-4.

23. Guo E, Gupta M, Deng J, Park Y-J, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. Journal of Medical Internet Research. 2024;26:e48996. doi: 10.2196/48996.

24. Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. Journal of Medical Internet Research. 2024;26:e52758. doi: 10.2196/52758.

25. Tran V-T, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. Annals of internal medicine. 2024;177(6):791-9. doi: https://doi.org/10.7326/M23-3389.

26. Al-Marridi AZ, Bensaid A, Ulde SM, Khwaileh T. ReviewGenie: a novel automated system for systematic reviews—an exploratory study in speech and language disorders. Systematic Reviews. 2025;14(1):167. doi: 10.1186/s13643-025-02895-z.

27. Escaldelai FMD, Escaldelai L, Bergamaschi DP. Sistema “Apoio à Revisão Sistemática”: solução web para gerenciamento de duplicatas e seleção de artigos elegíveis. Revista Brasileira de Epidemiologia. 2022;25:e220030. doi: 10.1590/1980-549720220030.

28. Ravikanth M, Korra S, Mamidisetti G, Goutham M, Bhaskar T. An efficient learning based approach for automatic record deduplication with benchmark datasets. Scientific Reports. 2024;14(1):16254. doi: 10.1038/s41598-024-63242-1.

29. de Bruin J, Lombaers P, Kaandorp C, Teijema J, van der Kuil T, Yazan B, et al. ASReview LAB v. 2: Open-source text screening with multiple agents and a crowd of experts. Patterns. 2025. doi: 10.1016/j.patter.2025.101318.

30. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic reviews. 2016;5(1):210. doi: https://doi.org/10.1186/s13643-016-0384-4.

31. Valizadeh A, Moassefi M, Nakhostin-Ansari A, Hosseini Asl SH, Saghab Torbati M, Aghajani R, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. BMC Medical Research Methodology. 2022;22(1):160. doi: https://doi.org/10.1186/s12874-022-01631-8.

32. Van der Mierden S, Tsaioun K, Bleich A, Leenaars CH. Software tools for literature screening in systematic reviews in biomedical research. Altex. 2019;36(3):508-17. doi: https://doi.org/10.14573/altex.1902131.

33. Liu JJ, Ein N, Gervasio J, Easterbrook B, Nouri MS, Nazarov A, et al. Usability and agreement of the SWIFT-ActiveScreener systematic review support tool: Preliminary evaluation for use in clinical research. Plos one. 2024;19(11):e0291163. doi: 10.1371/journal.pone.0291163.

34. Mitchell E, Are EB, Colijn C, Earn DJ. Using artificial intelligence tools to automate data extraction for living evidence syntheses. Plos one. 2025;20(4):e0320151. doi: 10.1371/journal.pone.0320151.

35. Cao C, Arora R, Cento P, Budak A, Manta K, Farahani E, et al. Automation of systematic reviews with large language models. medRxiv. 2025:2025.06. 13.25329541. doi: 10.1101/2025.06.13.2532954.

36. Galli C, Gavrilova AV, Calciolari E. Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations. Information. 2025;16(5):378. doi: 10.3390/info16050378. PubMed PMID: doi:10.3390/info16050378.

37. Homiar A, Thomas J, Ostinelli EG, Kennett J, Friedrich C, Cuijpers P, et al. Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ mental health. 2025;28(1). doi: https://doi.org/10.1136/bmjment-2025-301762.

38. Insuk S, Boonpattharatthiti K, Booncharoen C, Chaipitak P, Rashid M, Veettil SK, et al. How well do ChatGPT and Claude perform in study selection for systematic review in obstetrics. Journal of Medical Systems. 2025;49(1):110. doi: https://doi.org/10.1007/s10916-025-02246-4.

39. Kim JK, Rickard M, Dangle P, Batra N, Chua ME, Khondker A, et al. Evaluating large language models for title/abstract screening: a systematic review and meta-analysis & development of new tool. Journal of Medical Artificial Intelligence. 2025;8:34. doi: 10.21037/jmai-24-408.

40. Lee K, Paek H, Ofoegbu N, Rube S, Higashi MK, Dawoud D, et al. A4SLR: an agentic AI-assisted systematic literature review framework to augment evidence synthesis for HEOR and HTA. Value in Health. 2025. doi: 10.1016/j.jval.2025.08.002.

41. Oami T, Okada Y, Nakada T-a. Performance of a large language model in screening citations. JAMA network open. 2024;7(7):e2420496. doi: 10.1001/jamanetworkopen.2024.20496.

42. Oami T, Okada Y, Nakada T-a. Optimal large language models to screen citations for systematic reviews. Research Synthesis Methods. 2025:1-17. doi: 10.1017/rsm.2025.10014.

43. Trad F, Yammine R, Charafeddine J, Chakhtoura M, Rahme M, El-Hajj Fuleihan G, et al. Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. BMC medical research methodology. 2025;25(1):130. doi: https://doi.org/10.1186/s12874-025-02583-5.

44. Clark J, Barton B, Albarqouni L, Byambasuren O, Jowsey T, Keogh J, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods. 2025:1-19. doi: 10.1017/rsm.2025.16.

45. Flemyng E, Noel-Storr A, Macura B, Gartlehner G, Thomas J, Meerpohl JJ, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. SAGE Publications Sage CA: Los Angeles, CA; 2025. p. cl2. 70074.

46. Gartlehner G, Nussbaumer-Streit B, Hamel C, Garritty C, Griebler U, King VJ, et al. Responsible Integration of Artificial Intelligence in Rapid Reviews: A Position Statement From the Cochrane Rapid Reviews Methods Group. Cochrane Evidence Synthesis and Methods. 2025;3(6):e70063. doi: 10.1002/cesm.70063.

47. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. 2025;43(2):1-55. doi: 10.1145/3703155.

48. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interactive journal of medical research. 2025;14(1):e59823. doi: i:10.2196/59823.

49. Tonmoy S, Zaman S, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:240101313. 2024. doi: 10.48550/arXiv.2401.01313.

50. Motzfeldt Jensen M, Brix Danielsen M, Riis J, Assifuah Kristjansen K, Andersen S, Okubo Y, et al. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. PLoS One. 2025;20(1):e0313401. doi: 10.1371/journal.pone.0313401.

51. Kusa W, Lipani A, Knoth P, Hanbury A. An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. Intelligent Systems with Applications. 2023;18:200193. doi: 10.1016/j.iswa.2023.200193.

52. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74. doi: 10.2307/2529310.

53. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82. doi: 10.11613/BM.2012.031.

54. Remus Ballotin V, Carneiro Ferreira PL, Tonin de Almeida V, da Silva Borges C, Tonin de Almeida H, Giovanardi Pandolfo da Silva R, et al. Residual gastric content and aspiration-related outcomes in perioperative patients treated with semaglutide: A systematic review and meta-analysis. World J Gastrointest Surg. 2026. doi: https://doi.org/10.4240/wjgs.121449.

55. Huo B, Collins G, Chartash D, Thirunavukarasu A, Flanagin A, Iorio A, et al. Reporting guideline for chatbot health advice studies: The CHART statement. Artif Intell Med. 2025;168:103222. Epub 2025/08/03. doi: 10.1016/j.artmed.2025.103222. PubMed PMID: 40753040.

56. Chang S, Tang Y, Wang M, Zhu S, Tan X, Fan X, et al. Impact of glucagon-like peptide-1 receptor agonists (GLP-1 RAs) on increased residual gastric content in patients with and without concurrent colonoscopy: a retrospective case–control study. Journal of Clinical Medicine. 2026;15(6):2121.

57. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PloS one. 2020;15(1):e0227742. doi: 10.1371/journal.pone.0227742.

Supplementary material

S1 File. Supplementary appendix. This file contains the title and abstract screening audit, exploratory full-text AI module audit, workload and cost estimation, and completed CHART checklist, including Tables S1–S9.

Fig 1. Automated workflow and validation structure. The workflow normalized bibliographic data, applied exact and fuzzy deduplication with predefined safeguards, screened titles and abstracts using protocol-derived eligibility gates, and used a conservative machine-learning rescue step for borderline records. An exploratory AI module was used for full-text prioritization. Automated decisions and audit outputs were compared with protocol-based reviewer decisions from the completed systematic review.

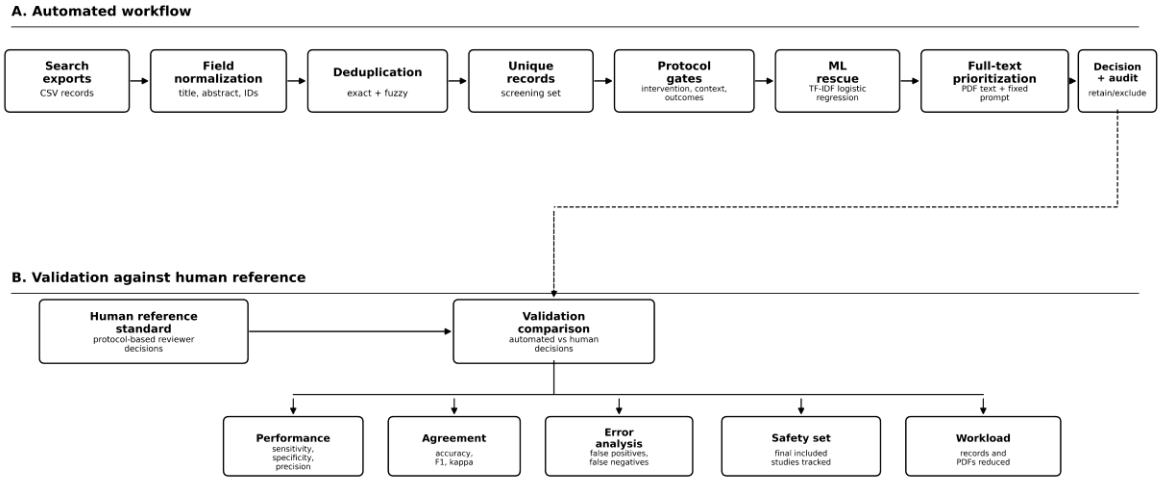


Fig 2. CHART-style methodological diagram for the exploratory generative-AI full-text module. The diagram summarizes model identification, API access, prompt development, query strategy, performance evaluation, and reproducibility details for the AI-assisted full-text prioritization module. The module was evaluated against protocol-based human full-text decisions and was not used to replace human assessment or modify the reference standard.

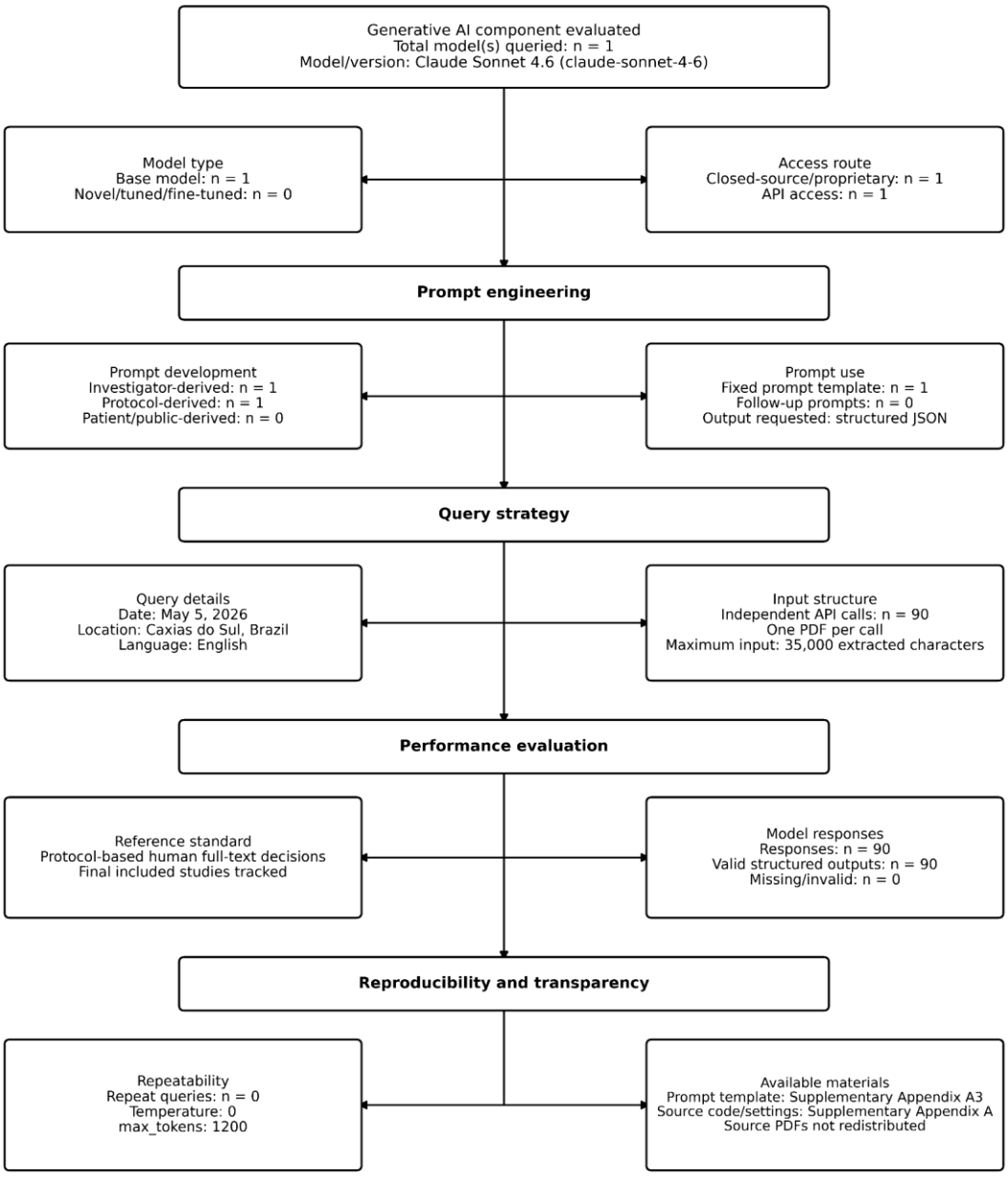
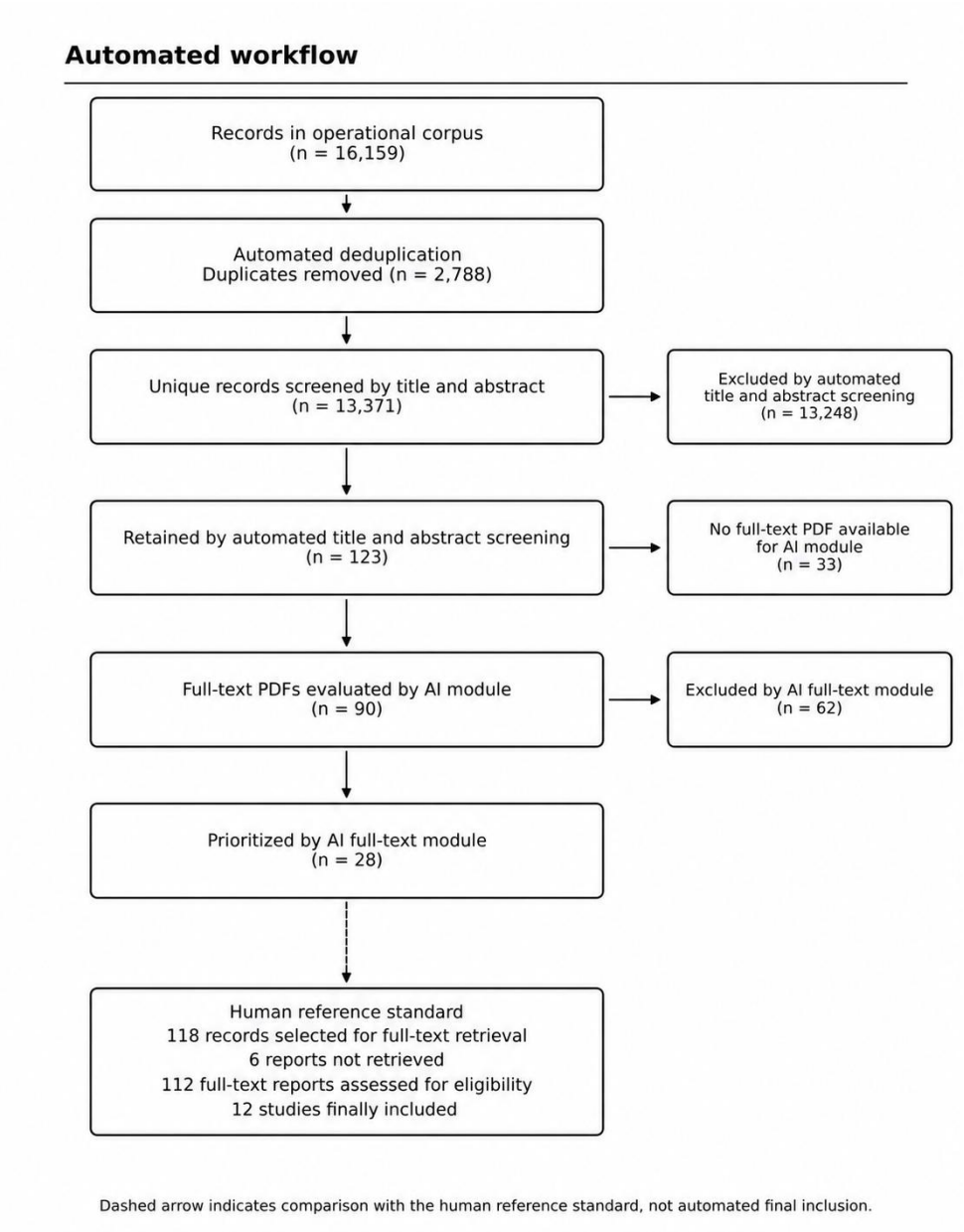


Fig 3. PRISMA-style flow of the automated workflow and human reference standard. The figure shows the operational corpus, automated duplicate removal, title and abstract screening, PDF availability for the full-text AI module, AI-based full-text prioritization, and comparison with the human reference standard. The human reference standard comprised 118 records selected for full-text retrieval, six reports not retrieved, 112 full-text reports assessed for eligibility, and 12 studies finally included. The dashed arrow indicates comparison with the human reference standard, not automated final inclusion.



APÊNDICE F – ARTIGO CIENTÍFICO SOBRE SEMAGLUTIDA E DESFECHOS PERIOPERATÓRIOS

World Journal of Gastrointestinal Surgery

Submit a Manuscript: https://www.f6publishing.com

World J Gastrointest Surg ; 0(0): 121449

DOI: 10.4240/wjgs.121449

ISSN 1948-9366 (online)

META-ANALYSIS

Residual gastric content and aspiration-related outcomes in perioperative patients treated with semaglutide: A systematic review and meta-analysis

Vinicius Remus Ballotin, Pedro Lucas Carneiro Ferreira, Valentina Tonin de Almeida, Carolina da Silva Borges, Henrique Tonin de Almeida, Rodrigo Giovanardi Pandolfo da Silva, Daniel Volquind, Luciano da Silva Selistre

Peer review: Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's classification

Scientific quality: Grade B, Grade C, Grade C

Novelty: Grade C, Grade C, Grade C

Creativity or innovation: Grade C, Grade C, Grade C

Scientific significance: Grade B, Grade C, Grade C

P-Reviewer: Gökdere OG, Assistant Professor, MD, Türkiye; Pappachan JM, Editor, FRCP, MD, MRCP, Professor, Senior Researcher, United Kingdom

Received: March 25, 2026

Revised: April 23, 2026

Accepted: June 3, 2026

Processing time: 142 Days and 18.8 Hours

Vinicius Remus Ballotin, Department of Health Sciences, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

Pedro Lucas Carneiro Ferreira, Valentina Tonin de Almeida, Carolina da Silva Borges, Henrique Tonin de Almeida, Rodrigo Giovanardi Pandolfo da Silva, School of Medicine, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

Daniel Volquind, Department of Anesthesia and Pain Medicine, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

Luciano da Silva Selistre, Department of Nephrology and Biostatistics, University of Caxias do Sul, Caxias do Sul, Brazil, University of Caxias do Sul, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil

ORCID number: Vinicius Remus Ballotin 0000-0002-2659-2249; Pedro Lucas Carneiro Ferreira 0000-0002-3872-830X; Valentina Tonin de Almeida 0000-0002-7748-6374; Carolina da Silva Borges 0009-0006-6605-7677; Henrique Tonin de Almeida 0009-0008-0242-740X; Rodrigo Giovanardi Pandolfo da Silva 0009-0000-7434-7734; Daniel Volquind 0000-0002-8298-823X; Luciano da Silva Selistre 0000-0002-0152-0636.

Co-first authors: Vinicius Remus Ballotin and Pedro Lucas Carneiro Ferreira.

Corresponding author: Vinicius Remus Ballotin, MD, Principal Investigator, Researcher, Department of Health Sciences, University of Caxias do Sul, Campus-Sede Rua Francisco Getúlio Vargas, Caxias do Sul 95070-560, Rio Grande do Sul, Brazil. vrballotin@ucs.br

Abstract

BACKGROUND

Semaglutide, a long-acting glucagon-like peptide-1 receptor agonist, is increasingly prescribed for type 2 diabetes and obesity. Its perioperative safety remains uncertain due to concerns regarding delayed gastric emptying, increased residual gastric content (RGC), and possible aspiration-related complications.

AIM

To determine whether perioperative semaglutide use is associated with RGC, pulmonary aspiration, digestive symptoms, pro-cedure discontinuation, and

gastroparesis *versus* controls.

METHODS
We conducted a systematic review and meta-analysis in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses and Cochrane guidelines. Seven electronic databases (Scopus, Web of Science, EMBASE, MEDLINE/PubMed, Latin American and Caribbean Health Sciences Literature, Cochrane Library, and OpenGrey) were searched from inception through April 2026. The methodological quality of observational studies was assessed using the Newcastle-Ottawa Scale, and the certainty of evidence for each outcome was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation approach.

RESULTS
Twelve observational studies (49651 patients) were included; no randomized trials were identified. Methodological quality was generally high (Newcastle-Ottawa Scale 8-9), although residual confounding and selection bias remain inherent limitations of observational evidence. Semaglutide use was associated with increased risks of RGC [relative risk (RR) = 3.49, 95% confidence interval (CI): 1.78-6.83; moderate certainty)] and digestive symptoms (RR = 5.65, 95%CI: 3.61-8.84; moderate certainty). For pulmonary aspiration (RR = 1.46, 95%CI: 0.29-7.24; very low certainty), procedure discontinuation (RR = 1.74, 95%CI: 0.59-5.11; low certainty), and gastroparesis (RR = 1.43, 95%CI: 0.71-2.87; low certainty), the pooled estimates were imprecise and not statistically significant, and clinically important differences could not be excluded. Heterogeneity regarding RGC was substantial (I^2 = 87.8%), which limits the interpretability of the pooled estimate.

CONCLUSION
Current observational evidence suggests that perioperative semaglutide use increases RGC and digestive symptoms, but clinical significance remains uncertain; aspiration-related outcomes require cautious interpretation and prospective standardized studies.

Key Words: Semaglutide; Glucagon-like peptide-1 receptor agonists; Residual gastric content; Pulmonary aspiration; Perioperative management; Gastric ultrasound

Core Tip: Perioperative semaglutide treatment was associated with an increased risk of residual gastric content and digestive symptoms. Evidence regarding pulmonary aspiration, gastroparesis, and procedure discontinuation remains uncertain because the available data are observational and imprecise. Prospective studies using standardized fasting, perioperative, and outcome definitions are needed.

Citation: Remus Ballotin V, Carneiro Ferreira PL, Tonin de Almeida V, da Silva Borges C, Tonin de Almeida H, Giovanardi Pandolfo da Silva R, Volquind D, da Silva Selistre L. Residual gastric content and aspiration-related outcomes in perioperative patients treated with semaglutide: A systematic review and meta-analysis. *World J Gastrointest Surg* 2026; 0(0): 121449
URL: https://www.wjgnet.com/1948-9366/full/v0/i0/121449.htm
DOI: https://dx.doi.org/10.4240/wjgs.121449

INTRODUCTION

Glucagon-like peptide (GLP)-1 receptor agonists (GLP-1RAs) have become central in treatment for obesity and type 2 diabetes[1,2]. According to the *World Obesity Atlas 2025*, more than 1 billion people worldwide are living with obesity[3]. The 11th edition of the International Diabetes Federation's Diabetes Atlas (2025) estimates that 589 million adults aged 20-79 years were living with diabetes in 2024, and more than 90% of these cases were type 2 diabetes[4].

Semaglutide, a long-acting GLP-1RA, is now widely used in both metabolic and obesity care. In addition to enhancing glucose-dependent insulin secretion and suppressing glucagon release, GLP-1 receptor agonism delays gastric emptying [5,6]. Experimental and clinical data indicates that this effect is mediated by relaxation of the gastric fundus, increased gastric compliance, inhibition of antral contractility, and increased pyloric tone[7]. Although this effect may be attenuated over time due to tachyphylaxis, delayed gastric emptying may persist in some patients and has raised concern about residual gastric content (RGC), regurgitation, and pulmonary aspiration during anesthesia or procedural sedation[8-10].

These potential complications could increase perioperative morbidity and prolong recovery. Current recommendations for the perioperative management of GLP-1RAs are based on limited evidence, primarily from case reports and small observational studies[6,11,12]. Therefore, this systematic review and meta-analysis was conducted to determine whether preoperative risk among adult patients undergoing surgery or procedural sedation is higher with or without semaglutide use. The primary outcome was RGC. Secondary outcomes included digestive symptoms, pulmonary aspiration, procedure discontinuation, and gastroparesis.

MATERIALS AND METHODS

Protocol and registration

This study was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-Analysis guidelines and the Cochrane handbook[13,14]. The systematic review protocol was registered with the International Prospective Register of Systematic Reviews, maintained by York University, approval No. CRD42024609601.

Search strategy and study selection

A search from inception to April 2026 was conducted in seven electronic databases: Scopus, Web of Science, EMBASE, MEDLINE (PubMed), Latin American and Caribbean Health Sciences Literature, Cochrane Library for Systematic Reviews, and Opengray.eu. Studies retrieved from these databases were systematically and independently reviewed by two researchers (Vinicius Remus Ballotin and Pedro Lucas Carneiro Ferreira), and the reference lists of the retrieved studies were submitted for manual search. Divergences in study selection were resolved by a third researcher (Carolina da Silva Borges). The full search strategy is described in the Supplementary material.

Eligibility criteria

The inclusion criteria were as follows: Adults aged ≥ 18 years; patients who received preoperative semaglutide treatment, regardless of therapy duration, before surgery; and randomized controlled trials and observational studies (including cohort, case-control, and cross-sectional designs) reporting at least one of the following perioperative outcomes: (1) RGC (primary outcome); (2) Digestive symptoms; (3) Pulmonary aspiration; (4) Procedure discontinuation; and (5) Gastroparesis.

The exclusion criteria were as follows: Significant comorbidities that independently increase the risk of perioperative complications (*e.g.*, severe cardiovascular or respiratory diseases); case reports, case series, reviews, and expert opinions that do not present original data or comparisons between groups; non-English language studies without readily available translated versions; and studies with significant missing data that impede appropriate analysis and interpretation of outcomes.

Methodological quality assessment

Study quality was assessed using the Newcastle-Ottawa Scale (NOS)[15,16], depending on type. Two researchers (Remus Ballotin V and Carneiro Ferreira PL) conducted independent assessments, and their results were compared to evaluate interrater reliability. Divergences were discussed with a third researcher (Tonin de Almeid V) until 100% consensus was reached. The NOS is an assessment tool that examines the quality of observational studies across selection, comparability, and outcomes, allowing for the classification of risk of bias and internal validity.

Certainty of evidence assessment

The certainty of evidence for each outcome was evaluated using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach[17,18]. For each outcome, the overall certainty of evidence was rated as high, moderate, low, or very low, according to GRADE guidelines.

Data extraction

Two researchers (Tonin de Almeida H, Giovanardi Pandolfo da Silva R) independently extracted qualitative and quantitative data using a predefined form. Any discrepancies were reviewed by a third researcher (Remus Ballotin V). The extracted qualitative data included the: (1) Study aim; (2) Databases searched; (3) Population, interventions, and outcomes; and (4) Authors' conclusions. The quantitative data extracted comprised: (1) The number of patients; (2) Population characteristics (*e.g.*, number of participants, percentage of females, and mean age); and (3) Effect-size results for perioperative outcomes comparing the semaglutide group with the control group. When an eligible study enrolled a broader GLP-1RA population, only data from the semaglutide-specific subgroup were extracted for this review. If semaglutide-specific population characteristics or outcome data were not separately reported, the corresponding authors were contacted to request disaggregated data. Studies from which semaglutide-specific data could not be isolated were excluded from the quantitative synthesis. This approach was taken to preserve exposure specificity and improve comparability across studies by ensuring that pooled estimates reflected semaglutide-specific rather than mixed GLP-1RA effects. When overlapping populations were suspected, only the most complete or most recent report was retained.

Statistical analysis

Data were synthesized through a meta-analysis of studies reporting perioperative outcomes associated with preoperative semaglutide use. All outcomes were extracted as the number of events and total participants in each group, allowing effect estimates to be standardized as risk ratios (RR). For pulmonary aspiration, the sample size required to detect a small absolute difference in risk was additionally estimated using a standard power calculation in G*Power 3.1.9.7. When at least two studies reported comparable outcomes, pooled estimates were calculated using both fixed-effect and random-effects models. Given the expected clinical and methodological heterogeneity across studies, the random-effects model was considered the primary analysis, whereas fixed-effect estimates were used as complementary analyses. Fixed-effects analyses employed the Mantel-Haenszel method. Random-effects meta-analyses used an inverse-variance approach with the Paule–Mandel estimator for between-study variance (τ^2) and the Q-profile method for τ^2 confidence intervals (CI). To provide more conservative estimates under heterogeneity, 95%CIs for pooled effects were calculated with the Knapp-

Hartung adjustment. Statistical heterogeneity was assessed using the I^2 statistic, and subgroup analyses were performed according to study design. Publication bias was evaluated visually using funnel plots and statistically with Egger's test. The analyses were conducted in R Studio 2024.12.1, with forest plots generated for visual representation. In sparse-event analyses, zero cell frequencies were calculated by the software using a continuity correction of 0.5, being applied only to calculate individual study estimates. In the sensitivity analyses, studies at higher risk of bias were excluded, and leave-one-out influence analyses were used.

Data and resource availability

The data underlying this study are available in the Figshare repository, accessible at https://doi.org/10.6084/m9.figshare.30450836[19].

RESULTS

Study selection

The database search yielded 16159 records. After removing duplicates, 13377 titles and abstracts were screened, of which 13259 were excluded. A total of 112 full-text articles were assessed for eligibility, of which 100 were excluded, leaving 12 studies in the final analysis. The reasons for exclusion are summarized in the Preferred Reporting Items for Systematic Reviews and Meta-Analysis flowchart (Figure 1) and detailed in the Supplementary material.

Quality assessment

NOS scores ranged from 8 to 9 (maximum of 9), with an initial interrater agreement of 92.3%. Most achieved maximum scores in the selection and outcome domains, reflecting representative cohorts, reliable exposure measurement, and valid outcome assessment. Minor reductions in total score were generally caused by inadequate follow-up or comparability. Given their observational design, these studies remain at inherent risk of residual confounding and selection bias, despite generally high NOS scores.

Characteristics of the included studies

Overall, 12 studies were included, comprising ten cohort studies, of which one was a prospective study[11] and two were case-control studies[20,21]. Most studies were published in 2024, except for one published before and two published after 2024[10,21,22]. Most of the studies were conducted in the United States, followed by Brazil, Canada, and Belgium. The largest study included 23940 participants[23], whereas the smallest included 88[22]. The percentage of female participants could not be assessed in three studies[11,20,23], and in one study, data were available only for the control group[11]. The mean age ranged from 45 years to 71 years in the intervention groups and from 39 years to 72 years in the control groups. The reported procedures included upper digestive endoscopy, colonoscopy, and elective surgeries, such as joint pro-sthesis implantation, cardiovascular surgery, and digestive tract surgery. Many studies did not report the semaglutide holding interval[8,21,23-25]. One study classified RGC content as small, medium, or large[26]. Two studies did not report RGC data[12,23]. Apart from RGC, the most commonly reported outcomes were pulmonary aspiration, digestive symptoms, procedure discontinuation, and gastroparesis. Only one study reported on 30-day mortality, postoperative length of stay, and pneumonia[12]. Further details are provided in Tables 1 and 2.

Population

The final analysis included 49651 patients, of whom 11022 (22.2%) were reported as semaglutide users. No overlapping patient populations were identified among the included studies. The mean age in the semaglutide group was 56.7 (SD ± 7.41) years, which was similar to that of the control group at 56 (SD ± 8.95) years. The semaglutide and control groups were 45.5% and 43.9% female, respectively. The mean body mass index was 31.58 (SD ± 4.54) for the semaglutide group and 29.2 (SD ± 3.28) for the control groups. Type 2 diabetes was reported in 4631 (42.3%) patients in the intervention group and 15253 patients (40.8%) in the control group. Most patients in both groups were classified as having American Society of Anesthesiologists physical status I or II; however, several studies did not report American Society of Anesthesiologists classification data. The mean duration of semaglutide use was 149 (SD ± 127) days. The perioperative drug holding interval was reported for most patients, being < 21 days in most cases when data were available.

Meta-analysis

RGC was reported in 239 of 1690 semaglutide users (14.1%) and 322 of 6106 controls (5.2%), corresponding to an absolute risk difference of 8.9 percentage points and a pooled RR of 3.49 (95%CI: 1.78-6.83). There was considerable heterogeneity among the studies (I^2 = 87.8%, P < 0.0001), with τ^2 = 0.72 (95%CI: 0.26-2.83), indicating substantial variability in effect sizes. The forest plot in Figure 2A summarizes the results[2,8,10-12,20-26].

Egger's test for funnel plot asymmetry found no significant evidence of publication bias (t = 1.33, degrees of freedom = 8, P = 0.21). In the regression model used for this test, residual heterogeneity remained substantial (τ^2 = 7.51), which suggests that the true effects differed between studies beyond sampling error. Influence analysis using a leave-one-out approach revealed that the overall effect estimate remained robust, with pooled RRs ranging from 2.48 to 4.56 after sequential exclusion of individual studies; all differences remained statistically significant (P < 0.0001). This indicates that no single study disproportionately influenced the overall findings. Details in Supplementary Figure 1.

Table 1 Baseline characteristics of patients in the included studies (reported data only) (total *n* = 49651), *n* (%)/mean ± SD

Variable	Semaglutide	Control
Patients	11022 (22.2)	38629 (77.8)
Sex (female)	2514 (45.5)	8785 (43.9)
Age (mean years)	56.7 ± 7.41	56 ± 8.95
BMI (mean)	31.58 ± 4.54	29.2 ± 3.28
ASA		
I	69 (0.6)	313 (0.8)
II	214 (1.94)	1094 (2.8)
III	8 (0.07)	184 (0.4)
IV	0 (0)	1 (0.002)
V	0 (0)	0 (0)
Comorbidities		
Type 2 diabetes	4631 (42.3)	15253 (40.8)
GERD	917 (8.3)	3313 (8.8)
Opioid use	1501 (13.7)	5712 (15.2)
Abdominal surgery	3	51
Cirrhosis	15	11
HbA1c (mean)	6.48 ± 0.58	6.42 ± 0.67
	47 mmol/mol	47 mmol/mol
Duration (mean day)	149 ± 127	-
Dose (mean mg)	1.33 ± 0.23	-
Fasting (mean hours)		
Solids	12.89 ± 3.28	12.83 ± 3.27
Clear fluids	6.51 ± 3.14	7.66 ± 3.8

Values are based on available reported data. Not all studies reported every baseline variable; thus, denominators may differ between characteristics and from those used in the pooled effect-size calculations for specific outcomes (*e.g.*, residual gastric content and aspiration). BMI: Body mass index; ASA: American Society of Anesthesiologists Physical Status classification; GERD: Gastroesophageal reflux disease; HbA1c: Glycated hemoglobin.

Digestive symptoms occurred in 27 of 156 semaglutide users (17.3%) and 43 of 1342 controls (3.2%), corresponding to an absolute risk difference of 14.1 percentage points and a pooled RR of 5.65 (95%CI: 3.61-8.84). Conversely, pulmonary aspiration was rare in both groups, occurring in 15 of 6222 semaglutide users (0.24%) and 80 of 21519 controls (0.37%), yielding a highly imprecise pooled RR of 1.46 (95%CI: 0.29-7.24). The forest plot in Figure 2B summarizes these results[2,8,10,23,24]. Details in Supplementary Figure 2.

Procedures were discontinued in 84 of 5979 semaglutide users (1.4%) and in 100 of 18776 controls (0.5%), with an RR of 1.74 (95%CI: 0.59-5.11). Gastroparesis was reported in 20 of 663 semaglutide users (3.0%) and 81 of 2824 controls (2.9%), with an RR of 1.43 (95%CI: 0.71-2.87). These estimates are based on relatively few events and remain highly uncertain, with wide CIs compatible with both no effect and clinically important differences. Secondary outcomes were not often reported. Detailed event counts and study-level estimates are provided in the Supplementary Figures 3 and 4.

Subgroup analysis

Given the high heterogeneity, subgroup analyses were conducted to explore the effect of semaglutide on RGC by study design. The pooled RR across all included studies was significantly high (random-effects model RR = 3.49, 95%CI: 1.78-6.83; P = 0.0008), with considerable heterogeneity (I^2 = 87.8%).

When stratified by study design, retrospective cohort studies (n = 7) showed a significant association (RR = 3.45, 95%CI: 1.28-9.35), although with high heterogeneity (I^2 = 91.5%). The single prospective cohort study reported a similar effect size (RR = 2.65, 95%CI: 1.46-4.79). The two case-control studies showed a larger effect estimate in the common-effect model (RR = 4.06, 95%CI: 2.15-7.66). Statistical tests for subgroup differences indicated a lack of significant variation between study designs (random-effects model P = 0.73), suggesting that the association between semaglutide use and increased RGC is consistent regardless of study type (Supplementary Figure 5).

Table 2 Characteristics of the included studies

Ref.	Country	Study type	Inclusion criteria	Exclusion criteria	Fasting	Procedure indication	Semaglutide interruption	Increased RGC definition	Anesthesia details	Extracted data source	NOS score
Santos *et al* [26], 2024	Brazil	Retrospective cohort	≥ 18 years-old, presenting for elective esophagogastroduodenoscopy	Gastric volvulus/frank outlet obstruction/active esophageal/gastric/duodenal bleeding, ASA physical status ≥ IV, recent (≤ 2 months) abdominal surgery, emergency procedures, EGD combined with surgical procedures, chronic renal/Liver disease, achalasia, Zenker's diverticulum, linitis plastica, multiple myeloma, systemic collagenosis, amyloidosis, pregnancy, chronic opioid use, drug addiction, use of vasoactive agents, intensive care patients, preoperative use of medication that affect gastric emptying (tricyclic antidepressants, opioids, pro-kinetics, histamine H2-receptor antagonists) other than semaglutide, and incomplete medical records. Patients using GLP-1-Ras other than semaglutide and/or oral/daily semaglutide were also excluded	≥ 2 hours for clear fluids, ≥ 8 hours for solids and fluids with residue	General indications for EGD included dysphagia, odynophagia, persistent abdominal pain, intractable and/or chronic gastro-esophageal reflux	Preoperative interruption intervals were categorized based on its ~7-day half-life as ≤ 7 days, 8-14 days, 15-21 days, and > 21 days	Any amount of solid content from the esophagus to the pylorus, or > 0.8 mL/kg of fluid content (given its higher risk of bronchoaspiration) as measured from the aspiration/suction canister	Premedication was not routinely administered. The sedation/anesthetic procedure was at the discretion of the attending anesthesiologist and generally consisted of propofol titration to maintain spontaneous ventilation (under supplemental oxygen *via* nasal cannula), with occasional tracheal intubation. Intraoperative monitoring included sphygmomanometer, electrocardiography, pulse, and capnography	Main group	High
Phan *et al* [24], 2024	United States	Retrospective cohort	Patients on confirmed GLP1-Ras receiving an EGD between 2021 and 2023	NA	NA	NA	NA	NA	NA	Subgroup	High
Nersessian *et al*[2], 2024	Brazil	Retrospective cohort	All patients aged ≥ 18 years presenting for elective surgery between July and December 2023 at the Sao Luiz Hospital/Rede D'Or, Sao Paulo, Brazil	Type 2 diabetes; hiatal hernia; previous gastric surgery (gastrectomy, Roux-en-Y gastrojejunostomy, gastric band, fundoplication); ASA physical status ≥ 3; BMI > 40 kg/m²; recent (within 2 months) abdominal surgery; preoperative semaglutide interruption interval > 10 days; chronic renal and/or liver disease; achalasia; Zenker's diverticulum; linitis plastica; multiple myeloma; systemic lupus erythematosus and other collagenosis; pregnancy; chronic opioid use; drug addiction; preoperative use of medication known to affect gastric emptying other than	≥ 2 hours for clear fluids and ≥ 8 hours for solids and fluids with residue	NA	Patients were grouped according to time since their last dose of semaglutide, either within 1-7 days or 8-10 days	Any amount of solid content or > 1.5 mL/kg of clear fluids	NA	Main group	High

				semaglutide (*e.g.*, tricyclic antidepressants and opioids); and GLP-1 receptor agonists other than semaglutide							
Gu *et al* [20], 2024	United States	Case-control	Adult patients on semaglutide who underwent an EGD between August 2022 and August 2023	Patients were excluded if they were in the intensive care unit at time of EGD, previously on GLP-1 agonist therapy, previously diagnosed with gastroparesis, or underwent EGD for acute gastrointestinal bleeding	NA	NA	NA	NA	NA	Main group	High
Zaffar *et al* [8], 2024	United States	Retrospective cohort	Of 2578 EGDs performed on adults (age 18-89 years) under deep sedation/general anesthesia between August 2022 and August 2023 were included in the study after simple random sampling	NA	NA	NA	NA	RGC found during EGD	Patients under deep sedation/general anesthesia	Subgroup	High
Silveira *et al*[10], 2023	Brazil	Retrospective cohort	All patients ≥ 18 years-old presenting for elective diagnostic upper endoscopy were eligible	Exclusion criteria were: Gastric outlet obstruction, gastric volvulus, frank/active esophageal/gastric/duodenal bleeding, ASA physical status ≥ IV, recent (≤ 2 months) abdominal surgery, emergency endoscopic procedures, urethral erosion combined with other/surgical procedures, chronic renal and/or liver disease, achalasia, Zenker's diverticulum, linitis plastica, multiple myeloma, systemic collagenosis, amyloidosis, pregnancy, chronic opioid use, drug addiction, use of vasoactive agents, patients admitted to the intensive care unit, preoperative use/ingestion of medication known to affect gastric emptying (*e.g.*, tricyclic antidepressants, opioids, pro-kinetics, histamine H2-receptor antagonists) other than semaglutide, and incomplete medical records. Patients using GLP-1 agonists other than semaglutide were also excluded	≥ 2 hours for clear fluids, and ≥ 8 hours for solids and fluids with residue	Elective diagnostic upper endoscopy	The time intervals of semaglutide interruption in patients with and without increased residual gastric content were 10 (6-15) and 11 (7.75-12.5) days, respectively (P = 0.67)	Any amount of solid content from the esophagus to the pylorus, or > 0.8 mL/kg of fluid content as measured from the aspiration/suction canister	The sedation/anesthetic procedure was at the discretion of the anesthesiologist	Main group	High
Korlipara *et al*[25], 2024	United States	Retrospective cohort	Adults undergoing EGD at Weill Cornell from January 2018 to March 2023, with or without semaglutide	NA	NA	NA	NA	Defined as "retained gastric contents" – no specific numeric threshold given	NA	Main group (author-provided)	High

Alkabbani *et al*[23],	United States	Retrospective cohort	Patients using GLP-1 receptor agonists or SGLT-2 inhibitors	Not clearly listed; based on claims data filters and diagnostic codes	NA	NA	NA	Implicitly assessed *via* diagnostic codes	NA		Subgroup	High

2024			undergoing upper endoscopy					for aspiration and procedure discontinuation			
Welk *et al* [12], 2024	Canada	Retrospective cohort	Patients ≥ 66 years with type 2 diabetes who underwent elective surgery under general or spinal anesthesia between February 2020 and March 2023	Emergency surgery, multiple surgeries on the same day, incomplete or inconsistent data	Exclusion criteria included emergency surgery, multiple surgeries on the same day, and incomplete or inconsistent data	Various elective surgeries (*e.g.*, joint replacement, cardiovascular, digestive, *etc.*)	None – patients were on active semaglutide therapy at the time of surgery	Not directly assessed – the outcome was postoperative pneumonia as an indirect marker of aspiration (RGC was neither visualized nor measured)	General anesthesia (58.6%) or spinal anesthesia (41.2%); data on type of induction, airway management, and medications were not included in the involved database	Main group	High
Ukwade *et al*[21], 2025	United States	Retrospective	EGD procedures performed between September 1, 2022, and October 30, 2023	Patients who were pregnant, incarcerated, under the age of 18, patients with a BMI > 50, and patients who were hospitalized. Patients were also excluded if they met the ASA physical status classification of ASA 4 (patients with severe systemic disease threatening life) or ASA 5	Midnight fast before EGD; morning medications allowed with small sips of water. Those who consumed any food are rescheduled as, per our institutional guidelines, since a fast of 8 hours is XXXrequire for any food other than clear liquids and a fast of 2 hours is XXXrequire for any clear liquids prior to an EGD	NA	NA	Defined as mentioning food or fluid in the endoscopy report, which was based on the endoscopist's clinical judgment	NA	Subgroup	High
Vlaeminck *et al*[22], 2026	Belgium	Prospective cohort	Adult patients receiving semaglutide treatment – regardless of dose, administration route, frequency, or therapeutic indication - who were scheduled for elective surgery under general anesthesia	Patients were not included if they declined to participate; had a contraindication to gastric ultrasound (*i.e.*, previous gastric surgery or hiatal hernia); comorbidities known to delay gastric emptying (scleroderma; systemic lupus erythematosus; hypothyroidism; Parkinson's	European Society of Anaesthesia and Intensive Care fasting guidelines (*i.e.*, > 2 hours for liquids and > 6 hours for	NA	The ASA recommended GLP-1 RA withholding period (*i.e.*, 1 week if administered weekly and 1 day if	A patient was considered to have a "full stomach" or a "positive" gastric ultrasound if solid gastric content was visible in any position or if the calculated gastric	NA	Main group	High

	disease; cerebral palsy; and multiple sclerosis; or were unable to assume the right lateral decubitus position	solid foods)	administered daily)	volume in the right lateral decubitus position exceeded

ASA: American Society of Anesthesiologists; BMI: Body mass index; EGD: Esophagogastroduodenoscopy; GLP-1: Glucagon-like peptide 1; NOS: Newcastle-Ottawa Scale; NA: Not available; RGC: Residual gastric content; SGLT-2: Sodium glucose co-transporter 2.

Subgroup analysis by procedure type demonstrated that semaglutide use was associated with an increased risk of RGC across all categories. For endoscopy, the pooled random-effects RR was 2.71 (95%CI: 0.63-11.64; I^2 = 92.9%), which did not reach statistical significance. In elective procedures, semaglutide users had a significantly higher risk (RR = 4.36, 95%CI: 0.4-46.98; I^2 = 74%). Conversely, one study assessing combined endoscopy and colonoscopy found a significantly reduced risk [odds ratio (OR) = 0.41, 95%CI: 0.23-0.73][25]. The overall pooled analysis across procedure types yielded an RR of 3.49 (95%CI: 1.78-6.83; I^2 = 87.8%). Although the test for subgroup differences was significant in the fixed-effect model (P < 0.001), it was not significant in the random-effects model (P = 0.54), indicating that between-study variability might explain the observed differences between procedure types (Supplementary Figure 6).

In a very exploratory meta-regression, the pulmonary aspiration rate did not significantly modify the association between semaglutide use and RGC (P = 0.82). The analysis included only four studies and was underpowered, with unstable coefficient estimates with wide CIs.

Sensitivity analysis

A leave-one-out sensitivity analysis was conducted to evaluate the robustness of the meta-analysis findings on RGC associated with semaglutide use. Sequential exclusion of each study showed that the overall pooled RR remained significant across all analyses, with common-effect model RRs ranging from 2.48 to 4.56 (all P < 0.0001). Likewise, random-effects model estimates remained significant, ranging from 3.03 to 4.25.

Heterogeneity remained substantial throughout the sensitivity analyses, with I^2 values consistently above 63.2%, mostly exceeding 89.1%, indicating persistent between-study variability not explained by the exclusion of any single study. The τ^2 estimates ranged from 0.35 to 0.82, confirming this residual heterogeneity.

These results revealed that no single study disproportionately influenced the overall effect estimate, confirming the strength of the association between semaglutide use and increased RGC.

GRADE assessment

Certainty of evidence, assessed using the GRADE approach, ranged from moderate to very low across outcomes. Semaglutide use was associated with a higher RGC risk (RR = 3.49, 95%CI: 1.78-6.83; moderate certainty) despite substantial heterogeneity. Digestive symptoms were also more frequent among semaglutide users (RR = 5.65, 95%CI: 3.61-8.84; moderate certainty), supported by a large, consistent effect size. Evidence for gastroparesis was limited (RR = 1.43, 95%CI: 0.71-2.87; low certainty), as CIs spanned potential benefit and harm. Pulmonary aspiration remained highly uncertain due to very few events and wide CIs (RR = 1.46, 95%CI: 0.29-7.24; very low certainty). Similarly, procedure discontinuation showed an imprecise association (RR = 1.74, 95%CI: 0.59-5.11; low certainty). Details are provided in Table 3.

Table 3 Summary of findings - semaglutide and perioperative outcomes

Outcome	Participants (total, studies)	Events (SEMA)	Events (control)	Relative effect (RR, 95%CI)	Certainty of evidence (GRADE)
Residual gastric content	7796 (10 studies)	239	322	3.49 (1.78-6.83)	●●○○ moderate[1]
Digestive symptoms	1498 (2 studies)	27	43	5.65 (3.61-8.84)	●●○○ moderate[2]
Gastroparesis	3487 (2 studies)	20	81	1.43 (0.71-2.87)	●○○○ low[3]
Pulmonary aspiration	27741 (6 studies)	15	80	1.46 (0.29-7.24)	○○○○ very low[4]
Procedure discontinuation	24755 (2 studies)	84	100	1.74 (0.59-5.11)	●○○○ low[5]

[1]Residual gastric content: Downgraded one level for serious inconsistency (I^2 = 90%, high heterogeneity).
[2]Digestive symptoms: Observational evidence, upgraded one level for large and consistent effect size; final rating = moderate.
[3]Gastroparesis: Downgraded for serious imprecision [wide confidence interval (CI), few events].
[4]Pulmonary aspiration: Downgraded two levels for inconsistency (variation in effect direction) and very serious imprecision (extremely wide CI, rare outcome).
[5]Procedure discontinuation: Downgraded for inconsistency and imprecision (heterogeneous effects, wide CI).
SEMA: Semaglutida; RR: Relative risk; CI: Confidence interval; GRADE: Grading of Recommendations Assessment.

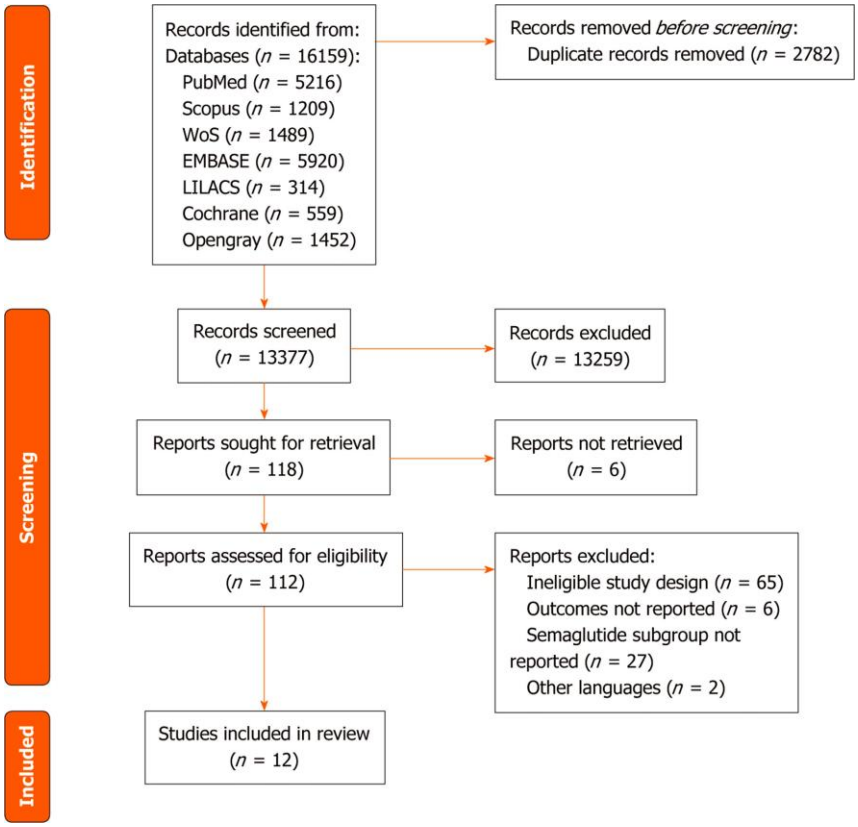


Figure 1 Preferred reporting items for systematic reviews and meta-analysis flowchart.

DISCUSSION

In this systematic review and meta-analysis of observational studies, preoperative semaglutide use was associated with increased risks of RGC and digestive symptoms. In contrast, the associations with pulmonary aspiration, procedure discontinuation, and gastroparesis remained uncertain because these outcomes were infrequently reported, with few events and wide CIs. Overall, the available evidence suggests that semaglutide is associated with intermediate perioperative markers of delayed gastric emptying, whereas evidence for rarer but clinically more consequential outcomes remains limited.

In these 12 observational studies, preoperative semaglutide use was associated with a 3.5-fold higher risk of increased RGC (RR = 3.49, 95%CI: 1.78-6.83) and a 5.6-fold higher risk of digestive symptoms (RR = 5.65, 95%CI: 3.61-8.84), with

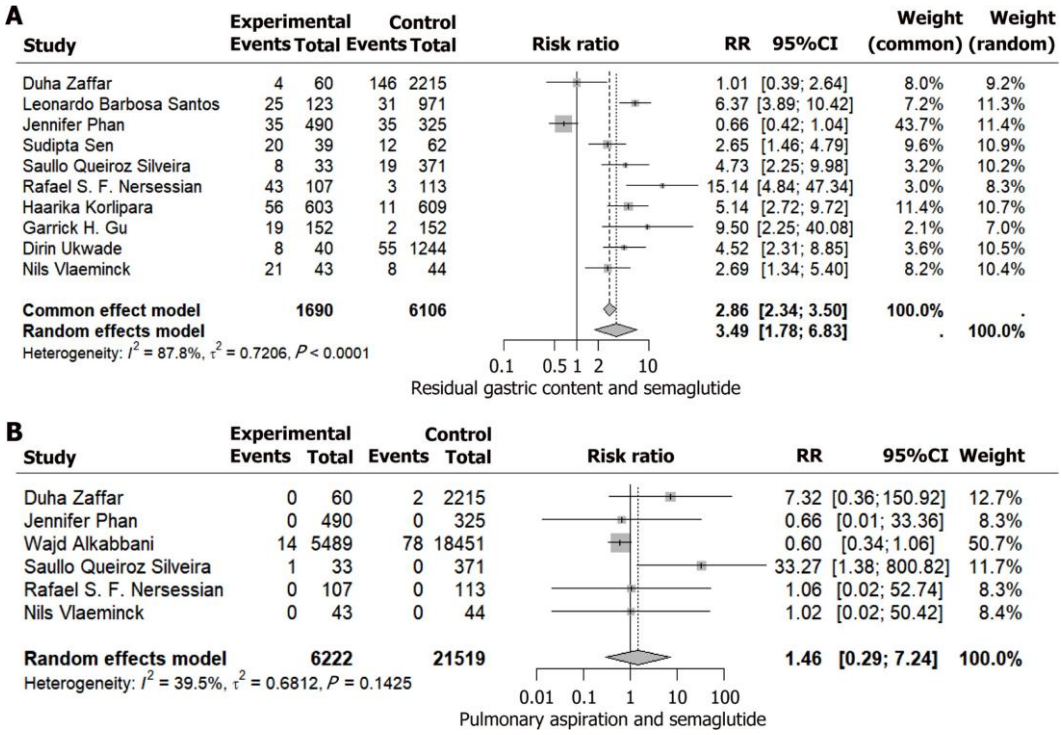


Figure 2 Forest plot for residual gastric content and pulmonary aspiration in relation to semaglutide use. A: Residual gastric content; B: Pulmonary aspiration. RR: Relative risk; CI: Confidence interval.

moderate-certainty evidence. However, substantial heterogeneity limits the interpretability of the pooled effect size estimates. Conversely, the associations with gastroparesis, procedure discontinuation, and pulmonary aspiration were highly uncertain due to few events and wide CIs, rated low to very low certainty.

Most semaglutide users in the included cohorts were overweight or obese middle-aged patients who were undergoing elective endoscopy or surgery, often for weight loss or metabolic indications[12,24,26]. As such, our findings are most applicable to this growing population of relatively stable outpatients.

The absolute risk increase of 8.9 percentage points for RGC suggests that approximately one additional patient in every 11 perioperative semaglutide users will have RGC compared with nonusers. Similarly, the 14.1 percentage point increase in digestive symptoms corresponds to roughly one additional symptomatic patient for every 7 semaglutide users.

Evidence suggests that tachyphylaxis may occur with prolonged therapy, particularly regarding its effect on gastric emptying[2,10,23]. In the included studies, the mean administered dose was 1.33 mg and the mean treatment duration was 149 days. This duration indicates that most patients were transitioning between the titration and maintenance phases of therapy. However, perioperative patients may present at different phases of dose escalation[26,27], which could influence gastrointestinal tolerance and the risk of RGC.

Although semaglutide offers substantial metabolic benefits, its long half-life and effects on gastric motility raise perioperative concerns, as delayed gastric emptying may persist in some patients even after 1 week of suspension[2,10,27].

Consistent with gastric ultrasonography reports, up to 90% of patients using GLP-1RAs have RGC despite ≥ 8 hours of fasting[1]. Definitions of increased RGC varied across studies, with thresholds ranging from 0.8 mL/kg to 1.5 mL/kg[2,9,10,11,26]. This definitional heterogeneity may have contributed to between-study variability.

Some studies have reported that a clear liquid diet in the 24 hours before procedures may help reduce RGC in patients using GLP1-RAs[1,8]. A study found same-day colonoscopy to be protective against RGC (OR = 0.41, 95%CI: 0.23-0.73), likely reflecting the effect of prolonged liquid fasting protocols[25]. Other observational studies have reported similar protective association, with one showing an OR of 0.34 (95%CI: 0.23-0.52)[1] and another demonstrating a further reduction when upper endoscopy was combined with colonoscopy (prevalence ratio 0.25, 95%CI: 0.16-0.39) compared with upper endoscopy alone[10].

Pulmonary aspiration under anesthesia, although rare (estimated incidence 1 in 3000-7000 elective procedures), involves substantial morbidity and remains the leading cause of anesthesia-related mortality[11,28]. Nearly half of the patients required intensive or high-dependency care, over one-third required mechanical ventilation, and mortality reached 6%-7%. It accounted for up to 50% of anesthesia-related mortality and 17% of major airway events, with death or permanent severe injury documented in more than 50% of reported cases[29-31].

Given this biologically plausible mechanism and the catastrophic consequences of aspiration, even a modest increase in risk would be clinically significant[32,33]. Consequently, prospective studies investigating pulmonary aspiration as a primary outcome, requires an impractically large sample size to achieve sufficient statistical power[9].

To contextualize the rarity of aspiration events, the sample size required to detect a small absolute difference in risk was estimated, assuming a baseline incidence of 0.05%[23], a reduction to 0.03%, and two-sided α of 0.05. Under these

assumptions, approximately 300000-400000 participants would be needed to achieve conventional power, whereas our meta-analysis included only 11022 semaglutide users and 38629 controls. This large discrepancy highlights that the available data are substantially underpowered for aspiration analysis and helps explain the wide CIs around our pooled estimate.

Although aspiration-focused trials are unlikely to be feasible because of the rarity of the outcome, randomized studies assessing intermediate endpoints, such as ultrasound-assessed RGC under different fasting or semaglutide-holding strategies, are needed. Current evidence therefore relies mainly on observational studies[6].

Although most included studies achieved high NOS scores, these ratings should be interpreted cautiously because all included studies were observational and, thus, remain vulnerable to residual confounding, selection bias, and unmeasured differences between groups. The NOS is a useful tool for assessing key domains such as selection, comparability, and outcome assessment[15], but high scores do not eliminate the fundamental limitations of nonrandomized designs. Accordingly, the available evidence should be interpreted as associative rather than causal, and pooling these studies does not overcome the underlying risk of bias.

The original protocol also included a dose–response/time-since-last-dose meta-analysis to examine the relationship between interruption duration and outcomes, but this was not feasible because interruption timing was inconsistently reported and rarely comparable across studies[2,10,26]. Only a minority of studies reported interruption intervals, and these definitions were heterogeneous[2,10,12,26]. Therefore, our data cannot determine whether a 7-day interruption is superior to shorter or longer intervals, and this approach should be prospectively tested rather than adopted as a fixed standard.

Substantial heterogeneity was observed in the pooled analysis for RGC and remained present across sensitivity analyses. Although exploratory subgroup analyses by study design and procedure type did not identify statistically robust between-group differences under the random-effects model, these analyses suggest that several factors may have contributed to between-study variability. These include differences in procedure type, baseline patient characteristics, definitions of RGC and digestive symptoms, semaglutide interruption intervals, and perioperative management strategies. Importantly, many of these variables were incompletely or inconsistently reported, which limited more granular subgroup and meta-regression analyses. Therefore, the pooled estimates should be interpreted cautiously, and future studies should use more standardized definitions and exposure reporting to improve comparability across cohorts.

This study has several limitations. First, important clinical data, such as the interval of semaglutide interruption, American Society of Anesthesiologists physical status, and perioperative fasting duration, were not consistently available across the included studies, which limits our ability to perform more granular subgroup analyses. Second, definitions of RGC and digestive symptoms varied between studies, contributing to heterogeneity. Third, the outcomes of greatest clinical concern - pulmonary aspiration and procedure discontinuation - were rarely reported, resulting in wide CIs and low certainty of evidence. Fourth, no randomized controlled trials were identified; all the included studies were observational, so residual confounding cannot be excluded, even though most studies scored highly on the NOS. Fifth, although studies enrolling broader GLP-1RA populations were only included when semaglutide-specific subgroup data could be isolated, differences in subgroup reporting and disaggregation across studies may still have affected comparability.

Finally, aspiration risk depends not only on gastric volume but also on whether full stomach precautions, such as rapid-sequence induction or gastric ultrasound, were applied, and this information was absent or inconsistently reported across included studies. Whether clinicians already aware of semaglutide-related risk selectively modified their practice also cannot be determined, which could have biased our estimates toward the null.

Consequently, our findings cannot establish that semaglutide increases RGC or aspiration, only that these events were observed more frequently in semaglutide users in the available cohorts. Although pooling such studies improves precision, it does not overcome the fundamental limitations of nonrandomized designs.

Future studies should systematically evaluate additional outcomes beyond the primary endpoints, including length of hospital stay, postoperative pain, 30-day readmission rates, and delayed recovery. Standardized definitions, validated measurement tools, weekly doses, therapy duration, and consistent follow-up intervals are essential to ensure comparability across studies and better characterize the broader effect of preoperative semaglutide on surgical recovery.

Existing guidelines address the perioperative management of GLP-1RA users differently. Multisociety, Brazilian, and European Society of Anaesthesiology and Intensive Care guidelines all endorse some combination of selective drug interruption, extended clear-fluid intake, and gastric ultrasonography, but acknowledge that the evidence base remains limited. A comparative summary of these recommendations is provided in Table 4[5,34,35].

CONCLUSION

Current observational evidence suggests that perioperative semaglutide treatment may be associated with increased RGC and digestive symptoms. However, RGC is a surrogate marker rather than a direct clinical endpoint, and its increase should not be interpreted as direct evidence of increased pulmonary aspiration risk. Evidence regarding pulmonary aspiration, procedure discontinuation, and gastroparesis remains uncertain because of sparse event data, wide CIs, and low to very low certainty of evidence. Given the observational design of the included studies, substantial heterogeneity, and absence of randomized trials, these findings should be interpreted cautiously. Prospective studies using standardized definitions, fasting protocols, and perioperative management strategies are needed to clarify the clinical significance of these associations.

Table 4 Comparative recommendations for perioperative management of glucagon-like peptide-1 receptor agonists

Guideline/source	Suspension timing	Dietary measures	Use of gastric ultrasound	Risk stratification/other notes
Multisociety Clinical Practice Guidance (ASA and others, 2024)[5]	No universal discontinuation. If high-risk: Daily formulations - hold on day of surgery; weekly formulations - hold 7 days prior. Duration beyond these intervals unknown	Consider ≥ 24 hours liquid diet if concern for delayed emptying	May be used when retained gastric content is a concern	High-risk factors: Dose escalation, high doses, weekly formulations, active gastrointestinal symptoms, comorbid dysmotility. Avoid discontinuation only in obesity (bias concern)
Brazilian Position Statement (SBD/SBA/ABESO, 2025)[34]	7 days for long-acting; 1 day for short-acting, in patients at increased aspiration risk or within 12 weeks of dose escalation/instability	Liquid diet 24 hours before + 8-12 hours fasting	POCUS is recommended whenever available	Emphasis on selective suspension plus diet + ultrasound
ESAIC Guideline (2025) [35]	≥ 7 days for weekly; ≥ 14 days if for obesity. Hold daily formulations on day of surgery	24 hours clear-fluid diet	Strongly encouraged; treat all GLP-1RA users as "full stomach" if content present	More conservative: Notes 1-week hold may not normalize gastric emptying; RSI if urgent

ASA: American Society of Anesthesiologists; ABESO: Brazilian Association for the Study of Obesity and Metabolic Syndrome; ESAIC: European Society of Anaesthesiology and Intensive Care; POCUS: Point-of-care ultrasound; RSI: Rapid-sequence induction; SBA: Brazilian Society of Anesthesiology; SBD: Brazilian Diabetes Society.

ACKNOWLEDGEMENTS

Assistance with the article

The authors gratefully acknowledge the support of colleagues from the Department of Anesthesiology and the Department of Nephrology and Biostatistics, University of Caxias do Sul, for their valuable input and assistance during this project.

Prior presentation

Preliminary results from this study were presented in abstract form at the American Society of Anesthesiologists Annual Meeting 2025, held on October 11-15, 2025, in San Antonio, Texas (Abstract #7227 - https://www.abstractsonline.com/pp8/#!/21028/presentation/7227).

Guarantor statement

Vinicius Remus Ballotin is the guarantor of this work and had full access to all data, taking responsibility for the integrity of the data and accuracy of the analyses.

REFERENCES

1 **Firkins SA**, Yates J, Shukla N, Garg R, Vargo JJ, Lembo A; Cleveland Clinic Obesity Medicine and Bariatric Endoscopy Working Group. Clinical Outcomes and Safety of Upper Endoscopy While on Glucagon-Like Peptide-1 Receptor Agonists. *Clin Gastroenterol Hepatol* 2025; **23**: 872-873.e3 [*RCA*] [PMID: 38574832 DOI: 10.1016/j.cgh.2024.03.013] [FullText]

2 **Nersessian RSF**, da Silva LM, Carvalho MAS, Silveira SQ, Abib ACV, Bellicieri FN, Lima HO, Ho AM, Anjos GS, Mizubuti GB. Relationship between residual gastric content and peri-operative semaglutide use assessed by gastric ultrasound: a prospective observational study. *Anaesthesia* 2024; **79**: 1317-1324 [*RCA*] [PMID: 39435967 DOI: 10.1111/anae.16454] [FullText]

3 **World Obesity Federation**. World Obesity Atlas 2025. [cited 24 March 2026]. Available from: https://www.worldobesity.org/resources/resource-library/world-obesity-atlas-2025

4 **International Diabetes Federation**. IDF Diabetes Atlas 11th Edition 2025. Brussels, Belgium: International Diabetes Federation, 2025

5 **Kindel TL**, Wang AY, Wadhwa A, Schulman AR, Sharaiha RZ, Kroh M, Ghanem OM, Levy S, Joshi GP, LaMasters T; Representing the American Gastroenterological Association, American Society for Metabolic and Bariatric Surgery, American Society of Anesthesiologists, International Society of Perioperative Care of Patients with Obesity, and the Society of American Gastrointestinal and Endoscopic Surgeons. Multi-society clinical practice guidance for the safe use of glucagon-like peptide-1 receptor agonists in the perioperative period. *Surg Endosc* 2025; **39**: 180-183 [*RCA*] [PMID: 39370500 DOI: 10.1007/s00464-024-11263-2] [FullText] [Full Text(PDF)]

6 **Klonoff DC**, Kim SH, Galindo RJ, Joseph JI, Garrett V, Gombar S, Aaron RE, Tian T, Kerr D. Risks of peri- and postoperative complications with glucagon-like peptide-1 receptor agonists. *Diabetes Obes Metab* 2024; **26**: 3128-3136 [*RCA*] [PMID: 38742898 DOI: 10.1111/dom.15636] [FullText] [Full Text(PDF)]

7 **Jalleh RJ**, Plummer MP, Marathe CS, Umapathysivam MM, Quast DR, Rayner CK, Jones KL, Wu T, Horowitz M, Nauck MA. Clinical Consequences of Delayed Gastric Emptying With GLP-1 Receptor Agonists and Tirzepatide. *J Clin Endocrinol Metab* 2024; **110**: 1-15 [*RCA*] [PMID: 39418085 DOI: 10.1210/clinem/dgae719] [FullText] [Full Text(PDF)]

8 **Zaffar D**, Hamza M, Borissov MB, Dy KC, Hwang SY, Hsieh P, Tsai J, Lee DU, Kim RE. Su1100 Association Between Perioperative GLP-1 Agonist Use And Post Operative Complications In Patients Undergoing Upper Endoscopy. *Gastroenterology* 2024; **166**: S-653 [DOI:

10.1016/s0016-5085(24)01955-3] [FullText]

9 **Mendes FF**, Carvalho LIM, Lopes MB. Glucagon-Like Peptide-1 agonists in perioperative medicine: to suspend or not to suspend, that is the question. *Braz J Anesthesiol* 2024; **74**: 844538 [*RCA*] [PMID: 38944239 DOI: 10.1016/j.bjane.2024.844538] [FullText]

10 **Silveira SQ**, da Silva LM, de Campos Vieira Abib A, de Moura DTH, de Moura EGH, Santos LB, Ho AM, Nersessian RSF, Lima FLM, Silva MV, Mizubuti GB. Relationship between perioperative semaglutide use and residual gastric content: A retrospective analysis of patients undergoing elective upper endoscopy. *J Clin Anesth* 2023; **87**: 111091 [*RCA*] [PMID: 36870274 DOI: 10.1016/j.jclinane.2023.111091] [FullText]

11 **Sen S**, Potnuru PP, Hernandez N, Goehl C, Praestholm C, Sridhar S, Nwokolo OO. Glucagon-Like Peptide-1 Receptor Agonist Use and Residual Gastric Content Before Anesthesia. *JAMA Surg* 2024; **159**: 660-667 [*RCA*] [PMID: 38446466 DOI: 10.1001/jamasurg.2024.0111] [FullText] [Full Text(PDF)]

12 **Welk B**, McClure JA, Carter B, Clarke C, Dubois L, Clemens KK. No association between semaglutide and postoperative pneumonia in people with diabetes undergoing elective surgery. *Diabetes Obes Metab* 2024; **26**: 4105-4110 [*RCA*] [PMID: 38860419 DOI: 10.1111/dom.15711] [FullText]

13 **Moher D**, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; **6**: e1000097 [*RCA*] [PMID: 19621072 DOI: 10.1371/journal.pmed.1000097] [FullText] [Full Text(PDF)]

14 **Higgins JPT**, Thomas J, Chandler J, Cumpston M, Li T, PageVivian A Welch MJ. Cochrane Handbook for Systematic Reviews of Interventions. NJ: John Wiley & Sons, 2019

15 **Luchini C**, Stubbs B, Solmi M, Veronese N. Assessing the quality of studies in meta-analyses: Advantages and limitations of the Newcastle Ottawa Scale. *World J Meta-Anal* 2017; **5**: 80-84 [*RCA*] [DOI: 10.13105/wjma.v5.i4.80] [FullText] [Full Text(PDF)]

16 **Wells GA**, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Jan, 2000. [cited 24 March 2026]. Available from: https: https://www.researchgate.net/publication/261773681_The_Newcastle-Ottawa_Scale_NOS_for_Assessing_the_Quality_of_Non-Randomized_Studies_in_Meta-Analysis

17 **Granholm A**, Alhazzani W, Møller MH. Use of the GRADE approach in systematic reviews and guidelines. *Br J Anaesth* 2019; **123**: 554-559 [*RCA*] [PMID: 31558313 DOI: 10.1016/j.bja.2019.08.015] [FullText]

18 **Guyatt GH**, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; **336**: 924-926 [*RCA*] [PMID: 18436948 DOI: 10.1136/bmj.39489.470347.AD] [FullText]

19 **Ballotin VR**. A Multimodal Approach to Optimizing Semaglutide Suspension Timing Before Surgery: A Systematic Review and Meta-Analysis. Nov 2, 2025. [cited 24 March 2026]. Available from:https://figshare.com/articles/online_resource/_b_A_Multimodal_Approach_to_Optimizing_Semaglutide_Suspension_Timing_Before_Surgery_A_Systematic_Review_and_Meta-Analysis_b_/30450836

20 **Gu GH**, Pauplis C, Devuni D, Krishnarao A. Sa1967 The Impact Of Semaglutide On Retained Gastric Contents During Endoscopy. *Gastroenterology* 2024; **166**: S-599 [DOI: 10.1016/s0016-5085(24)01844-4] [FullText]

21 **Ukwade D**, Hernandez LJ, Modi Z, Raza HS, Siegel M, Nwachukwu D, Sarraf P, Memon A, Booth M, Cooper K, Boulay BR, Mutlu EA. GLP-1 receptor agonist increase retained gastric contents on EGD and same-day colonoscopy reduces this risk. *Front Med (Lausanne)* 2025; **12**: 1638981 [*RCA*] [PMID: 41001395 DOI: 10.3389/fmed.2025.1638981] [FullText] [Full Text(PDF)]

22 **Vlaeminck N**, Van de Putte P, Dekeyser M, Baert N, Wallyn A, Vernieuwe L, Smitz C, Wouters K, Van Cauwenberghe J, Saldien V. Gastric ultrasound in patients receiving semaglutide: a prospective, multicentre, matched control study. *Anaesthesia* 2026 [*RCA*] [PMID: 41631344 DOI: 10.1111/anae.70129] [FullText]

23 **Alkabbani W**, Suissa K, Gu KD, Cromer SJ, Paik JM, Bykov K, Hobai I, Thompson CC, Wexler DJ, Patorno E. Glucagon-like peptide-1 receptor agonists before upper gastrointestinal endoscopy and risk of pulmonary aspiration or discontinuation of procedure: cohort study. *BMJ* 2024; **387**: e080340 [*RCA*] [PMID: 39438043 DOI: 10.1136/bmj-2024-080340] [FullText] [Full Text(PDF)]

24 **Phan J**, Chang P, Issa D, Turner R, Dodge J, Westanmo A, Karna R, Olive L, Bahdi F, Aldzhyan V, Bilal M, Tielleman T. 885 Glucagon-Like Peptide 1 Receptor Agonists (GLP1-RA) Prior To Endoscopy Have Low Gastric Retained Contents: A Multicentered Study. *Gastroenterology* 2024; **166**: S-208 [DOI: 10.1016/s0016-5085(24)00970-3] [FullText]

25 **Korlipara H**, Kumar S, Yeung M, Buckholz A, Jamison J, Gonzalez A, Krisko T, Sharaiha R, Newberry C. Semaglutide Is An Independent Risk Factor For Retained Gastric Contents And Recommendation For Repeat Upper Endoscopy: Results From A Real-World Population Cohort Study. *Gastrointest Endosc* 2024; **99**: AB147 [DOI: 10.1016/j.gie.2024.04.938] [FullText]

26 **Santos LB**, Mizubuti GB, da Silva LM, Silveira SQ, Nersessian RSF, Abib ACV, Bellicieri FN, Lima HO, Ho AM, Dos Anjos GS, de Moura DTH, de Moura EGH, Vieira JE. Effect of various perioperative semaglutide interruption intervals on residual gastric content assessed by esophagogastroduodenoscopy: A retrospective single center observational study. *J Clin Anesth* 2024; **99**: 111668 [*RCA*] [PMID: 39476514 DOI: 10.1016/j.jclinane.2024.111668] [FullText]

27 **Bergmann NC**, Davies MJ, Lingvay I, Knop FK. Semaglutide for the treatment of overweight and obesity: A review. *Diabetes Obes Metab* 2023; **25**: 18-35 [*RCA*] [PMID: 36254579 DOI: 10.1111/dom.14863] [FullText] [Full Text(PDF)]

28 **Sakai T**, Planinsic RM, Quinlan JJ, Handley LJ, Kim TY, Hilmi IA. The incidence and outcome of perioperative pulmonary aspiration in a university hospital: a 4-year retrospective analysis. *Anesth Analg* 2006; **103**: 941-947 [*RCA*] [PMID: 17000809 DOI: 10.1213/01.ane.0000237296.57941.e7] [FullText]

29 **Kluger MT**, Culwick MD, Moore MR, Merry AF. Aspiration during anaesthesia in the first 4000 incidents reported to webAIRS. *Anaesth Intensive Care* 2019; **47**: 442-451 [*RCA*] [PMID: 31438719 DOI: 10.1177/0310057X19854456] [FullText]

30 **Warner MA**, Meyerhoff KL, Warner ME, Posner KL, Stephens L, Domino KB. Pulmonary Aspiration of Gastric Contents: A Closed Claims Analysis. *Anesthesiology* 2021; **135**: 284-291 [*RCA*] [PMID: 34019629 DOI: 10.1097/ALN.0000000000003831] [FullText]

31 **Robinson M**, Davidson A. Aspiration under anaesthesia: risk assessment and decision-making. *Cont Educ Anaesth Crit Care Pain* 2014; **14**: 171-175 [DOI: 10.1093/bjaceaccp/mkt053] [FullText]

32 **DiBardino DM**, Wunderink RG. Aspiration pneumonia: a review of modern trends. *J Crit Care* 2015; **30**: 40-48 [*RCA*] [PMID: 25129577 DOI: 10.1016/j.jcrc.2014.07.011] [FullText]

33 **Studer P**, Räber G, Ott D, Candinas D, Schnüriger B. Risk factors for fatal outcome in surgical patients with postoperative aspiration pneumonia. *Int J Surg* 2016; **27**: 21-25 [*RCA*] [PMID: 26804349 DOI: 10.1016/j.ijsu.2016.01.043] [FullText]

34 **Marino EC**, Negretto LAF, Ribeiro RS, Momesso D, Feitosa ACR, Toyoshima MTK, da Silva Junior JC, Vencio S, Lauria MW, de Sá JR, Malerbi DA, Valente F, Leite SAO, Amaral DEO, Guimarães GMN, da Cunha Leal P, Lopes MB, Salles LCB, de Araújo Azi LMT, Fonseca

AG, Carvalho LIM, Coelho FF, Halpern B, Valerio CM, Trujilho FR, Brandão ACA, Lyra R, Bertoluci M. Perioperative screening and management of hyperglycemia: a joint position statement from the Brazilian Diabetes Society (SBD), the Brazilian Society of Anesthesiology (SBA) and the Brazilian Association for the Study of Obesity and Metabolic Syndrome (ABESO). *Diabetol Metab Syndr* 2026; **18**: 91 [*RCA*] [PMID: 41761348 DOI: 10.1186/s13098-025-02060-5] [FullText] [Full Text(PDF)]

35 **Lamperti M**, Romero CS, Guarracino F, Cammarota G, Vetrugno L, Tufegdzic B, Lozsan F, Macias Frias JJ, Duma A, Bock M, Ruetzler K, Mulero S, Reuter DA, La Via L, Rauch S, Sorbello M, Afshari A. Preoperative assessment of adults undergoing elective noncardiac surgery: Updated guidelines from the European Society of Anaesthesiology and Intensive Care. *Eur J Anaesthesiol* 2025; **42**: 1-35 [*RCA*] [PMID: 39492705 DOI: 10.1097/EJA.0000000000002069] [FullText]

FOOTNOTES

Specialty type: Gastroenterology and hepatology

Country of origin: Brazil

Author contributions: Remus Ballotin V and Carneiro Ferreira PL performed the literature search, study screening, and data extraction; they both contributed equally to this article and are the co-first authors of this manuscript; Vinicius RB and Luciano SS conducted the statistical analyses and drafted the manuscript; da Silva Borges C and Tonin de Almeida H assisted with study selection, data verification, and evidence quality assessment; da Silva Borges C, Tonin de Almeida H, and Giovanardi Pandolfo da Silva R supported data extraction, tabulation, and figure preparation; Remus Ballotin V, Volquind D, and da Silva Selistre L conceptualized and designed the study; Volquind D, and da Silva Selistre L supervised the methodology, contributed to data interpretation, and critically revised the manuscript; and all authors read and approved the final version of the manuscript.

AI contribution statement: All the contents of the document responding to the review comments were written by humans. ChatGPT is only used for language polishing and grammar correction.

Supported by Graduate Support Program for Community Higher Education Institutions Scholarship, granted by the Brazilian Government Agency Coordination for the Improvement of Higher Education Personnel.

Conflict-of-interest statement: All the authors report no relevant conflicts of interest for this article.

PRISMA 2009 Checklist statement: The authors have read the PRISMA 2009 Checklist, and the manuscript was prepared and revised according to the PRISMA 2009 Checklist.

S-Editor: Bai Y
L-Editor: Filipodia
P-Editor: Zhao YQ

APÊNDICE G – REGISTRO DO PROTOCOLO DA REVISÃO SISTEMÁTICA: PROSPERO (INTERNATIONAL PROSPECTIVE REGISTER OF SYSTEMATIC REVIEWS)

NIHR | National Institute for Health and Care Research

PROSPERO
International prospective register of systematic reviews

Optimizing Semaglutide Suspension Timing Pre-Surgery: A Systematic Review and Meta-Analysis

Vinicius Remus Ballotin, Luciano da Silva Selistre

REVIEW TITLE AND BASIC DETAILS

Review title
Optimizing Semaglutide Suspension Timing Pre-Surgery: A Systematic Review and Meta-Analysis

Review objectives
In patients scheduled for surgery (P), does the administration of pre-operative semaglutide (I) compared to patients who have not been exposed to semaglutide (C) impact the incidence of increased residual gastric content, bronchial aspiration, and complications related to airway management (O) during the perioperative period (T)?

Keywords
Anesthesia, Obesity, Perioperative, Residual gastric content, Semaglutide, Type 2 diabetes

SEARCHING AND SCREENING

Searches
Searches will be running on the electronic databases Scopus, Web of Science, MEDLINE (PubMed), Embase, BIREME (Biblioteca Regional de Medicina), LILACS (Latin American and Caribbean Health Sciences Literature), Cochrane Library for Systematic Reviews, SciELO (Scientific Electronic Library Online), Embase and Opengray.eu. Languages were restricted to English, Spanish and Portuguese. Any date of publication will be included.

Study design

Randomized controlled trials (RCTs) and observational studies (including cohort, case-control, and cross-sectional designs)

ELIGIBILITY CRITERIA

Condition or domain being studied

The healthcare domain being studied in this systematic review is the management of patients with type 2 diabetes, obesity or overweight with related health conditions undergoing surgical procedures. Semaglutide, a GLP-1 receptor agonist, is commonly prescribed for these patients to improve glycemic control and support weight loss. However, its impact on perioperative outcomes, such as increased residual gastric content, bronchial aspiration, and complications related to airway management, is not well understood.

Patients with type 2 diabetes and/or obesity are at an elevated risk for perioperative complications due to factors such as delayed gastric emptying and potential airway challenges. These complications can lead to increased morbidity and prolonged recovery times.

Additionally, an important aspect of this review will be to explore the ideal timing for the withdrawal of semaglutide prior to surgery, as this remains an unknown variable that could significantly influence patient outcomes. By addressing these concerns, the review aims to enhance understanding of the safe management of semaglutide in the perioperative setting and improve clinical guidelines for these high-risk patients.

Population

Inclusion Criteria

1. Adults aged 18 years and older.
2. Patients who have received preoperative semaglutide treatment, regardless of the duration of therapy prior to surgery.
3. Randomized controlled trials (RCTs) and observational studies (including cohort, case-control, and cross-sectional designs) that report perioperative-related outcomes, including:

a. Residual gastric content
b. Incidence of bronchial aspiration
c. Complications related to airway management during the perioperative period
d. Withdrawal Semaglutide Period
e. Overall perioperative complications.

Exclusion Criteria

1. Patients with significant comorbidities that independently elevate the risk of perioperative complications (e.g., severe cardiovascular or respiratory diseases).
2. Patients with contraindications to semaglutide treatment.
3. Studies that focus exclusively on other GLP-1 receptor agonists or diabetes medications without including semaglutide.
4. Case reports, case series, reviews, and expert opinions that do not present original data or comparisons between groups.
5. Non-English language studies will be excluded unless translated versions are readily available.
6. Studies with significant or missing data that impede proper analysis and interpretation of

outcomes.

Intervention(s) or exposure(s)
Patients exposed to semaglutide

Comparator(s) or control(s)
Patients not exposed to semaglutide

Context
The inclusion criteria for this study focus on adults aged 18 years and older who have received preoperative semaglutide treatment, irrespective of the duration of therapy prior to surgery. Eligible studies must consist of randomized controlled trials (RCTs) and observational designs, including cohort, case-control, and cross-sectional studies, that report perioperative outcomes such as residual gastric content, incidence of bronchial aspiration, complications related to airway management during the perioperative period, the ideal time to withdrawal Semaglutide, and overall perioperative complications. In contrast, the exclusion criteria eliminate patients with significant comorbidities that could independently elevate the risk of perioperative complications, such as severe cardiovascular or respiratory diseases. Additionally, patients with contraindications to semaglutide treatment will be excluded. Studies focusing solely on other GLP-1 receptor agonists or diabetes medications, without including semaglutide, are also ineligible. Furthermore, case reports, case series, reviews, and expert opinions that do not present original data or comparisons between groups will not be considered for this analysis. By adhering to these criteria, the study aims to ensure a focused examination of the impact of preoperative semaglutide treatment on perioperative outcomes, thereby providing reliable insights into its efficacy and safety in surgical settings.

OUTCOMES TO BE ANALYSED

Main outcomes
The pre-specified primary outcomes of this review focus on essential perioperative metrics that evaluate the impact of preoperative semaglutide treatment. These metrics include residual gastric content, defined as the volume of gastric contents present at the induction of anesthesia, measured in milliliters (mL) immediately before anesthesia begins. The incidence of bronchial aspiration, which indicates the occurrence of gastric contents entering the bronchial tree, will be assessed based on clinical signs and confirmed through imaging during the intraoperative period. Furthermore, complications related to airway management, including adverse events such as failed intubation, will be reported by the studies. The withdrawal semaglutide period will monitor the duration and management of semaglutide cessation prior to surgery, with measurements taken from the withdrawal date to the surgery date. Lastly, overall perioperative complications, encompassing any adverse events occurring throughout the perioperative period, will be categorized as minor or major based on clinical assessments and surgical records. By systematically measuring these outcomes, the review aims to provide valuable insights into the efficacy and safety of preoperative semaglutide treatment in surgical patients.

Additional outcomes

The review includes several pre-specified additional outcomes that complement the main outcomes. These additional outcomes encompass length of hospital stay, defined as the duration from surgery to discharge, which will be measured in days and assessed upon patient discharge. Postoperative pain levels will be evaluated using a standardized pain scale (such as the Visual Analog Scale) at specified intervals post-surgery (e.g., 24 hours, 48 hours, and upon discharge), providing insights into patient recovery experiences. The review will also assess quality of life (QoL) using validated instruments such as the EQ-5D or SF-36, with measurements taken preoperatively and at follow-up intervals to gauge any changes following surgery. Additionally, the review will examine readmission rates within 30 days post-surgery, defined as any unplanned return to the hospital, which will be documented through hospital records. Finally, the incidence of delayed recovery will be noted, characterized by any prolonged recovery exceeding expected timeframes based on the type of surgery. By including these additional outcomes, the review aims to provide a comprehensive evaluation of the implications of preoperative semaglutide treatment on patient recovery and overall health.

DATA COLLECTION PROCESS

Data extraction (selection and coding)

A systematic literature search will be conducted from inception to November 2024 in nine electronic databases: Scopus, Web of Science, Embase, MEDLINE (PubMed), BIREME (Biblioteca Regional de Medicina), LILACS (Latin American and Caribbean Health Sciences Literature), SciELO (Scientific Electronic Library Online), Cochrane Library for Systematic Reviews, and Opengray.eu (the full search strategy will be described in the Appendix). Studies retrieved from the databases will be systematically reviewed by two independent researchers, and the reference lists of the retrieved studies will undergo a manual search. Any divergences in study selection will be resolved by a third researcher.

Risk of bias (quality) assessment

The quality of the included meta-analyses will be assessed using either the updated version of GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) or the Newcastle-Ottawa Scale (NOS), depending on the type of study. Two independent researchers will conduct the assessments, and their results will be compared to evaluate inter-rater reliability. Any divergences will be discussed with a third researcher until 100% consensus is reached. The GRADE scale is an assessment tool that evaluates the quality of evidence and the strength of recommendations across various research areas, considering five main domains that influence confidence in the results. Similarly, the Newcastle-Ottawa Scale (NOS) is an assessment tool that examines the quality of observational studies across three domains: selection, comparability, and outcomes, allowing for the classification of studies in terms of risk of bias and internal validity.

PLANNED DATA SYNTHESIS

Strategy for data synthesis

In this review, the synthesis of data will primarily involve conducting a meta-analysis for the identified studies that report on perioperative outcomes related to preoperative semaglutide treatment. We will employ a random-effects model for the meta-analysis, as this model accounts for variations among study populations and interventions, providing a more generalized estimate of the treatment effect. The random-effects model will be particularly useful given the anticipated diversity in study designs, populations, and methodologies.

To explore statistical heterogeneity among the studies, we will utilize the I^2 statistic, which quantifies the percentage of variation across studies that is attributable to heterogeneity rather than chance. An I^2 value of 25% will be considered low, 50% moderate, and 75% high heterogeneity. In addition, we will perform subgroup analyses based on variables such as study design (RCT vs. observational), type of surgery, and duration of semaglutide treatment to assess the impact of these factors on outcome variability. If significant heterogeneity is detected, we will also explore potential sources using meta-regression techniques.

All analyses will be conducted using the RStudio, which will facilitate the integration of diverse study results and provide forest plots for visual representation of the findings. Sensitivity analyses will be performed to determine the robustness of the results by excluding studies with high risk of bias. By employing these specific methods, we aim to provide a comprehensive and reliable synthesis of the evidence on the impact of preoperative semaglutide treatment on perioperative outcomes.

Analysis of subgroups or subsets

In this review, we plan to investigate several predefined subgroups to assess how various factors influence perioperative outcomes associated with preoperative semaglutide treatment. First, we will differentiate between randomized controlled trials (RCTs) and observational studies (cohort, case-control, cross-sectional) to evaluate if study design impacts outcomes, using separate random-effects meta-analyses for each group. We will also categorize patients based on the type of surgical procedure (e.g., elective, emergency, gastrointestinal) and perform subgroup analyses to explore differences in outcomes across these categories. Additionally, we will assess the impact of the duration of semaglutide treatment prior to surgery by grouping patients into short-term (< 3 months), medium-term (3-6 months), and long-term (> 6 months) treatment categories. Age groups will also be analyzed, stratifying participants into younger adults (18-50 years), middle-aged adults (51-70 years), and older adults (71 years and above) to examine differential effects. Lastly, we will explore the influence of comorbid conditions by comparing outcomes between patients with significant comorbidities (e.g., cardiovascular disease) and those without. These subgroup analyses will utilize meta-regression techniques to enhance our understanding of how specific factors may modify the efficacy and safety of preoperative semaglutide treatment.

REVIEW AFFILIATION, FUNDING AND PEER REVIEW

Review team members

- Mr Vinicius Remus Ballotin, University of Caxias do Sul
- Luciano da Silva Selistre, University of Caxias do Sul

Review affiliation
Universidade de Caxias do Sul

Funding source
The authors declare that there are no funding sources or sponsors for this study

Named contact
Vinicius Remus Ballotin. 1130 Francisco Getúlio Vargas StreetZIP Code 95070-560 - Caxias do Sul - University of Caxias do Sul - Postgraduate Program in Health Sciences
viniballotin@gmail.com

TIMELINE OF THE REVIEW

Review timeline
Start date: 11 November 2024. End date: 01 June 2025

Date of first submission to PROSPERO
04 November 2024

Date of registration in PROSPERO
16 November 2024

AVAILABILITY OF FULL PROTOCOL

Availability of full protocol
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

CURRENT REVIEW STAGE

Publication of review results 1 change
Results of the review will be published in English.

Preprint publication
https://doi.org/10.4240/wjgs.121449

Journal publication
Not yet published in a journal but will be in future.

Stage of the review at this submission 1 change

Review stage	Started	Completed
Pilot work	✓	✓
Formal searching/study identification	✓	✓
Screening search results against inclusion criteria	✓	✓

Review stage	Started	Completed
Data extraction or receipt of IPD	✓	✓
Risk of bias/quality assessment	✓	✓
Data synthesis	✓	✓

Review status
The review is completed.

ADDITIONAL INFORMATION

PROSPERO version history 1 change

- Version 1.4, published 29 Aug 2026
- Version 1.3, published 24 Mar 2025
- Version 1.2, published 24 Mar 2025
- Version 1.1, published 16 Nov 2024
- Version 1.0, published 16 Nov 2024

Review conflict of interest
None known

Country
Brazil

Medical Subject Headings
Glucagon-Like Peptides; Humans; Hypoglycemic Agents; Incidence; Patients; semaglutide

Revision note 1 change
The manuscript has been completed and submitted to the medical journal.

Disclaimer
The content of this record displays the information provided by the review team. PROSPERO does not peer review registration records or endorse their content.

PROSPERO accepts and posts the information provided in good faith; responsibility for record content rests with the review team. The guarantor for this record has affirmed that the information provided is truthful and that they understand that deliberate provision of inaccurate information may be construed as scientific misconduct.

PROSPERO does not accept any liability for the content provided in this record or for its use. Readers use the information provided in this record at their own risk.

Any enquiries about the record should be referred to the named review contact