ESCOLA BAHIANA DE MEDICINA E SAÚDE PÚBLICA

PROGRAMA DE GRADUAÇÃO EM MEDICINA

REBECA MENEZES DE OLIVEIRA LIMA

FREQUÊNCIA DE CONCLUSÕES POTENCIALMENTE FALSO-POSITIVAS EM ENSAIOS CLÍNICOS RANDOMIZADOS

Salvador - BA

2022

REBECA MENEZES DE OLIVEIRA LIMA

FREQUÊNCIA DE CONCLUSÕES POTENCIALMENTE FALSO-POSITIVAS EM ENSAIOS CLÍNICOS RANDOMIZADOS

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Medicina da Escola Bahiana de Medicina e Saúde Pública para aprovação parcial no 4º ano de Medicina.

Orientação: Dr. Diego Ribeiro Rabelo.

Salvador – BA

2022

AGRADECIMENTOS

Esse trabalho de conclusão de curso signficou muito para mim e não poderia ter sido concretizado sem a ajuda e contribuição de determindas pessoas. Gostaria de agradecer ao Professor Diego Rabelo, por toda a paciênca e pelos ensinamentos que ultrapassaram os limites acadêmicos. Gostaria de agradecer também à minha amiga/irmã Luisa Covre que me acompanhou durante todo processo e me ajudou nas momentos mais difícies - sem ela nada disso poderia ter sido feito. Além disso, gostaria de agradecer a todos meus amigos da faculdade e à minha família, que se mostraram uma fonte de apoio importante quando precisei.

“If you torture your data long enough, they will tell you whatever you want to hear”

- MILLS (1993)

RESUMO

Introdução: Na pesquisa científica existem dois tipos de erros que interferem na validade interna do artigo: erro sistemático e erro aleatório, sendo este subdivido em erro do tipo I (falso-positivo) e erro do tipo II (falso-negativo). Nesse contexto, atualmente, existe uma tendência dos instrumentos de validade interna de avaliarem diretamente apenas a presença de erro sistemático, sendo o erro aleatório, muita vezes, negligenciado. Esse cenário é extremamente preocupante no que tange à presença ensaios clínicos randomizados potencialmente falso-positivos, pois prejudicam a prática clínica. **Objetivo:** Examinar a frequência de ensaios clínicos randomizados potencialmente falso-positivos em periódicos de grande relevância científica. **Metodologia:** Trata-se de um estudo metacientífico que incluiu ensaios clínicos publicados entre 2020-2021 em revistas de grande relevância científica (British Medical Association, Journal of the American Medical Association, LANCET e New England Journal of Medicine). Os critérios de exclusão foram: ensaios clínicos negativos, não randomizados, pragmáticos, de não inferioridade, de equivalência e ensaios clínicos duplicados. O ensaio clínico foi considerado falso-positivo se tivesse a presença de análise de desfecho secundário, análise de subgrupo e/ou análise interina na conclusão, ou seja, variáveis que aumentam a probabilidade de erro do tipo I. **Resultados:** Foram coletados 601 ensaios clínicos, 125 foram excluídos, resultando em 476 ensaios clínicos incluídos e destes, 295 apresentaram uma conclusão positiva e foram incluídos neste estudo. Dessa amostra, foram encontrados 76 (25%) ECs com resultados potencilamente falso-positivos na conclusão. Destes, 40 (52,6%) concluíram resultados por meio de análises de desfechos secundários, 33 (43,4%) por análises interinas, e 3 (4%) tanto por análises de desfechos secundários quanto análises interinas. **Conclusão:** Foi encontrada uma notória proporção de ensaios clínicos randomizados falso-positivos. Mais estudos na área metacientífica são necessários para aprofundar a evidência nesse âmbito.

.

ABSTRACT

Introduction: In scientific research, there are two types of errors that interfere with the internal validity of the article: systematic error and random error, which is subdivided into type I error (false-positive) and type II error (false-negative). In this context, currently, there is a tendency for internal validity instruments to directly assess only the presence of systematic error, with random error often being neglected. This scenario is extremely worrying regarding the presence of potentially false-positive randomized clinical trials, as they impair clinical practice. **Objective:** To examine the frequency of potentially false-positive randomized clinical trials in journals of great scientific relevance. **Methodology:** This is a meta-scientific study that included clinical trials published between 2020-2021 in highly relevant scientific journals (British Medical Association, Journal of the American Medical Association, LANCET and New England Journal of Medicine). Exclusion criteria were: negative clinical trials, non-randomized, pragmatic, non-inferiority, equivalence and duplicate clinical trials. The clinical trial was considered false positive if it had the presence of secondary outcome analysis, subgroup analysis and/or interim analysis at conclusion, that is, variables that increase the probability of type I error. **Results:** 601 clinical trials were collected, 125 were excluded, resulting in 476 clinical trials included and of these, 295 had a positive conclusion and were included in this study. From this sample, 76 (25%) ECs with potentially false-positive results were found. Of these, 40 (52.6%) concluded results by means of secondary outcome analyses, 33 (43.4%) by interim analyses, and 3 (4%) by both secondary outcome and interim analyses. **Conclusion:** A remarkable proportion of potentially false-positive randomized clinical trials was found. More studies in the meta-scientific area are needed to deepen the evidence in this area.

SUMÁRIO

1 INTRODUÇÃO

Ensaios clínicos randomizados (ECRs), quando adequadamente desenhados, conduzidos e relatados, constituem-se importantes estudos quanto à avaliação de intervenções na área de saúde[1]. No entanto, sem o devido rigor metodológico e sem a transparência adequada, a validade dos seus resultados está mais sujeita aos dois tipos de erros: sistemáticos e aleatórios[2]. O erro sistemático, ou viés, é uma fonte de variação que distorce os achados do estudo para uma determinada direção e o erro aleatório é um resultado equivocado devido ao acaso, com igual probabilidade de distorcer as aferições em ambas as direções. O erro aleatório pode ser subdividido em erro do tipo I (falso-positivo) ou erro do tipo II (falso-negativo). O erro do tipo I está relacionado ao problema das múltiplas comparações (desfechos secundários, análises de subgrupo e análises interinas)[3].

Apesar da importância de entender, lidar e reconhecer os erros nos estudos, instrumentos de validade interna envolvem cuidados mais direcionados a fim de diminuir o erro sistemático muitas vezes deixam de lado cuidados com os erros aleatórios[4]. Porém, como os processos aleatórios estão onipresentes, mesmo se tivermos um estudo com poucos erros sistemáticos, os resultados ainda estão sujeitos aos erro aleatórios e, assim, os autores não devem colocar nas suas conclusões resultados com potencial erro do tipo I, como aqueles provinientes de múltilplas comparações. A presença de estudos com conclusões que apresentam resultados potencialmente falsos é extremamente problemática, principalmente no que tange aos resultados falso-positivos, uma vez que eles disseminam rapidamente informações falsas sobre tratamentos, diagnósticos e medicamentos promissores, influenciando diretamente na prática clínica[5].

Assim, será que os erros aleatórios foram esquecidos e existe uma alta frequência de conclusões com resultados com alta probabilidade de erro do tipo I? Ainda são escassos estudos meta-epidemiológicos que analisam erros aleatórios, e os que analisam, não chegaram a pesquisar a frequência de estudos potencialmente falso-positivos. Estudos futuros nessa linha, caso demonstrem alta frequência de ECRs falso-positivos, refletiriam uma possível falha nas ferramentas metodológicos que diminuem o impacto dos erros aleatórios nas pesquisas, mostrando a necessidade de criação de ferramentas mais acuradas. Além disso, seria um incentivo

para adoção de critérios mais rígidos por parte dos periódicos, diminuindo o número de publicações falsas.

2 OBJETIVO

2.1 Objetivo geral

Examinar a frequência de ensaios clínicos potencialmente randomizados falso-positivos em periódicos de grande relevância científica.

2.2 Objetivo específico

Identificar variáveis na conclusão de ensaios clínicos randomizados que predisponham alto risco de erro do tipo I.

3 REVISÃO DE LITERATURA

3.1) Erros amostrais: sistemáticos e aleatórios

A estatística pode ser classificada em três categorias: descritiva, comparativa e inferencial[2]. A estatística inferencial é o tipo de estatística capaz de extrapolar os resultados observados de uma amostra para sua respectiva população alvo[6], sendo frequentemente utilizada para comparar diferenças entre grupos de tratamento em um estudo. Uma vez que é praticamente impossível realizarmos um estudo com toda a população, a maioria dos estudos é realizada com uma amostra de indivíduos retirada de uma população alvo definida e, em seguida, utilizamos a estatística inferencial para extrapolar seus resultados.

Como a amostra é inerentemente imprecisa, dois tipos de erros amostrais podem interferir na inferência estatística, ou seja, a diferença entre o padrão de dados da amostra em relação a população pode ocorrer por: um erro sistemático (afastamento da verdade devido ao viés) ou um erro aleatório (afastamento da verdade devido ao acaso), que é dividido em erro do tipo I e erro do tipo II. O erro sistemático é uma fonte de variação que distorce as aferições do estudo para uma determinada direção, enquanto o erro aleatório é uma fonte de variação com probabilidade igual de distorcer as aferições do estudo em ambas as direções[3]. O erro sistemático pode ser evitado através de uma metodologia correta, no entanto, o erro aleatório não pode ser evitado, pode-se apenas diminuir o seu impacto.

3.2) Erros aleatórios

Mesmo se o erro sistemático for excluído, a amostra ainda está sujeita ao erro aleatório, pelos desígnios do acaso. Por isso, o pesquisador deve sempre se perguntar: qual a probabilidade de observar esse padrão de dados por puro acaso? Existem duas formas de estimar o papel do acaso na nossa amostra: pela avaliação da hipótese (podendo ser realizada através do método de Fisher, método de Neyman-Pearson e método bayesiano) ou pela estimativa[7]. O método de Fisher está relacionado ao famoso valor de p, e o método de Neyman-Peason ao teste de hipóteses. Apesar de serem métodos diferentes, os pesquisadores costumam utilizá-los de forma híbrida: o teste de hipóteses durante o delineamento e o valor de p para testar a hipótese nula do estudo[8].

Nesse sentido, o valor de p, chamado de valor de probabilidade, foi introduzido por Karl Pearson em 1900, mas foi popularizado por Ronald Fisher, em 1925. Segundo este, o valor de p é definido como a probabilidade de se obter um resultado igual ou mais extremo que o observado, dado que a hipótese nula é verdadeira[9], sendo esse resultado qualquer medida de associação entre duas variáveis[10]. Podemos dizer, então, que o valor de p reflete se os dados observados na amostra são compatíveis com a hipótese nula[8], sendo que quanto menor o valor p, mais improvável é essa hipótese nula testada e vice e versa. O valor de p é a medida de associação mais utilizada na pesquisa biomédica, segundo um estudo que avaliou a frequência com que ele relatado nos resumos e no texto completo dos artigos durante 25 anos[11].

Já o método de Neyman-Pearson está relacionado ao teste de hipóteses. Esse teste foi formalizado por Jerzy Neyman e Egon Pearson em 1930, ao introduzir o conceito de hipótese alternativa, valor alfa, valor beta e de erro do tipo I e do tipo II[12]. Assim, para teste de hipóteses, existem quatro possibilidades da relação entre o resultado do teste e a realidade. Duas dessas quatro possibilidades levam a conclusões corretas: a) rejeita-se a hipótese nula no estudo e no mundo real ela é falsa e b) aceita-se a hipótese nula no estudo e no mundo real ela é verdadeira. Porém, as outras duas levam a conclusões erradas: c) erro tipo I e d) erro tipo II[13].

O erro do tipo I, ou falso-positivo, corresponde a rejeição incorreta da hipótese nula (rejeita-se a hipótese nula no estudo, quando na realidade ela é verdadeira). Já o erro do tipo II, ou falso negativo, corresponde a aceitação incorreta da hipótese nula (aceita-se a hipótese nula no estudo, quando na realidade ela é falsa)[8]. A probabilidade máxima de cometer um erro do tipo I é denominado valor alfa (ou nível de significância estatística) e, convencionalmente, o valor atribuído a ele é de 5%. Já a probabilidade máxima de cometer um erro do tipo II é denominado valor beta e, convencionalmente, o valor atribuído a ele é de 0,2. E, por fim, a probabilidade máxima de rejeitar a hipótese nula quando ela é falsa, ou seja, o complemento de beta ($1 - \beta$), é denominado poder estatístico.

Mas, qual a diferença entre o valor alfa e o valor de p? O valor alfa, segundo o teste de hipóteses, é o valor limite contra o qual medimos o valor de p durante a análise dos dados – ou seja, comparamos o valor de p com o valor alfa. Considera-se, geralmente, que um valor de $p < 0,05$ é suficiente para rejeitar a hipótese nula

(resultado estatisticamente significante) e um valor de p > 0,05 é insuficiente para rejeitar a hipótese nula (resultado estatisticamente não significante)[2].

3.4) Erro do tipo I e as múltiplas comparações

Em geral, os erros do tipo I são mais sérios do que os do tipo II – consideramos que ver um efeito quando não há um é pior que não ver um efeito em um estudo[2]. Existem diversos fatores que influenciam e aumentam a probabilidade do erro do tipo I em um estudo, sendo os principais aqueles relacionados ao problema das múltiplas comparações (multiplicidade). Nesse sentido, idealmente, os estudos devem apresentar um pequeno número de testes de significância estatística, uma vez que quanto maior o número de testes, maior a probabilidade de que pelo menos um ou mais destes testes rejeite a hipótese nula devido ao acaso, ou seja, a probabilidade de erro tipo I aumenta conforme o número de comparações feitas, levando a conclusões falsas[1]. Entre as diversas situações que envolvem as múltiplas comparações, as principais são: a) supervalorização de desfechos secundário; b) análises de subgrupo e c) análises interinas[14,15]

Durante o delineamento do estudo, os pesquisadores devem especificar o desfecho primário do estudo (aquele com o qual mais se preocupam) e os desfechos secundários. Esses desfechos devem ser pré-especificados no protocolo do estudo. O problema da multiplicidade pode surgir quando pesquisadores supervalorizam tais desfechos secundários – o que aumenta a probabilidade de encontrar um resultado positivo pelo puro acaso.

Existem diversas formas de supervalorização de desfechos secundários, entre elas: a dragagem de dados (em que pesquisadores analisam múltiplos desfechos secundários e só reportam os resultados estatisticamente significantes), mudança de desfechos primário (quando os pesquisadores mudam o desfecho primário entre o desenho do estudo e a publicação[13]), fenômeno de spin (que pode ser definido como qualquer ato que deturpa os resultados do estudo no intuito de destacar/enfatizar o efeito benéfico de uma determinada intervenção/procedimento ou mesmo implicar um efeito benéfico falso[16], seja nos métodos, resultados ou conclusão do estudo[17]) e, comparações múltiplas para uma única variável de resultado (o que ocorre em estudos multi-braços), dentre outras.

Nesse sentido, um estudo de 2004 identificou que de 102 ensaios clínicos, 62% deles tiverem pelo menos 1 desfecho primário foi alterado, introduzido ou omitido[18]. Outro estudo de 2010 avaliou a natureza e a frequência de spin em ensaios clínicos com resultados negativos de desfechos primários - 69 de 72 apresentavam spin[19]. Já em 2014, um estudo avaliou ensaios multi-braços publicados em quatro revistas de alto impacto e descobriu que entre 39, apenas 46% realizaram ajustes de multiplicidade[20].

As análises de subgrupo podem ser definidas como qualquer avaliação de efeito de tratamento para um específico subgrupo do estudo, a fim de identificar alguma modificação de efeito (chamada de fenômeno de interação ou heterogeneidade) em determinada característica da população, como sexo, idade ou presença de fator de risco[3]. Essas análises podem fornecer informações úteis (quando justificadas, planejadas, reportadas e interpretadas da forma correta), porém, devem ser desencorajadas - tanto porque aumentam a probabilidade de um resultado falso-positivo, quanto por conta da raridade da modificação de efeito.

Em um estudo de 2000, foram revisados 50 relatórios de ensaios clínicos de revistas de alto impacto - 70% deles relataram análises de subgrupo, sendo que 40% destes fizeram pelo menos 6 análises, menos da metade realizaram os devidos testes de interação a maioria não ofereceu informações sobre se as análises foram predefinidas ou post hoc[21]. Outro estudo de 2006, revisou 63 relatórios de ensaios clínicos cardiovasculares – 39 relataram uma análise de subgrupo e 29 relataram no mínimo 5, apenas 11 realizaram testes de interação e apenas 14 foram predefinidos3. Por fim, um estudo de 2007 revisou 97 ensaios clínicos de uma única revista de alto impacto, e encontrou os mesmos problemas[22].

As análises interinas são análises provisórias feitas durante a realização do estudo. A ideia da análise interina é examinar os resultados à medida em que os dados se acumulam, a fim de mudar o delineamento do estudo, como por exemplo, interrompê-lo[23]. As análises e a decisão de interromper o estudo devem ser feitas por um grupo independente de pessoas, chamado de Comitê Independente de Monitoramento e Segurança (DSMB) e procedimentos estatísticos para lidar com o problema da multiplicidade devem ser utilizados, como o Haybittle-Peto, O'Brien-Fleming e The Pocock. Apesar desses cuidados, uma revisão sistemática de 2005 avaliou a epidemiologia e a qualidade dos relatórios de 143 ensaios clínicos

randomizados (ECR) que foram encerrados por benefício aparente - 92 de 143 falharam em reportar as principais informações devidas (número e intervalo de análises interinas planejadas, como chegaram à decisão de interromper e os procedimentos estatísticos utilizados)[20].

Existem quatro motivos principais que levam à interrupção precoce de um ensaio clínico: 1) se o ensaio mostrar efeitos adversos graves; 2) se novas informações surgirem durante a realização do estudo, respondendo à pergunta principal ou levantando problemas de segurança; 3) quando o ensaio é interrompido depois de análise de futilidade e 4) ao encontrar um benefício aparente (estudo truncado)[24]. No entanto, a truncagem de estudos é problemática, não só porque aumenta a probabilidade de encontrar um resultado falso positivo, mas porque superestima a magnitude do efeito testado. Essa superestimação da magnitude de efeito foi comprovada em uma revisão sistemática de 2010 que comparou 91 estudos truncados com seus respectivos equivalentes não truncados e encontrou que o risco relativo dos truncados foi 29% menos do que o risco relativo dos não truncados[25].

4 METODOLOGIA

4.1 Desenho do estudo

Trata-se de um estudo metacientífico, do tipo primário, individuado, observacional, prospectivo e descritivo.

4.2 Amostra

Foram incluídos ensaios clínicos publicados em 2020 e 2021 em periódicos de grande relevância científica, sendo eles, o British Medical Association (BMJ), Journal of the American Medical Association (JAMA), LANCET e New England Journal of Medicine (NEJM). A inclusão dos artigos foi realizada por duas pesquisadores de forma independente através da seleção manual no site de cada periódico. Em seguida, os ensaios clínicos foram reservados em uma pasta no Mendeley e depois transferidos para o Rayyan, onde foram aplicados os critérios de exclusão. Depois, os ECs coletados foram classificados em positivos e negativos de acordo com a conclusão de cada autor e somente os positivos foram analisados quanto ao erro do tipo I. Além disso, os desacordos foram resolvidos por consenso ou julgados por um terceiro pesquisador.

4.3 Critérios de inclusão e de exclusão

- Critérios de inclusão: ensaios clínicos presentes nos periódicos BMJ, JAMA, LANCET e NEJM no período de 2020-2021;

- Critérios de exclusão: ensaios clínicos negativos, não randomizados, pragmáticos, de não inferioridade, de equivalência e ensaios clínicos duplicados.

4.4 Instrumentos utilizados:

- Mendeley: utilizado para transferir os ensaios clínicos para o rayyan, bem como para o gerenciamento de referências utilizadas;

- Rayyan: utilizado para aplicação dos critérios de exclusão;

- Excel: utilizado para o armazenamento e gerenciamento dos dados coletados.

4.5 Variáveis de caracterização

- Especialidade da área médica: variável qualitativa nominal politômica, com base no ramo da área médica que o trabalho aborda;
- Presença de financiamento: variável qualitativa dicotômica, por meio da análise da declaração de financiamento pelos autores;
- Estudo multicêntrico: variável dicotômica, coletada com a base na quantidade de centros de saúde em que a pesquisa foi realizada, devendo ser maior que um.
- Tipo de intervenção: variável qualitativa nominal politômica, de acordo com a estratégia utilizada no grupo tratamento.

4.6 Variáveis principais

O ensaio clínico foi considerado como potencialmente falso positivo se tiver a presença da alta probabilidade de erro do tipo I na conclusão do artigo, variável dicotômica definida como a presença de pelo menos uma das seguintes variáveis na conclusão:

- Presença da análise de desfecho secundário, categorizada em “sim” ou “não” para presença de conclusão positiva baseada em desfecho secundário ou quando o artigo não define o desfecho primário e o secundário;

- Presença da análise de subgrupo, categorizada em “sim” ou “não” para presença de conclusão positiva baseada em análise de subgrupo;

- Presença da análise interina, categorizada em “sim” ou “não” para presença de conclusão positiva baseada em análise interina.

4.6 Variáveis exploratórias:

- Tipo de conclusão, variável qualitativa nominal – se refere à maneira como o autor lida com o erro do tipo I na conclusão: exploratória (autor reconhece o potencial erro do tipo I devido à múltipla comparação) ou confirmátoria (autor não reconhece o erro do tipo I e confirma o resultado de forma prejudical);

- Positivação pós desfecho primário negativo - categorizada em “sim” ou “não” para presença de conclusão positiva baseada em desfecho secundário pós resultado do

desfecho primário negativo;

- Justificativa para análise interina, variável qualitativa nominal – se refere à justificativa do autor ter concluído a análise interina. O autor pode pausar o estudo por superioridade (eficácia), apresentar um resultado positivo preliminar positivo antes do estudo acabar, pausar um estudo por segurança ou ainda pausar o estudo devido ao advento de outro estudo que respondeu o seu objetivo.

- Ausência do comitê de monitoramento, categorizada em “sim” ou “não” para ausência do comitê de monitoramento (responsável por gerenciar as análises interinas do ensaio clínico)

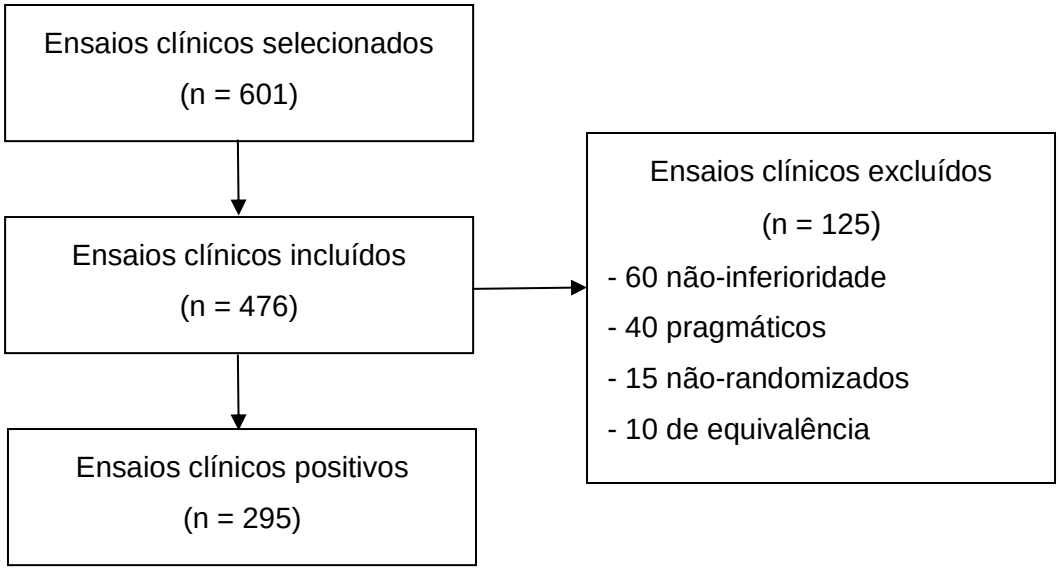
4.6 Aspectos éticos

Nesse estudo metacientífico, ou seja, que tem artigos como objeto de pesquisa, por não incluir seres humanos, aspectos éticos não são aplicados.

5 RESULTADOS

A busca inicial nos periódicos BMJ, JAMA, LANCET e NEJM entre os anos de 2020 e 2021, resultou em 601 ensaios clínicos. Entretanto, 125 artigos foram excluídos por terem critérios de exclusão, resultando em 476 ensaios clínicos incluídos. Destes, 295 apresentaram uma conclusão positiva e foram incluídos neste estudo (figura 1).

Figura 1 - (Seleção dos ensaios clínicos randomizado positivos)



Fonte: elaborado pelos autores, 2022.

5.1 Caracterização dos ensaios clínicos positivos

A maioria dos ensaios clínicos positivos encontrados foram da área de oncologia (18%), infectologia (17%) e cardiologia (15%). Um total de 223 (76%) dos estudos foram financiados por empresas privadas e trazer o número (60%) multicêntricos. Ademais, 225 (77%) deles analisaram os efeitos de intervenções medicamentosas, 57 (19%) procedimentos e 13 (4%) cirurgias **(Tabela 1).**

Tabela 1 - Características dos EC publicados positivos publicados na BMJ, JAMA, LANCET e NEJM entre os anos 2020 e 2021 (n = 295).

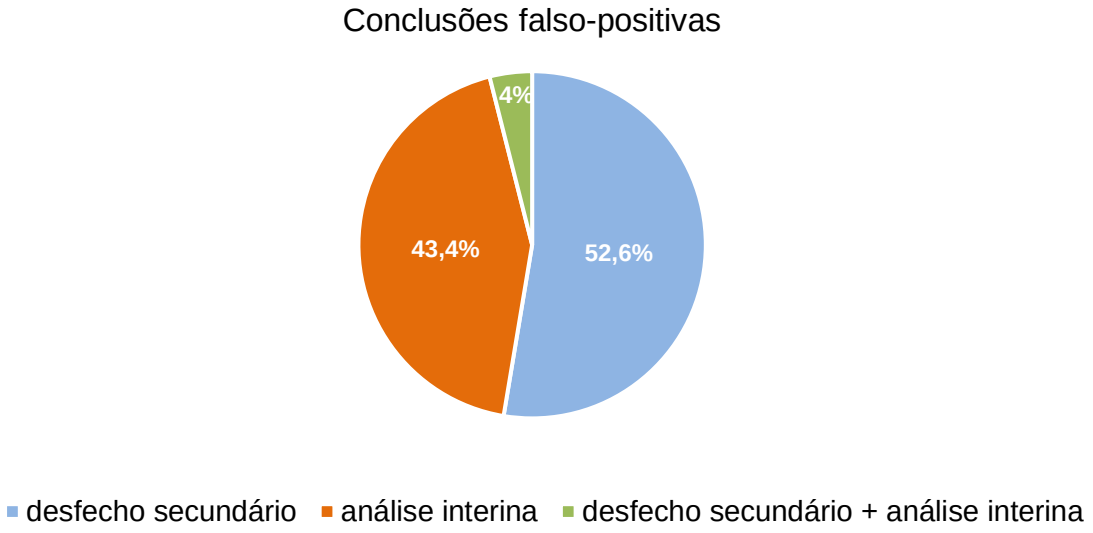
Variável analisada	Valor encontrado
Especialidade	
Oncologia	53 (18%)
Infectologia	50 (17%)
Cardiologia	44 (15%)
Outros	148 (50%)
Tipo de intervenção	
Medicamento	225 (77%)
Procedimento	57 (19%)
Cirurgia	13 (4%)
Presença de financiamento	223 (76%)
Estudo multicêntrico	177 (60%)

Fonte: elaborado pelos autores, 2022.

5.2 ECs potencialmente falso-positivos e tipos de múltipla comparação

Dos 295 ensaios clínicos positivos incluídos para análise, foram encontrados 76 (25%) ECs potencialmente falso-positivos. Destes, 40 (52,6%) concluíram por meio de análises de desfechos secundários, 33 (43,4%) por análises interinas, e 3 (4%) tanto por análises de desfechos secundários quanto análises interinas. **(Gráfico 1).**

Gráfico 1 – Tipo de múltipla comparação encontrada nas conclusões dos ensaios clínicos falso-positivos (n = 76).



Fonte: elaborado pelos autores, 2022.

5.3 Análise detalhada dos tipos de múltiplas comparações

Em relação aos 40 EC que utilizaram a estratégia de análises de desfechos secundários, 14 (35%) utilizaram a análise de desfechos secundários após a constatação de desfechos primários negativos e um deles (2%) mudou o desfecho primário durante o estudo.

Quanto aos 33 EC que concluíram análises interinas, 20 (61%) interromperam o estudo por superioridade, 7 (21%) por resultados preliminares positivos, 3 (9%) por segurança e 3 (9%) por advento de publicação de estudos com a resposta à pergunta do estudo (9%). Além disso, todas essas análises interinas foram feitas através de um comitê de monitoramento.

Os que concluíram tanto por meio de análises interinas como desfechos secundários, 2 (70%) pararam por superioridade e 1 (30%) foi um resultado preliminar. Além disso todos tiveram a presença do comitê de monitoramento e nenhum utilizou a análise de desfechos secundários após a constatação de desfechos primários

negativos.

Por fim, de uma maneira geral, dos 76 EC potencialmente falso-positivos, 62 (82%) concluíram seus resultados de forma confirmatória e 14 (18%) de forma exploratória.

6 DISCUSSÃO

O presente estudo teve como objetivo examinar a frequência de ensaios clínicos randomizados potencialmente falso-positivos presentes em periódicos de grande relevância científica da área de saúde, no período de 2020-2021. Foi encontrada uma frequência de 76 ensaios clínicos potencialmente falso-positivos (25%) e a análise detalhada dos tipos de múltipla comparação indicou má prática em pesquisa.

Foi possível observar, portanto, que, apesar da sua invisibilidade, o erro aleatório, assim está contribuindo para a já preocupante diminuição da confiabilidade dos ensaios clínicos da área da saúde. Essa alta proporção de falso-positivos pode estar relacionada ao viés de publicação, que se configura como uma tendência por parte dos períodicos e pesquisadores em publicar evidências científicas positivas ao invés daquelas negativas e, dessa forma, isso pode estimular a utilização de más práticas na intenção de positivar, intencionalmente ou não, os resultados. E de fato, de acordo com o epidemiologista John Ioannidis, muitas publicações do ecossistema científico são falsas, refutadas por provas posteriores. Todo esse cenário, infelizmente, tem consequências negativas para a prática clínica e para os pacientes.

Quanto à supervalorização de desfechos secundários, sabe-se que o desfecho primário é o guia para o delineamento do estudo e as conclusões devem girar em torno dele, porém, da presente amostra de ensaios clínicos falso-positivos, a maioria deles concluiu desfechos secundários positivos. Sabendo, então, que o conceito de desfechos não está claro para os pesquisadores e periódicos, faz-se necessário reforçar a função do desfecho secundário, que é a de compreender a veracidade do desfecho primário.

Houve também uma grande quantidade de ensaios clínicos que supervalorizaram análises interinas, sendo que a maioria deles pausou o estudo por superioridade e alguns artigos pausaram o estudo por segurança. Sabe-se que estudos truncados podem ser perigosos porque podem superestimar os efeitos do tratamento. Assim, conclui que a interrupção precoce de um estudo por aparente superioridade de um tratamento deve ser evitada.

7 CONCLUSÃO

Foi realizado um estudo metacientífico descritivo para analisar a frequência de ensaios clínicos randomizados falso-positivos presentes em periódicos de grande relevância científica da área de saúde, no período de 2020-2021. Foi encontrada uma notória proporção de ensaios clínicos randomizados com resultados potencialmente falso-positivos, ou seja, com conclusões que supervalorizam múltiplas comparações. Mais estudos na área meta-científica são necessários para aprofundar a evidência nesse âmbito.

REFERÊNCIAS

1. Last JM. A dictionary of epidemiology. Vol. 15, International Journal of Epidemiology. 1986. 277 p.
2. H. Fletcher R, Fletcher SW, Fletcher GS. Essentials of Clinical Epidemiology. 2014. 273 p.
3. CUMMINGS SR, BROWNER WS, GRADY DG, NEWMAN TB. Delineando a Pesquisa Clínica.
4. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340(7748):698–702.
5. Ioannidis JPA. Why most published research findings are false. Get to Good Res Integr Biomed Sci. 2018;2(8):2–8.
6. Moyaho A, Beristain-Castillo E. Experimental design: Basic concepts. Encycl Anim Behav. 2019;471–9.
7. Pek J, Van Zandt T. Frequentist and Bayesian approaches to data analysis: Evaluation and estimation. Psychol Learn Teach. 2020;19(1):21–35.
8. Mark DB, Lee KL, Harrell FE. Understanding the role of P values and hypothesis tests in clinical research. JAMA Cardiol. 2016;1(9):1048–54.
9. Petrie A. Medical Statistics Fourth Edition. 211 p.
10. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: An explanation for new researchers. Clin Orthop Relat Res. 2010;468(3):885–92.
11. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting P values in the biomedical literature, 1990-2015. JAMA - J Am Med Assoc. 2016;315(11):1141–8.
12. Kyriacou DN. The enduring evolution of the P value. JAMA - J Am Med Assoc. 2016;315(11):1113–5.
13. Motulsky H. Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking [Internet]. 1995. 284–290 p. Available from: http://www.intuitivebiostatistics.com/
14. Schulz KF, Grimes DA. Multiplicity in randomised trials II: Subgroup and interim analyses. Lancet. 2005;365(9471):1657–61.
15. Schulz KF, Grimes DA. Multiplicity in randomised trials I: Endpoints and treatments. Lancet. 2005;365(9470):1591–5.
16. Alves CL, Costa GG da, Segundo J de DB, Helal L. Spin: modificações na redação científica que escondem fragilidades metodológicas com impacto social negativo. J Evidence-Based Healthc. 2020;2(1):97–105.
17. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A. 2018;115(11):2613–9.
18. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. J Am Med Assoc. 2004;291(20):2457–65.
19. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA - J Am Med Assoc. 2010;303(20):2058–64.
20. Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. J Clin Epidemiol. 2008;61(3):241–6.

21. Louisa M, Sugiarti L, Kurniawan SV, Wanandi SI. Subgroup analysis and other (mis)uses of baseline data in clinical trials Susan. Adv Sci Lett. 2017;23(7):6838–40.
22. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. N Engl J Med. 2007;357(21):2189–94.
23. Illigens FF and BMW. Critical Thinking in Clinical Research.
24. Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects. Jama. 2010;303(12):1180–7.
25. Mueller PS et al. Annals of Internal Medicine Ethical Issues in Stopping Randomized Trials Early Because of. Ann Intern Med. 2007;146:878–81.