保健医療科学　2025　Vol.74 No.3 p.313－326

＜解説＞

CHecklist for statistical Assessment of Medical Papers (the CHAMP statement)：詳細と解説

富樫慎太郎[1)]，野口泰司[2)]，小山田隼佑[3,4)]，山口拓洋[3,4)]，白岩健[1)]，福田敬[1)]

[1)] 国立保健医療科学院 保健医療経済評価研究センター
[2)] 国立長寿医療研究センター研究所 老年学・社会科学研究センター 老年社会科学研究部
[3)] 東北大学大学院医学系研究科 医学統計学分野
[4)] 東北大学病院 臨床試験データセンター

CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration

TOGASHI Shintaro[1)], NOGUCHI Taiji[2)], OYAMADA Shunsuke[3,4)], YAMAGUCHI Takuhiro[3,4)], SHIROIWA Takeru[1)], FUKUDA Takashi[1)]

[1)] Center for Outcomes Research and Economic Evaluation for Health, National Institute of Public Health
[2)] Department of Social Science, Center for Gerontology and Social Science, Research Institute, National Center for Geriatrics and Gerontology
[3)] Division of Biostatistics, Tohoku University Graduate School of Medicine
[4)] Clinical Research Data Center, Tohoku University Hospital

抄録

医学およびスポーツ科学研究における統計の誤った使用は一般的であり，ヘルスケアに有害な結果をもたらす可能性がある．医学論文の多くの著者，編集委員，査読者は，統計に関する専門知識を有しておらず，もしくは医学研究に正しい統計を適用することの重要性を十分に理解していない可能性がある．医学論文における統計報告に関するガイドラインは存在する一方で，論文を査読する際に評価するための，より一般的かつ共通して見られる統計的側面に関するチェックリストが必要とされている．本稿では，研究の計画と実施，データ解析，報告とプレゼンテーション，解釈に関連する30項目で構成される，「医学論文の統計評価のためのチェックリスト（CHecklist for statistical Assessment of Medical Papers, CHAMP）」を提案する．CHAMPの主な目的は，医学論文の統計評価の際に編集委員と査読者を支援することとしているが，医学研究における統計の利用に関する著者や読者のプラクティスを改善するために役立つ参考資料になると考えている．編集委員と査読者には，投稿論文を評価する際にCHAMPを参照することを強く推奨する．また著者も，医学研究の統計手法と報告の妥当性を担保するためにCHAMPを活用することができ，読者は手元の論文の統計評価を強化するためにCHAMPを利用可能である．

Abstract

Misuse of statistics in medical and sports science research is common and may lead to detrimental consequences to healthcare. Many authors, editors and peer reviewers of medical papers will not have expert knowledge of statistics or may be unconvinced about the importance of applying correct statistics in medical research. Although there are guidelines on reporting statistics in medical papers, a checklist on the more general and commonly seen aspects of statistics to assess when peer-reviewing an article is needed. In this

連絡先：富樫慎太郎
〒351-0197　埼玉県和光市南2-3-6
2-3-6 Minami, Wako, Saitama 351-0197, Japan.
Tel: 048-458-6142/ E-mail: togashishintaro.42@gmail.com
[令和7年2月25日受理]

article, we propose a CHecklist for statistical Assessment of Medical Papers (CHAMP) comprising 30 items related to the design and conduct, data analysis, reporting and presentation, and interpretation of a research paper. While CHAMP is primarily aimed at editors and peer reviewers during the statistical assessment of a medical paper, we believe it will serve as a useful reference to improve authors' and readers' practice in their use of statistics in medical research. We strongly encourage editors and peer reviewers to consult CHAMP when assessing manuscripts for potential publication. Authors also may apply CHAMP to ensure the validity of their statistical approach and reporting of medical research, and readers may consider using CHAMP to enhance their statistical assessment of a paper.

(accepted for publication, February 25, 2025)

〈訳者解説〉

本資料は，BMJ Publishing Group が発行する British Journal of Sports Medicine誌のCHAMP statementの日本語訳版である．CHAMP statementは，論文査読における統計的側面を総合的に評価するためのチェックリストであり，1986 年にBMJ誌に掲載された投稿論文評価用チェックリストを基に，幅広い文献レビューと複数の統計専門家の経験を集約して大幅に改訂されたものである．本資料の原版は，スポーツ科学誌に掲載されているものの，その適用範囲は特定の分野に限定されるものではない．また，統計の非専門家が論文査読者として活用することを主な目的としているが，研究実施者，論文執筆者，および読者が統計手法を適切に理解し活用する際の参考資料としても有用である．これらのことから，我が国の研究者にとってもその内容を把握し，活用することは査読者のみならず研究実施者の視点でも非常に重要だと思われる．

翻訳について，可能な限り平易な日本語となるように努めたつもりである．ただし，なるべく原版を尊重するように努力したため，翻訳調になっているところや英語のまま記載しているところも多く，それらは訳者らの能力不足に由来する．訳された日本語の文章に訳者らの「解釈」が含まれることから，我々の解釈が読者の理解を促すために有用であればよいが，そうでないと感じた場合は必ず原版および参考文献を参照することをお願いしたい．

なお，CHAMP statementを作成した著者らは，Lancet誌やBMJ誌といった影響力の高い医学雑誌の統計専門編集委員を務め，数多くの論文の統計的側面の査読経験を有している．加えて，3 名の著名な統計専門家を外部専門家として招き，その専門的知見をCHAMP statementに反映している．さらに，2024 年 2 月 17 日号のLancet誌には，CHAMP statementの筆頭著者による「医学研究における統計報告に関する 10 項目のガイダンス」（Lancet. 2024;403(10427):611–612）が掲載された．本ガイダンスでは，医学研究でよく遭遇する統計的な不備とそれを回避するための簡単な改善策が示されており，本文書と併せて参照されることをお勧めする．

〈訳者謝辞〉

本翻訳実施にあたってご協力いただいたMohammad Ali Mansournia教授（Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences）およびBMJ Publishing Groupに感謝する．

医学論文の統計評価のためのチェックリスト（CHecklist for statistical Assessment of Medical Papers, CHAMP）

医学およびスポーツ科学研究において，統計の誤った使用や誤った方法論の適用は，信頼性に欠ける，もしくは誤った結論につながる可能性がある．不備のある方法論によって得られた結果は，公衆衛生，患者管理，アスリートのパフォーマンスに望ましくない影響を及ぼす可能性がある[1]．残念ながら，研究計画，統計解析，結果の報告と解釈における誤りは医学雑誌では共通してみられる問題であり[2, 3]，医学論文の質に疑問を生じさせる[4]．

過去数十年間で，特にインパクトファクターの高い雑誌では，適切な方法論が優先されてきた．これは，査読プロセスに，より多くの統計専門の編集委員や他の方法論の専門家（疫学者など）が関与するようになったことから示されている．加えて，調査・研究のステークホルダーは，著者や査読者の教育，オンラインリソースの提供，ガイドラインの作成（および拡張），定例の学術集会での方法論に関するコンテンツの提供など，統計学，疫学，方法論の教育への投資を強化するよう促されてきた[5]．また，報告ガイドライン（例えば，CONSORT，STROBE，STARD，REMARK，TRIPODなど）の遵守にも重点が置かれるようになった[6–10]．

そういった状況にも関わらず，多くの医学およびスポーツ科学誌では，査読プロセスに統計専門家が関与していない．これは残念なことである．なぜなら著者，編集委員，査読者の統計知識が不十分であったり，医学研究における正しい統計の重要性に関する理解に乏しかったりする場合，基本的な統計の誤りが起こりやすいからである．特に，臨床系雑誌において投稿論文の統計の利用が体系的に評価されることはめったにない[11, 12]．したがって，論文が科学雑誌に掲載された後でも，統計的なデザインや解析が適切であるかどうかに加えて，結論が正当化されるかどうかについて，細心の注意を払

表1　医学論文の統計評価のためのチェックリスト（CHAMP）

項目			
研究計画と実施			
1. 研究のゴール、研究目的、研究デザイン、研究対象者の明確な記述	Yes	Unclear	No
2. アウトカム、曝露/治療と共変量、およびそれらの測定方法の明確な記述	Yes	Unclear	No
3. 研究デザインの妥当性	Yes	Unclear	No
4. サンプルサイズについての明確な説明と正当性	Yes	Unclear	No
5. 事前に計画された研究計画の違反(design violations)の報告と受け入れの程度	Yes	Unclear	No
6. 論文と既報のプロトコルの一貫性	Yes	Unclear	No
データ解析			
7. 統計手法の正確かつ完全な説明	Yes	Unclear	No
8. 妥当な統計手法の適用と仮定の明示	Yes	Unclear	No
9. 治療効果または治療と他の共変量との交互作用の適切な評価	Yes	Unclear	No
10. 相関および関連性に関する統計検定の正しい使用	Yes	Unclear	No
11. 連続変数の適切な取り扱い	Yes	Unclear	No
12. 信頼区間にとりえない値の有無	Yes	Unclear	No
13. RCT におけるベースライン時点の背景の適切な群間比較	Yes	Unclear	No
14. 交絡の正確な評価と調整	Yes	Unclear	No
15. データによって裏付けられていないモデルの外挿が避けられている	Yes	Unclear	No
16. 欠測データの適切な取り扱い	Yes	Unclear	No
報告とプレゼンテーション			
17. データの適切かつ正確な記述	Yes	Unclear	No
18. 記述的結果として発生(発症)指標・信頼区間の提示と、分析的結果として関連指標・信頼区間・p 値の提示	Yes	Unclear	No
19. 各群の信頼区間を個別に提示されるのではなく、群間の比較/対比の信頼区間が提示されている	Yes	Unclear	No
20. 結果の選択的報告と p-hacking が避けられている	Yes	Unclear	No
21. 効果量、検定統計量、p 値について、適切で一貫した正確な数値の報告	Yes	Unclear	No
22. 今後実施される可能性がある、メタ解析に含めることができる十分な数値結果の提示	Yes	Unclear	No
23. 図表の適切な提示	Yes	Unclear	No
解釈			
24. p 値とともに関連指標と 95%信頼区間に基づいて結果が解釈され、p 値が大きいことを「効果がない証拠」ではなく「決定的でない結果」として正しく解釈されている	Yes	Unclear	No
25. 研究結果の解釈において事後的な検出力分析ではなく信頼区間が使用されている	Yes	Unclear	No
26. 発生(発症)または関連の指標が正しく解釈されている	Yes	Unclear	No
27. 因果を関連と相関から区別されている	Yes	Unclear	No
28. 解釈において、事前に規定された解析の結果と探索的解析の結果が区別されている	Yes	Unclear	No
29. 研究方法論の限界についての適切な議論	Yes	Unclear	No
30. 統計解析によって裏付けられた結論のみを導き出し、標的集団以外の対象に対して結果を一般化していない	Yes	Unclear	No

い，論文の内容を慎重に読む必要がある．高ランクの雑誌に掲載された研究報告も，査読プロセスでは特定されなかった方法論的または統計的な不備から免れているわけではない．一部の雑誌では，査読プロセスで統計家を起用する（統計専門の査読者や編集委員として）ことによって，このような問題を軽減しようとしている一方で，専門の統計査読者を起用できない場合，科学論文の方法論的または統計的内容を評価するためのガイドラインが役立つだろう[5, 13, 14].

医学論文における統計の報告方法に関するガイドラインは存在する[15, 16]が，我々は査読プロセスにおいて投稿論文の統計的側面を判断するための総合的なチェックリストを提案する．すべてを網羅することは不可能である一方で，医学およびスポーツ科学の研究論文で使用される統計手法を，より広範に評価するための基本的なチェックリストがあると便利であると考える．そこで，既存のチェックリスト[17]を大幅に改訂した上で，我々は，投稿論文の査読を支援するために，計画あるいはデザイン，解析，報告，解釈の段階における 30 項目から構成される「医学論文の統計評価のためのチェックリスト（CHecklist for statistical Assessment of Medical Papers, CHAMP; 表1）」について記述する[18].

チェックリストの開発と説明

チェックリストの 30 項目は，BMJ誌投稿論文評価用の既存チェックリスト[17]，幅広い文献レビュー，および様々な医学雑誌に投稿された多数の論文の統計的内容を査読した著者らの集団での経験に基づいて選択された．（原版の）筆頭著者がチェックリストの草案を作成し，共著者が項目の追加もしくは削除を提案し，全著者が最終版を承認した．外部専門家からは本資料に対する幅広いコメントを受けた．その専門家らは謝辞に記載されている．我々のチェックリストは，医学統計の全ての側面を網羅することを意図しておらず，また網羅することもできていない．むしろ，臨床研究で共通して直面する重要な問題に焦点を当てている．したがって，研究論文の査読中によく遭遇する統計的課題のみがCHAMPに含まれた．我々のチェックリストを活用するには，統計に関する基礎知識が求められる．しかしながら，各項目について簡単な説明を備えており，詳細については関連する参考文献を引用している．最初の 6 項目は研究計画と実施に関するもの，7–16 項目はデータ解析に関するもの，17–23 項目は報告とプレゼンテーションに関するもの，24–30 項目は結果の解釈に関するものである

1–6 項目：研究デザインと研究実施

1： 研究のゴール，研究目的，研究デザイン，研究対象者の明確な記述

研究のゴール，研究目的，研究デザイン，研究対象集団（study population）と標的集団（target population）は，明確に記述しなければならない．これは，編集委員と読み手が研究における内的妥当性と外的妥当性（一般化可能性）を判断するためである．

研究のゴールを明確にすることは，科学分野によらず，質の高い科学の前提である．例えば，研究のゴールは 3 つの分類がよく用いられている．1 つ目は記述，2 つ目は予測，3 つ目は因果推論である．予測は，アウトカムが発生するリスクが高い人は“誰”であるかを特定することに相当する．一方で，因果推論は“なぜ”アウトカムが発生するのかを説明する試みである（例えば，因果効果の検討）[5, 19].

研究目的は研究の背景理論を記述し，注目している特定の研究疑問を指し示す．例えば，Heated water-based exercise（HEx）trialというランダム化比較試験Randomized controlled trial（RCT）の目的は，治療抵抗性高血圧を有する患者を対象として，24 時間自由行動下血圧測定値に対する温水ベースでの運動療法の効果を検討すること[20]としている．研究目的は通常，背景（Introduction）セクション内で，理論的根拠（rationale）が記載された後に提示される．

研究デザインは研究の種類を指し，方法（Methods）セクションに記述する[21]．一般的な研究デザインの例として，RCTや観察研究（コホート研究・症例対照研究・横断研究）がある[22]．研究デザインは，詳細に記載すべきである．特に，RCTにおけるランダム割付方法，コホートにおける追跡期間，症例対照研究における対照群の選択方法，横断研究におけるサンプリング方法は十分に説明する[6, 7]．原則として，他の研究者がその研究を正確に再現するために，研究デザインは十分に説明しなければならない．

研究対象集団は源泉集団（source population）のうちデータが実際に収集される集団を指す一方で，標的集団（target population）は研究結果を一般化しようとする集団を指す．これら 2 つの集団の関係性は，選択（組み入れ）基準と除外基準を用いて特徴付けられており，一般化可能性を検討するために重要である．HEx試験を例にすると，研究対象者は 5 年以上の治療抵抗性高血圧を有している 40–65 歳の人々に限定している[20]．臨床試験と観察研究の両方において重要なことは，原集団の何割が研究対象となったかを知ることである．例えば，原集団は，ある期間にある疾患で病院に入院した全ての患者を包含したが，解析用データセットはその集団の 50% であることが発生し得る．理由は様々であり，患者から同意を得られなかった，測定できなかった，および患者が脱落となった等のようなものが挙げられる．HE x 試験を例とすると，著者等は高血圧を有する患者 125 名をスクリーニングした結果，治療抵抗性高血圧の選択（組み入れ）基準に該当した 32 名を特定した．これらの情報は，いくつかの視点で重要な意味があり，具体的には研究の一般化可能性，温水ベースでの運動療法を誰に提供できるか，および臨床上の実践性にどの程度の影響があるか，が挙げられる．

2： アウトカム，曝露/治療と共変量，およびそれらの測定方法の明確な記述

統計解析で考慮される全ての変数は，アウトカム，曝露/治療，予測因子と潜在的な交絡因子/中間因子/効果修飾因子を含めて（Box 1），論文内に明確に記述する．加えて，それぞれの変数の測定方法と測定時点も明確に示す．もし研究のゴールが観察研究を用いて因果推論（“なぜ”アウトカムが発生するのかを説明する試み）を行うことである場合は，著者らが因果関係についての仮定を因果ダイアグラム（causal diagram）で示すべきである[23–25]．例示として，変形性関節症を有する患者を対象とした，身体活動による身体機能と膝関節痛に関する影響を評価したコホート研究を取り上げる[26]．曝露である身体活動は高齢者における身体活動スケールを用いて測定し，アウトカムである身体機能は20m歩行時間を用いて，自己報告型の膝関節痛の測定はWestern Ontario and McMaster Universities Osteoarthritis Indexを 用 いて，それぞれ測定した．抑うつ症状が潜在的な交絡変数であるとみなされ，うつ病自己評価尺度（Center for Epidemiologic Studies Depression Scale: CES-D）を用いて測定した．全ての変数はベースライン時および3回の追跡調査時に測定され，加えて，関心がある効果（effect）を推定するために，研究対象集団において想定されるcausal diagramに基づく因果推論の手法を用いた[26]．

3： 研究デザインの妥当性

研究デザインは，妥当であり，かつ研究結果にバイアスをもたらすことなく，研究疑問に合致しているべきである．症例対照研究の例では，対照群が症例の源泉集団（source population: 訳者注，ケースが生み出される集団）を十分に代表しているかどうかについて，編集委員が評価可能でなければならない．また臨床試験の例では，1つ（もしくはそれ以上）の対照群があるかどうかを明確にする．もし対照群がある場合は，患者が介入と対照へランダム割付されているかどうか，もしランダム割付であれば，ランダム割付の詳細な方法と割付の隠匿が適切かどうかを明確に記述する．

4： サンプルサイズについての明確な説明と正当性

サンプルサイズを明確に正当化するセクションを設ける[27]．サンプルサイズ計算が正当化される場合，サンプルサイズのセクションには，算出に使用した値の選択，算出の根拠となったアウトカム，臨床的に意味（意義）のある最小の効果量（minimum clinically important effect size）などに関する明確な根拠（参考文献による裏付け）とともに，推定値の再現が可能な程度に詳細に記述する[28, 29]．例えば，典型的なサンプルサイズ計算の目的は，以下の2点である．1つ目は発生（発症）指標（例えば，リスク）もしくは関連指標（例えば，リスク比）の推定について十分な精度[30, 31]が含まれていること，2つ目は統計的仮説検定を行う場合は真の効果（例えば，真の差）を検出するために十分な検出力を保証することである．ここで考慮すべき点は，脱落，追跡不能や無回答，デザイン効果（例えば，クラスタリングによるもの）である．予測モデルの開発と検証に関するサンプルサイズ計算に関するガイダンスは，すでに既報[32–34]がある．

5： 事前に計画された研究計画の違反（design violations）の報告と受け入れの程度

研究計画の違反（Design violations）は，研究実施において頻繁に発生する．例としては，質問紙調査における不返答，前向き研究における打ち切り（追跡不可もしくは競合リスク発生）[35]，およびRCTにおける介入不遵守が挙げられ，論文中に明確に報告されなければならない[36, 37]．研究デザインの妥当性の視点では，設計違反の受け入れの程度を評価する．評価の視点の例としては，「観測された無回答や打ち切りの割合はあまりにも高いか？」「データ欠測の理由は何であるか？」「これらの程度はその研究の科学的目標の到達のために受け入れられるかどうか？」が挙げられる．

6： 論文と既報のプロトコルの一貫性

査読者は，研究において重要な特徴（サンプルサイズ，主要・副次・探索的アウトカム，解析方法を含む）について，既報のプロトコル（関連する場合は登録情報）との矛盾を特定する．

7–16項目：データ解析

7： 統計手法の正確かつ完全な説明

論文内の方法セクションにおいて，統計手法は独立した項目として記述する．統計査読者が研究目的への適合性と完全性を判断できるように，記述的統計手法および分析的統計手法の両方を十分に記述する必要がある．

8： 妥当な統計手法の適用と仮定の明示

統計手法の妥当性は，いくつかの仮定に依存する．例えば，独立した2群の平均値の差の検定であるt検定には3つの仮定が求められる．これは，観測値の独立性，正規性，分散均一性[38]の3つである．もう1つの例は，χ^2検定であり，期待度数は全て1以上でなければならず，期待度数が5を下回るセルが20%以下であることも挙げられる．これらの統計的な仮定は，文脈に応じて判断されるか，もしくは正規確率プロットを用いて正規性の仮定を確認するなど，適切な方法を用いて評価するべきである[39]．もし，いくつかの仮定が明確に違反している場合は，代替となる統計的仮説検定を適用する．注意すべき点は，いくつかの統計的仮説検定定は仮定の軽度～中程度の違反に対して頑健であることである．t検定を例にすると，正規性の欠如や分散の不均一性は，必ずしもt検定を無効にするものではないが，一方でアウトカム変数の独立性の欠如はその結果が妥当ではないことを意味する[40]．独立したt検定は，順序変数（例えば，

0,1,2,3 のような値の変数）やサンプル数が 20 例である場合でも有効であるが，最適ではないことが分かっている[41].

推定されたオッズ比（OR），リスク比および率比のような比の推定値において，帰無仮説から離れる方向に偏った結果が導出されることは，実践の視点では重要だが，しばしば無視される．このバイアスはスパースデータバイアスとして知られており，スパースデータ（まばらなデータ）によって増幅される[42]．スパースデータのサインは非現実的な大きい比の点推定値と信頼限界であり，これらの結果はまばらなデータに起因して導出される．例えば，非伝染性疾患におけるOR>10は，スパースデータバイアスの警告サインである可能性を考慮すべきである．極端に言えば，空のセルは無限大の不合理なOR推定値につながりうる，これは分離（separation）として知られる事象である[43]．スパースデータバイアスを減少させるために，罰則化やベイズ法などの特別な統計手法を適用する必要がある[43, 44]．統計解析におけるその他の重要な検討事項は，(1) 相関が想定されるデータの分析における相関を考慮すること（例えば，縦断研究における繰り返し測定[45]，クラスターランダム化試験[46]，複雑な健康調査[47]）; (2) マッチされた症例対照研究やコホート研究におけるマッチングを考慮すること[48–50]; (3) 解析における複数の群に関する順序付けを考慮すること; (4) 生存時間データ解析における打ち切りを考慮すること; (5) RCTの解析におけるアウトカムのベースライン値を調整すること[28]; (6) 集団寄与分画（population attributable fraction）の正しい計算と解釈[51, 52]; (7) 予測モデルを作成する際に，縮小化と罰則化を用いて，オーバーフィッティングを調整すること[53, 54]; (8) ネットワークメタ解析において同質性と一貫性の仮定を評価すること[55].

9: 治療効果または治療と他の共変量との交互作用の適切な評価

治療効果と潜在的な交互作用を評価するために，適切な統計的仮説検定を使用する．各群の信頼区間が重なり合っているかどうかのみで評価することは誤解を招く可能性がある[56–58]．したがって，その治療群の信頼区間の比較を治療効果の統計的仮説検定の代替として使用すべきではない．さらに，共変量の各水準（例えば，男性と女性）での治療効果のp値を比較することは，治療と共変量の間の交互作用の検定の代替として使用すべきではない．例えば，男性でp値<0.05，女性でp値>0.05と観測された場合，性別が効果修飾因子であると誤って結論付けられる可能性がある[59]．同様に，サブグループの信頼区間が重なり合った場合，効果修飾がないと結論付けることはできない[60].

10: 相関および関連性に関する統計検定の正しい使用

相関および関連性に関する統計的仮説検定の誤用は頻繁にある．例えば，2 つ以上の測定方法を比較する研究において，相関を用いて 2 つの測定方法の一致性を評価するべきではない[61]．この理由として，もしXとYの 2 つの測定方法が完全に相関しているとしても，XがYの 2 倍である条件では，一致性は乏しいからである．同様に，平均値の差の検定（例えば，対応のあるt検定）を適用した時にp値は十分に大きい値になることから，2 つの測定方法が十分に一致しているとは推論することはできない．実際，差の分散が大きいことは一致性が低いことを示すが，対応のあるt検定の結果としてp値が大きくなる可能性を高め，その結果として 2 つの測定方法が一致しているに見えてしまう[1].

11: 連続変数の適切な取り扱い

査読者は，連続変数を二値化またはカテゴリー化している研究に注意すべきであり，こういった取り扱いは一般的に避けられるべきである[62]．連続変数を二値/カテゴリー化し，それをモデル内のカテゴリカル変数として使用すると，バイアス，統計的非効率性，残差交絡が生じる可能性がある．連続変数は連続変数のまま保持し，線形性の仮定が正しくない可能性がある場合は，その関数形式を検討する．連続変数型の予測因子を取り扱う分析アプローチとして，多項式の使用や回帰スプラインの適用が挙げられる[62–65].

12: 信頼区間にとりえない値の有無

有効な信頼区間はとりえない値を含まない．例えば，割合に対する単純なWald信頼区間（$p \pm 1.96\sqrt{p(1-p)/n}$）は，pが 0 または 1 に近い場合には妥当ではなく，割合の可能な範囲（$0 \leq p \leq 1$）を超える負の値を生じることがある[66]．このような状況に対応するために，Wilsonのスコア区間またはAgresti-Coull区間が適用されることがある[6].

13: RCTにおけるベースライン時点の背景の適切な群間比較

RCTにおいて，群間での背景の差異は，偶然（または未報告のバイアス）によるものである．p値を報告しても意味をもたないため，査読者はベースライン時の背景の統計的仮説検定定に注意する必要がある[67]．どの背景（予後因子）が調整に含まれるかの意思決定は，プロトコルで事前に計画され，p値ではなく研究課題に関する臨床知見に基づいて行う．ベースライン時点の背景における群間の差異は，差異の大きさによって特定され，結果の解釈に与える潜在的な影響の観点から議論されるべきである．

14: 交絡の正確な評価と調整

健康科学研究の重要なゴールの 1 つとして，因果関係を推論することが挙げられる．ここでの関心は，ある曝露がアウトカムに与える因果効果である．観察研究だけ

でなく，RCT（小規模から中規模のサンプルサイズ）を含む因果関係を評価するための研究を脅かす主なバイアスの原因は，交絡である[68–71]．交絡はデザイン段階（例えば，限定やマッチングを通じて）や解析段階（例えば，回帰モデル，標準化，または傾向スコア法を用いて）で制御することができる[72–74]．交絡因子の選択は，しばしば因果ダイアグラム（causal diagram）で表されるような因果関係に関する事前知識に基づくべきであり[23, 75–77]，p値（例えば，ステップワイズ法の使用）に基づくべきではない．ステップワイズ法のような自動化された統計手法では，交絡因子と，解析で調整すべきではない中間因子（mediators）や合流点（colider）などの他の共変量を区別できない．さらに，ステップワイズ法は交絡因子とアウトカムとの関連のみに基づいており，交絡因子と曝露との関連を無視している．したがって，ステップワイズ法は交絡因子の選択には使用すべきではない．実際には，多くの交絡因子（および曝露とアウトカム）は時間によって変化しており[78, 79]，時間依存性交絡を適切に調整するための因果推論の方法（causal methods）を適用するべきである[80, 81]．同様に，新しい変数の予後効果（prognostic effect）を評価する研究では，既存の予後因子に対する調整がルーチン的に行われるべきであり，既存の因子の変数選択は一般的には必要とされない[53].

15: データによって裏付けられていないモデルの外挿が避けられている

多くの健康科学研究において関心が向けられるゴールは，回帰モデルを用いて１つ以上の説明変数からアウトカムを予測することである．統計モデルは説明変数の観測されたデータの範囲内でのみ有効であり，その範囲外の人々に対して予測を行うことはできない．これはモデルの外挿として知られている[82]． コホート研究において，Body Mass Index（BMI）と血圧（BP）の間に，以下の式に基づく線形関係が見出されたと仮定する:

$$BP = A + B * (BMI)$$

この場合，切片Aは解釈できない．なぜなら，それはBMIがゼロの人の期待される血圧値に対応しているからである．解決策として，BMIを中心化し，中心化変数（BMI―平均BMI）をモデルに含めることが挙げられる．これによって，新しい切片は集団の平均BMIを持つ人の期待される血圧値を指すようになる．

もう一つの例では，RCTにおいて以下の線形関係が成り立つと仮定する:

$$BP = A + B * (TRT) + C * (BMI) + D * (TRT * BMI)$$

この場合，TRTは治療（1：介入，0：プラセボ）を表し，TRT*BMIは治療とBMIの交互作用項である．このモデルでは，パラメータBは単独では解釈できない．理由としてBMIがゼロの人における２つの群間の血圧の平均差を表すことが挙げられる． この例でも解決策としてBMIを中心化し，中心化済みBMIと，TRTと中心化済みBMIの交互作用項をモデルに含めることが挙げられる．これによって，*B'*（新しいモデルにおけるTRTの係数）は，分析対象集団の平均的なBMIを持つ人の血圧の平均差を指すようになる．

16: 欠測データの適切な取り扱い

欠測データの取り扱いに用いた方法は，欠測データに関する仮定（Missing completely at random，Missing at random，Missing not at random）に関係して記述され，正当化されるべきであり，必要に応じて感度分析を行われなければならない．欠測データ[83]は選択バイアスをもたらす可能性があり，多重代入法[84]や逆確率重み付け法[85]などの適切な方法を用いて取り扱うべきである．単純な方法（例えば，complete case analysis，観測データの平均値を用いた単一代入法，Last Observation Carried Forward（LOCF），欠測指標（missing indicator method））は，一般的に統計的に妥当ではなく，深刻なバイアスにつながる可能性がある[86].

17-23項目：報告とプレゼンテーション

17: データの適切かつ正確な記述

平均と標準偏差（SD）は，ある程度対称的な分布を持つ連続変数のデータを要約する．標準誤差（SE）をSDの代わりに使用することは適切ではない[87]．データの記述においてはSDを使用し，パラメータの推定にはSEを使用する[88]．また，「平均±SD」と報告することは多くの場合適切ではないため（これは単にデータの約68％が含まれる範囲を意味するため），「平均（SD）」と報告すべきである[1]．データが大きく歪んでいる場合は，中央値と四分位範囲（IQR）の方がより有益な要約統計量である．正の変数の平均/SD比が２未満である場合は，データが歪んでいる可能性があるため注意する必要がある[89]．カテゴリーデータは，その数（n）とパーセンテージとして要約する[90]．また，コホートデータの場合，追跡期間は中央値やIQRの要約を報告する．

18：記述的結果として発生（発症）指標・信頼区間の提示と，分析的結果として関連指標・信頼区間・p値の提示

発生（発症）指標の点推定値（例えば，有病率，リスク，発生率）は，記述を目的とした結果として95%信頼区間とともに報告する[90]．関連指標の点推定値（例えば，OR，リスク比，率比）は，分析を目的とした結果として95%信頼区間とp値とともに報告する[91, 92].

19：各群の信頼区間を個別に提示されるのではなく，群間の比較/対比の信頼区間が提示されている

RCTのような分析的研究では，信頼区間は各群についてではなく群間の比較/対比として示すべきである[6].

上記の血圧の例では[20]，著者は各グループ内（治療群と対照群）の治療前後の血圧の平均値の差に対して，それぞれ95%信頼区間を報告した．しかし，研究目的は治療群と対照群を比較することであるため，血圧のグループ間の平均値の差について95%信頼区間を示すべきであった．

20：結果の選択的報告とp-hackingが避けられている

行われた全ての統計解析は，その結果に関わらず報告する必要がある．p-hackingは，望ましいp値（値が大きい方向でも小さい方向でも）を生成するためにデータを操作することであり，避けなければならない[93–95]．読者または査読者は，選択的報告やp-hackingを評価することが難しい可能性があるが，通常，目的に記載されているよりも多くの解析が行われている場合や，方法でより多くの変数が使用されているにも関わらず統計的に有意な結果のみが提示されている場合は，この手がかりとなる可能性がある．

21：効果量，検定統計量，p値について，適切で一貫した正確な数値の報告

p値は，たとえ0.05より大きい場合でも，1または2桁の有効数字でその実数を報告する必要がある（$p>0.05$は不適切であり，$p=0.09$や$p=0.28$と報告する）．$p<0.05$のように「統計的有意性」にだけ焦点を当て，p値を二分すべきでない[96–98]．また，「0.000」や「NS」といった表現を使用しない．それでも，小数点以下の桁数が多すぎる数値表記は避けるべきである（通常，0.001未満のp値は<0.001と表記しても問題ない）[99, 100]．またサンプルサイズが100よりはるかに小さい場合，小数点以下のパーセンテージを表示することには意味がない．

22：今後実施される可能性がある，メタ解析に含めることができる十分な数値結果の提示

RCTや観察研究のメタ解析は，健康科学研究において高いエビデンスレベルを提供する．個々の研究で，今後実施される可能性があるメタ解析に寄与する数値結果を提示することは特に重要である．フォローアップ後のスコアとベースラインからの変化スコアは，RCTで治療効果を推定するために適用できる2つのアプローチである[101]．フォローアップスコアのメタ解析では，介入群と対照群の2つのグループにおける介入後の平均値とSDが必要だが，変化スコアのメタ解析を行うためにはベースラインからの差の平均とSDが必要となる．しかし，実際の論文では介入前後の平均値とSDのみが報告されることが多い．各群の差の平均は，平均値の差から計算可能だが，差のSDの計算には介入前と介入後のSDに加えて，グループ毎のベースラインとフォローアップ後のスコアの相関係数の推定値が必要である．

23：図表の適切な提示

表や図は，効果的なデータ表示であり，適切に取り扱われるべきである[102–105]．図は，変数の種類に基づいて選択し，適切にスケーリングする必要がある．エラーバーグラフは，平均値と信頼区間を表示するために使用できる．代わりに平均値にSEバーが重ねられた棒グラフを示すことは不適切である（ダイナマイト・プランジャー・プロットと呼ばれる[105]）．表は，それ自体で成立する必要があり，ラベル，単位，値など十分な詳細を含める必要がある．

24-30項目：解釈

24：p値とともに関連指標と95%信頼区間に基づいて結果が解釈され，p値が大きいことを「効果がない証拠」ではなく「決定的でない結果」として正しく解釈されている

研究結果は，p値だけでなく，平均差と95%信頼区間などの適切な関連性の尺度の推定値に照らして解釈すべきである．「治療効果がない」という帰無仮説を検定する場合，p値は帰無仮説と検定に使用した全ての仮定が正しいとして，統計的関連性が観察されたものと同程度かそれ以上に極端になる確率である．null以外の効果量のp値も計算できる．点推定値は$p=1.00$になるという意味でデータに最も適合する効果量であり，95%CIは$p>0.05$になるという意味でデータと合理的に適合する効果量の範囲を示している[97]．結果の臨床的重要性と統計的信頼性は，p値が閾値を超えているかどうかだけでなく，95%信頼区間と正確なp値の両方によって判断する必要がある[28, 106]．$p>0.05$を効果がないと解釈することは誤りであり，結果が決定的でないことを表している[107, 108]．また，効果が重要でないという証拠にはならない（エビデンスがないことは，エビデンスが存在しないことを表すわけではない）．重要でないことを推論するには，信頼区間内のすべての効果量が重要でないとみなされる必要がある[97]．

25：研究結果の解釈において事後的な検出力分析ではなく信頼区間が使用されている

検出力を観察された研究結果に関係するものとして解釈することは妥当ではない[109–111]．検出力は，サンプルサイズの計算など，実際の研究が始まる前の研究の合理性とデザインの一部として扱われるべきである．検出力は，その後の観察結果を正しく説明するものではない．例えば，研究の検出力が高く，大きいp値が観察されたにも関わらず，帰無仮説よりも対立仮説を支持する結果もある[111]．結果の精度は信頼区間を使用して評価すべきである．

26：発生（発症）または関連の指標が正しく解釈されている

発生（発症）と関連の指標を正しく解釈することは極

めて重要である．オッズ比は，よく誤った解釈の例となる．イベントが稀であればリスク比に近似できるが，概念的には同じではなく，イベントが多い場合はもはや近似することはできない[112, 113]．曝露群のリスクが60%，非曝露群のリスクが40%の研究では，オッズ比（2.25）をリスク比（1.5）として解釈することは，かなりの誤りが生じる．横断研究における有病率は別の例だが，誤って「リスク」と呼ばれることがある．

27：因果を関連と相関から区別されている

効果（effect），関連（association），相関（correlation）などの用語の正しい使用には注意が必要である．関連性（association）は独立ではないという意味で，因果（効果：effect）を意味するものではない．因果効果（causal effect）の推定には，結果の前の曝露の測定（時間的関係性）と交絡の調整が必要である．相関関係（correlation）は，2つの変数間の単調な関連性を指す．したがって，相関関係がないということは関連性がないことを意味するわけではない．

28：解釈において，事前に規定された解析の結果と探索的解析の結果が区別されている

事前に規定されプロトコルに記載された解析（a priori）から得られた結果は，データドレッジ（データ導出または事後分析）後に得られた結果よりもはるかに信頼性が高い．

29：研究方法論の限界についての適切な議論

研究デザインと解析の方法論の限界について議論する必要がある．理想的には，バイアスパラメータに確率分布を仮定し，モンテカルロ感度分析またはベイズ分析を使用して，バイアスを確率的に考慮する確率的バイアス分析を実行して，制御されていない交絡（例えば，未測定の交絡因子），選択バイアス（例えば，アウトカムデータの欠測），測定バイアス（例えば，曝露の測定誤差）を調整する必要がある[114–116]．論文著者らは，バイアスの主な原因とそれが研究結果に与える影響について，少なくとも定性的に議論をする必要がある[117, 118]．

30：統計解析によって裏付けられた結論のみを導き出し，標的集団以外の対象に対して結果を一般化していない

研究の解釈は，結果についてだけでなく，対象集団や，デザインと解析における限界にも基づかなければならない[82]．例えば，研究が女性を対象に行われた場合，必ずしも男性と女性の集団に一般化できるわけではない．

結論

医学研究において適切な統計と方法論は非常に重要な役割を果たす．我々は，著者が医学研究を実施し報告する際にCHAMPを遵守すること，また編集委員と査読者が投稿論文の評価する際にCHAMPを遵守することを強く推奨する．我々は基本的な項目しか取り上げていないが，研究や統計モデルの種類ごと（例えば，RCT，予測モデルなど）に，統計の専門知識を必要とする固有の問題がある．チェックリストの中には明確な答えがない項目もあり，論文の統計を評価する際に主観が入り込む余地があることを認識している．さらに，チェックリストの設問は等しく重要というわけではなく，濃淡がある．例えば，研究計画に重大なエラーがある論文は，データの解析方法に関係なく統計的に受け入れられないが，プレゼンテーションの側面はチェックリストの他の要素ほど重要ではない．経験豊富な統計家による統計査読は，チェックリストよりも研究論文の統計を査読するために適した方法であることに留意することが重要である．我々はCHAMPが医学研究における統計の活用において著者の実践向上に役立ち，医学論文の統計評価の際に編集委員と査読者にとって有用かつ手軽な参考資料となることを願っている．

Box 1 用語

関連性（Association）：2つの変数間の関係を示す，統計的依存関係．

関連性の尺度：絶対的または相対的な変数間の関連性の尺度．絶対的な関連性の尺度は，リスク差，率差などの発生(発症)の尺度の差である．相対的な関連性の尺度は，リスク比，率比，オッズ比などの発生(発症)尺度の比の指標である．

因果ダイアグラム（causal directed acyclic graph（DAG））：矢印で関連付けられたノードを含み，次の2つの性質を持つ図．(i) 2つの変数間に矢印がない場合は，直接的な因果関係がないことを意味する．(ii) 変数の任意のペアに共通するすべての原因がグラフに含まれる．

合流点（Colider）：2つの変数の共通効果である変数．

効果修飾因子：曝露が結果に与える影響を修飾する変数．

交絡因子：曝露と結果の共通原因の経路にある変数．

交絡：曝露と結果の共通原因によって生じるバイアス．

相関：単調（全体的に非増加または完全に非減少）な関連性．

データドレッジング（データフィッシング）：統計的に有意であると示すことができるパターンを見つけるためにデータ解析を誤用すること．

デザイン効果：サンプリングスキームからの推定量の分散と，同じサンプルサイズでの単純ランダムサンプリングからの推定量の分散の比率．

効果（因果効果）：潜在アウトカム（反事実アウトカム）のフレームワークでは，Aを変更するとBが変化するユニットが少なくとも1つある場合，ユニットの母集団でAはBに対して（因果的）効果があると言う．

線形性の仮定：評価すべき定量的な予測因子を含めることによって課される回帰モデルの基礎となる仮定．

媒介因子（Mediator）：曝露によって影響を受け，アウト

カムにも影響を与える変数.

帰無仮説：統計的仮説検定で想定される仮説で，母集団内の２つの変数間に関連性がないことにしばしば相当する.

発生（発症）指標：リスク（発生（発症）割合），率比，有病割合などの疾患頻度の尺度.

スパースデータバイアス：スパースデータの結果として発生するバイアスで，効果量の推定値の過大につながる.

〈原版謝辞〉

本論文の初稿に対して貴重なコメントをいただいたSander Greenland氏，Stephen Senn氏，Richard Riley氏に感謝する.

参考文献

[1] Altman DG. Practical statistics for medical research. New York: CRC press; 1990.

[2] Thiese MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. Biochem Med. 2015;25:5–11.

[3] Thiese MS, Walker S, Lindsey J. Truths, lies, and statistics. J Thorac Dis. 2017;9:4117–4124.

[4] Altman DG. The scandal of poor medical research. BMJ. 1994;308:283–284.

[5] Nielsen RO, Shrier I, Casals M, et al. Statement on methods in sport injury research from the 1st methods matter meeting, Copenhagen, 2019. Br J Sports Med. 2020;54:941.

[6] Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10:28–55.

[7] Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med. 2007;147:W163-194.

[8] Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138:W1-12.

[9] Altman DG, McShane LM, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Med. 2012;10:51.

[10] Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-73.

[11] Altman DG. Statistical reviewing for medical journals. Stat Med. 1998;17:2661–2674.

[12] Goodman SN, Altman DG, George SL. Statistical reviewing policies of medical journals: caveat lector? J Gen Intern Med. 1998;13:753–756.

[13] Nielsen RØ, Shrier I, Casals M, et al. Statement on methods in sport injury research from the 1st methods matter meeting, Copenhagen, 2019. J Orthop Sports Phys Ther. 2020;50:226–233.

[14] Verhagen E, Stovitz SD, Mansournia MA, et al. BJSM educational editorials: methods matter. Br J Sports Med. 2018;52:1159–1160.

[15] Lang TA, Altman DG. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Handbook, European association of science editors. 2013.

[16] Assel M, Sjoberg D, Elders A, et al. Guidelines for reporting of statistics for clinical research in urology. Eur Urol. 2019;75:358–367.

[17] Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. Br Med J. 1986;292:810–812.

[18] Mansournia MA, Collins GS, Nielsen RO, et al. CHecklist for statistical Assessment of Medical Papers: the CHAMP statement. Br J Sports Med. 2021;55:1002–1003.

[19] Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: A classification of data science tasks. Chance. 2019;32:42–49.

[20] Guimaraes GV, de Barros Cruz LG, Fernandes-Silva MM, et al. Heated water-based exercise training reduces 24-hour ambulatory blood pressure levels in resistant hypertensive patients: a randomized controlled trial (HEx trial). Int J Cardiol. 2014;172:434–441.

[21] Centre for Evidence-Based Medicine. Study design.

https://www.cebm.net/2014/04/study-designs/
[22] Machin D, Campbell MJ. The design of studies for medical research. Chichester: John Wiley & Sons Ldt; 2005.
[23] Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest. 2020;158:S21–S28.
[24] Stovitz SD, Verhagen E, Shrier I. Distinguishing between causal and non-causal associations: Implications for sports medicine clinicians. Br J Sports Med. 2019;53:398–399.
[25] Etminan M, Nazemipour M, Candidate MS, et al. Potential biases in studies of acid-suppressing drugs and COVID-19 Infection. Gastroenterology. 2021;160:1443–1446.
[26] Mansournia MA, Danaei G, Forouzanfar MH, et al. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis: Analysis with marginal structural models. Epidemiology. 2012;23:631–640.
[27] Machin D, Campbell MJ, Tan SB, et al. Sample sizes for clinical laboratory and epidemiology studies. Oxford: John Wiley & Sons Ltd; 2018.
[28] Mansournia MA, Altman DG. Invited commentary: Methodological issues in the design and analysis of randomised trials. Br J Sports Med. 2018;52:553–555.
[29] Cook JA, Julious SA, Sones W, et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. BMJ. 2018;363:k3750.
[30] Bland JM. The tyranny of power: Is there a better way to calculate sample size? BMJ. 2009;339:b3985.
[31] Rothman KJ, Greenland S. Planning study size based on precision rather than power. Epidemiology. 2018;29:599–603.
[32] Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
[33] Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. Stat Med. 2019;38:1276–1296.
[34] Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I - continuous outcomes. Stat Med. 2019;38:1262–1275.
[35] Jungmalm J, Bertelsen ML, Nielsen RO. What proportion of athletes sustained an injury during a prospective study? censored observations matter. Br J Sports Med. 2020;54:70–71.
[36] Nielsen RO, Bertelsen ML, Ramskov D, et al. Randomised controlled trials (RCTs) in sports injury research: Authors-please report the compliance with the intervention. Br J Sports Med. 2020;54:51–57.
[37] Edouard P, Steffen K, Navarro L, et al. Methods matter: Instrumental variable analysis may be a complementary approach to intention-to-treat analysis and as treated analysis when analysing data from sports injury trials. Br J Sports Med. 2021;55:1009–1011.
[38] Mansournia MA, Nazemipour M, Naimi AI, et al. Reflection on modern methods: Demystifying robust standard errors for epidemiologists. Int J Epidemiol. 2021;50:346–351.
[39] Altman DG, Bland JM. Statistics notes: The normal distribution. BMJ. 1995;310:298.
[40] Senn S. The t-test tool. Signif (Oxf). 2008;5:40–41.
[41] Heeren T, D'Agostino R. Robustness of the two independent samples t-test when applied to ordinal scaled data. Stat Med. 1987;6:79–90.
[42] Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. BMJ. 2016;352:i1981.
[43] Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic regression: Causes, consequences, and control. Am J Epidemiol. 2018;187:864–870.
[44] Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Stat Med. 2015;34:3133–3143.
[45] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. New Jersey: John Wiley & Sons Inc; 2012.
[46] Mansournia MA, Altman DG. Some methodological issues in the design and analysis of cluster randomised trials. Br J Sports Med. 2019;53:573–575.
[47] Korn EL, Graubard BI. Analysis of health surveys. John Wiley & Sons; 2011.
[48] Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. Int J Epidemiol. 2013;42:860–869.
[49] Mansournia MA, Jewell NP, Greenland S. Case-control matching: effects, misconceptions, and recommendations. Eur J Epidemiol. 2018;33:5–14.
[50] Greenland S, Jewell NP, Mansournia MA. Theory and methodology: essential tools that can become dangerous belief systems. European journal of epidemiology. 2018;33:503–506.
[51] Mansournia MA, Altman DG. Population attributable fraction. BMJ. 2018;360:k757.
[52] Khosravi A, Nielsen RO, Mansournia MA. Methods matter: Population attributable fraction (PAF) in sport and exercise medicine. Br J Sports Med. 2020;54:1049–

1054.
[53] Riley RD, van der Windt DA, Croft P, et al. Prognosis research in healthcare: concepts, methods and impact. Oxford University Press; 2019.
[54] Steyerberg EW. Clinical prediction models. Springer; 2019. https://link.springer.com/book/10.1007/978-3-030-16399-0
[55] Doosti-Irani A, Nazemipour M, Mansournia MA. What are network meta-analyses (NMAs)? a primer with four tips for clinicians who read NMAs and who perform them (methods matter series). Br J Sports Med. Epub ahead of print September 18, 2020. DOI: 10.1136/bjsports-2020-102872
[56] Bland JM, Peacock JL. Interpreting statistics with confidence. Obstet Gynaecol. 2002;4:176–180.
[57] Austin PC, Hux JE. A brief note on overlapping confidence intervals. J Vasc Surg. 2002;36:194–195.
[58] Mittal N, Bhandari M, Kumbhare D. A tale of confusion from overlapping confidence intervals. American journal of physical medicine & rehabilitation. 2019;98:81–83.
[59] Matthews JN, Altman DG. Statistics notes. interaction 2: compare effect sizes not P values. BMJ. 1996;313:808.
[60] Knol MJ, Pestman WR, Grobbee DE. The (mis)use of overlap of confidence intervals to assess effect modification. Eur J Epidemiol. 2011;26:253–254.
[61] Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327:307–310.
[62] Andersen PK, Skovgaard LT. Regression with linear predictors. Springer; 2010.
[63] Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. John Wiley & Sons; 2008.
[64] Harrell FE Jr. Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015. https://link.springer.com/book/10.1007/978-3-319-19425-7
[65] Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. Stat Med. 2013;32:3788–3803.
[66] Mardani M, Rahnavardi M, Rajaeinejad M, et al. Crimean-Congo hemorrhagic fever among health care workers in Iran: A seroprevalence study in two endemic regions. Am J Trop Med Hyg. 2007;76:443–445.
[67] Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994;13:1715–1726.
[68] Suzuki E, Tsuda T, Mitsuhashi T, et al. Errors in causal inference: An organizational schema for systematic error and random error. Ann Epidemiol. 2016;26:788-793. e1.
[69] Greenland S, Mansournia MA. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. Eur J Epidemiol. 2015;30:1101–1110.
[70] Mansournia MA, Higgins JPT, Sterne JAC, et al. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiology. 2017;28:54–59.
[71] Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. Epidemiology. 2015;26:466–472.
[72] Almasi-Hashiani A, Nedjat S, Mansournia MA. Causal methods for observational research: a primer. Arch Iran Med. 2018;21:164–169.
[73] Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J. 2011;32:1704–1708.
[74] Gharibzadeh S, Mohammad K, Rahimiforoushani A, et al. Standardization as a tool for causal inference in medical research. Arch Iran Med. 2016;19:666–670.
[75] Nielsen RO, Bertelsen ML, Møller M, et al. Training load and structure-specific load: applications for sport injury causality and data analyses. Br J Sports Med. 2018;52:1016–1017.
[76] Nielsen RO, Bertelsen ML, Møller M, et al. Methods matter: Exploring the "too much, too soon" theory, part 1: causal questions in sports injury research. Br J Sports Med. 2020;54:1119–1122.
[77] Nielsen RO, Simonsen NS, Casals M, et al. Methods matter and the "too much, too soon" theory (part 2): What is the goal of your sports injury research? Are you describing, predicting or drawing a causal inference? Br J Sports Med. 2020;54:1307–1309.
[78] Nielsen RO, Bertelsen ML, Ramskov D, et al. Time-to-event analysis for sports injury research part 1: time-varying exposures. Br J Sports Med. 2019;53:61–68.
[79] Nielsen RO, Bertelsen ML, Ramskov D, et al. Time-to-event analysis for sports injury research part 2: time-varying outcomes. Br J Sports Med. 2019;53:70–78.
[80] Mansournia MA, Etminan M, Danaei G, et al. Handling time varying confounding in observational research. BMJ. 2017;359:j4587.
[81] Mansournia MA, Naimi AI, Greenland S. The implications of using Lagged and baseline exposure terms in longitudinal causal and regression models. Am J Epidemiol. 2019;188:753–759.
[82] Altman DG, Bland JM. Generalisation and extrapola-

tion. BMJ. 1998;317:409–410.
[83] Altman DG, Bland JM. Missing data. BMJ. 2007;334:424.
[84] Vickers AJ, Altman DG. Statistics notes: missing outcomes in randomised trials. BMJ. 2013;346:f3438.
[85] Mansournia MA, Altman DG. Inverse probability weighting. BMJ. 2016;352:i189.
[86] Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009;338:b2393.
[87] Altman DG, Bland JM. Standard deviations and standard errors. BMJ. 2005;331:903.
[88] Campbell MJ. Statistics at square one. John Wiley & Sons; 2021. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119402350
[89] Altman DG, Bland JM. Detecting skewness from summary information. BMJ. 1996;313:1200.
[90] Nielsen RO, Debes-Kristensen K, Hulme A, et al. Are prevalence measures better than incidence measures in sports injury research? Br J Sports Med. 2019;53:396–397.
[91] Nielsen RO, Bertelsen ML, Verhagen E, et al. When is a study result important for athletes, clinicians and team coaches/staff? Br J Sports Med. 2017;51:1454–1455.
[92] Pourahmadi M, Koes BW, Nazemipour M, et al. It is time to change our mindset and perform more high-quality eesearch in low back pain. Spine. 2021;46:69–71.
[93] Stovitz SD, Verhagen E, Shrier I. Misinterpretations of the "p value": a brief primer for academic sports medicine. Br J Sports Med. 2017;51:1176–1177.
[94] Windt J, Nielsen RO, Zumbo BD. Picking the right tools for the job: opening up the statistical toolkit to build a compelling case in sport and exercise medicine research. Br J Sports Med. 2019;53:987–988.
[95] Nielsen RO, Chapman CM, Louis WR, et al. Seven sins when interpreting statistics in sports injury science. Br J Sports Med. 2018;52:1410–1412.
[96] McShane BB, Gal D, Gelman A, et al. Abandon statistical significance. Am Stat. 2019;73:235–245.
[97] Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–350.
[98] Rothman KJ , Greenland S , Lash TL. Precision and statistics in epidemiologic studies. In: Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
[99] Altman DG, Bland JM. Presentation of numerical data. BMJ. 1996;312:572.
[100] Kordi R, Mansournia MA, Rostami M, et al. Troublesome decimals; a hidden problem in the sports medicine literature. Scand J Med Sci Sports. 2011;21:335–336.
[101] Higgins J WG. Cochrane handbook for systematic reviews of interventions. 2023. https://training.cochrane.org/handbook/current (accessed 2024-07-12)
[102] Schriger DL, Sinha R, Schroter S, et al. From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the British Medical Journal. Ann Emerg Med. 2006;48:750–756, 756.e1–21.
[103] Morris TP, Jarvis CI, Cragg W, et al. Proposals on Kaplan-Meier plots in medical research and a survey of stakeholder views: KMunicate. BMJ Open. 2019;9:e030215.
[104] Vickers AJ, Assel MJ, Sjoberg DD, et al. Guidelines for reporting of figures and tables for clinical research in urology. Eur Urol. 2020;78:97–109.
[105] Freeman JV, Walters SJ, Campbell MJ. How to display data. Wiley; 2009.
[106] Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. John Wiley & Sons; 2008.
[107] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305–307.
[108] Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol. 2020;20:244.
[109] Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55:19–24.
[110] Bacchetti P. Peer review of statistics in medical research: the other problem. BMJ. 2002;324:1271–1273.
[111] Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. Ann Epidemiol. 2012;22:364–368.
[112] Janani L, Mansournia MA, Nourijeylani K, et al. Statistical issues in estimation of adjusted risk ratio in prospective studies. Arch Iran Med. 2015;18:713–719.
[113] Talebi SS, Mohammad K, Rasekhi A, et al. Risk ratio estimation in longitudinal studies. Arch Iran Med. 2019;22:46–49.
[114] Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer; 2011.
[115] Greenland S, Lash TL. Bias analysis. In: Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
[116] Lash TL, Fox MP, MacLehose RF, et al. Good prac-

tices for quantitative bias analysis. Int J Epidemiol. 2014;43:1969–1985.
[117] Altman DG, Bland JM. Uncertainty beyond sampling error. BMJ. 2014;349:g7065.
[118] Altman DG, Bland JM. Uncertainty and sampling error. BMJ. 2014;349:g7064.