原稿作成日: 2017 年 3 月 31 日

症例数の設計：

信頼できるエビデンスを得るために症例数は計画時に必ず決めておく

〈教材提供〉
AMED 支援「国際誌プロジェクト」　提供

無断転載を禁じます

草案
新谷歩　大阪市立大学医学研究科医療統計学講座教授
加葉田大志朗　大阪市立大学医学研究科医療統計学講座特任助教

査読
大門貴志　兵庫医科大学医療統計学教授
角間辰之　久留米大学バイオ統計センター教授
市川家國　信州大学特任教授
山本紘司　大阪市立大学大学院医学研究科医療統計学講座准教授
石原拓磨　大阪市立大学大学院医学研究科医療統計学講座特任助教

無断転載禁止

目次

無断転載禁止

はじめに

ここまでの単元で、研究に参加した人（あるいはマウスなど）の数が統計的記述や統計的推測に影響を与えることは説明してきました。この、研究に参加した研究対象者の数のことを慣例的に「症例数」と呼びます。本単元では、多くの研究者にとって大きな関心事となる、この症例数について説明します。研究計画時に症例数を設計することの必要性、そのための基本的な考え方などをここで学習します。症例数の設計を行う際には、その計算のために統計ソフトウェア、インターネットのサイトなどを利用する必要があります。その実施方法についてはビデオで説明します。

学習目標

本単元を通じてあなたが修得を目指すものは：

- 症例数の設計の必要性を知る
- 症例数の設計の基本的な考え方を知る
- 基本的な症例数設計を行うことができる

P 値の落とし穴

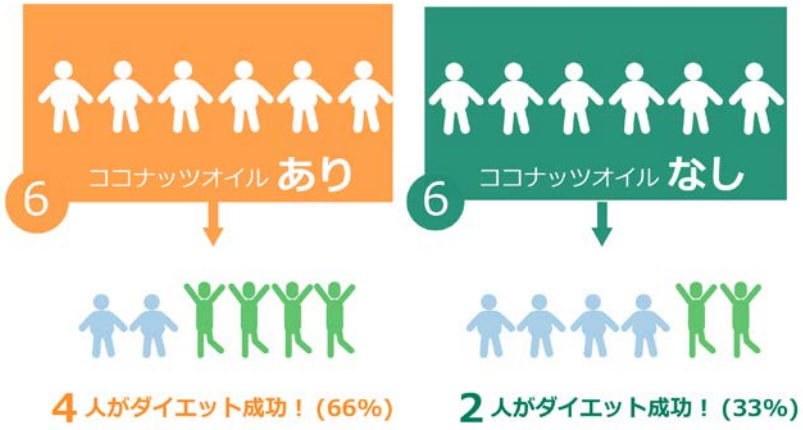
P 値に最も影響するもの

「ココナッツオイルの摂取がダイエットに効果がある」という仮説を証明するのに、以下の 3 つの研究を実施したとしましょう。この 3 つの研究では、同じココナッツオイルを用いたので、本来その効果は同じであると考えることができます。ここでは、いずれの研究でもココナッツオイルを摂取した群で 6 分の 4 がダイエットに成功し、摂取しなかった群では 6 分の 2 しか成功しなかったとしましょう。この 6 分の 4（66%）と 6 分の 2（33%）が 3 つの研究それぞれで、統計的に有意な差であるかどうかを調べてみましょう。

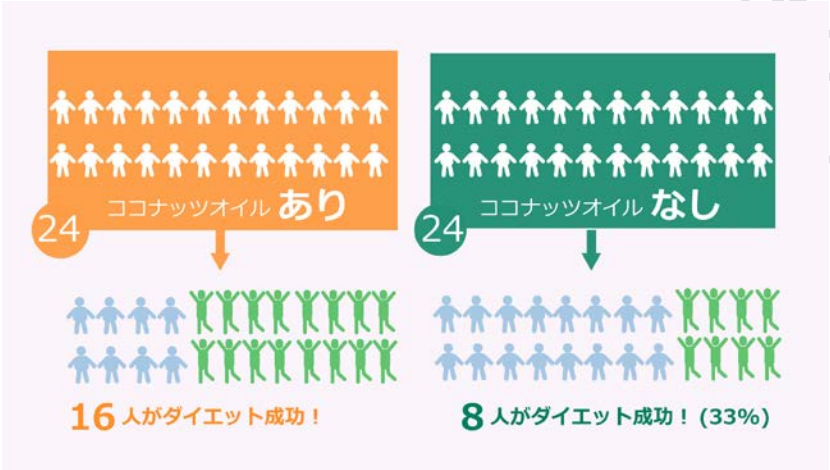
研究例 1 では、症例数がココナッツオイルを摂取した群と摂取しなかった群でそれぞれ 6 人でした。研究 2 ではそれぞれ 24 人、研究 3 においてはそれぞれ 36 名でした。各研究で得られる P 値を見てみましょう。

無断転載禁止

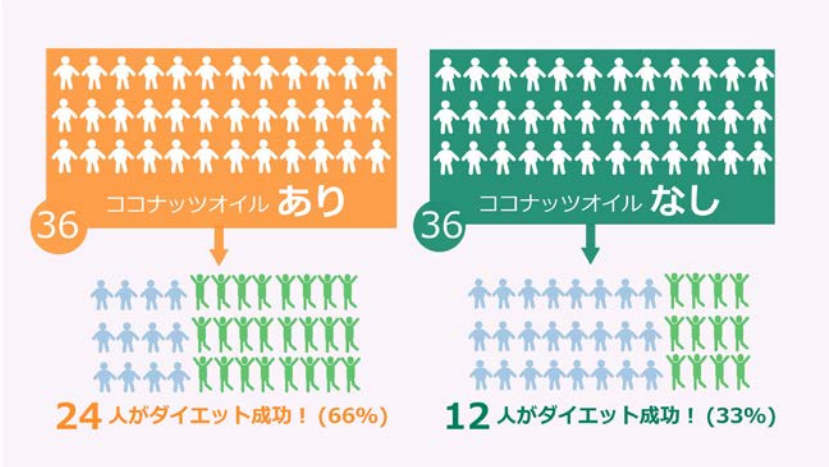
研究例 1



研究例 2



研究例 3

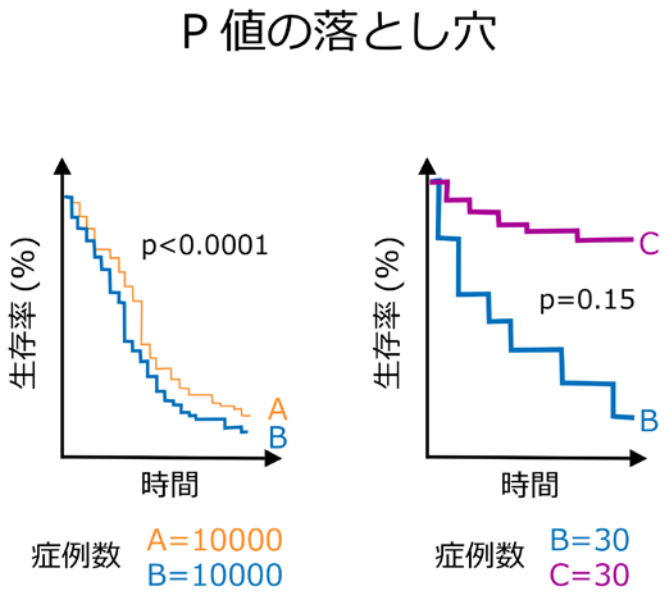


ここでの例に関して P 値を定義しますと、P 値は、ココナッツオイルに全くダイエット効果がない場合（つまり、摂取群も非摂取群もダイエットに成功する人の割合が同じとき）でも全く偶然に 66%と 33%の差、またはそれ以上の差が得られる確率の実現値となります。また、ココナッツオイルに全くダイエット効果がないことが真実だとしますと、このもとでそのようなデータが得られたときに、効果があると判断してしまうとすると、P 値は「そのような誤りを犯す確率の実現値を与える」ともいえます。P 値は研究 1 では 0.56、研究 2 では 0.04、研究 3 ではほぼ 0 となりました。

3 つの研究ではココナッツオイルの効果の指標であるダイエット成功者の割合は摂取群が 66%、非摂取群が 33%、といずれの研究も同じでした。異なっている点は症例数です。ダイエット成功者の割合について摂取群が 66%、非摂取群が 33%という差が見られるのは、ココナッツオイルに全くダイエットの効果が実際はない場合でも、症例数が各群 6 人だと 2 回に 1 回程度、各群 24 人になれば 100 回に 4 回程度、各群 36 人まで増やせば、その頻度はさらに激減するのです。つまり、真にダイエット効果がないとき、偶然に効果に差が見られる、あるいはそのように判断してしまう確率は、症例数を増やせば増やすほど小さくなります。すなわち、データの信頼性は、データが増えれば増えるほど上がることが分かります。

P 値の落とし穴

研究を開始するときには、本当に臨床的に意味のある差があるときに統計的に有意な差が得られる（かつ本当に臨床的に意味がある差がないときには、そのような示唆を得ることのできる）最低限の症例数を設計する必要があります。症例数は多ければ多いほど統計解析から得られる結果の精度が上がるので、科学的妥当性は高くなります。例えば、日常診療で記録されたカルテから研究用にデータを抽出したとします（このような研究を、既存試料を用いた観察研究といいます）。このときに頑張って過去 10 年分のデータをコンピュータから抽出したとしても、患者さんに直接的な影響はほとんどありません。従って、データの量は多ければ多いほど良いのですが、この時留意すべき点は、データの量が多すぎると解析の精度が上がりすぎて臨床的に意味のない差でも統計的有意差が確認されてしまうことです。科学的にみればデータの



無断転載禁止

量が多すぎて精度が上がることは問題ありません。しかし臨床的に意味のない差を、統計的に有意な差があるからといって、あたかも臨床的に意味があるかのように報告することには大きな問題があります。確認された群間のアウトカムの違いが、臨床的に意味があるかどうかを見定めたうえで、統計的に意味があるかどうかを議論する必要があります。下記に、臨床的な意義は小さいけれど症例数が多いために統計的には有意な差と判断される例（左）と、臨床的に意味のある差が出ているけれど症例数が少ないために統計的には有意な差とは言えない例(右)を紹介しておきます。

症例数が多い方が解析結果の精度が上がり、科学性が増しますが、研究のために新たにデータを収集する際には、日常診療の幅を超えて患者さんに負担（侵襲）がかかる場合が多く発生します。安全性が担保されていない試験薬に必要以上に多くの患者さんを暴露させることはできません。したがって、研究を実行するに当たっては、倫理的な理由や研究にかかる費用や労力などを十分に考慮して、研究対象者の数は必要最低限とすることが原則です。これは対象が人以外の動物の場合でも同様です。動物だからといって必要以上に多くの動物を危険にさらすことは許されません。

症例数は研究計画時に設計すべき

統計的に有意な差があるという判断をもって、有効性が確認されたと判断するような検証的な研究においては、研究計画時に症例数を決定します。それは、本当に臨床的に意味のある差がある場合には統計的に有意な差が得られる（また、逆に本当に臨床的に有意な差がないときには、そのような統計的示唆を得ることのできる)、最低限必要な症例数の設計です。P 値が 0.0001 でも､0.001 でも､あるいは 0.049 でも統計的に有意な差があると認識するのであれば、統計的に有意な差を検出できる最低限必要な数とは、P 値が 5%をちょうど下回るために必要な症例数ということになります。

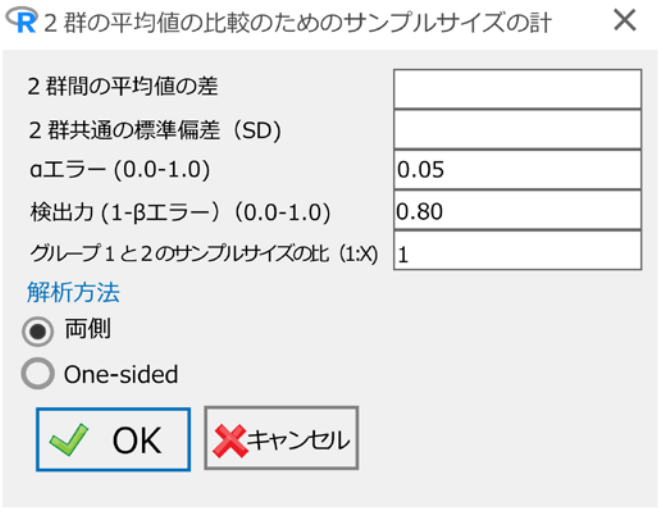
P 値を研究途中で何回も計算して有意差が確認されたときに研究対象者の組み入れをストップさせてはどうか、と考えてしまうかもしれませんが、それは厳禁です。「多重性の問題：研究の事前計画の重要性」の単元で学習したように、何度も何度も P 値を計算すると、間違って統計的に有意な差として認識されやすくなります。その対策として研究の途中で実施した統計解析の回数が多ければ多いほど、有意とする P 値の水準をより厳しく設定する必要がでてきますから、結果的に個々の解析に対しては統計的に有意な差は得られにくくなっていきます。したがって、研究途中で何度も P 値を計算することは通常ご法度です。研究計画時には、P 値の補正を含めて､どのように解析を行うかを中間解析のやり方､として決めておく必要があります。中間解析を計画しない研究では、研究開始時に決めた症例数に到達し、データを固定して、仮説の証明が可能となります。

研究計画時に必要な症例数をしっかりと設計することを心がけましょう。研究実施計画書に症例数を予め記載することなく、上述のように研究途中で出てきたデータを用いて密かに何度も P 値を点検することはご法度です。

症例数の設計のための計算は、最近では無料の統計ソフトウェア、インターネットのサイトなどを使えば簡単に行うことができます。その際、例えば、以下の点を考慮するとよいでしょう。

1. アウトカムは連続変数か 2 値変数か？
2. 1 つの群の中で比較を行うのか、2 つの群の間で比較を行うのか？
3. 2 群比較の場合、比較群は対応があるか、ないか？

例えば、「新薬を投薬された群と既存薬を投薬された群で平均血圧を比べる」という場合、アウトカムは血圧なので連続変数ですし、二つの異なる群のアウトカムの平均を比較します。以下は最終解析をスチューデントの t 検定で行うことを想定した場合に、EZR という無料の統計ソフトウェアで症例数設計を行う場合の画面です。



ここでは、画面内の以下の 5 つの項目に想定される値を入力します。

◆ 2 群間の平均値の差：期待される効果を表します。
◆ 2 群共通の標準偏差：データのばらつきを表します。
◆ α エラー：有意水準または第 I 種の過誤（1 型エラー）確率を表します。
◆ 検出力（1 マイナス β エラー）：1 から第 II 種の過誤（2 型エラー）確率を引いたものを表します。
◆ グループ 1 と 2 の症例数の比（1：X）：グループ間で症例数が異なることもあります。

それぞれの項目で留意すべき事項などを以下に示します。

(1) **2 群間の平均値の差**

治療効果が大きいものほど統計的に有意な差が得られやすいので、この項目の値を大きく設定するほど研究に必要な症例数は少なくなります。しかし、症例数設計を行う段階で治療効果を過剰に大きく期待していたものの、研究結果では期待どおりの効果が得られず、統計的に有意な差が得られなくなることがよくあります。そうした場合は、症例数設計を行う段階で、治療効果を過剰に大きく見積もってしまったがために、必要な症例数を少なく設計してしまったといえます。このようなことを回避するために、症例数設計を行う際の効果の見積もりは、先行文献、先行試験、臨床的観点から適切に行う必要があります。事前に全く予想がつかない場合は、最低限これくらいの差であれば臨床的に意味があると考えられる差を見積もりとすることでも構いません。効果の見積もりに関しては、研究計画時に最善の努力をすることが大切です。ランダム化臨床試験を計画する際に観察研究か

無断転載禁止

ら治療効果を見積もる場合は注意が必要です。通常ランダム化臨床試験では、対照群（例えばプラセボを処方されている研究対象者）にもプラセボ効果など心理的作用によって症状の改善が表れてしまうことがあるため、群間の効果の違いは観察研究で観測されるものよりも小さくなることが多いので、効果の見積もりはできるかぎり慎重に行う必要があります。

(2) 2 群共通の標準偏差

データのばらつきの指標である標準偏差を大きく想定した場合には、必要症例数は多くなります。ここでの 2 群共通の標準偏差は、各群について関心の対象となるアウトカムの標準偏差が先行文献や先行試験から得られる場合、症例数で重み付けを行って平均することで得ることができます。あるいは、必要な症例数の設計にあたり安全策を取りたい場合は、2 群のうち標準偏差の大きな方を採用するやり方もあります。

(3) α エラー（有意水準、第 I 種の過誤確率）

有意水準とは帰無仮説を棄却できる基準を示します。例えば P 値が 5%未満であれば統計的に有意だと判断する場合は、有意水準は 5%となります。有意水準は解析前に定めておく必要があり、多重性の問題などがない限り平均値の両側の 5%を使用することが一般的です。EZR では両側の有意水準がデフォルトで用いられています。ここで有意水準を小さくするほど、帰無仮説を棄却するためには P 値が小さくなる必要があり、多くの症例数が必要になります。

片側の有意水準を使用した場合は、必要症例数は両側より少なくなります。両側か片側かの選択に当たっては、単元「検定と P 値：統計的エビデンスとは」で述べたように、非劣性試験など特別な場合を除いて両側を使用するように心がけて下さい。

(4) 検出力（1 マイナス β エラー）

「検出力」とは、調べたい治療に本当に効果があるときにその効果があると判断できる確率のことを意味します。例えばある治療に本当に効果があるとき、100 人の研究者が、この治療の効果を確かめる同じ研究を行った場合に、100 人のうち 80 人の研究者に効果があると判断したとき、検出力は 80%であると表現します。逆に検出力が 30%の研究とは、本当に効果のある治療でも、100 人の研究者が同じ研究を行った場合、せいぜい 30 人に効果があるとまでしか言えないということです。それ故、検出力の低い研究は避けたいものです。したがって、検出力は通常 80%または 90%と比較的高い値に設定する必要があります。100%からこの検出力を差し引いた値（検出力が 80%であれば 20%）は「第 II 種の過誤確率（β エラー）」と呼ばれます。これは検出力とは相反する確率ですから、ある治療に本当に効果がある時に「効果がない」と誤って判断してしまう確率です。この第 II 種の過誤確率が大きくなってしまうと、本当に効果がある治療を見逃す確率が大きくなってしまいます。例えば、第 II 種の過誤確率を 50%と設定した場合には、本当は効果があるのに半分はそれを見逃してしまうため、多くの研究が無駄になってしまいます。そのため、第 II 種の過誤確率は 20%未満、つまり検出力は 80%以上が望ましいとされています。なお、検出力を大きくするほど多くの症例数が必要になります。

(5) グループ 1 と 2 の症例数の比（1：X）

全体の症例数が同じ場合は、通常比較群の数を 1 対 1 でそろえたほうが検出力は高くなり、必要症例数は少なくなります。しかし、研究によっては、新薬をより多くの人で試したい場合など、新薬群の症例数を既存薬群の例数の 2 倍、3 倍にすることも可能です。新薬群

は数が増やせないが既存薬群は増やすことができる場合は、既存薬群の症例数を新薬群の症例数の 2 倍、3 倍にすることも可能です。

この 5 つの要素を考慮して、症例数を設計します。このとき、研究で設定した主要評価項目に対する最終解析との整合性を意識すべきです。すなわち、症例数設計は、主要評価項目とその解析方法に沿って行う必要があります。例えば、割合を 2 群で比較する検定に沿って症例数を設計したものの、最終解析は t 検定で解析することは避けるべきです。検定の方法については単元 13 で勉強します。

解析に用いられた症例数と研究に参加した症例数の食い違い

解析に用いられた症例数と研究に参加した症例数の食い違い

ある一定期間、研究対象者を追跡して観察するような研究では、途中で追跡ができなくなってしまう対象者も出てくる恐れがあります。このような現象を「脱落」などと表現します。脱落によって解析対象者数が減ることが予想される場合には、この脱落例数を上乗せして必要最低限な症例数を設計する必要があります。

除かれた標本の表記

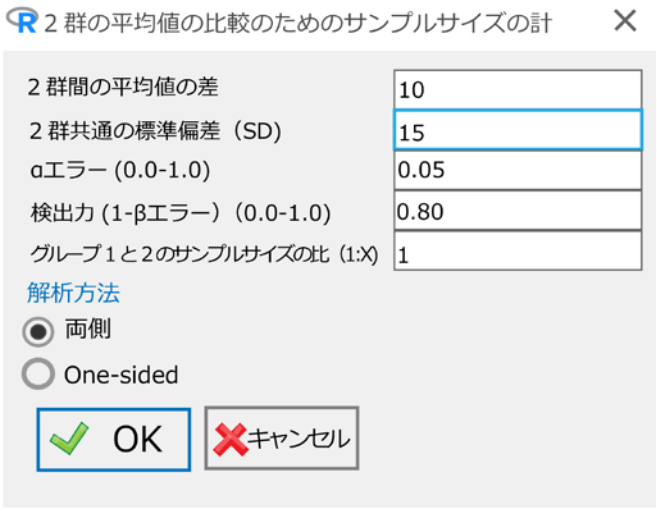
上記のように研究をする上で、何らかの理由でデータが収集されず欠損することはよくあることです。多くの国際誌の持つチェックリストでは、それぞれのデータについての欠損値の数を記述しておくことが望ましいとされています。

症例数計算の例

それでは、以下の研究に対して必要な症例数を計算してみましょう。

新薬を投薬された群と既存薬を投薬された群で平均血圧を比較する研究を計画しているとしましょう。
2 群の平均血圧の差を 10mmHg、血圧の共通標準偏差を 15、有意水準を両側の 5%、検出力は 80%を想定し、新薬群と既存薬群の症例数は同じだけ組み入れるとしましょう。

無断転載禁止



EZR の計算結果

	仮定
2 群間の平均値の差	10
標準偏差	15
α エラー	0.05
	両側検定
検出力	0.8
N2 と N1 のサンプルサイズの比	1
必要サンプルサイズ	計算結果
N1	36
N2	36

上述の値を統計ソフトウェアに入力し、計算を実行しますと、必要症例数は各群 36 人と算出されます。約 10%の脱落例が予想されるのであれば、各群 40（=36/0.9）人を登録することを計画することになります。

求められる症例数の設計

国際誌が提示する論文投稿チェックリストでは、どのように症例数を設定したのか、その妥当性を含め記載することを求めています。最近ではほとんどの倫理審査委員会において、介入研究のみならず観察研究においても、どのような基準で症例数を計算したのかを記載することが求められます。一度設計した症例数は、研究の途中で原則変えることはできません。したがって、症例数設計は、研究計画時に慎重に行うことが必要になります。

無断転載禁止

この単元に関係するビデオ教材

症例数計算　対応のない２群の平均値の比較
検出力計算　対応のない２群の平均値の比較

本単元は日本医療研究開発機構:研究公正高度化モデルである「医系国際誌が規範とする研究の信頼性にかかる倫理教育プログラム」（略称：AMED 国際誌プロジェクト）によって作成された教材です。作成および査読等に参加した専門家の方々の氏名は、冒頭に掲載されています。

無断転載禁止

この単元に関係する国際誌におけるチェックポイントをいくつか紹介します。
（内容は解釈を助けるために一部意訳している部分もあります）

①Nature
（http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2018.02.11）
②New England Journal of Medicine （http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2018.02.11）
③Science （http://www.sciencemag.org/authors/science-editorial-policies; visited 2018.02.11）
④The EMBO Journal （http://emboj.embopress.org/authorguide#embargopolicy; visited on 2018.02.11）
⑤JAMA （http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2018.02.11）

①Nature

- 研究開始時の症例数と、それぞれの解析で使用された症例数が明記されていること
- 症例数計算の方法や妥当性について記載すること
- 全ての解析において、解析の対象とした集団について記載すること
- データの除去を行なった場合にはその手順の記載と理由を説明すること
- 各解析間で含まれる症例数が異なる場合には、その理由を明記すること

④The EMBO Journal

- 症例数が小さい場合、正しい統計的な手法が用いられるべきであり、その妥当性もあわせて明記する必要がある。
- 複雑な実験手技が必要であるために独立な研究対象から多くのデータをとることが困難であることも想定される。しかし統計解析においては症例数が非常に小さい場合には、統計的に有意と言える基準を超えることができないことも懸念される。そのような小症例数（症例数が 5 例未満のような場合）には、実際の各観測値についても図示化しておくことを推奨する。また症例数が小さい場合には、利用した仮説検定の妥当性についても説明する必要がある。また少数の研究対象から反復してデータが得られている場合についても、統計解析には利用可能である。しかしその反復測定の内容などについては詳細に記述しておくべきである。

⑤JAMA

- 観察研究では研究対象となった症例の数を記載しておく。無作為化試験においては無作為化された症例数を記載すること。その際には途中で抜け落ちた症例など追跡不可能な症例の数も記載すること。
- 無作為化比較試験では検出力や症例数の計算についての記載が必須である（EQUATOR Network の CONSORT のガイドラインを参照）。観察研究においては、対象となる症例数が固定されている場合には検出力計算は必須ではない。しかし症例数を研究者が設定したのであれば、その正当性について記載すべきである。通常これらの検出力・症例数計算の手順については統計手法の章の先頭におく。

無断転載禁止