편집인들이 알아야 할 의학통계

이 윤 석

동국대 의학전문대학원 마취통증의학과

The Korean Journal of Anesthesiology

2015년 12월 5일

편집인이 알아야 할?

- 우수한 통계심사위원을 가진다.

* 범주형의 결과를 분석할 때, 사건의 발생이 희소하고 작은 수의 표본이 사용되었을 경우에는 반드시 정확성검정이나 적절히 보정된 점근적검정법을 사용해야 한다. 표준적인 카이제곱검정과 비율 비교는 충분히 큰 표본수와 사건이 전제되었을 때 사용될 수 있다.
* 여러 개 군집 중에서 두 군집의 결과를 비교할 때 해당하는 통계 분포 안에서 반드시 제1 통계오류를 통제하여야 한다. 정규분포인 자료의 경우라면 t검정이 적절할 것이며, 정규분포가 아니더라도 표본수가 아주 크다면 t검정의 결과는 강건성을 가진다. 표본수가 작고 이전의 연구들을 통해 정규분포가 아니라고 알려진 자료라면 비모수통계법을 쓰는 것이 적절하다.
* 결과는 과학적으로 의미를 가지는 정밀성까지만 유효숫자를 표기한다. 가령 오즈비(odds ratio)처럼 연관의 정도를 표기할 때엔 유효숫자는 두 자리까지이다.
* 신뢰구간 같은 불확실성의 표기는 일관적이어야 한다. 그림 표현도 포함된다.
* 비열등시험처럼 단측검정을 한 경우가 아니라면 모든 P값은 양측검정의 값이어야 한다. 0.01이 넘는 P값은 소수점 두 자리까지 표기하며, 0.01과 0.001 사이의 값은 소수점 세 자리까지, 0.001 미만인 값은 $P < 0.001$로 표기해야 한다. 분석 단계에서 정지규칙(stopping rule)을 쓰는 경우, 유전자 스크리닝 연구의 경우는 예외로 둔다.
* 흔히 원고의 첫 표로 등장하는 무작위시험의 치료군들을 비교하는 표에서 군-간의 통계적으로 유의한 차이는($P < 0.05$) 표의 각주에서 표기해야 하고 정확한 P값의 표기는 본문에 해야 한다. 표 안에 P값을 적지 말라.
* 무작위임상시험의 원고라면 CONSORT 양식에 따른 플로차트를 그리거나 CONSORT 체크리스트에 포함된 정보를 포함해도 좋다. 원고의 양이 문제가 된다면 별첨할 수도 있다. CONSORT 진술서, 체크리스트, 플로차트는 당 웹사이트에서 참고할 수 있다.

* 신뢰구간 같은 불확실성의 표기는 일관적이어야 한다. 그림 표현도 포함된다.

* 신뢰구간 같은 불확실성의 표기는 일관적이어야 한다. 그림 표현도 포함된다.
* 비열등시험처럼 단측검정을 한 경우가 아니라면 모든 P값은 양측검정의 값이어야 한다. 0.01이 넘는 P값은 소수점 두 자리까지 표기하며, 0.01과 0.001 사이의 값은 소수점 세 자리까지, 0.001 미만인 값은 $P < 0.001$로 표기해야 한다. 분석 단계에서 정지규칙(stopping rule)을 쓰는 경우, 유전자 스크리닝 연구의 경우는 예외로 둔다.
* 흔히 원고의 첫 표로 등장하는 무작위시험의 치료군들을 비교하는 표에서 군-간의 통계적으로 유의한 차이는($P < 0.05$) 표의 각주에서 표기해야 하고 정확한 P값의 표기는 본문에 해야 한다. 표 안에 P값을 적지 말라.

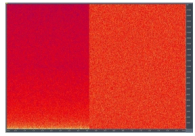
* 신뢰구간 같은 불확실성의 표기는 일관적이어야 한다. 그림 표현도 포함된다.
* 비열등시험처럼 단측검정을 한 경우가 아니라면 모든 P값은 양측검정의 값이어야 한다. 0.01이 넘는 P값은 소수점 두 자리까지 표기하며, 0.01과 0.001 사이의 값은 소수점 세 자리까지, 0.001 미만인 값은 $P < 0.001$로 표기해야 한다. 분석 단계에서 정지규칙(stopping rule)을 쓰는 경우, 유전자 스크리닝 연구의 경우는 예외로 둔다.
* 흔히 원고의 첫 표로 등장하는 무작위시험의 치료군들을 비교하는 표에서 군-간의 통계적으로 유의한 차이는 ($P < 0.05$) 표의 각주에서 표기해야 하고 정확한 P값의 표기는 본문에 해야 한다. 표 안에 P값을 적지 말라.
* 여러 개 군집 중에서 두 군집의 결과를 비교할 때 해당하는 통계 분포 안에서 반드시 제1 통계오류를 통제하여야 한다. 정규분포인 자료의 경우라면 t검정이 적절할 것이며, 정규분포가 아니더라도 표본수가 아주 크다면 t검정의 결과는 강건성을 가진다. 표본수가 작고 이전의 연구들을 통해 정규분포가 아니라고 알려진 자료라면 비모수통계법을 쓰는 것이 적절하다.

편집인과 심사위원
편집인과 저자

Scientists Perturbed by Loss of Stat Tools to Sift Research Fudge from Fact

The journal *Basic and Applied Social Psychology* recently banned the use of p-values and other statistical methods to quantify uncertainty from significance in research results

By Regina Nuzzo | April 16, 2015



Many researchers have labored under the misbelief that the p-value gives the probability that their study's results are just pure random chance.
Credit: Lentilucho/Wikipedia

Psychology researchers have recently found themselves engaged in a bout of statistical soul-searching. In apparently the first such move ever for a scientific journal the editors of *Basic and Applied Social Psychology* announced in a February editorial that researchers who submit studies for publication would not be allowed to use a common suite of statistical methods, including a controversial measure called the p-value.

These methods, referred to as null hypothesis significance testing, or NHST, are deeply embedded into the modern scientific research process, and some researchers have been left wondering where to turn. "The p-value is the most widely known statistic," says biostatistician Jeff Leek of Johns Hopkins University. Leek has estimated that the p-value has been used at least three million scientific papers. Significance testing is so popular that, as the journal editorial itself acknowledges, there are no widely accepted alternative ways to quantify the uncertainty in research results—and uncertainty is crucial for estimating how well a study's results generalize to the broader population.

Banning the P values [1]

"The null hypothesis significance testing procedure is logically invalid, and so it seems sensible to eliminate it from science," says psychologist David Trafimow of New Mexico State University in Las Cruces, editor of the journal.

[1] Regina Nuzzo. Scientists Perturbed by Loss of Stat Tools to Sift Research Fudge from Fact. Scientific American. April 16, 2015

Korean Journal of Anesthesiology

Effects of a fentanyl-propofol mixture on propofol injection pain: a randomized clinical trial

Nurcan Kizilcik, Ferdi Menda, Sevgi Bilgen, Ozgül Keskin, and Ozge Koner

Department of Anesthesiology and Reanimation, Yeditepe University School of Medicine, Istanbul, Turkey

Background: Propofol injection pain is a common problem that can be very distressing for patients. We compared the effects of injection with saline followed by injection with a fentanyl-propofol mixture, injection with fentanyl followed by a propofol injection, and injection with saline followed by propofol alone on propofol injection pain.

Methods: The patients were assigned randomly to one of three groups. A rubber tourniquet was placed on the forearm to produce venous occlusion for 1 min. Before anesthesia induction, group C (control, n = 50) and group M (fentanyl-propofol mixture, n = 50) received 5 ml of isotonic saline, while group F (fentanyl, n = 50) received 2 μg/kg of fentanyl. After the tourniquet was released, groups C and F received 5 ml of propofol and group M received 5 ml of a mixture containing 20 ml of propofol and 4 ml of fentanyl. At 10 s after the study drugs were given, a standard question about the comfort of the injection was asked of the patient. We used a verbal rating scale to evaluate propofol injection pain. Statistical analyses were performed with Student's t-tests and Fisher's exact tests; $P < 0.05$ was considered to indicate statistical significance.

Results: The demographic data were similar among the groups. In group M, the number of patients reporting propofol injection pain was significantly lower than in groups F and C (both $P < 0.001$). No patient in group F or M experienced severe pain, whereas 24 patients (48%) had severe pain in group C (both $P < 0.001$).

Conclusions: This study shows that a fentanyl-propofol mixture was more effective than fentanyl pretreatment or a placebo in preventing propofol injection pain.

Key Words: Fentanyl, Injection pain, Propofol.

P값은 정확히 무엇을 말하는가?

Pearson and Nyman said,

- With defining *null hypothesis*[2] *as well as an* alternative hypothesis [3], when $P < \alpha$, we can draw a conclusion that "the null hypothesis can be rejected beside of a luck."
- So, "P < 0.05" means exclusion of a luck, not any big size of effect.

[2] a hypothesis that there is no effect.

[3] a hypothesis that the effect is greater than zero.

왜 $P < 0.05$인가?

그런데 P는?

- 차이의 크기와 자유도(표본수)의 함수이다.
- 차이의 크기는 불변하므로[4] P값은 자유도(표본수)에 의해서 좌우된다.
- 그렇다면 P값은 다음처럼 해석되어야 한다: "$P < 0.05$는 내가 가진 표본으로 설명할 수 있다는 것"을 뜻한다.[5]

[4]일반적인 감기의 유병 기간이 x_1 일이라고 하고, 감기 신약을 복용한 환자의 유병기간이 x_2 일이라고 할 때, 차이$(x_1 - x_2)$는 제각기 고정되어 있다.

[5]내가 가진 표본수로 신약이 감기의 유병기간을 줄인다고 설명할 수 있다.

P value is difficult to interpret without statistical power. [6]

Significance testing is the most commonly used method for evaluating statistical hypotheses in genetic studies, including GWASs and exome sequencing studies. However, as described above, *P values are difficult to interpret without some consideration of statistical power*, as an insignificant test can result both from the absence of effect and from inadequate statistical power. Consideration of statistical power is therefore important not only in the design of efficient genetic studies but also in the interpretation of statistical findings.

[6] Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved; 2014;15:335–46.

아, P > 0.05. Still Not Significant

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

- (barely) not statistically significant (p=0.052)
- a barely detectable statistically significant difference (p=0.073)
- a borderline significant trend (p=0.09)
- a certain trend toward significance (p=0.08)
- a clear tendency to significance (p=0.052)
- a clear trend (p<0.09)
- a clear, strong trend (p=0.09)
- a considerable trend toward significance (p=0.069)
- a decreasing trend (p=0.09)
- a favourable statistical trend (p=0.09)
- a little significant (p<0.1)
- a margin at the edge of significance (p=0.0608)
- a marginal trend (p=0.09)
- a marginal trend toward significance (p=0.052)
- a marked trend (p=0.07) a mild trend (p<0.09) a moderate trend toward significance (p=0.068) a near-significant trend (p=0.07) a negative trend

표현 금지

귀무가설의 논리에는 기각과 기각 실패만 존재한다.
중간이나 근접 상황은 존재하지 않아서 다음의 표현들은 금지되어 있다.

1. "$\cdots$ showed nearly significant differences $\cdots$ $(P = 0.052)$"
2. "$\cdots$ showed trend toward significance $\cdots$ $(P = 0.07)$"
3. "$\cdots$ approached significance $\cdots$ $(P = 0.066)$"

정언명제

전칭긍정명제(H_0): A는 B이다.(모든 A는 B이다.)

전칭부정명제(H_1): A는 B가 아니다.(어떤 A도 B가 아니다.)

특칭긍정명제: 어떤 A는 B이다.
특칭부정명제: 어떤 A는 B가 아니다.

---

* 귀무가설의 기각은 '전칭긍정명제의 거부'를 뜻한다. 논리학에서 말하는 '전칭부정명제를 채택'하는 것이다.
* 하지만 기각의 실패는 '전칭부정명제를 거부하지 못했음'만을 뜻하므로 '채택한 것이 아니다,' 특칭긍정명제와 특칭부정명제의 가능성을 여전히 내포하고 있다.

채택의 논리

	표본을 통한 의사 결정	
	H_0 채택	H_0 기각
모집단의 사실 여부		
H_0 = true	correct	Type I error (α)
H_0 = false	Type II error (β)	correct

* 검정력 (power; $1 - \beta$) 을 0.8 이상으로 유지하기 위해서는 최소표본수 설정이 필수적이다.
* 최소표본수는 연구 고안 단계에서 ('방법'에 기술), 검정력은 분석 단계에서 ('결과'에 기술) 정립한다.

Anesthesia Awareness and the Bispectral Index

Michael S. Avidan, M.B., B.Ch., Lini Zhang, M.D., Beth A. Burnside, B.A., Kevin J. Finkel, M.D., Adam C. Searleman, B.S., Jacqueline A. Selvidge, B.S., Leif Saager, M.D., Michelle S. Turner, B.S., Srikar Rao, B.A., Michael Bottros, M.D., Charles Hantler, M.D., Eric Jacobsohn, M.B., Ch.B., and Alex S. Evers, M.D.

ABSTRACT

BACKGROUND

Awareness during anesthesia is a serious complication with potential long-term psychological consequences. Use of the bispectral index (BIS), developed from a processed electroencephalogram, has been reported to decrease the incidence of anesthesia awareness when the BIS value is maintained below 60. In this trial, we sought to determine whether a BIS-based protocol is better than a protocol based on a measurement of end-tidal anesthetic gas (ETAG) for decreasing anesthesia awareness in patients at high risk for this complication.

METHODS

We randomly assigned 2000 patients to BIS-guided anesthesia (target BIS range, 40 to 60) or ETAG-guided anesthesia (target ETAG range, 0.7 to 1.3 minimum alveolar concentration [MAC]). Postoperatively, patients were assessed for anesthesia awareness at three intervals (0 to 24 hours, 24 to 72 hours, and 30 days after extubation).

RESULTS

We assessed 967 and 974 patients from the BIS and ETAG groups, respectively. Two cases of definite anesthesia awareness occurred in each group (absolute difference, 0%; 95% confidence interval [CI], −0.56 to 0.57%). The BIS value was greater than 60 in one case of definite anesthesia awareness, and the ETAG concentrations were less than 0.7 MAC in three cases. For all patients, the mean (±SD) time-averaged ETAG concentration was 0.81±0.25 MAC in the BIS group and 0.82±0.23 MAC in the ETAG group (P=0.10; 95% CI for the difference between the BIS and ETAG groups, −0.04 to 0.01 MAC).

From the Department of Anesthesiology, Washington University School of Medicine, St. Louis. Address reprint requests to Dr. Avidan at Washington University School of Medicine, 660 S. Euclid Ave., Campus Box 8054, St. Louis, MO 63110, or at avidanm@wustl.edu.

N Engl J Med 2008;358:1097-108.

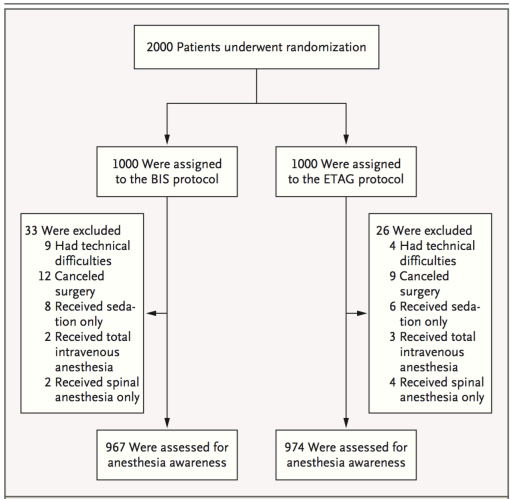


Figure 1. Trial Enrollment.

BIS denotes bispectral index, and ETAG end-tidal anesthetic gas.

On the basis of the accounts given by the patients and the information in the anesthesia records, an investigator who was unaware of the

included in the maintenance period. Every trace was analyzed for sustained 30-second periods of BIS values above the threshold of 60 or ETAG concentrations below the threshold of 0.7 MAC during the maintenance period. Periods with missing data were excluded from the analysis.

STATISTICAL ANALYSIS

The primary outcome of the study was a decrease in definite anesthesia awareness in the BIS group as compared with the ETAG group. The anticipated incidence of anesthesia awareness was 1% for the ETAG group, on the basis of the incidence rates reported for patients at high risk for anesthesia awareness,[3-5] and 0.1% for the BIS group, on the basis of previous studies.[3,21] A total of 940 patients would be required in each group to detect this 0.9% difference with a one-tailed alpha of 0.05 and a power of 80% with the use of Fisher's exact test. Confidence intervals for absolute risk reduction were calculated with the use of Newcombe's method without continuity correction.[22] There was no interim analysis. The chi-square test, Fisher's exact test, an unpaired t-test, and an unpaired Mann–Whitney test were used for other comparisons between groups. Intention-to-treat analysis was planned. Agreement among the experts who were assessing anesthesia awareness was quantified with the use of a two-way, ran-

74.5% of patients who did not have anesthesia awareness. The low mean BIS values in the BIS group could reflect the unwillingness of the an- the protocols.

This trial has some important limitations. Although the trial did not demonstrate a reduction

in anesthesia awareness, with 95% confidence intervals for absolute risk reduction of definite anesthesia awareness of −0.56 to 0.57%, the results remain consistent with a clinically significant number needed to treat in order to benefit of 179 and a clinically significant number needed to treat in order to harm of 175 with the BIS protocol. This study is also subject to some concerns common to all studies of anesthesia awareness: the diagnosis of anesthesia awareness may be subjective, the awareness interview may be invalid because repeated questioning may induce false memories, and it may be difficult to distinguish between memories of events in the operating room and events in the intensive care unit. It is encouraging that there was good agreement among the three assessors, who were unaware of the treatment assignments, and it was unnecessary to refer any decision to a fourth assessor.

Anesthesia awareness cannot predictably be prevented in all patients with the BIS monitoring protocol used in this study. When a potent volatile anesthetic gas was administered, a structured protocol based on the BIS was not shown to be superior to a protocol based on ETAG concentrations for preventing anesthesia awareness. Reliance on BIS technology[24] may provide patients and health care practitioners with a false sense of security about the reduction in the risk of anesthesia awareness. If BIS monitoring were routinely applied to all patients in the United States receiving general anesthesia,[7] the cost of disposable electrodes alone would exceed $360 million annually. Our study was unable to demonstrate superiority of a BIS-guided protocol over an ETAG-guided protocol for preventing anesthesia awareness and does not provide support for the additional cost of BIS monitoring as part of standard

Capnography During Deep Sedation with Propofol by Nonanesthesiologists: A Randomized Controlled Trial [7]

RESULTS: From April 2010 to January 2011, 427 patients were enrolled. In the capnography group, 206 patients and in the standard care group, 209 patients were analyzed. The percentage of patients with a hypoxemic episode was 25.7% (53 of 206) in the capnography group and 24.9% (52 of 209) in the standard care group, resulting in an absolute difference of 0.8% (−7.5 to 9.2%).
CONCLUSIONS: We were unable to confirm an additive role for capnography in preventing hypoxemia during elective nonanesthesiologist-administered propofol (monotherapy) sedation in healthy women in whom supplemental oxygen is not routinely administered. Based on the confidence interval, the benefit of adding capnography is at most an absolute hypoxemia reduction of 7.5%, suggesting that adding it in this practice setting to the routine monitoring strategy does not necessarily improve patient safety in daily practice.

[7] van Loon K, van Rheineck Leyssius AT, van Zaane B, Denteneer M, Kalkman CJ. Capnography During Deep Sedation with Propofol by Nonanesthesiologists: A Randomized Controlled Trial. Anesth Analg. 2014;119.

Capnography During Deep Sedation with Propofol by Nonanesthesiologists: A Randomized Controlled Trial [8]

RESULTS: From April 2010 to January 2011, 427 patients were enrolled. In the capnography group, 206 patients and in the standard care group, 209 patients were analyzed. The percentage of patients with a hypoxemic episode was 25.7% (53 of 206) in the capnography group and 24.9% (52 of 209) in the standard care group, resulting in an absolute difference of 0.8% (−7.5 to 9.2%).
CONCLUSIONS: We were unable to confirm an additive role for capnography in preventing hypoxemia during elective nonanesthesiologist-administered propofol (monotherapy) sedation in healthy women in whom supplemental oxygen is not routinely administered. Based on the confidence interval, the benefit of adding capnography is at most an absolute hypoxemia reduction of 7.5%, suggesting that adding it in this practice setting to the routine monitoring strategy does not necessarily improve patient safety in daily practice.

[8] van Loon K, van Rheineck Leyssius AT, van Zaane B, Denteneer M, Kalkman CJ. Capnography During Deep Sedation with Propofol by Nonanesthesiologists: A Randomized Controlled Trial. Anesth Analg. 2014;119.

Research Article

Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think

Statistical Reform Lessons From Medicine

Fiona Fidler,[1,2] Neil Thomason,[2] Geoff Cumming,[1] Sue Finch,[1] and Joanna Leeman[1]

[1]La Trobe University, Melbourne, Australia, and [2]The University of Melbourne, Melbourne, Australia

ABSTRACT—*Since the mid-1980s, confidence intervals (CIs) have been standard in medical journals. We sought lessons for psychology from medicine's experience with statistical reform by investigating two attempts by Kenneth Rothman to change statistical practices. We examined 594* American Journal of Public Health *(AJPH) articles published between 1982 and 2000 and 110* Epidemiology *articles published in 1990 and 2000. Rothman's editorial instruction to report CIs and not* p *values was largely effective: In* AJPH*, sole reliance on* p *values dropped from 63% to 5%, and CI reporting rose from 10% to 54%;* Epidemiology *showed even stronger compliance. However, compliance was superficial: Very few authors referred to CIs when discussing results. The results of our survey support what other research has indicated: Editorial policy alone is not a sufficient mechanism for statistical reform. Achieving substantial, desirable change will require further guidance regarding* seek lessons for psychology. We investigated reporting of confidence intervals (CIs) and effect size, because these are highly desirable and now recommended in psychology (American Psychological Association, APA, 2001).

LITTLE INTERDISCIPLINARY DISCUSSION OF NHST

In their 1970 anthology, *The Significance Test Controversy*, Morrison and Henkel noted that there had been little interdisciplinary exchange concerning NHST issues, even between closely related disciplines. Their book contained chapters from both psychologists and sociologists. Until that time, scrutiny of NHST in the two disciplines had been "parallel but quite independent" (Morrison & Henkel, 1970b, p. 182). Unfortunately, there has been little exchange since.

In the mid-1990s, the APA and the American Psychological Society held symposia to discuss banning NHST from psychology journals.

교육, 편집인의 의무[9]

Editorial policy alone is not a sufficient mechanism for statistical reform. Achieving substantial, desirable change will require further guidance regarding use and interpretation of CIs and appropriate effect size measures.

[9] Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think: Statistical Reform Lessons From Medicine. Psychological Science. 2004;15:119–26.

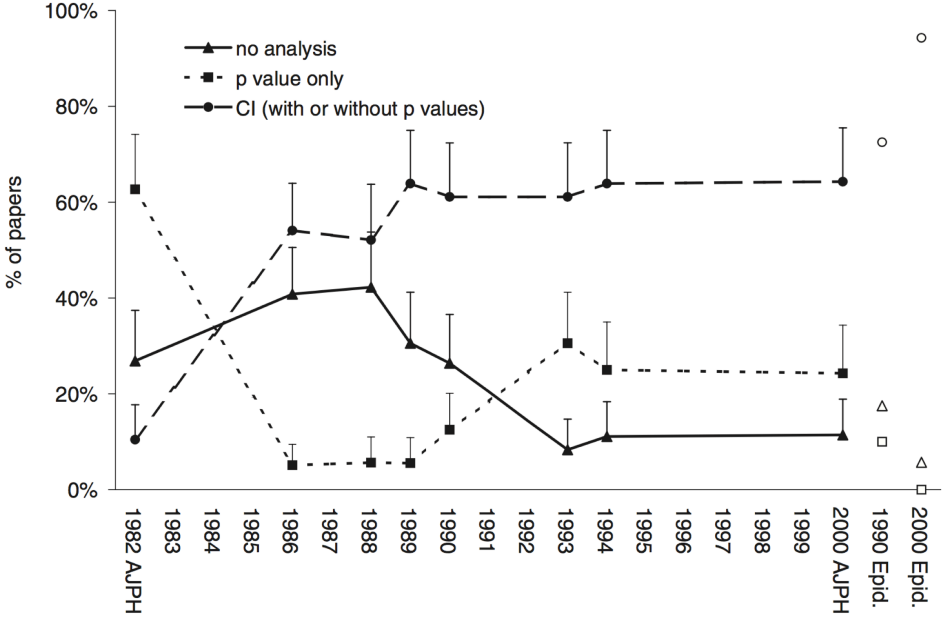


Fig. 1. Percentage of *American Journal of Public Health* (*AJPH*) and *Epidemiology* (*Epid.*) articles reporting only *p* values (and no confidence intervals, CIs), at least one CI, or neither. Error bars are upper-half 95% CIs. The values for *Epidemiology* are shown by open symbols corresponding to the closed symbols for *AJPH*.

정력적인 편집인[10]

"All references to statistical hypothesis testing and statistical significance should be removed from the paper. *I ask that you delete p values as well as comments about statistical significance*. If you do not agree with my standards (concerning the inappropriateness of significance tests), you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere."

* American Journal of Public Health: 19th in the catogory 'Public Health' of SCImago

[10] Kenneth Rothman, Vice-Editor, American Journal of Public Health, 1986

Expressing one of CI or P, or Both in NEJM

927 Articles Published in 2008-2013

	CI	Both	P	Nothing	Sum
2008	14 (12%)	33 (28%)	48 (41%)	23 (19%)	118
2009	19 (13%)	39 (26%)	50 (33%)	42 (28%)	150
2010	12 (10%)	37 (30%)	47 (38%)	29 (23%)	125
2011	14 (9%)	52 (32%)	53 (32%)	45 (27%)	164
2012	17 (9%)	59 (33%)	71 (40%)	32 (18%)	179
2013	25 (13%)	53 (28%)	69 (36%)	44 (23%)	191

피인용 수[11] vs. CI 포함 여부

927 Articles Published in 2008-2013

Year	no CI	CI	차이 (95% CI)
2008	441 (362)	516 (404)	(-217, 135)
2009	322 (342)	587 (448)	(-252, 41)
2010	288 (300)	342 (287)	(-100, 125)
2011	536 (267)	542 (250)	(-152, 186)
2012	147 (102)	184 (154)	(-101, -3)
2013	144 (129)	151 (148)	(-61, 24)

[11] No. of citation dataset was retrieved from Web of Science at 3-Dec-2015.

# $P < 0.05$ as against a luck

Luck accumulates when comparisons repeat.

So, limit of luck ($\alpha = 0.05$) should be re-adjusted.[12]

[12] Or P should be re-adjusted.

# $P < 0.05$ as against a luck

Luck accumulates when comparisons repeat.

- So, limit of luck ($\alpha = 0.05$) should be re-adjusted.
- If comparisons repeat three times, α should be, as one of easiest adjustment technique, divided by three $= \frac{0.05}{3} = 0.017$.[13]

[13] Or, *P* is multiplied by 3

# $P < 0.05$ as against a luck

Luck accumulates when comparisons repeat.

- So, limit of luck ($\alpha = 0.05$) should be re-adjusted.
- If comparisons repeat three times, α should be, as one of easiest adjustment technique, divided by three $= \frac{0.05}{3} = 0.017$.[14]
- Always maintain the overall $\alpha = 0.05$.

[14] Or, *P* is multiplied by 3

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 23, 2006 VOL. 354 NO. 8

Glucosamine, Chondroitin Sulfate, and the Two in Combination for Painful Knee Osteoarthritis

Daniel O. Clegg, M.D., Domenic J. Reda, Ph.D., Crystal L. Harris, Pharm.D., Marguerite A. Klein, M.S., James R. O'Dell, M.D., Michele M. Hooper, M.D., John D. Bradley, M.D., Clifton O. Bingham III, M.D., Michael H. Weisman, M.D., Christopher G. Jackson, M.D., Nancy E. Lane, M.D., John J. Cush, M.D., Larry W. Moreland, M.D., H. Ralph Schumacher, Jr., M.D., Chester V. Oddis, M.D., Frederick Wolfe, M.D., Jerry A. Molitor, M.D., David E. Yocum, M.D., Thomas J. Schnitzer, M.D., Daniel E. Furst, M.D., Allen D. Sawitzke, M.D., Helen Shi, M.S., Kenneth D. Brandt, M.D., Roland W. Moskowitz, M.D., and H. James Williams, M.D.

ABSTRACT

BACKGROUND

Glucosamine and chondroitin sulfate are used to treat osteoarthritis. The multicenter, double-blind, placebo- and celecoxib-controlled Glucosamine/chondroitin Arthritis Intervention Trial (GAIT) evaluated their efficacy and safety as a treatment for knee pain from osteoarthritis.

METHODS

We randomly assigned 1583 patients with symptomatic knee osteoarthritis to receive 1500 mg of glucosamine daily, 1200 mg of chondroitin sulfate daily, both glucosamine and chondroitin sulfate, 200 mg of celecoxib daily, or placebo for 24 weeks. Up to 4000 mg of acetaminophen daily was allowed as rescue analgesia. Assignment was stratified according to the severity of knee pain (mild [N=1229] vs. moderate to severe [N=354]). The primary outcome measure was a 20 percent decrease in knee pain from baseline to week 24.

RESULTS

The mean age of the patients was 59 years, and 64 percent were women. Overall, glucosamine and chondroitin sulfate were not significantly better than placebo in reducing knee pain by 20 percent. As compared with the rate of response to placebo (60.1 percent), the rate of response to glucosamine was 3.9 percentage points higher (P=0.30), the rate of response to chondroitin sulfate was 5.3 percentage points higher (P=0.17), and the rate of response to combined treatment was 6.5 percentage points higher (P=0.09). The rate of response in the celecoxib control group was 10.0 percent-

From the University of Utah School of Medicine, Salt Lake City (D.O.C., C.G.J., A.D.S., H.J.W.); the Hines Veterans Affairs Cooperative Studies Program Coordinating Center, Hines, Ill. (D.J.R., H.S.); the Clinical Research Pharmacy Coordinating Center, Albuquerque, N.M. (C.L.H.); the National Center for Complementary and Alternative Medicine, National Institutes of Health, Bethesda, Md. (M.A.K); the University of Nebraska Medical Center, Omaha (J.R.O.); Case Western Reserve University, Cleveland (M.M.H., R.W.M.); the Indiana University School of Medicine, Indianapolis (J.D.B., K.D.B.); the Hospital for Joint Diseases, Rheumatology and Medicine, New York (C.O.B.); Cedars-Sinai Medical Center, Los Angeles (M.H.W.); the University of California, San Francisco (N.E.L.); the Presbyterian Hospital of Dallas, Dallas (J.J.C.); the University of Alabama, Birmingham (L.W.M.); the Hospital of the University of Pennsylvania, Philadelphia (H.R.S.); the University of Pittsburgh (C.V.O.); the Arthritis Research and Clinical Centers, Wichita, Kans. (F.W.); the Virginia Mason Medical Center

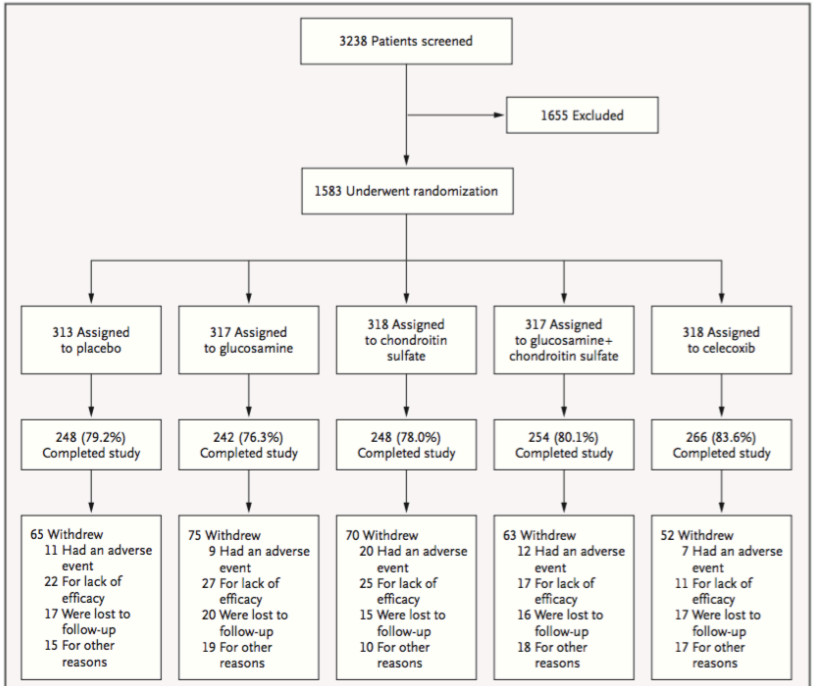


Figure 1. Enrollment and Outcomes.

ment for osteoarthritis.[17] A response was classified as an improvement in pain or function of at least 50 percent and a decrease of at least 20 mm on the visual-analogue scale for pain or function or the occurrence of at least two of the following: a decrease in pain of at least 20 percent and at least 10 mm on the visual-analogue scale; an improvement in function of at least 20 percent and a decrease of at least 10 mm on the visual-analogue scale; and an increase in the patient's global assessment score by at least 20 percent and at least 10 mm on the visual-analogue scale. Since we prospectively collected data on each component, the OMERACT–OARSI response rate is also reported.

PRODUCT SELECTION

Our study was conducted under an investigational new drug application, and the study agents were subject to pharmaceutical regulation by the Food and Drug Administration (FDA). The Cooperative Studies Program Clinical Research Pharmacy Coordinating Center, a facility licensed by the FDA, used a vendor-certification program to evaluate available commercial products and raw materials in order to select the suppliers of glucosamine and chondroitin sulfate. Donated or purchased ingredients were tested for purity, potency, and quality. Certificates of analysis were obtained for the agents, and Drug Master Files were on file with the FDA. Capsules containing 250 mg of glucosamine hydrochloride, 200 mg of sodium chondroitin sulfate, the two in combination, and matching placebo were manufactured, distributed, and placed on a shelf-life–stability program throughout the study at the Pharmacy Coordinating Center. In addition, 200-mg capsules of celecoxib after an overnight fast. In patients with diabetes at enrollment, fasting blood glucose and glycosylated hemoglobin levels were monitored. A test for fecal occult blood (Hemoccult, Beckman Coulter) was performed at the visit at week 24. Medication was withdrawn from patients in whom diabetes or gastrointestinal bleeding developed, and the patients were referred for further evaluation.

STATISTICAL ANALYSIS

An absolute increase in the response rate of 15 percent, as compared with the rate in the placebo group, was considered to indicate a clinically meaningful treatment effect. We estimated that 1588 patients would need to be enrolled to provide the study with a statistical power of 85 percent to detect one or more clinically meaningful differences between the placebo group and the glucosamine group, the chondroitin sulfate group, and the combined-treatment group, assuming a rate of response of 35 percent in the placebo group and a withdrawal rate of 20 percent. Pairwise comparisons of the glucosamine group, the chondroitin sulfate group, and the combined-treatment group with the placebo group were made with the use of a two-sided chi-square test with an α value of 0.017 for each comparison (overall α value, 0.05). A side comparison between celecoxib and placebo also used an α value of 0.017. The data and safety monitoring board reviewed study performance and safety data annually but did not conduct interim monitoring of the primary outcome. Analysis of the primary outcome measure was conducted according to the intention to treat.

Analyses of the secondary outcome measures followed the pairwise-comparison plan described

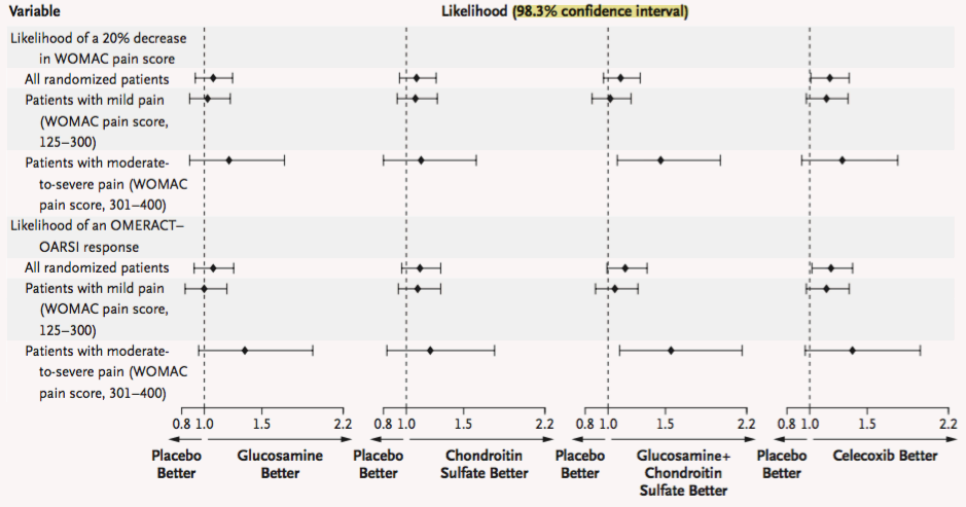


Figure 2. Pairwise Comparisons of the Overall Likelihood of a Response.

Scores for the pain subscale of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) can range from 0 to 500, with higher scores indicating more pain. A response according to the guidelines of the Outcome Measures in Rheumatology Clinical Trials and Osteoarthritis Research Society International (OMERACT–OARSI) was classified as an improvement in function or pain of at least 50 percent and a decrease of at least 20 mm on the visual-analogue scale for pain or function or the occurrence of at least two of the following: a decrease in pain of at least 20 percent and at least 10 mm on the visual-analogue scale; an improvement in function of at least 20 percent and a decrease of at least 10 mm on the visual-analogue scale; and an increase in the patient's global assessment score by at least 20 percent and at least 10 mm on the visual-analogue scale.

$P = 0.xx$ or $P = 0.xxx$

0.01이 넘는 P값은 소수점 두 자리까지 표기하며, 0.01과 0.001 사이의 값은 소수점 세 자리까지, 0.001 미만인 값은 $P < 0.001$로 표기해야 한다. 분석 단계에서 정지규칙(stopping rule)을 쓰는 경우, 유전자 스크리닝 연구의 경우는 예외로 둔다.

- 소수점 아래 두 자리까지 적는다.
- 소수점 아래 세 자리까지 when $P < 0.01$ and $P > 0.001$
- $P < 0.001$ when $P = 0.000x$.

비모수검정 (nonparametric test)

정규분포인 자료의 경우라면 t검정이 적절할 것이며, 정규분포가 아니더라도 표본수가 아주 크다면 t검정의 결과는 강건성을 가진다. 표본수가 작고 이전의 연구들을 통해 정규분포가 아니라고 알려진 자료라면 비모수통계법을 쓰는 것이 적절하다.

모수	비모수
t test	Mann-Whitney U test
paired t test	Wilcoxon signed rank test
ANOVA	Kruskal-Wallis test
RMANOVA	Friedman test
Pearson's product moment correlation	Spearman's rank correlation

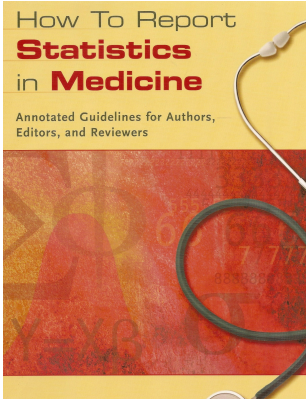
비모수검정 (nonparametric test)

표현

비모수검정으로 얻은 통계 결과는 모수적으로 표현될 수 없다. 즉, 평균과 표준편차로 표현될 수 없다.

How to Report Statistics in Medicine

Tom Lang & Michelle Secic, American College of Physicians, 2006, Philadelphia.



1st Edition in 1997 (367 pages)
2nd Edition in 2006 (490 pages)

Uniform Requirement for Manuscripts · · ·

International Committee of Medical Journal Editors.

Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication

Updated April 2010

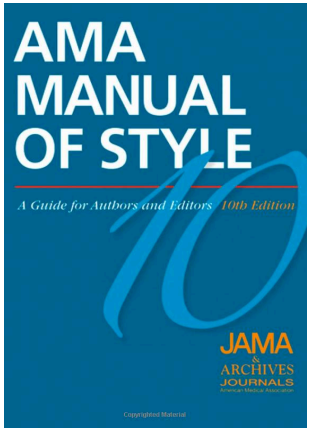
Publication Ethics: Sponsorship, Authorship, and Accountability

International Committee of Medical Journal Editors

The following information is available to be viewed/printed in Adobe Acrobat pdf format.

AMA Manual of Style

American Medical Associations



Oxford University Press, 10th edition, 2007
(1032 pages)

유효숫자

관찰값은 측정된 정밀도, 유효숫자만큼만 기재한다.

유효숫자

관찰값은 측정된 정밀도, 유효숫자만큼만 기재한다.

$\cdots$ muscular relaxation reached maximum at
$3\frac{3}{4}$ minutes from intravenous administration of $\cdots$

유효숫자 하나 더 - 관찰값으로부터 유도된 값

* 1 mmHg까지 측정된 혈압

$\cdots$ Mean blood pressure at 3 minutes after endotracheal intubation was 75.4 (12.2) mmHg , $\cdots$

* 1 mmHg까지 측정된 CO_2 농도

$\cdots$ The difference between the ET_{CO_2} and Pa_{CO_2} was 10.4 ± 8.9 mmHg in group Y and 4.6 ± 3.9 mmHg in group O ($P < 0.05$). $\cdots$

유효숫자의 포기 - 지나친 정밀도는 독자들을 혼란시킨다

1. Cardiac out increased after aortic valvular replacement from 2.412 to 3.137 .
2. Cardiac out increased after aortic valvular replacement from 2.41 to 3.14 .
3. Cardiac out increased after aortic valvular replacement from 2.4 to 3.1 .

빈도, 비율

* 제1 원칙 : 분자(관찰수)와 분모(표본수)를 모두 기재한다.
* 제2 원칙 : 표본수가 20 이하일 때에는 백분율을 계산하지 않는다.

Of the 18 patients having malignant tumors, only 3 died.
Here,

1. The ratio of dead to alive was 3:18 in patients having malignant tumors.
2. The proportion of death is 3/18 in $\cdots$
3. Death occurred in 3 of 18 patients.

명백한 통계 용어

형용사 significant는 언제나 '통계적으로 유의한 차이를 보이는'의 뜻으로만 써야 한다. 다음처럼 그저 '중요한'으로 써서는 안 된다.

This finding is significant in everyday anesthesia.

이런 단어 중에는 'random', 'normal', 'correlation', 'sample'이 더 있다.

표

제목 - 구체적일수록 좋다

Table 1. Patient Characteristics

표

제목 - 구체적일수록 좋다

Table 1. Patient Characteristics

Table 1. Characteristics of 567 Patients Having Malignant Melanoma

표

제목 - 구체적일수록 좋다

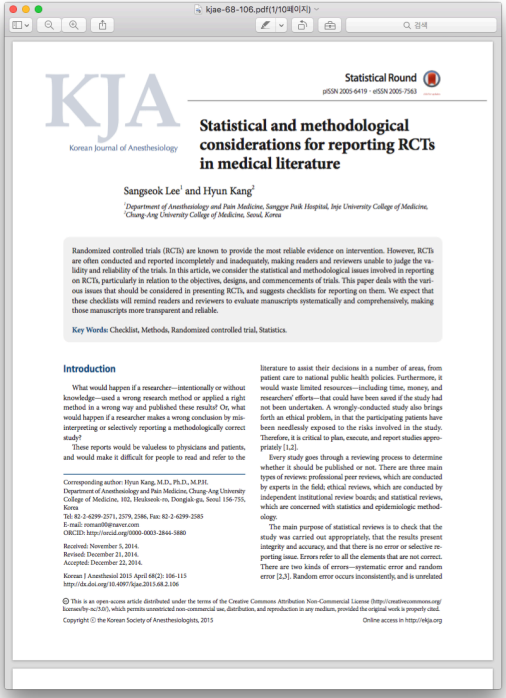
Table 2. Causes of Death

표

제목 - 구체적일수록 좋다

Table 2. Causes of Death

Table 2. Leading Causes of Cancer Death in Korea, 2008



KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 · eISSN 2005-7563

Statistical and methodological considerations for reporting RCTs in medical literature

Sangseok Lee[1] and Hyun Kang[2]

[1]*Department of Anesthesiology and Pain Medicine, Sanggye Paik Hospital, Inje University College of Medicine,*
[2]*Chung-Ang University College of Medicine, Seoul, Korea*

Randomized controlled trials (RCTs) are known to provide the most reliable evidence on intervention. However, RCTs are often conducted and reported incompletely and inadequately, making readers and reviewers unable to judge the validity and reliability of the trials. In this article, we consider the statistical and methodological issues involved in reporting on RCTs, particularly in relation to the objectives, designs, and commencements of trials. This paper deals with the various issues that should be considered in presenting RCTs, and suggests checklists for reporting on them. We expect that these checklists will remind readers and reviewers to evaluate manuscripts systematically and comprehensively, making those manuscripts more transparent and reliable.

Key Words: Checklist, Methods, Randomized controlled trial, Statistics.

Introduction

What would happen if a researcher—intentionally or without knowledge—used a wrong research method or applied a right method in a wrong way and published these results? Or, what would happen if a researcher makes a wrong conclusion by misinterpreting or selectively reporting a methodologically correct study?

These reports would be valueless to physicians and patients, and would make it difficult for people to read and refer to the literature to assist their decisions in a number of areas, from patient care to national public health policies. Furthermore, it would waste limited resources—including time, money, and researchers' efforts—that could have been saved if the study had not been undertaken. A wrongly-conducted study also brings forth an ethical problem, in that the participating patients have been needlessly exposed to the risks involved in the study. Therefore, it is critical to plan, execute, and report studies appropriately [1,2].

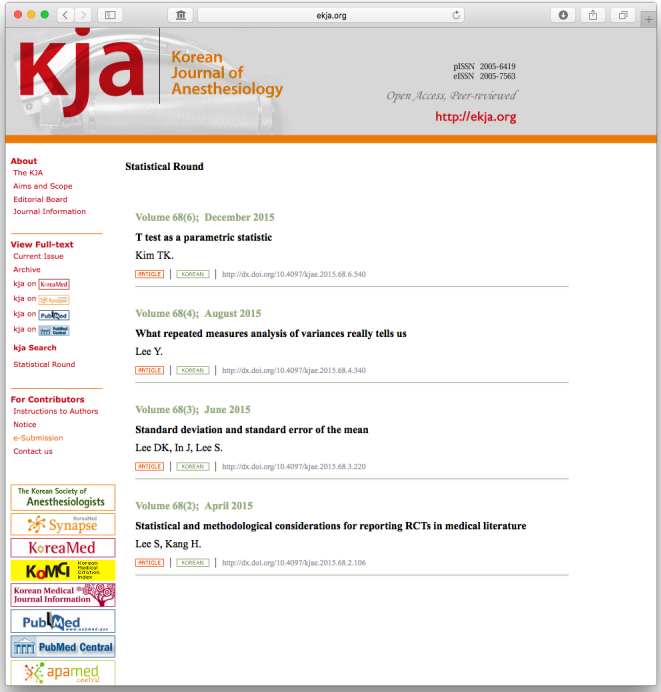
Every study goes through a reviewing process to determine whether it should be published or not. There are three main types of reviews: professional peer reviews, which are conducted by experts in the field; ethical reviews, which are conducted by independent institutional review boards; and statistical reviews, which are concerned with statistics and epidemiologic methodology.

The main purpose of statistical reviews is to check that the study was carried out appropriately, that the results present integrity and accuracy, and that there is no error or selective reporting issue. Errors refer to all the elements that are not correct. There are two kinds of errors—systematic error and random error [2,3]. Random error occurs inconsistently, and is unrelated

Corresponding author: Hyun Kang, M.D., Ph.D., M.P.H.
Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine, 102, Heukseok-ro, Dongjak-gu, Seoul 156-755, Korea
Tel: 82-2-6299-2571, 2579, 2586, Fax: 82-2-6299-2585
E-mail: romarc00@naver.com
ORCID: http://orcid.org/0000-0003-2844-5880

Received: November 3, 2014.
Revised: December 21, 2014.
Accepted: December 22, 2014.

Korean J Anesthesiol 2015 April 68(2): 106-115
http://dx.doi.org/10.4097/kjae.2015.68.2.106


ekja.org
Korean Journal of Anesthesiology
pISSN 2005-6419
eISSN 2005-7563
Open Access, Peer-reviewed
http://ekja.org
About
The KJA
Aims and Scope
Editorial Board
Journal Information
View Full-text
Current Issue
Archive
kja on KoreaMed
kja on Synapse
kja on PubMed
kja on PubMed Central
kja Search
Statistical Round
For Contributors
Instructions to Authors
Notice
e-Submission
Contact us
The Korean Society of Anesthesiologists
Annotated Synapse
KoreaMed
KoMCI
Korean Medical Journal Information
PubMed
PubMed Central
apamed central
Statistical Round
Volume 68(6); December 2015
T test as a parametric statistic
Kim TK.
ARTICLE | KOREAN | http://dx.doi.org/10.4097/kjae.2015.68.6.540
Volume 68(4); August 2015
What repeated measures analysis of variances really tells us
Lee Y.
ARTICLE | KOREAN | http://dx.doi.org/10.4097/kjae.2015.68.4.340
Volume 68(3); June 2015
Standard deviation and standard error of the mean
Lee DK, In J, Lee S.
ARTICLE | KOREAN | http://dx.doi.org/10.4097/kjae.2015.68.3.220
Volume 68(2); April 2015
Statistical and methodological considerations for reporting RCTs in medical literature
Lee S, Kang H.
ARTICLE | KOREAN | http://dx.doi.org/10.4097/kjae.2015.68.2.106

For Contributors
Instructions to Authors
Notice
e-Submission
Contact us