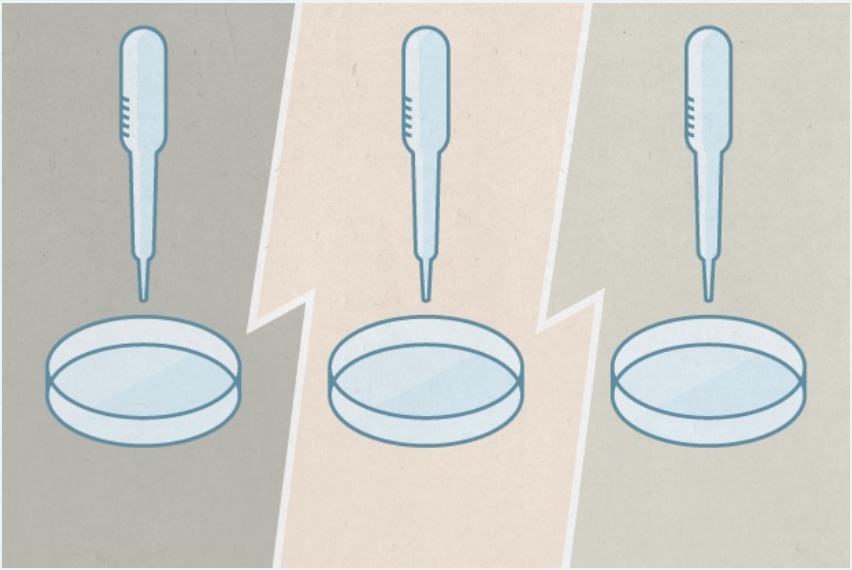
dongguk UNIVERSITY

이윤석

동국의대 의료원(고양) 마취통증의학과

논문 작성 시 흔히 보는 통계 오류

» *2017* 의편협 논문 아카데미



과학논문의 재현불가(irreproducibility)

(1) 결과에 대한 박약한 숙고, (2) 연구디자인의 복잡화, (3) 통계분석의 복잡화, (4) 연구성과에 대한 압박

주된 쟁점

- 모수검정과 비모수검정
- 자료의 독립성 문제(짝지은 자료인가 아닌가)
- P값 해석

모수검정(parametric test)의 조건

표본의 측정 단위가 간격척도(interval scale)이고, 정규분포인 모집단으로부터 추출됐음이 ('정규성') 전제되어야 한다.

정규성검정(normality test)[1]

- Kolmogorov-Smirnov test
- Shapiro-Wilks test
-
- Graphing technique (QQ plot)

[1]이 검정의 귀무가설은 '표본의 분포는 정규분포와 같다,' 즉 test $P < \alpha$에서 귀무가설을 기각한다(=정규분포와 다르다).

비모수검정 (nonparametric test)

요약

- 표본의 분포에 관한 특수한 분포가 전제되어 있지 않고,
- 관찰된 값 그 자체로부터가 아니라 변환된 자료 (순위 또는 빈도) 로 통계량을 계산한다.
- 모수 정보 (관찰된 값 그 자체) 가 손실되므로 모수검정의 조건이 만족되는 경우에는 가급적 피한다.
- 모수자료와 비모수자료는 서로 변환할 수 있지만 정보의 손실은 불가피하다.

두 개의 평균(또는 중간값)

Propofol 군에 20명의 피험자, Thiopental 군에 20명의 피험자를 배치하여 지정된 약제 투여 후 혈압과 삽관환경(1–4)을[2] 측정하였다.

	Propofol	Thiopental	
Mean Blood Pressure (mmHg)	94.1 (5.9)	96.2 (14.5)	t-test
Median Intubation Condition (IQR)	1 (0)	2 (1)	Mann-Whitney test

[2]Intubation Condition: 1 = excellent, 2, 3, and 4 = worst

두 개의 평균(또는 중간값): 짝지은(=비독립적인) 자료의 문제

20명의 단일군 피험자에게 esmolol 투여 전후의 심박수와 진정지수를[3] 측정하였다.

	Esmolol 투여 전	Esmolol 투여 후	
Heart rate (bpm)	98 (11)	76 (5)	paired t-test
Median Sedation Score (IQR)	3 (2)	2 (1)	Wilcoxon Signed Rank test

[3]Sedation Score: 0 = unawakable to 4 = spontaneous eye opening

눈에 잘 보이지 않는 짝지은 자료

짝지은 자료는 비단 시간적일(temporal) 뿐 아니라 공간적(spatial), 방법론적일(methodological) 수 있다.

공간적으로(spatially) 짝지은 자료

50명의 피험자에서 왼팔과 오른팔에서 측정한 수축기혈압을 비교한다.

	Left Arm	Right Arm
SBP (mmHg)	132.6 ± 14.3	130.7 ± 14.5

방법론적으로(methodologically) 짝지은 자료

20명의 피험자에서 lighted stylet을 써서 기관삽관을 할 때 경추의 신전 각도를 측정한 뒤, 다시 안정시킨 후 일반후두경을 이용하여 기관삽관을 할 때 경추의 각도를 측정하여 비교한다.

	Lighted Stylet	Laryngoscopy
Occiput-C1 Motion (°)	5.5 (4.3)	9.7(4.5)

짝지음(반복측정)의 최소단위

- 일반적인 의학연구에서 짝지음의 최소단위는 '피험자'이다.
- 연구의 환경에 따라서 최소단위가 '벤치'이거나 '제품 일련번호 묶음'이거나 '의료기관'이 될 수도 있다.
- 짝지음의 단위 내에서 발생하는 분산을 '개체-내 분산'으로 부르며, 필요에 따라 추리통계의 대상이 될 수도, 정규화의 대상으로 보아 통제될 수도 있다.
- 연구의 환경, 분석의 목표 등에 맞추어 최소단위를 일치시켜서 해석해야 심사자와 편집인을 만족시킬 수 있다.

세 개 이상의 평균(중간값)의 비교

이전의 예제에서 Propofol 군, Thiopental 군 외에 Etomidate 군이 추가된 연구이다.

	Propofol	Thiopental	Etomidate	
Mean Blood Pressure (mmHg)	94.1 (5.9)	96.2 (14.5)	101.1 (4.3)	ANOVA
Median Intubation Condition (IQR)	1 (0)	2 (1)	2 (2)	Kruskal-Wallis test

ANOVA 결과 P값이 유의하게 계산되었다면?

전체 비교의 P값이다. 즉 Propofol vs. Thiopental, 또는 Propofol vs. Etomidate, 또는 Thiopental vs. Etomidate 가운데 어느 한 조합 이상이 유의한 차이를 보인다는 뜻이다.

세 개 이상의 평균(중간값)의 비교 2

구체적으로 어느 조합에서 차이가 있는지 알고자 한다면 다중비교(multiple comparisons)를 통한다.

다중비교의 전략은 간단하다.

1. 비교하고자 하는 조합을 애초의 실험설계에서 선언해야 하고,
2. 목표로 하는 유의도(α; 흔히 0.05)를 조합의 개수(n)로 나누어 보정된 유의도를 설정한다. ($\alpha_{adj} = \alpha/n$)
3. 다중비교로 나온 P값과 보정된 유의도를 비교한다. $P < \alpha_{adj}$?
4. 95% 신뢰구간을 계산할 때도 보정된 유의도의 값을 취한다.

세 개 이상의 평균(중간값)의 비교: 비독립적인 측정의 경우

여기서도 시간적, 공간적, 방법론적인 반복측정이 모두 적용된다.

시간적인 반복측정 자료

	Baseline	1 month	5 months	
Potassium (mEq/L)	5.5 (0.7)	4.9 (0.8)	4.5 (0.3)	RMANOVA
Pain Score by 11-NRS	8 (2)	6 (3)	3 (2)	Friedman test

공간적인 반복측정 자료

	Radial artery	Femoral artery	Dorsalis pedis
MBP (mmHg)	83.2 (3.8)	100.4 (4.9)	92.4 (10.0)

반복측정분산분석(RMANOVA)의 전제

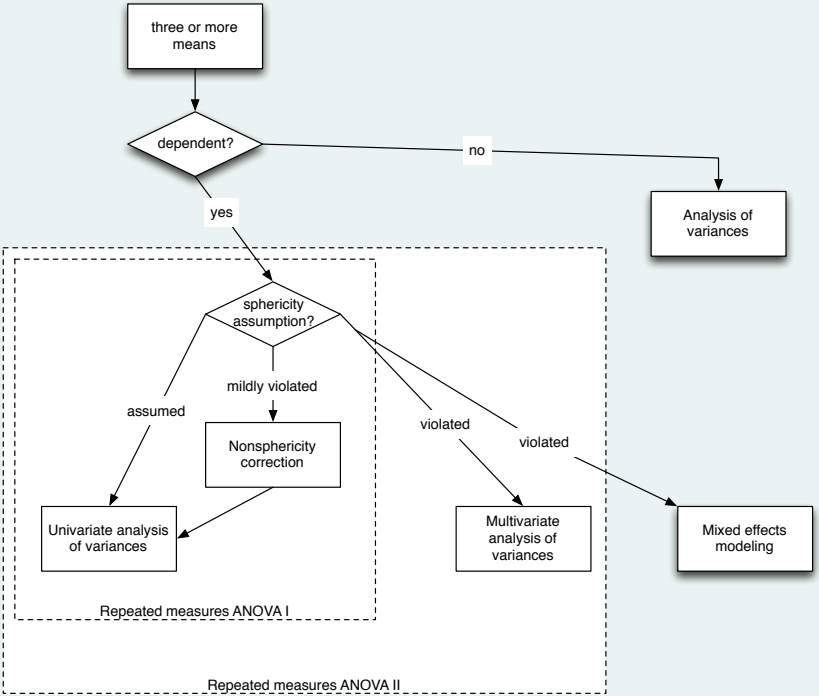
분석기법의 저변에 아주 복잡한 전제가 내포되어 있다.

1. 정규정 전제
2. 구형성(sphericity; circularity) 전제: Mauchly's test $P > 0.05$
3. 정규성을 충족하지 않으면 아예 다른 방법으로 가고,
4. 구형성을 충족하지 못하면:
   - 비구형성보정(nonsphericity adjustment)을 하거나
   - 다변량분석(MANOVA)이나 혼합효과모형분석(MEM)으로 간다.

Summary of 95 KJA Articles using RMANOVA

From: *Park et al*. Korean J Anesthesiol. 2016; 69: 97-9.

Assumptions	Correctly described	Others
Normality test	8 (9%)	87 (91%)
Sphericity assumption	2 (2%)	93 (98%)


three or more means
dependent?
no
Analysis of variances
yes
sphericity assumption?
assumed
mildly violated
violated
violated
Nonsphericity correction
Univariate analysis of variances
Multivariate analysis of variances
Mixed effects modeling
Repeated measures ANOVA I
Repeated measures ANOVA II

요약

Glantz의 Primer of Biostatistics로부터 부분 인용

	모수적 방법	비모수적 방법
측정 단위	간격척도	순서척도
두 개의 군	t-test	Mann-Whitney test
짝지은 두 개의 군	paired t-test	Wilcoxon signed rank test
세 군 이상	ANOVA	Kruskal-Wallis test
반복측정된 세 군 이상	RMANOVA, MANOVA, MEM	Friedman test

모수적 방법과 비교해서 비모수적 방법은 P값 외의 강건한 통계량(효과크기 등)을 산출하기 어려우므로 설계 단계에서 모수적 방법을 고집하는 것이 바람직하다.

빈도 자료

빈도 자료 분석의 기본은 Chisquare test.

2x2 수표

	술후 오심구토	없음
대조군	36	23
치료군	7	52

rx2 수표

	대조군	치료군
하복부 수술	34	31
하지 수술	33	39
회음부 수술	33	30

Chisquare test의 한계

- 대전제: 각 셀의 관찰값이 독립적
- 세부 조건(2x2 수표 기준으로)
  1. 총 표본수는 20 이상
  2. 최소 기대돗수는 5 이상
  3. 어느 한 셀의 관찰도수도 0이 아니어야 한다.
- 세부 조건이 맞지 않으면 Fisher의 정확도검정(Fisher's exact test)으로 간다.
- 대전제가 맞지 않으면?

빈도 자료의 분석: 비독립적일 때

총 90명의 피험자에서 광봉을 이용하여 삽관한 뒤 다시 맹목 기관삽관을 하여 성공빈도를 측정하였다.

McNemur test

	광봉 성공	광봉 실패
맹목 성공	63	1
맹목 실패	23	3

표를 잘못 그리면?

	맹목 삽관	광봉 삽관
성공	64	86
실패	26	4

앞 장의 수표를 이처럼 잘못 그려서 분석하면 잘못된 결과가 나온다.

검정 결과 해석의 문제(일명 P값의 문제)

'미국통계학회의 성명서: P값을 올바르게 사용하기' (Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108)

P값은 정확히 무엇을 말하는가?

Pearson and Nyman said,

- With defining *null hypothesis*[4] *as well as an alternative hypothesis* [5], 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.

[4] a hypothesis that there is no effect.

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

왜 $P < 0.05$인가?

- 아무도 모른다.
- 약속이다.

그런데 P는?

- 차이의 크기와 자유도(표본수)의 함수이다.
- 차이의 크기는 불변하므로[6] P값은 자유도(표본수)에 의해서 좌우된다.
- 그렇다면 P값은 다음처럼 해석되어야 한다: "$P < 0.05$는 내가 가진 표본으로 차이를 설명할 수 있다는 것"을 뜻한다.[7]

> P = 0.04의 결과를 가진 약제 d와 P = 0.02의 결과를 가진 약제 D에서, D가 d보다 약효가 좋은가?

아니다. 약효는 P값이 아니라 차이의 크기에서 해석해야 한다. ("Moving from tests to estimates" by *Greeland et al.* Eur J Epidemiol 2016; 31: 337–50)

---

[6] 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. 우리를 혼란스럽게 만드는 것은 무질서도(분산으로 나타남)이며, 표본수가 커지면 무질서도는 줄어든다.

[7] 내가 가진 표본수에서, 신약이 감기의 유병기간을 줄인다고 설명할 수 있다.

통계음성적결과(Negative result)의 문제

P값으로 우리가 할 수 있는 일은 $P < 0.05$일 때 귀무가설을 기각하는 것이다. 그렇다면 $P > 0.05$라면?

- 기각할 수 없는 것이 채택한다는 뜻은 아니다. (복잡한 논리적인 함의가 있다.)
- 간단히 말해서 귀무가설을 채택하기 위해서 필요한 통계량은 $P > 0.05$가 아니라 Power > 80%이다. (섬세한 연구에서는 90%를 잡기도 한다.)
- 따라서, $P > 0.05$라고 해서 '두 평균은 같다(= 치료의 효과가 없다).'라고 결론을 내릴 수 없다.

	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; $(1-\beta)\times 100$)을 80% 이상으로 유지하기 위해서는 최소표본수 설정이 필수적이다(효과의 크기는 고정되어 있으므로).
* 최소표본수와 검정력은 연구 고안 단계에서('방법'에 기술) 정립하고, 최소표본수를 계산하지 않은 연구에서는 분석 단계에서 ('결과'에 기술) 검정력을 계산한다.

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 definite trend (p=0.08)
- 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 (p=0.09)

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.

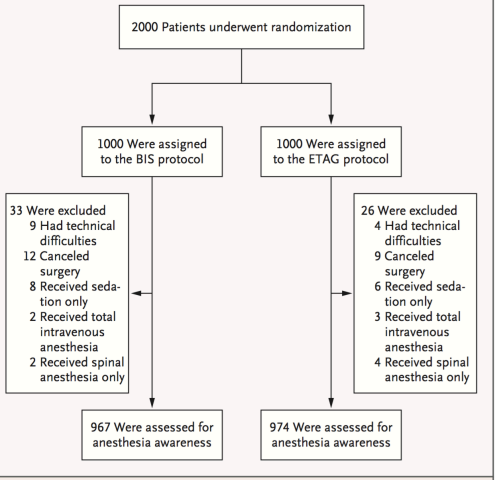


Figure 1. Trial Enrollment.

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 anthe 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

반복 비교에 따른 유의수준 보정의 문제

대전제 하나의 가설 아래 수집된 자료에서 통계검정을 할 때마다 통계의 제I종 오류가 팽창한다(inflation of α error).

유의도 0.05로 설정한 하나의 자료에서 단 한 번의 통계 검정을 시행하면 $\alpha = 0.05$지만 두 번 시행하면 $\alpha = 0.05 + 0.05$, 세 번 시행하면 $\alpha = 0.05 + 0.05 + 0.05$ 식으로 차츰 낮아지면서 제I종 오류가 팽창한다.

- 이를 막기 위해서는 개별유의도를 처음부터 보정해서 써야 한다. 세 번 분석할 경우에는 $\alpha_{adj} = 0.05/3$가 되며, 총 유의도는 개별 보정유의도의 총합과 같다.
- 어떤 통계분석법을 적용하는지에 무관하게 모든 분석법의 P값을 해석할 때 고려해야 한다.

Table 2. Cervical Spine Motion at 3 Cervical Segments

	LALI (n =20)	CLI (n =20)	Mean difference (98.33% CI)	*P* Value
Occiput-C1 (°)	5.6 (4.3)	9.3 (4.5)	−3.8 (−7.2 to −0.3)	.007*
C1-C2 (°)	5.9 (3.1)	6.0 (3.3)	−0.1 (−2.6 to 2.5)	.911
C2-C5 (°)	1.5 (3.9)	1.7 (2.6)	−0.2 (−2.8 to 2.5)	.795

Values are mean (SD).

Abbreviations: CLI, conventional lightwand intubation; LALI, laryngoscope-assisted lightwand intubation.

*Statistically significant after multiple comparisons ($P < .05/3$).

그림: From: *Kim et al.* Anesth Analg 2017; 125: 485–90.

Randomness in selecting journal names was assured using dplyr package (dplyr: A Grammar of Data Manipulation. Hadley Wickham and Romain Francois., R package version 0.5.0). Throughout the data acquisition and analyses; API procedures for web scraping, data handling, graphing, and statistical analyses were powered by R software version 3.3.2 (R: A language and environment for statistical computing.; R Foundation for Statistical Computing., Vienna, Austria) added on GNU Emacs version 25.1.1 (Free Software Foundation, Inc., Boston, MA, USA; 2016). Linear mixed-effects models were constructed using the lme4 package (lme4: R package for linear mixed-effects models. Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker, R package version 1.0.+) (9), with maximum likelihood method. Since the authors planned 2 inferential tests separately on 2 dependent variables (acceptance and lead lag), each inference was targeted to α value of 0.025, keeping overall α value, 0.05. So, CIs in this report were within 97.5%, as well. Subsidiary *P* values were attained by performing the likelihood ratio test against a null model. The journal names were set in italicized International Organization for Standardization (ISO)-abbreviaharbored in 2013, the more the journal inclined to reduce the lag. On the contrary, based on the lag in 2013, faster journals, recording the acceptance lag < 100 days, failed to reduce the acceptance lag (*J GYNECOL ONCOL*, *YONSEI MED J*, *ALLERGY ASTHMA IMMUNOL RES*) with the exception of the *KOREAN J ANESTHESIOL* and the *J CLIN NEUROL*.

The overall lead lag was 123.0 (63.0, 236.0) days. Between the slowest (44.0 [27.0, 62.0] days) and the faster one (323.0 [269.0, 372.0] days), there was an 8-fold difference (Fig. 4). Eight journals managed to reduce the lead lag, while another 2 failed to reduce it. Like the acceptance lag, changes in the lead lag seemed to be linked to the track record of 2013. Among 5 journals with shorter lead lag in 2013 (the lag < 100 days), 3 journals (*ANN LAB MED*, *J GYNECOL ONCOL*, and *J KOREAN MED SCI*) managed to reduce the lead lag.

Modeling

The year of publication did not significantly affect the acceptance lag (χ^2 [*df* = 1] = 0.22, *P* = 0.640), and supposedly shortening it by about 1.4 (97.5% CI, –5.2 to 8.0) days/year, while the

그림: From: *Lee et al*. J Korean Med Sci 2017; 32: 1235–42.

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

유의수준 보정의 문제 2: 일차끝점과 이차끝점

유의수준 보정은 연구의 일차끝점(primary endpoints)에만 더한다.

> R statistical software version 3.2.3 "Wooden Christmas- Tree" (R Foundation for Statistical Computing, Vienna, Austria) was used for the whole process of data analyses. The α value adjustment with Bonferroni correction was done to compensate for multiple comparisons within primary outcomes. The α value was adjusted to .0166 instead of .05. The P values were compared with this adjusted α value in interpreting primary outcome measures. Otherwise, P values <.05 were deemed to indicate statistical significance. (From: *Kim et al*. Anesth Analg 2017; 125: 485–90.)

엄격한 편집인들

- 하나의 가설로 모은 자료에서 산출하는 모든 P값은 보정된 유의수준에서 비교되어야 한다. (심지어는 중간분석과 정규성검정까지!)
- 대책:
  1. 중간분석과 정규성검정에 추리통계를 가하지 않는다.
  2. 환자들의 인구학적 특성은 군-간의 비교를 하지 않는다.
  3. 이차끝점을 두지 않는다.

다시, 재현성 위기 (Irreproducibiblity Crisis)

대부분의 저널에서 요구하는바, 통계 방법과 결과는 독자들도 그대로 따라할 수 있도록 자세히 기술한다.

과학의 제분야에서 요구하는 엄격함을 지키고 있음에도 불구하고 현대의 의학연구는 재현성이 낮다 (*Peng*. Biostatistics 2009; 10: 405–8.)

이를 막는 최소한의 장치로서의 재현가능연구 (reproducible research):

1. 원시자료 (raw dataset) 와
2. 분석코드 (analysis code) 를 공개한다.

WORLD VIEW *A personal take on events*

Science, lies and video-taped experiments

Too many researchers make up or massage their data, says **Timothy D. Clark**. *Only stringent demands for proof can stop them.*

Late last month, a US physicist began a jail sentence for scientific fraud. Darin Kinion took funds for research on quantum computing but did not carry out the work he claimed; instead, he invented the data that the research supposedly produced.

Scientists like to think that such blatant dishonesty is rare, but I myself have witnessed several serious cases of scientific misconduct, from major data manipulation to outright fabrication. Most have gone unpunished — in fact, it has been disheartening to see the culprits lauded. It makes little sense for fraudsters to fabricate mediocre data. Their falsehoods generate outstanding stories, which result in high-profile publications and a disproportionately large chunk of the funding pie.

I have noticed a lesser-known motive for bad science in my field, experimental biology. As environmental change proceeds, there is great demand from the public and policymakers for simple stories that show the damage being done to wildlife. I occasionally meet scientists who argue

field: using a tank of flowing water to expose fish to environmental perturbations and looking for shifts in behaviour. It is trivial to set up a camera, and equally simple to begin each recorded exposure with a note that details, for example, the trial number and treatment history of the organism. (Think of how film directors use clapper boards to keep records of the sequence of numerous takes.) This simple measure would make it much more difficult to fabricate data and 'assign' animals to desired treatment groups after the results are known.

My colleagues and I are currently using this approach to record studies of how coral-reef fish respond to dissolved carbon dioxide. There would also be benefits for other disciplines, including social-psychology studies based on direct observations.

Sharing visual evidence is straightforward. Video files can be compressed and transferred without excessive loss of resolution. Files can then be uploaded to free data repositories (such as figshare or Zenodo) before

그림: From: Clark. Nature Worldview 2017; 542: 139

EDITORIAL

JKMS

https://doi.org/10.3346/jkms.2017.32.7.1051 • *J Korean Med Sci 2017; 32: 1051-1053*

Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors

Darren B. Taichman,[1] Peush Sahni,[2] Anja Pinborg,[3] Larry Peiperl,[4] Christine Laine,[5] Astrid James,[6] Sung-Tae Hong,[7] Abraham Haileamlak,[8] Laragh Gollogly,[9] Fiona Godlee,[10] Frank A. Frizelle,[11] Fernando Florenzano,[12] Jeffrey M. Drazen,[13] Howard Bauchner,[14] Christopher Baethge,[15] and Joyce Backus[16]

[1]Secretary, ICMJE, Executive Deputy Editor, *Annals of Internal Medicine*; [2]Representative and Past President, World Association of Medical Editors; [3]Scientific Editor-in-Chief, *Ugeskrift for Laeger* (*Danish Medical Journal*); [4]Chief Editor, *PLOS Medicine*; [5]Editor-in-Chief, *Annals of Internal Medicine*; [6]Deputy Editor, *The Lancet*; [7]Editor-in-Chief, *Journal of Korean Medical Science*; [8]Editor-in-Chief, *Ethiopian Journal of Health Sciences*; [9]Editor, *Bulletin of the World Health Organization*, Coordinator, WHO Press; [10]Editor-in-Chief, *The British Medical Journal* (*BMJ*); [11]Editor-in-Chief, *New Zealand Medical Journal*; [12]Editor, *Revista Médica de Chile* (*Medical Journal of Chile*); [13]Editor-in-Chief, *New England Journal of Medicine*; [14]Editor-in-Chief, *Journal of the American Medical Association* (*JAMA*) and the JAMA Network; [15]Chief Scientific Editor, *Deutsches Ärzteblatt* (*German Medical Journal*) & *Deutsches Ärzteblatt International*; [16]Representative and Associate Director for Library Operations, National Library of Medicine

그림: From: *ICMJE Editors*. J Korean Med Sci 2017; 32: 1051–3.

Give up your data to cure disease

Agus DB. Feb 6, 2016. New York Times

"의학은 뜻밖의 발견으로 발전한다. 과학자들은 기대하지도 않았거나 심지어는 쳐다보지도 않았던 곳에서 중요한 것을 찾아내곤 한다. 나는 모든 질병을 예방하고 치료하고 낫게 하는 데 필요한 약물과 치료법을 인간이 이미 다 가지고 있을지도 모른다고 자주 말하곤 한다. 다만 어느 때 어느 걸 얼마나 써야 할지 모를 뿐이다... 당신들이 모은 자료를 독점하지 말고 공유함으로써 다시 분석될 기회를 주어라(피험자들의 자발적 희생을 값지게 만들 수 있는 기회이다)."

Give Up Your Data to Cure Disease

By DAVID B. AGUS FEB. 6, 2016

The ICMJE Statement 2017

임상연구 자료의 공유(Clinical data sharing):

- 2018년 7월부터 ICMJE 회원 학술지에 제출하는 임상연구 논문은 자료공유에 관한 선언을 포함해야 한다.
- 2019년 1월부터 피험자 모집을 시작하는 임상연구는 연구등록 단계에서 자료공유 계획을 포함한다.

어떤 자료를 공유하는가(최대 범위)?

- 취득한 개별 피험자에서 연구를 위해 수집한 자료 전체(물론, deidentification 필요)
- 그밖에도 연구계획서, 통계분석 계획서, 피험자 동의서, 연구보고서, 분석코드

요약

1. 개별 분석의 전제와 조건을 숙지한다.
2. 낚시질(statistical fishing)을 하지 않는다.
3. 자료에 가할 추리통계의 수효를 사전에 결정해서 전체 오류를 통제한다.
4. P값보다는 효과의 크기로 결과를 해석한다.
5. 통계 방법과 결과를 상세히 기술한다(독자들이 따라할 수 있도록).
6. 원자료공유에 동참한다.

여러분들이 이 강의에서 기대했던 것들

1. 통계분석법 하나하나를 더 잘 수행했으면 좋겠다.
2. 내 통계결과를 제대로 해석하고 싶다.
3. 좋은 논문을 써서 유명 저널에 척척 게재하고 싶다.
4. 내가 쓴 논문은 환자 치료에 이득이 되어야 한다.

교훈. 좋은 통계와 좋은 임상 판단으로 가는 길

Michenfelder. Anesthesia and the brain. Churchill-Livingstone, 1988

Dr. Michenfelder는 지난 16년간 뇌동맥류수술 후 간질의 발생이 증가하고 있다는 도전에 직면하고 전체 1209명의 임상자료를 수집하여 분석과 분석과 분석을 거듭한 끝에 결론을 내린다.

- 임상적 느낌을 무시하지 말라.
- 쉽게 결론 내리지 말라.
- 반박을 과소평가하지 말라.
- 성급하게 내뱉지 말라.
- 동료들을 푸시하지 말라.
- 후향 자료를 신뢰하지 말라.
- 고민하고 또 고민하라.
- 조작하지 말라.
- 전문 통계학자와 협업하라.
- 연구계획 단계부터 통계학자와 협업하라.