KUCE 2020

대한비뇨의학회 통합학술대회

The Korean Urological Congress and Expo

상위저널 71%에서도 틀리는 통계:
임상연구 통계보고 가이드라인

SNUH 서울대학교병원

SEOUL NATIONAL UNIVERSITY HOSPITAL

Chang Wook Jeong M.D., Ph.D.

Department of Urology, Seoul National University Hospital &
Seoul National University College of Medicine

EUROPEAN UROLOGY 75 (2019) 358–367

available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Opinion – Editor's Choice

Guidelines for Reporting of Statistics for Clinical Research in Urology

Melissa Assel[a], *Daniel Sjoberg*[a], *Andrew Elders*[b], *Xuemei Wang*[c], *Dezheng Huo*[d], *Albert Botchway*[e], *Kristin Delfino*[e], *Yunhua Fan*[f], *Zhiguo Zhao*[h], *Tatsuki Koyama*[h], *Brent Hollenbeck*[i], *Rui Qin*[j], *Whitney Zahnd*[k], *Emily C. Zabor*[a], *Michael W. Kattan*[g], *Andrew J. Vickers*[a,*]

[a] *Memorial Sloan Kettering Cancer Center, New York, NY, USA;* [b] *Glasgow Caledonian University, Glasgow, UK;* [c] *The University of Texas, MD Anderson Cancer Center, Houston, TX, USA;* [d] *The University of Chicago, Chicago, IL, USA;* [e] *Southern Illinois University School of Medicine, Springfield, IL, USA;* [f] *University of Minnesota, Minneapolis, MN, USA;* [g] *Cleveland Clinic, Cleveland, OH, USA;* [h] *Vanderbilt University Medical Center, Nashville, TN, USA;* [i] *University of Michigan, Ann Arbor, MI, USA;* [j] *Janssen Research & Development, NJ, USA;* [k] *University of South Carolina, Columbia, SC, USA*

Eur Urol 2019;75:358-67

- 다음 중 올바른 통계 제시 법은?

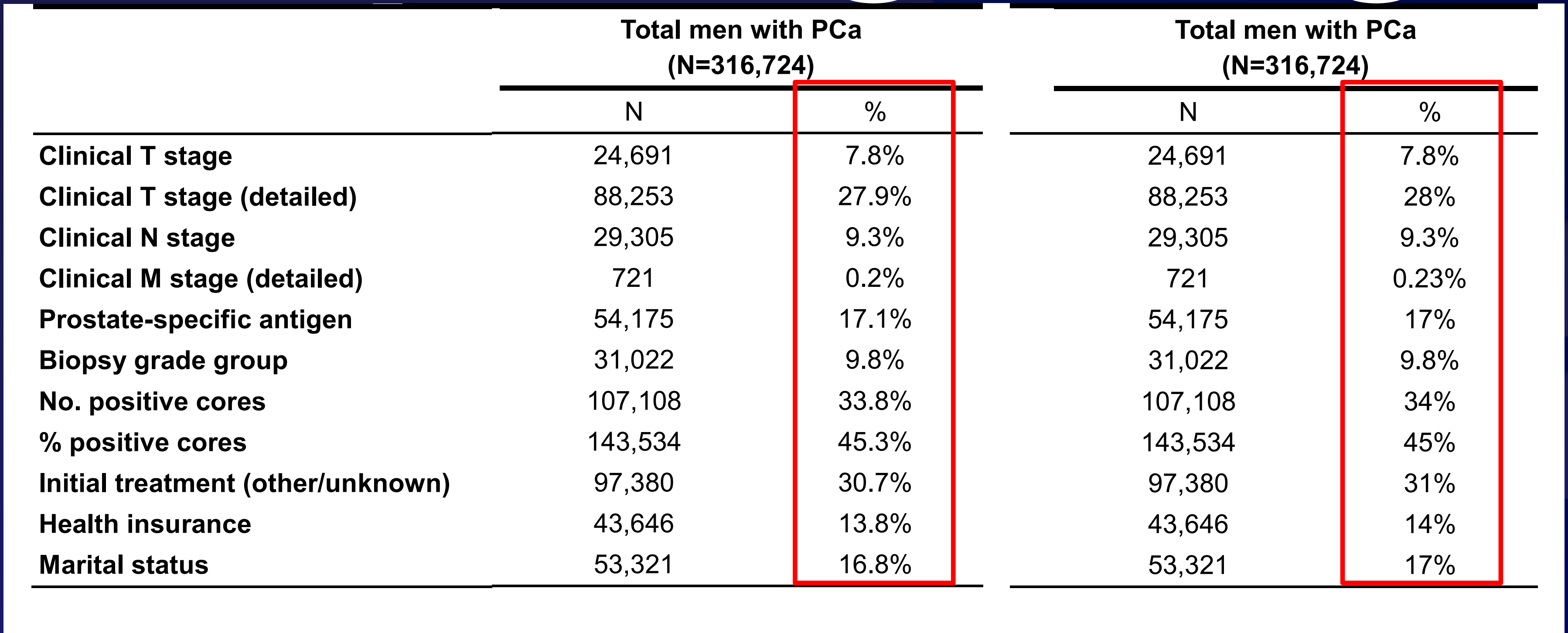
①

Age (mean: 72, S.D.: 2.1)
PSA (mean: 11, S.D.: 0.6)

②

Age (mean: 72, S.D.: 2.1)
PSA (median: 10, IQR:2.7-16)

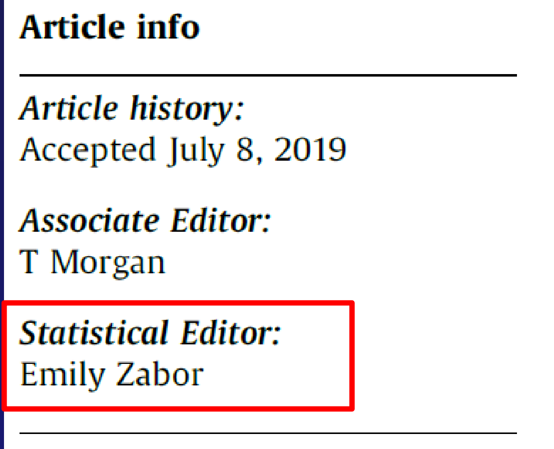
- 다음 중 올바른 분율 표기 방법은?



	① Total men with PCa (N=316,724)		② Total men with PCa (N=316,724)	
	N	%	N	%
Clinical T stage	24,691	7.8%	24,691	7.8%
Clinical T stage (detailed)	88,253	27.9%	88,253	28%
Clinical N stage	29,305	9.3%	29,305	9.3%
Clinical M stage (detailed)	721	0.2%	721	0.23%
Prostate-specific antigen	54,175	17.1%	54,175	17%
Biopsy grade group	31,022	9.8%	31,022	9.8%
No. positive cores	107,108	33.8%	107,108	34%
% positive cores	143,534	45.3%	143,534	45%
Initial treatment (other/unknown)	97,380	30.7%	97,380	31%
Health insurance	43,646	13.8%	43,646	14%
Marital status	53,321	16.8%	53,321	17%

Background

- **71% paper provided any statistical error in a single issue (August 2004) of 4 leading journals (Eur Urol, J Urol, BJU Int., Urology)**
- **2015, first guidelines in *European Urology***
- **2019, 2nd guidelines for Clinical Research in Urology**
- **Statistical editor reviews in an additional round**



Article info

Article history:
Accepted July 8, 2019

Associate Editor:
T Morgan

Statistical Editor:
Emily Zabor

J Urol 2005;174:1374-9, Eur Urol 2015;181-7, Eur Urol 2019;75:358-67

1. The golden rule

1.1. Break any of the guidelines if it makes scientific sense to do so

2. Reporting of design and statistical analysis

2.1. Follow existing reporting guidelines for the type of study you are reporting, such as CONSORT for randomized trials, ReMARK for marker studies, TRIPOD for prediction models, STROBE for observational studies, or AMSTAR for systematic reviews

2.2. Describe cohort selection fully

- The study cohort consisted of 1144 patients treated for BPH at our institution

→ The study cohort consisted of consecutive 1144 patients treated for BPH (IPSS>12) presenting March 2013 to December 2017 at our institution

- Exclusions should be described one by one with the number of patients omitted (Patients with prior surgery [n=43], allergies to 5-ARIs [n=12], and missing data on prostate volume [n=86] were excluded to give a final cohort for analysis of 1003 patients.)

2.3. Describe the practical steps of randomization in randomized trials

- Allocation concealment

2.4. The statistical methods should describe the study questions and the statistical approaches used to address each question

- "Mann-Whitney was used for comparisons of continuous variables and Fisher's exact for comparisons of binary variables."
  → X
- Statistical methods sections should lay out each primary study question separately

2.5. The statistical methods should be described in sufficient detail to allow replication by an independent statistician given the same data set

- Gleason grade was included in the model

→ Gleason grade group was included in four categories 1, 2, 3, and 4 or 5

3. Inference and *p* values

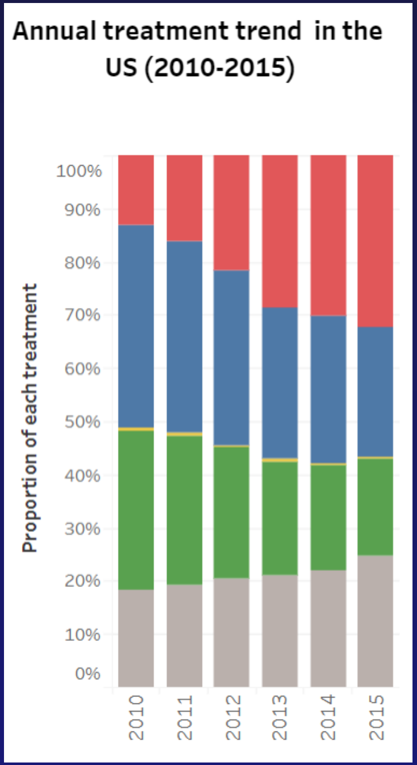
3.1. Do not accept the null hypothesis

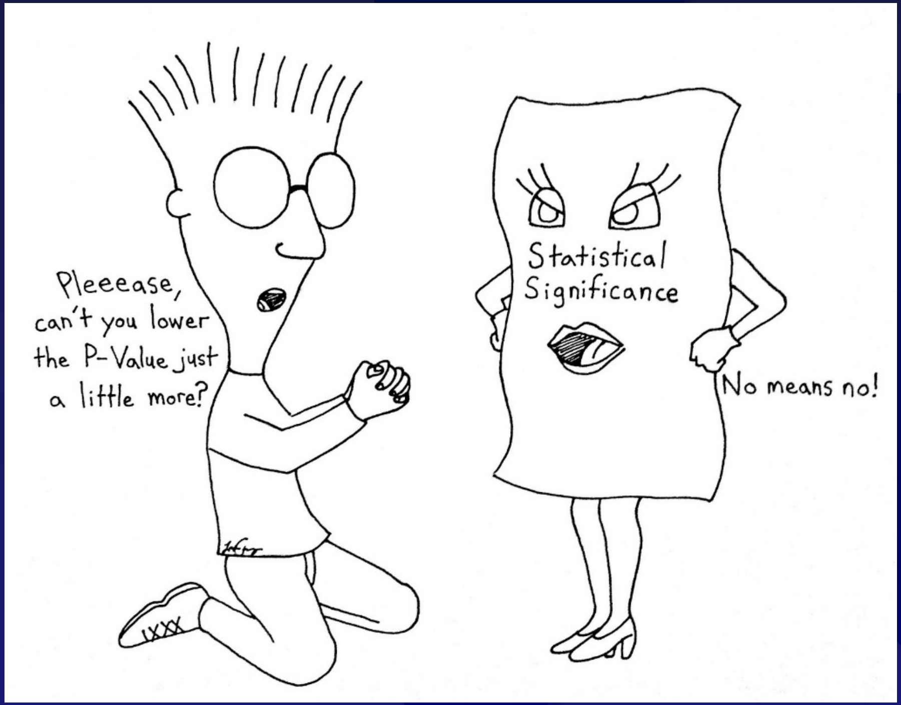
- Guilty or not guilty; Innocent X
- *P*>0.05, “the drug was ineffective”, “there was no difference between groups”, “response rates were unaffected” → X

→ “We did not see evidence of a drug effect”, “we were unable to demonstrate a difference between groups”, “there was no statistically significant difference in response rate”

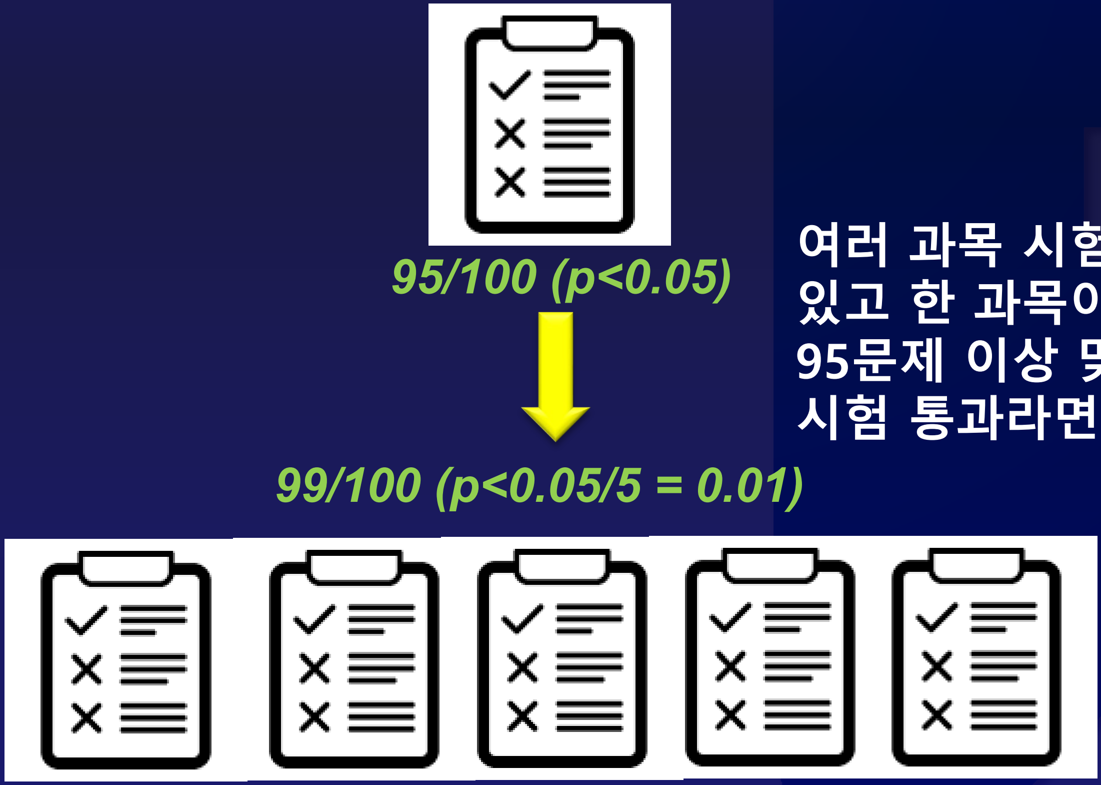
3.2. P values just above 5% are not a trend, and they are not moving

- “Differences between groups did not meet conventional levels of statistical significance.”





3.5. Take care to interpret results when reporting multiple p values (Bonferroni correction)

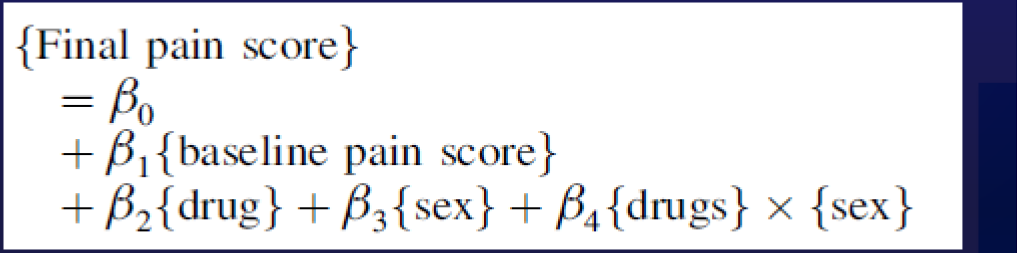


95/100 ($p<0.05$)

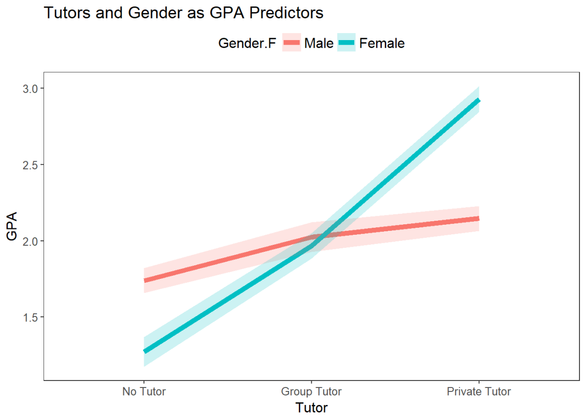
99/100 ($p<0.05/5 = 0.01$)

여러 과목 시험을 칠 수 있고 한 과목이라도 95문제 이상 맞으면 시험 통과라면...

3.7. Use interaction terms in place of subgroup analyses



$$\begin{aligned}\{\text{Final pain score}\} &= \beta_0 \\ &+ \beta_1\{\text{baseline pain score}\} \\ &+ \beta_2\{\text{drug}\} + \beta_3\{\text{sex}\} + \beta_4\{\text{drugs}\} \times \{\text{sex}\}\end{aligned}$$



3.8. When reporting p values, be clear about the hypothesis tested and ensure that hypothesis is a sensible one

- “Pain scores were higher in group 1 and similar in groups 2 and 3 (p = 0.02)”

→ T-test for 1 vs. 2+3 ?

→ ANOVA for 1 vs. 2 vs. 3 ?

4. Reporting of study estimates

4.1. Use appropriate levels of precision

- *p*=0.7345 → appreciable difference between 0.7344 and 0.7346
- 16.9% of 83 patients → precision is 200 times greater than CI (10-27%)

1. Report *p* values to a single significant figure unless the *p* value is close to 0.05 (say, 0.01-0.2), in which case, report two significant figures. Very low p values can be reported as p<0.001 or similar (Good example <0.001, 0.004, 0.045, 0.13, 0.3, 1)
2. Report percentages, rates, and probabilities to two significant figures (Good example 75%, 3.4%, 0.13%)
3. Do not report *p* values of 0
4. Do not give decimal places if a probability or proportion is 1 (Bad example 1.00 or 100.0%). The decimal places suggest that is possible to have, say, a *p* value of 1.05. (Mean number of pregnancy was 2.4 → O, 29% of women reported 1.0 pregnancy → X)
5. No need to report estimates to more than 3 significant figures.
6. HR and OR are normally reported to 2 decimal places, although this can be avoided for high odds ratios (18.2 rather than 18.17)

4.2. Avoid redundant statistics in cohort description

- 40% were men and 60% were women
- Do not describe combined whole cohort

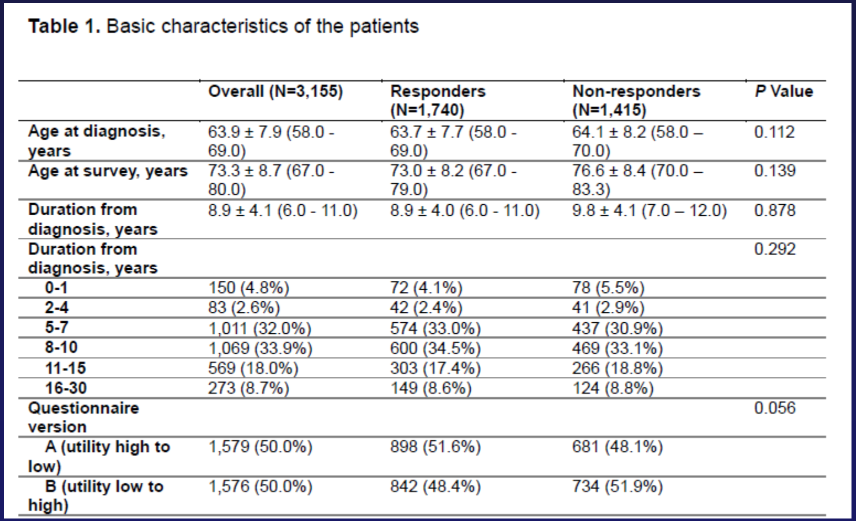


Table 1. Basic characteristics of the patients

	Overall (N=3,155)	**Responders (N=1,740)**	**Non-responders (N=1,415)**	***P* Value**
Age at diagnosis, years	63.9 ± 7.9 (58.0 - 69.0)	63.7 ± 7.7 (58.0 - 69.0)	64.1 ± 8.2 (58.0 – 70.0)	0.112
Age at survey, years	73.3 ± 8.7 (67.0 - 80.0)	73.0 ± 8.2 (67.0 - 79.0)	76.6 ± 8.4 (70.0 – 83.3)	0.139
Duration from diagnosis, years	8.9 ± 4.1 (6.0 - 11.0)	8.9 ± 4.0 (6.0 - 11.0)	9.8 ± 4.1 (7.0 – 12.0)	0.878
Duration from diagnosis, years				0.292
0-1	150 (4.8%)	72 (4.1%)	78 (5.5%)	
2-4	83 (2.6%)	42 (2.4%)	41 (2.9%)	
5-7	1,011 (32.0%)	574 (33.0%)	437 (30.9%)	
8-10	1,069 (33.9%)	600 (34.5%)	469 (33.1%)	
11-15	569 (18.0%)	303 (17.4%)	266 (18.8%)	
16-30	273 (8.7%)	149 (8.6%)	124 (8.8%)	
Questionnaire version				0.056
A (utility high to low)	1,579 (50.0%)	898 (51.6%)	681 (48.1%)	
B (utility low to high)	1,576 (50.0%)	842 (48.4%)	734 (51.9%)	

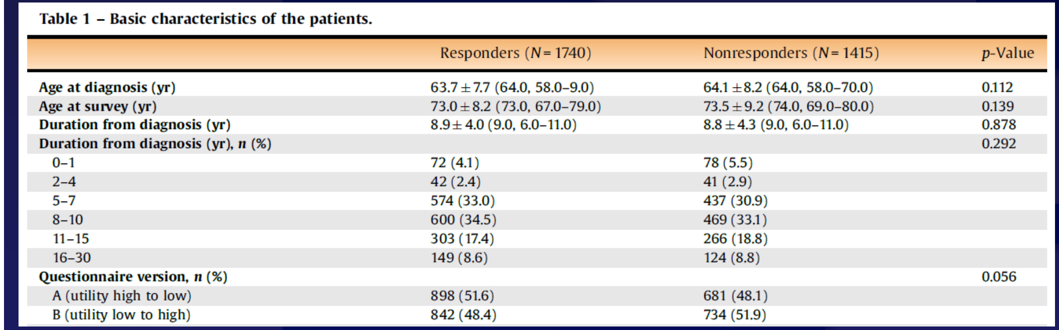


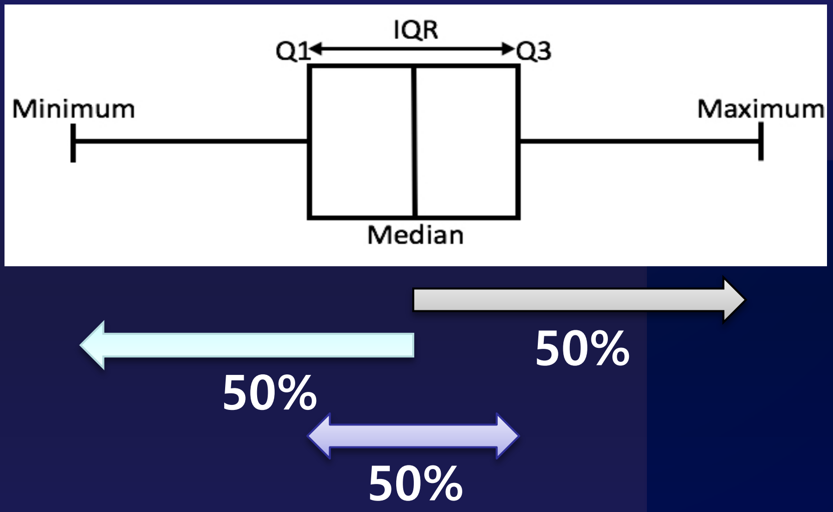
Table 1 – Basic characteristics of the patients.

	Responders (*N* = 1740)	Nonresponders (*N* = 1415)	*p*-Value
Age at diagnosis (yr)	63.7 ± 7.7 (64.0, 58.0–9.0)	64.1 ± 8.2 (64.0, 58.0–70.0)	0.112
Age at survey (yr)	73.0 ± 8.2 (73.0, 67.0–79.0)	73.5 ± 9.2 (74.0, 69.0–80.0)	0.139
Duration from diagnosis (yr)	8.9 ± 4.0 (9.0, 6.0–11.0)	8.8 ± 4.3 (9.0, 6.0–11.0)	0.878
Duration from diagnosis (yr), *n* (%)			0.292
0–1	72 (4.1)	78 (5.5)	
2–4	42 (2.4)	41 (2.9)	
5–7	574 (33.0)	437 (30.9)	
8–10	600 (34.5)	469 (33.1)	
11–15	303 (17.4)	266 (18.8)	
16–30	149 (8.6)	124 (8.8)	
Questionnaire version, *n* (%)			0.056
A (utility high to low)	898 (51.6)	681 (48.1)	
B (utility low to high)	842 (48.4)	734 (51.9)	

✗

○

4.3. For descriptive statistics, median and quartiles are preferred over means and SD; range should be avoided



4.4. Report estimates for the main study questions

- Authors should give an estimate of the difference between groups, and avoid giving only data on each group separately

- ORs or HRs, as well as reporting a p value

→ 5-year OS was 45% and 66%, respectively (p=0.04) → X

→ + Adjusted HR was 1.64 (p=0.02) → O

4.5. Report CI for the main estimates of interest

- Authors should generally report a 95% CI around the estimates relating to the key research questions, but not other estimates given in a paper.

- For instance, in a study comparing two surgical techniques, adverse event rates of 10% and 15%; however, the key estimate in this case is the difference between groups, so this estimate, 5%, should be reported along with a 95% CI (eg, 1–9%).
- CIs should not be reported for the estimates within each group (eg, adverse event rate in group A of 10%, 95% CI 7–13%).

4.6. Do not treat categorical variables as continuous

- **Gleason grade groups (1-5) is not a continuous variable!**

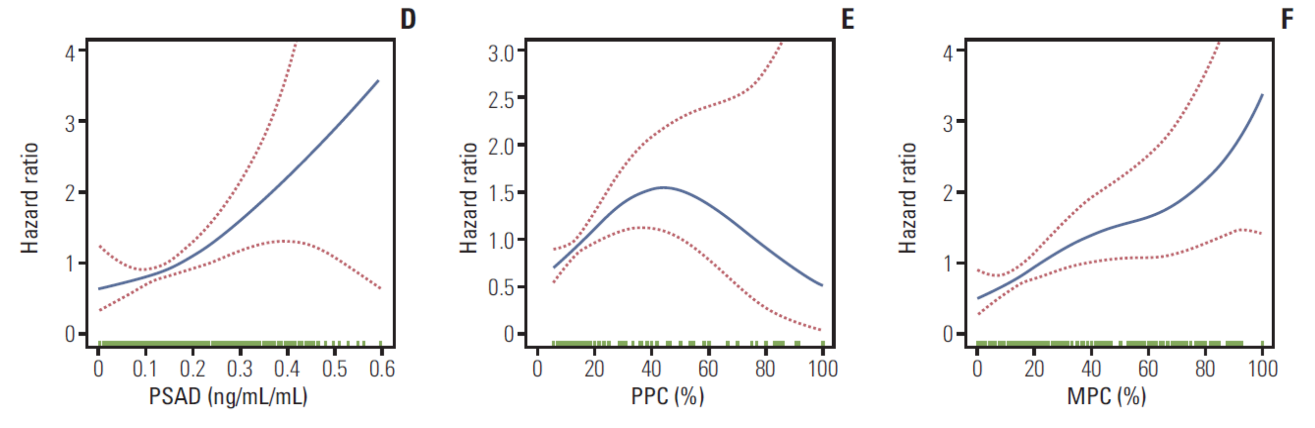
- **Proportion of each group → O, Mean Gleason score of 2.4 → X**

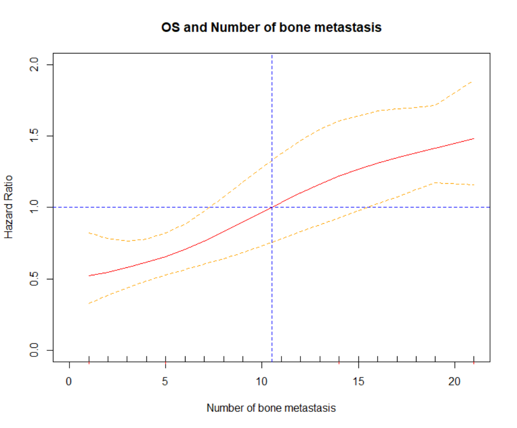
- **Should not be entered as continuous variable in regression models**

→ HR of 1.5 per 1-point increase in Gleason grade group → X

4.7. Avoid categorization of continuous variables unless there is a convincing rationale

4.9. The association between a continuous predictor and outcome can be demonstrated graphically, particularly by using nonlinear modeling





4.11. For time-to-event variables, report the number of events but not the proportion

- “Of 60 patients accrued,10 (17%) died.” → X (17% is meaningless)
- Standard statistical approach: To calculate probability
- “The risk of death being 60% by 5 yr” or “The median survival was 52 mo.”→ O

4.12. For time-to-event analyses, report median follow-up for patients without the event or the number followed without an event at a given follow-up time

- 예: 40년 된 소아암 환자 cohort에서 cure rate가 30% 였다면 median F/U duration은 수년 이내임.
- Median F/U duration + “312 patients have been followed without an event for a least 35 years.”

4.14. For time-to-event analyses, avoid reporting mean follow up or survival time, or estimates of survival in those who had the event

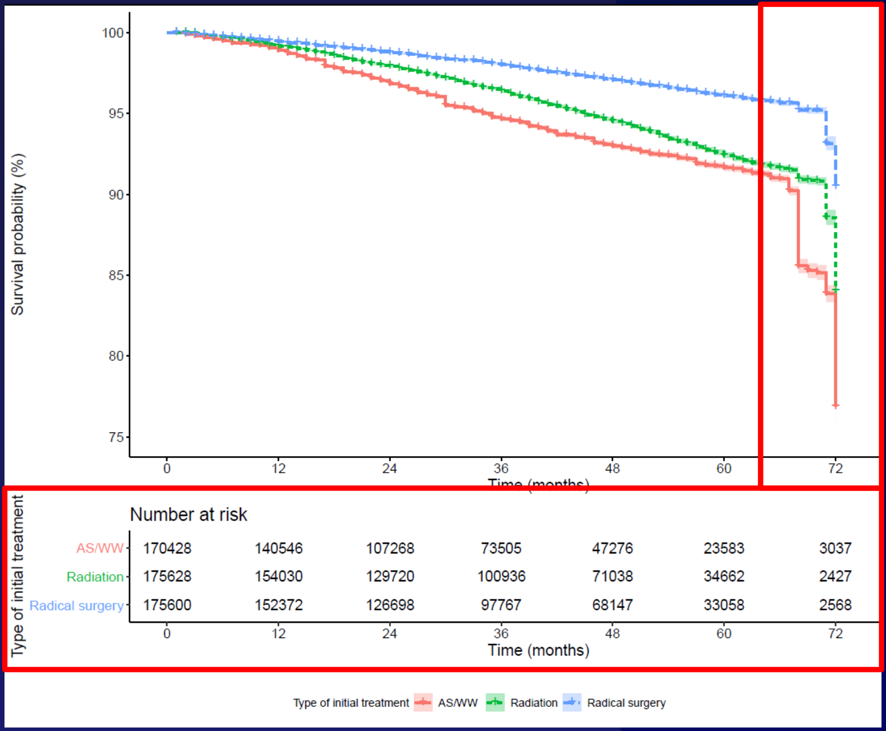
- All three estimates are problematic in the context of censored data.
- Bad example: 전체 평균 40개월의 추적관찰 기간 중, 재발한 환자들에서 재발까지의 평균 기간은 24개월이었다.

4.15. For time-to-event analyses, make sure that all predictors are known at time zero or consider alternative approaches such as a landmark analysis or time-dependent covariates

- PSA velocity (when?)
- A “landmark analysis” is often used when the variable of interest is generally known within a short and well-defined period of time, such as adjuvant therapy or chemotherapy response. In brief, the investigators start the clock at a fixed “landmark” (eg, 6 mo after surgery, patients who recur before 6 mo are excluded)

4.16. When presenting Kaplan-Meier figures, present the number at risk and truncate F/U when numbers are low

- A good rule of thumb is to truncate follow-up when the number at risk in any group falls below 5 (or even 10) as the tail of a Kaplan-Meier distribution is very unstable.



Type of initial treatment	0	12	24	36	48	60	72
AS/WW	170428	140546	107268	73505	47276	23583	3037
Radiation	175628	154030	129720	100936	71038	34662	2427
Radical surgery	175600	152372	126698	97767	68147	33058	2568

5. Multivariable models and diagnostic tests

5.1. Multivariable, propensity, and instrumental variable analyses are not a magic wand

5.4. Rescale predictors to obtain interpretable estimates

- Age: OR 1.02 (95% CI 1.01-1.02) → per 10 year of age OR 1.16 (95% CI 1.10-1.22)

5.5. Avoid reporting both univariate and multivariable analyses unless there is a good reason

5.6. Avoid ranking predictors in terms of strength

- Such rankings are not meaningful since it depends on how variables are coded.

5.9. Calibration should be reported and interpreted correctly

- Where a prespecified model is tested on an independent data set, calibration should be displayed graphically in a calibration plot.
- The Hosmer-Lemeshow test addresses an inappropriate null hypothesis and should be avoided.

5.10. Avoid reporting sensitivity and specificity for continuous predictors or a model

5.11. Report the clinical consequences of using a test or a model

- Authors are encouraged to choose illustrative cut-points and then report results in terms of clinical consequences

6. Conclusions and interpretation

6.1. Draw a conclusions, do not just repeat the results

- A statistically significant relationship was found between body mass index (BMI) and disease outcome. → X

- To make a recommendation for more aggressive treatment of patients with a higher BMI → O

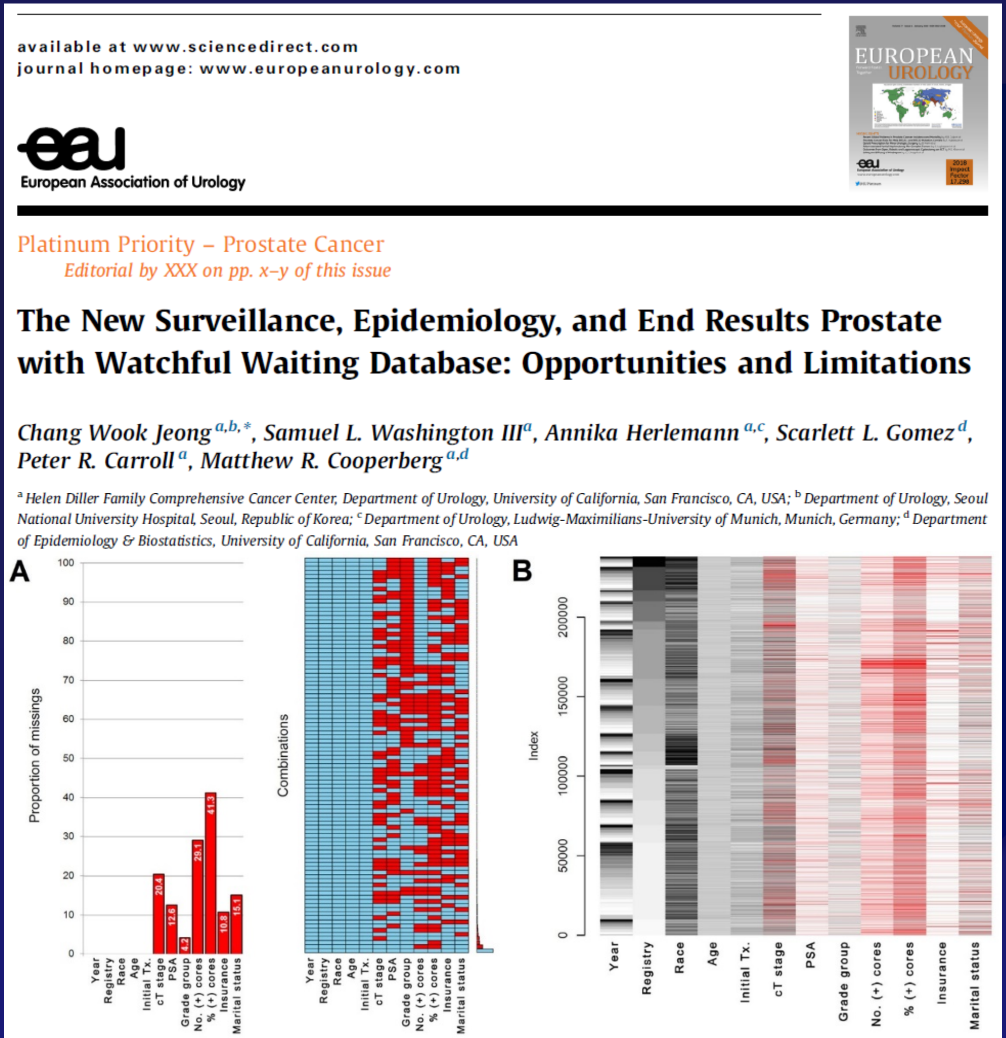
6.3. A statistically significant p value does not imply clinical significance

6.4. Avoid pseudolimitations such as "small sample size" and "retrospective analysis"; consider instead sources of potential bias and the mechanism for their effect on findings

- For instance, if a treatment or predictor is associated with a very large odds ratio, a large sample size might be unnecessary.
- Discussion of limitations should include both the likelihood and the effect size of possible bias.

6.5. Consider the impact of missing data and patient selection

- It is rare that complete data are obtained from all patients in a study



available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Priority – Prostate Cancer
Editorial by XXX on pp. x–y of this issue

The New Surveillance, Epidemiology, and End Results Prostate with Watchful Waiting Database: Opportunities and Limitations

Chang Wook Jeong [a,b,], Samuel L. Washington III[a], Annika Herlemann [a,c], Scarlett L. Gomez [d], Peter R. Carroll [a], Matthew R. Cooperberg [a,d]*

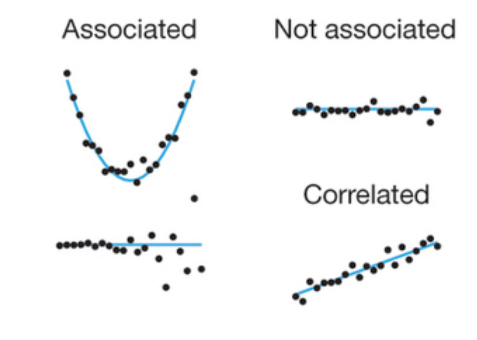
[a] Helen Diller Family Comprehensive Cancer Center, Department of Urology, University of California, San Francisco, CA, USA; [b] Department of Urology, Seoul National University Hospital, Seoul, Republic of Korea; [c] Department of Urology, Ludwig-Maximilians-University of Munich, Munich, Germany; [d] Department of Epidemiology & Biostatistics, University of California, San Francisco, CA, USA

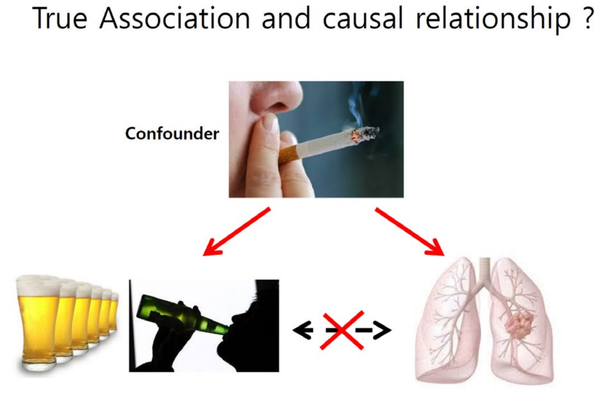
6.7. Do not confuse outcome with response among subgroups of patients undergoing the same treatment: patients with poorer outcomes may still be good candidates for that treatment

- **Patients with large tumors are more likely to recur after surgery than patients with small tumors, but that cannot be taken to suggest that resection is not indicated for patients with tumors greater than a certain size.**

6.8. Be cautious about causal attribution: correlation does not imply causation

- **연관성 (Association): 관계 a general relationship**
- **상관성 (Correlation): (선형)관계 a type of association**
- **인과성 (Causation): 원인 a cause**

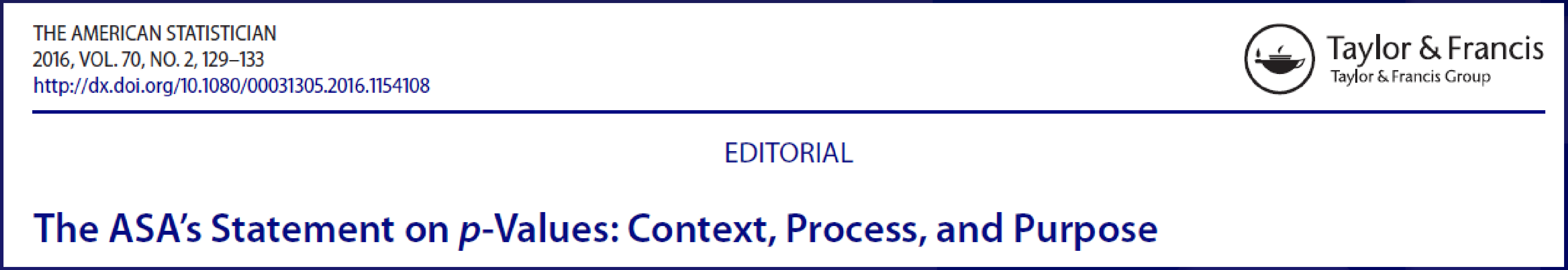




7. Use and interpretation of *p* values

The more general problem, which we address here, is that p values are often given excessive weight in the interpretation of a study. Indeed, studies are often classed by investigators into “positive” or “negative” based on statistical significance. Gross misuse of p values has led some to advocate banning

- In particular, we emphasize that a *p* value is just one statistic that helps interpret a study



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA’s Statement on *p*-Values: Context, Process, and Purpose

The American Statistician 2016;70:129-33

Thank you