KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 • eISSN 2005-7563

Assessment of P values for demographic data in randomized controlled trials

Eun Jin Ahn[1], Jong Hae Kim[2], Tae Kyun Kim[3], Jae Hong Park[4], Dong Kyu Lee[5], Sangseok Lee[6], Junyong In[7], and Hyun Kang[8]

Department of Anesthesiology and Pain Medicine, [1]Inje University Seoul Paik Hospital, Inje University College of Medicine, Seoul, [2]Daegu Catholic University School of Medicine, Daegu, [3]Yangsan Hospital, Pusan National University School of Medicine, Busan, [4]Inje University Haeundae Paik Hospital, Inje University College of Medicine, Busan, [5]Guro Hospital, Korea University School of Medicine, Seoul, [6]Sanggye Paik Hospital, Inje University College of Medicine, Seoul, [7]Dongguk University Ilsan Hospital, Goyang, [8]Chung-Ang University College of Medicine, Seoul, Korea

In a large number of randomized controlled trials, researchers provide P values for demographic data, which are commonly reported in table 1 of the article for the purpose of emphasizing the lack of differences between or among groups. As such, the authors intend to demonstrate that statistically insignificant P values in the demographic data confirm that group randomization was adequately performed. However, statistically insignificant P values do not necessarily reflect successful randomization. It is more important to rigorously establish a plan for statistical analysis during the design and planning stage of the study, and to consider whether any of the variables included in the demographic data could potentially affect the research results. If a researcher rigorously designed and planned a study, and performed it accordingly, the conclusions drawn from the results would not be influenced by P values, regardless of whether they were significant. In contrasts, imbalanced variables could affect the results after variance controlling, even though whole study process are well planned and executed. In this situation, the researcher can provide results with both the initial method and a second stage of analysis including such variables. Otherwise, for brief conclusions, it would be pointless to report P values in a table simply listing baseline data of the participants.

Keywords: Baseline; Bias; Characteristics; Demographic data; Difference; P value; Randomization; Randomized controlled trial; Variable.

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 06973, Korea
Tel: 82-2-6299-2586, Fax: 82-2-6299-2585
Email: roman00@naver.com
ORCID: https://orcid.org/0000-0003-2844-5880

Received: November 30, 2018.
Revised: December 4, 2018.
Accepted: December 5, 2018.

Korean J Anesthesiol 2019 April 72(2): 130-134
https://doi.org/10.4097/kja.d.18.00333

Pre-occupation with P values for Baseline Characteristics

The statistical tests used to compare baseline data in randomized controlled trials have remained questionable since 1990. In studies investigating baseline balance in clinical trials, Roberts and Torgerson [1], and Senn [2] suggested that significance tests to detect baseline differences are inappropriate. In 1990, 41% of randomized controlled trial (RCT)s reported inadequate comparisons of baseline characteristics [3]. One-half of the trials published in 1997 assessed imbalances between treatment groups using significant tests [4]. In addition, there are different

rules among the major journals and the CONSORT guidelines for reporting P values comparing baseline characteristics. The *New England Journal of Medicine* mandates statistical tests with P values for baseline characteristics.[1] Otherwise, CONSORT 2010 discourages statistical tests of baseline characteristics with the following comment: "Such significance tests assess the probability that observed baseline differences could have occurred by chance; however, we already know that any differences are caused by chance. Tests of baseline differences are not necessarily wrong, just illogical."[2]

In several studies designed and performed as a RCT, baseline data, such as demographics, medical history, vital signs and measurements, are usually collected. There are several reasons for collecting baseline data. First, the data provide information about the characteristics of the included patients. Second, baseline data show that the groups are well balanced by comparing groups, especially with critical variables that may significantly influence the results. Third, subgroup analyses may be performed on selected patient characteristics, which can also influence the results. Finally, covariate adjustment may be used to account for particular baseline factors.

Randomization is performed to avoid systematic errors that may occur during group assignment [5]. However, randomization cannot always prevent imbalances between two groups; more specifically, statistically significant differences in baseline data could occur by chance after randomization.

For example, if an RCT has 2 groups and 30 subjects, including both sexes, in each group, when randomization is conducted in such a study, the probability of a statistically significant difference (i.e., $P < 0.05$) in the sex variable is 0.0519. Of all studies published in the *Korean Journal of Anesthesiology* (*KJA*) between 2010 and 2017, 58 reported a P value for the sex variable (Table 1). Assuming these 58 studies have same number of groups and subjects in each group, the probability of all studies to reporting a statistically insignificant difference is $(1–0.0519)^{58}$. This value is 0.045451 and is statistically significant ($P < 0.05$). Moreover, 2.6 of 58 studies are expected to reveal a statistically significant difference in the sex variable. From 8 variable categories, 318 variables reported P values (Table 1). From these 318 variables, only 9 reported statistically significant difference. However, assuming 318 variables have same number of groups and subjects, and also the same probability of a statistically significant difference between groups, the possibility of reporting a statistical difference ≤ 9 is 0.004812.

The most inappropriate scientific point is that the null hypothesis against randomization is never proven during statistical analysis of baseline variables. That is, P values presented to contend the balanced baseline parameters have not enough evidences to reject the null hypothesis, imbalanced variables between groups.

The P value may also be partially influenced by sample size. In a small study, the P value may not reach statistical significance, even when there is a clinically relevant difference in a given baseline characteristic. For example, as shown in Table 2 [6], patient ages are not statistically different (8.1 [3.4] versus 10.0 [3.9]; $P = 0.052$). However, an almost 2-year gap in a group of pediatric patients could have a meaningful effect on the result in a clinical situation. Furthermore, a larger number of subjects would increase the possibility of obtaining smaller P values. Therefore, the authors suggest that the P values usually reported in Table 1 have no practicable meaning, but do afford the chance to incorrectly interpret the results of studies.

P values Published in the KJA

A total of 312 RCTs were published in the *KJA* from 2010 to 2017, and were reviewed. In most studies, patient baseline characteristics, such as age, sex, American Society of Anesthesiologists (ASA) physical status classification, height, weight, body mass index (BMI), duration of anesthesia and duration of the operation, that fulfilled inclusion criteria, were described in the baseline tables. Therefore, the baseline tables in each article, in terms of these eight variables, were reviewed. As shown in Table 1, 82 of the 312 studies reported a P value when comparing ages between or among groups, while 58 reported P values for comparisons of sex between or among groups. Respectively, 31, 60, 67, 17, 31, and 35 of the 312 RCTs reported P values for

Table 1. P values Reported in Baseline Tables during the Analysis Period (2010–2017)

Variable	P value in baseline table	Number of studies with $P < 0.05$
Age	82/311 (26.4)	1
Sex	58/250 (23.2)	0
ASA	31/100 (31)	0
Height	60/253 (23.7)	2
Weight	67/291 (23.0)	3
Body mass index	17/44 (38.6)	0
Duration of anesthesia	31/103 (30.1)	1
Duration of surgery	35/117 (29.9)	2
Total	83/312 (26.6)	6

Data are presented as absolute number (%), in which the denominator represents the number of studies that reported each variable and the numerator represents number of studies that reported P value for the variable. ASA: American Society of Anesthesiologists.

[1] http://www.nejm.org/page/author-center/manuscript-submission [accessed November 21, 2018]

[2] http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data [accessed November 21, 2018]

comparisons of ASA, height, weight, BMI, duration of anesthesia and surgery. Eighty-three (26.5%) studies reported P values in baseline data tables, among which 6 [6–11] reported P values that were statistically significant (i.e., < 0.05). Descriptions such as "similar" or "comparable" in the article were not considered as significant. Among the 6 studies that reported P values, variables that demonstrated significant differences were controlled for 2 investigations. In the study by Shin et al. [10], groups were divided according to age. In the study by Kim et al. [9] groups were divided according to the type of surgery, which could have possibly caused a difference in the duration of the operation and anesthesia between the groups.

The number of baseline variables varied widely, from 0 to 23 (Table 3). One study [12] did not report baseline characteristics of the included patients. In this study, the baseline table presented information only about the assessment of intubation conditions, including ease of laryngoscopy, vocal cord position and vocal cord movement, among others. More than one-half (57.7%) of the included studies reported 5 to 9 baseline variables.

How to Improve the Assessment of Balance in Baseline Characteristics of Clinical Trial Participants?

It would be ideal if studies recruited subjects with little-to-no heterogeneity between and among groups through rigorous methodology design, careful planning and study execution, and additionally, the inclusion of a large sample size. Furthermore, researchers could gather pilot patient data to determine whether they present any risk of bias or imbalance in baseline information between groups before designing and planning the study. However, these processes are often cost prohibitive, and involve large consumption of human and time resources. For these limitations, researchers could perform statistical interventions.

Statistical conclusions leave little room for doubt, but only if the baseline variables are well randomized and do not influence the results of the study. However, if there are any risks of influencing patient outcomes, these risks for bias on statistical outcomes may be of great concern [2].

How can variables that have the potential to affect study results be controlled? The researcher should review whether adequate and appropriate randomization has been performed. Randomization reduces risk for confounding by generating groups that are fairly comparable with regard to known and unknown confounding variables [13]. However, as mentioned above, randomization does not always prevent imbalance between two groups; therefore, to control imbalances in baseline data, several strategies can be applied.

First, restriction can eliminate variation in the confounder. Inclusion criteria could be restricted to a certain population of interest in the design and planning stages of the study [13]. For example, female is a risk factor for postoperative nausea and vomiting (PONV). If the researcher plans to study the relationship between a drug and PONV, the influence of this variable (i.e., sex) in the results may disappear by including only female

Table 3. Number of Baseline Variables (n) Compared in Randomized Controlled Trials during the Analysis Period (2010–2017)

Baseline variables compared	Trials
0	1
1–4	67
5–9	180
10–19	60
20–29	4
> 30	0
Total	312

Data are presented as number.

Table 2. Demographic Data and Postoperative Rescue Medications [6]

Characteristic	Group D (n = 30)	Group DP (n = 30)	P value
Sex (M/F)*	16/14	13/17	0.438
Age (yr)†	8.1 ± 3.4	10.0 ± 3.9	0.052
Weight (kg)†	27.3 ± 11.5	34.7 ± 15.9	0.042
Height (cm)†	127.5 ± 21.4	133.7 ± 20.5	0.257
Operation duration (min)†	119.3 ± 19.3	113.0 ± 26.9	0.299
Anesthetic duration (min)†	161.3 ± 23.3	165.1 ± 26.5	0.554
Recovery time (min)†	14.3 ± 8.4	12.1 ± 10.5	0.382
Postoperative analgesics/antiemetic			
Fentanyl consumption (μg/kg)†	10.7 ± 2.6	11.1 ± 2.0	0.534
Rescue analgesic needed‡	14 (46.7%)	8 (26.7%)	0.180
Rescue antiemetic needed‡	14 (46.7%)	13 (43.3%)	1

Data are presented as number or mean ± SD, number (%). Statistical analyses were performed using the *chi-squared test, †Student t-test, or ‡Fisher's exact test. Groups D and DP represent dexamethasone only, and dexamethasone and propofol treated-patients, respectively.

subjects. Furthermore, including only elderly patients or infants in studies can control the influence of age on the study result(s).

Second, stratified randomization could help to prevent confounding variables that cause bias by chance with the help of generating strata *before* randomization. For example, when patient age is anticipated to be a highly important factor that may affect the results, the age of the included patients can be stratified into several groups (e.g., group 1, 20–40; group 2, 40–60; group 3, 60–80; group 4, > 80 years of age). Stratification can help mitigate the level of confounding and produce groups in which the confounder does not vary [13]. However, stratification could also cause the size of subgroups to be smaller (data thinning).

Third, covariate adaptive randomization helps to prevent imbalance in important covariates that could affect study outcomes. Covariate adaptive randomization assigns new subjects to the treatment groups, taking into account the covariates of previously assigned subjects to the treatment groups [14].

Finally, statistical methods that adjust for possible covariates, such as analysis of covariance (ANCOVA) or multivariate analysis of covariance (MANCOVA), can be used. These methods, which adjust for a highly prognostic covariate, can improve precision. Covariates should be chosen on the basis of their possible correlation with variable and outcomes, regardless of whether the baseline data exhibit "imbalances in statistical tests." Covariates should be chosen during the design and planning stages of studies with thorough consideration. If chosen, a covariate must be adjusted for, regardless of whether imbalances are observed. Even if there is little imbalance, adjustment of covariates will result in smaller standard errors, tighter confidence intervals, and more powerful significance tests. However, if it performed without consideration during the design and planning stages of the study, typical methods for estimating standard errors will incorrectly assume that the investigators never controlled for the variable, regardless of the extent of imbalance observed. Additionally, sample size of the study should be calculated based on the planned method of statistical analysis [15].

Most importantly, all of the methods mentioned above should be concretely established at the design and planning stages of the study, and should be included in the statistical analysis plan. If statistical imbalance in baseline characteristics arises, and the researchers missed all the methods mentioned above, the researcher may be questioned by reviewers or readers whether the outcome has been influenced by the imbalance in baseline characteristics. In such circumstances, an additional/supplemental adjusted analysis can be performed, considering the imbalance in baseline characteristics. The researcher can imitate the adjustment for all variables that were identified as prognostic factors in advance. If both adjusted and unadjusted analyses yield the same results, interpretation is achieved without difficulty and the conclusions would be accepted without dispute. However, if adjusted and unadjusted analyses yield different results, there could be a debate about certain results and their proper interpretation. In such cases, the results from statistical methods that are pre-planned in the statistical analysis plan should be taken primarily [16]. If the study was well conducted according to the pre-planned statistical methodology, the authors suggest that imbalances in data could be considered to have been caused by chance. In addition, bias is not expected to be serious if investigators pre-plan the statistical method, analyze data according to pre-planned statistical method and be forthcoming and transparent in describing the limitations of the study in the discussion section [17].

In conclusion, the authors suggest that authors should try to apply strategies to control the imbalance for possible confounders in the design and planning stages of the study rather than reporting P values for baseline data in RCTs for the purposes of demonstrating that the randomization process was adequate.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Author Contributions

Eun Jin Ahn (Data curation; Investigation; Methodology; Writing – original draft; Writing – review & editing)
Jong Hae Kim (Data curation; Investigation; Writing – review & editing)
Tae Kyun Kim (Data curation; Investigation; Writing – review & editing)
Jae Hong Park (Data curation; Investigation; Writing – review & editing)
Dong Kyu Lee (Conceptualization; Data curation; Methodology; Writing – review & editing)
Sangseok Lee (Conceptualization; Data curation; Investigation; Writing – review & editing)
Junyong In (Data curation; Investigation; Writing – review & editing)
Hyun Kang (Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing)

ORCID

Eun Jin Ahn, https://orcid.org/0000-0001-6321-5285
Jong Hae Kim, https://orcid.org/0000-0003-1222-0054
Tae Kyun Kim, https://orcid.org/0000-0002-4790-896X

Jae Hong Park, https://orcid.org/0000-0003-0779-4483
Dong Kyu Lee, https://orcid.org/0000-0002-4068-2363
Sangseok Lee, https://orcid.org/0000-0001-7023-3668
Junyong In, https://orcid.org/0000-0001-7403-4287
Hyun Kang, https://orcid.org/0000-0003-2844-5880

References

1. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ 1999; 319: 185.
2. Senn S. Testing for baseline balance in clinical trials. Stat Med 1994; 13: 1715-26.
3. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990; 335: 149-53.
4. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-9.
5. Kim JH, Kim TK, In J, Lee DK, Lee S, Kang H. Assessment of risk of bias in quasi-randomized controlled trials and randomized controlled trials reported in the Korean Journal of Anesthesiology between 2010 and 2016. Korean J Anesthesiol 2017; 70: 511-9.
6. Kim J, Jang GD, Kim DS, Min KT. Small dose of propofol combined with dexamethasone for postoperative vomiting in pediatric Moyamoya disease patients: a prospective, observer-blinded, randomized controlled study. Korean J Anesthesiol 2013; 64: 127-32.
7. Choi HR, Cho JK, Lee S, Yoo BH, Yon JH, Kim KM. The effect of remifentanil versus N2O on postoperative pain and emergence agitation after pediatric tonsillectomy/adenoidectomy. Korean J Anesthesiol 2011; 61: 148-53.
8. Jeong SJ, Han JI, Baik HJ, Lee H, Lee GY, Kim JH. The effect of pyridostigmine on bispectral index during recovery from sevoflurane anesthesia. Korean J Anesthesiol 2011; 61: 460-4.
9. Kim GH, Ahn HJ, Kim HS, Bang SR, Cho HS, Yang M, et al. Postoperative nausea and vomiting after endoscopic thyroidectomy: total intravenous vs. balanced anesthesia. Korean J Anesthesiol 2011; 60: 416-21.
10. Shin YH, Kim MH, Lee JJ, Choi SJ, Gwak MS, Lee AR, et al. The effect of midazolam dose and age on the paradoxical midazolam reaction in Korean pediatric patients. Korean J Anesthesiol 2013; 65: 9-13.
11. Siddiqui KM, Ali MA, Ullah H. Comparison of spinal anesthesia dosage based on height and weight versus height alone in patients undergoing elective cesarean section. Korean J Anesthesiol 2016; 69: 143-8.
12. Jung W, Hwang M, Won YJ, Lim BG, Kong MH, Lee IO. Comparison of clinical validation of acceleromyography and electromyography in children who were administered rocuronium during general anesthesia: a prospective double-blinded randomized study. Korean J Anesthesiol 2016; 69: 21-6.
13. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench 2012; 5: 79-83.
14. Kang H. Random allocation and dynamic allocation randomization. Anesth Pain Med 2017; 12: 201-12.
15. Lee S, Kang H. Statistical and methodological considerations for reporting RCTs in medical literature. Korean J Anesthesiol 2015; 68: 106-15.
16. Altman DG. Covariate Imbalance, Adjustment for. Encycl Biostat 2005: 1273-8. Available from https://doi.org/10.1002/0470011815.b2a01015.
17. Permutt T. Testing for imbalance of covariates in controlled experiments. Stat Med 1990; 9: 1455-62.