The Blind Review: A quality assessment measure for review standardization of Institutional Review Boards

Sinyoung Park

Yonsei University Health System

Cho Rong Ahn

Yonsei University Health System

Yang Hee Noh

Yonsei University Health System

Sejung Park

Yonsei University College of Medicine

Chung Mo Nam

Yonsei University College of Medicine

Myoung Soo Kim

Yonsei University Health System

Seung Min Kim

Yongin Severance Hospital

Jin Seok Kim

Yonsei University Health System

Sun Young Rha (✉ rha7655@yuhs.ac)

Yonsei University Health System

Research article

Keywords: Institutional Review Board, Standardization, Quality Assurance

Posted Date: December 18th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-42889/v3>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Standardization of IRB reviews has become increasingly important with the rise in multinational trials. Though inconsistency is often inevitable because of varying opinions on ethics, standardizing and understanding the differences between review results is required to ensure that high IRB review quality is maintained. Therefore, we aimed to develop a quality assessment measure of IRB, named “blind review,” by reviewing the same research protocols followed by multiple IRB panels. We then analyzed the differences between the panels to understand the mechanism of IRB standardization.

Methods: Based on the Human research Protection Program (HRPP) Standard Operating Procedures (SOPs), eight blind review results from January 2010 to December 2018 at a single institution with multiple panels, using the Severance Hospital HRPP database as the source, were analyzed. The review scores ranged from 0 to 60 points, including good clinical practice (GCP) requirements and protocol issues. Panel agreement was estimated by observed multiple rater agreement. Differences between review scores according to member expertise and IRB member duration were analyzed using the Wilcoxon rank-sum test and Kruskal-Wallis test.

Results: The observed multiple raters’ agreement increased from 0.444 (95% CI: 0.167-1.000) in 2010 to 0.479 (95% CI: 0.271-0.708) in 2014-2018, as IRB reviewer experience increased. To analyze the review mechanism, three GCP requirements and three protocol issues were scored (range 0 to 60). The mean values for GCP requirements and protocol issues were 19.25 ± 8.21 and 18.40 ± 9.04 , respectively. The mean score of the panels in which experts participated ($n=16$, 28.13 ± 10.47) was higher than those of the control group ($n=32$, 25.16 ± 10.96) ($p=0.93$). According to IRB members’ experience, scores for the group whose career spanned less than 3 years was 25.0 ± 10.0 ($n=14$), those for the group whose career spanned 3-5 years was 26.3 ± 9.6 ($n=23$), and those for the group whose career spanned more than 5 years was 27.3 ± 14.2 ($n=11$). These results were statistically significant ($p=0.09$).

Conclusions: We suggest blind review as an effective measure for overseeing and ensuring IRB review quality and overall GCP compliance.

Background

The Institutional Review Board (IRB), also known as the Independent Ethics Committee or Ethical Review Board, plays an important role in protecting the rights and welfare of human research subjects [1]. IRBs review the consent process, recruitment procedure, compensation arrangement, and the scientific validity of submitted documents, as well as conduct risk and benefit analysis [2]. In recent decades, medical knowledge and research techniques have advanced remarkably. In response, laws and regulations related to clinical research ethics have become ever more complex [3, 4]. As multinational clinical trials increase in number, both domestic laws and international regulations are becoming more important, and whenever there is any discrepancy, the more rigorous of them should be followed in each IRB review [5]. Such changes to the research environment create problems for both IRB members and investigators. These

include unintentional non-compliance through incomprehension of the up-to-date regulations, which requires frequent good clinical practice training to resolve [6]. There have been discrepancies between IRB decisions regarding multicenter studies, even when following the same protocol [7]. These divergences may be due to different interpretations of Good Clinical Practice (GCP), different local regulatory requirements, or the individual characteristics of IRB members.

In 1982, Veatch demonstrated the inconsistency in IRB review results for the first time [8]. Goldman and Katz [9] also analyzed IRB review results using the same protocols. The inconsistency in IRB decisions is an inveterate concern, both for IRB members and investigators, presenting an additional burden for review. Several investigations have analyzed the inconsistencies in IRB panels' results [10-14]. Further, consistent IRB results are associated with high-quality IRB discussions and determinations, based on appropriate attention to important issues and fulfilling the expectation of researchers [15, 16]. However, because of the characteristics of the IRB and its reviews, including its diverse members and complex study designs, inconsistencies are inevitable. Lynch et al. [17] recently demonstrated that a broader concept of IRB and HRPP quality includes (1) effectiveness, (2) procedures and structures likely to promote effectiveness, and (3) features unrelated to effectiveness but nonetheless essential. Therefore, we must analyze why such inconsistencies occur [13], based on their details. However, research on review mechanisms or their associated factors is lacking. Most studies have focused only on inconsistencies using one or two protocols. Therefore, we developed "blind review" as a quality measure that compares the review results of different IRB panels following the same protocol to determine the degree of standardization as an empirical evaluation. We evaluated the effectiveness of this method, as well as factors affecting reviews, especially those relating to IRB members' experience. This study piloted the blind review and provides information on its effectiveness based on data from a single institution.

Methods

Blind review

We developed the blind review methodology as a part of quality assurance (QA) activities within human research protection programs (HRPPs) to analyze consistency among IRB panel decisions. In this process, one protocol selected for blind review was submitted to every IRB panel except to the one that originally conducted the review. Three HRPP staff members with more than two years' experience in the Human Research Protection Center (HPC) selected the protocol. Although there were no specific criteria for selection, we considered current research trends and specific issues, such as Artificial Intelligence software study, verbal consent, and germline mutation study, whose review guidelines are a work in progress. Thus, we included various research types to compare the decisions made by different IRB panels, and did not exclude poorly designed protocols. Identical protocols were submitted to each panel simultaneously via a convened meeting, and IRB members were not notified about the blind review. After the blind review, HPC QA staff assessed the results and shared them in the IRB member workshop. However, the investigators were not notified about the blind review results and thus had no obligation to

respond to any comments from the blind review. The IRB's decision regarding the protocol was made by the original review panel. A detailed description of the blind review process is shown in Figure 1.

Materials

As a QA measure, the HPC at Severance Hospital, Yonsei University College of Medicine, Seoul, Korea, has conducted a blind review annually, according to the HRPP standard operating procedures (SOPs). We retrospectively analyzed the blind review records of eight protocols conducted between January 2010 and December 2018. The first three results, from 2010, were pilot tests across four panels; the other five, one for each year from 2014 and 2018, were across seven panels and aligned with the SOP. No blind review was conducted between 2011 and 2013 as our institution had to develop an e-IRB system from the paper review process. We collected IRB member designation logs from the HPC database to assess IRB member factors, such as IRB years of experience and major specialty. We did not enroll any human research subjects, or use their personal data. Therefore, IRB approval was not required for this study.

Review criteria specified by the Severance Hospital IRB

Severance Hospital follows domestic regulations for research involving human participants, such as the Pharmaceutical Affairs Act, Bioethics and Safety Act, Korean GCP guidelines, and the Medical Device Act. However, if there are issues not specified in domestic laws, and if the following regulations are stricter than the ones specified in domestic laws, internal regulations will be provided by referring to International Council for Harmonization (ICH), GCP guidelines, US Department of Health and Human Services regulations, federal and local laws, regulations for Food and Drug Administration participants (21 CFR 50, 56, 312 and 812), etc. The criteria of IRB review and approval are as follows:

- (1) Approval: If the research activity conforms to the criteria for approval defined by the related law, and if modifying the research is not recommended.
- (2) Approved with modifications: The research activity meets the criteria for approval, and the modifications required by the IRB are only minor and of minimal risk.
- (3) Deferred: The study does not meet the criteria for approval as defined in the relevant laws and regulations, lacks sufficient information to conduct an adequate review, and/or if substantial revisions to the protocol are necessary.
- (4) Disapproved: The study is found to be unethical, without scientific or scholarly merit, and/or does not meet the criteria for approval as defined in the relevant laws and regulations.
- (5) Tabled: A study that cannot be reviewed at the meeting because of lack of time, lack of quorum, and/or extenuating circumstances.

Scoring method for IRB review results

To assess the review quality and standardization, HPC staff developed “essential review points,” based on GCP and HRPP guidelines, and “protocol review points” that should be detected by IRB members. These were to be discussed during the review. In detail, as a first step, the QA team staff analyzed the IRB minutes and set out the discussion points. The second step was to collect the IRB minutes of all panels and confirm the details of the discussion results. Similar to conducting a systemic review or meta-analysis study, we found the specific word and other similar words to ensure that all relevant discussions were included. Then, the QA team discussed and reached a consensus on the content of the discussion and whether it validated the blind review score. Thereafter, we developed a 3-point scale with the values 0, 5, and 10, with 10 indicating that both reviewers addressed each review point and 5 indicating that the primary reviewer addressed it but not the secondary reviewer. This was applied even when a wrong comment was made in the IRB review (it is important to know what to point out as well as make accurate comments). Thus, even if an accurate point was not made, if any IRB member discussed the topic, we marked 5 points for the question. A score of 0 was allocated when none of the IRB members discussed or mentioned the issue. For GCP requirements, we identified “risk and benefit analysis” (ICH-GCP 2.2), “determining the continuing review interval” (ICH-GCP 3.3.4), and “research resources” (ICH-GCP 3.1.3., 3.1.9) as mandatory review points based on GCP regulations [18]. In addition, we assessed whether the panels identified the issues to be discussed, as given information what is required. The consent process, vulnerability of participants, waivers for stored specimens, concerns regarding the confidentiality of participants, randomization methods, and the appropriateness of placebo procedures for the control group formed the majority of these considerations.

Statistical Analysis

Descriptive data were expressed as mean $\pm$ standard deviations. Differences between review scores, according to member expertise and IRB experience, were analyzed using the Wilcoxon rank-sum test and Kruskal-Wallis test [19, 20]. Since Fleiss’ kappa has paradoxically low values, despite the high proportion of agreement when the marginal totals of raters are imbalanced, the observed multiple rater agreement was used instead of Fleiss’ kappa to assess the agreement of IRB results among panels [21]. The observed multiple rater agreement was estimated by summing all pairwise agreement tables from the panels, and a 95% confidence interval (CI) was obtained by 1,000 instances of bootstrap resampling [22]. All statistical analyses were conducted using SPSS statistical software version 23 (SPSS Inc., Chicago, IL, USA), SAS software version 9.4 (SAS Institute Inc., Cary, NC), and R software version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria). A two-sided p-value <0.05 was considered statistically significant.

Results

Baseline Characteristics of protocols and IRB

We executed eight different blind review protocols, from March 2010 to October 2018. Among the eight protocols, 75% (six out of eight) were reviewed in the convened meeting and 25% (two out of eight) in an

expedited review originally prior to a blind review, based on review type criteria (Convened meeting/Expedited Review). These eight protocols were selected as the “blind review protocol,” and all of them were reviewed via a convened meeting despite the actual review process. Overall, single-center studies (75%) and interventional studies (62.5%) prevailed among protocols included in the blind review. The study population and notable points of discussion are described in Table 1.

Table 1. Baseline characteristics of protocols included in the blind review (n=8)

No	Year	Center type	Study type	Study population	Major discussion issue	Panels (n)
1	2010	Single Center	Non-intervention	Breast cancer	Retained tissue	4
2	2010	Multinational, Multicenter	Intervention (Phase 3)	Pulmonary hypertension	Extension study	4
3	2010	Single Center	Non-intervention	Epilepsy	Privacy for epilepsy	4
4	2014	Single Center	Non-intervention	Meniere's syndrome	Phone Screening	7
5	2015	Single Center	Intervention	Atrial fibrillation	Proper study design	7
6	2016	Single Center	Intervention	Sleep apnea syndrome	Sham procedure	7
7	2017	Domestic, Multicenter	Intervention (Phase 2)	Hepatocellular Carcinoma	Rationale for study	7
8	2018	Domestic, Multicenter	Intervention	End-stage Renal Disease	Artificial Intelligence	7

Agreement among the panels on each protocol

The IRB review results of different panels were scored on a five-point scale, with “Approved” =1, “Approved with modification” = 2, “Deferred” = 3, “Disapproved” =4, and “Tabled” =5. The correspondence of IRB results among panels according to the protocol is visualized in Figure 2. We analyzed IRB agreements across two groups—the pilot assessment with three protocols and the main assessment with five protocols. The observed multiple rater agreement was 0.444 (95% CI: 0.167-1.000) with protocols 1-3 in 2010, while an agreement of 0.448 (95% CI: 0.362-0.514) was observed between 2014 and 2018 with seven panels. To include recent blind review results, we extracted the results of panels 1-4 from all the reviewed protocols (pilot and main), which showed that agreement increased to 0.479 (95% CI: 0.271-0.708) from 0.444 (pilot) and 0.448 (main) (Table 2).

Table 2. Observed multiple raters’ agreement on Blind Review results

	Pilot Study (Protocol 1, 2, 3)	Main Study (Protocol 4, 5, 6, 7, 8)
Observed multiple raters’ agreement	0.444¶ (95% CI: 0.167, 1.000)	0.448† (95% CI: 0.362, 0.514)
Observed multiple raters’ agreement (Only with panels 1-4)	0.479¶ (95% CI: 0.271, 0.708)	

¶ Panels 1-4, † Panels 1-7

Scores and associated factors regarding IRB review results

The scores ranged from 0 to 30 points for each protocol, including GCP requirements and protocol issues, and thus the total score ranged from 0 to 60 points. We analyzed the mean score and standard deviation and compared these between protocols (Table 3). The mean values of the two categories, including GCP requirements and protocol-specific issues, were 19.25±8.21 and 18.40±9.04, respectively. In terms of GCP requirements, the scores gradually increased from the pilot studies in 2010 through to the recent group (Protocol 1= 15±12.25, Protocol 2=15±9.13 vs. Protocol 7=22.8±6.98, Protocol 8=20±8.16). Compared to GCP comments, the scores for the protocol review issue were similar to the average. A comparison of the scores produced by the different protocols is shown in Figure 3, using a graph with a standard deviation range. Detailed scores and review issues are described in Supplementary Tables 1 and 2.

Table 3. Comparison of the review scores according to the protocols

	Pilot study			Main study				
	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5	Protocol 6	Protocol 7	Protocol 8
GCP	15 ± 12.25	15 ± 9.13	17.5 ± 5.0	20 ± 6.45	23.6 ± 4.76	15.7 ± 11.34	22.8 ± 6.98	20 ± 8.16
Protocol	18.75 ± 10.30	15 ± 12.90	17.5 ± 9.57	25.71 ± 5.34	18.57 ± 10.69	17.14 ± 7.56	15.71 ± 6.07	17.14 ± 11.12
Total	33.75 ± 7.5	30 ± 16.83	35 ± 12.91	45.71 ± 4.5	42.14 ± 12.20	21.43 ± 12.15	38.57 ± 11.80	37.14 ± 17.04

(mean ± standard deviation)

To assess the impact of IRB members' characteristics on review results, we collected information on the experience of the IRB members who participated in the convened meeting. In addition, we also investigated whether there were any IRB members who had specific insights regarding the protocols that were applied in the blind review. For example, protocol 2 focused on pulmonary hypertension; thus, the IRB panel that featured a member who was familiar with cardiology was regarded as containing a member specialized in that protocol. We excluded the IRB members who had disclosed conflicts of interest regarding the protocol, as well as those panels that had no specialized members. The differences in the review results based on the inclusion/exclusion of expert IRB members are as follows (Table 4). The mean score of the panels with experts was higher than that of the control group, but this result was not statistically significant ($p=0.93$). According to the career duration of IRB members, review scores of the groups whose members' careers spanned less than 3 years or 3-5 years were shown to be generally low compared to the groups whose members' careers spanned more than 5 years ($p=0.09$).

Table 4. The difference in review results depending on the presence of experts on each IRB panel.

	Review score (mean ± SD)	p-value
Including/excluding expert members		
No expert participated	25.16 ± 10.96	0.93*
Expert participated	28.13 ± 10.47	
Duration of IRB members' career		
<3 years	25.0 ± 10.0	0.09¶
3-5 years	26.3 ± 9.6	
≥ 5 years	27.3±14.2	

*Wilcoxon rank-sum test (Mann-Whitney U test), ¶Kruskal-Wallis test, SD: Standard Deviation

Discussion

Inconsistencies in IRB decisions have been observed since the 1980s, largely due to an increase in the number of multicenter clinical trials [9, 14]; furthermore, the fact that IRB panels sometimes review the same protocol in different ways remains a major issue. IRBs aim to conform to the protocol and ensure sound research design, both scientifically and ethically. Thus, in terms of ethical considerations, differing opinions are an undisputed consequence. In addition, consistency among IRBs cannot ensure a correct decision, but major inconsistencies can affect the reliability of the IRB in terms of the wider research environment. Therefore, HRPP QA activities must verify whether IRBs identify inappropriate designs or miss essential review points.

This study analyzed the IRB minutes of panels at the same institution, following the same IRB regulations, to observe the consistency of review results. Several studies have shown that IRBs are relatively consistent in terms of their final decision but are often inconsistent regarding the reasons for their decisions [11, 12, 23, 24]. Taylor et al. [25] suggested a process for measuring the ethical quality of local IRBs. They defined seven points—scientific value, assessment of risk, benefit, acceptable risk/benefit ratio, fair selection of subject, adequately informed consent process, and adequate mechanisms for respecting enrolled subjects—while also confirming how many points were satisfied when one protocol was reviewed by multiple IRBs. A revised common rule was also announced recently regarding the use of a single IRB, which aims to reduce administrative costs. As of January 25, 2018, the National Institutes of Health (NIH) required the use of a single IRB record for most domestic NIH-funded multisite research studies [26]. There is growing interest in single IRBs, and a research study has developed a measuring process for IRB inconsistency called “mystery shopper” and a scoring system using a discussion theme [13].

Although many studies continue to find inconsistencies, it is still unclear why such inconsistencies occur in multiple IRBs. Our analysis sought to clarify this inconsistency by considering both the issue of discussion and the characteristics of IRB members. We assessed the consistency of seven IRB panel results within a single institution. Partial consistency was observed, although we have only analyzed the results in an exploratory manner. Agreement was shown to be poor, though we focused on the review process rather than on the review results. In addition, to resolve the disagreements in each panel, HPC provided IRB workshops to discuss the inconsistent blind review results and review process as an HRPP QA activity. In addition, to overcome the limitations of previous studies with a scoring system, HRPP staff in this study comprehensively discussed and determined the review points. The total score can be considered a direct measuring tool for IRB QA activity. Based on our data, compliance with the GCP requirements during an IRB review showed gradual improvement over time, while protocol-specific reviews were influenced by the characteristics of each protocol, making it relatively difficult to achieve consistency or standardization. This indicates that IRB reviews are commonly affected by specific protocols. However, both the observed multiple raters’ agreement and the review scores increased with an increase in the experience of panel members. Therefore, IRB members’ career spans and compatibility are factors that affect IRB review processes, and active and systematic training of IRB members is

recommended to improve the quality of IRB reviews. IRBs are not a permanent system—rather, they should be considered a regular interaction space for experts and non-experts. There is a significant need to support IRBs at the institutional level because laws change continuously and often become stricter. Our data show that the continued involvement of IRB members is essential to ensure and improve the quality of IRB reviews. Continuous monitoring of IRB reviews is necessary for agreed standardization, excluding extreme and odd results rather than viewing them as being of equal merit. We recommend that the blind review process be integrated into the HRPP's activities for assessing the quality of an IRB review. Notably, proxy outcomes with review standardization can lead to participant protection within the HRPP, while IRBs were originally instituted to secure participants' rights, welfare, and well-being. Our results can be applied practically while the regulatory compliance and proportionality of multiple IRBs is integrated into the advances made in IRB and HRPP effectiveness [17]. Furthermore, standard review results and IRB recommendations from many different IRBs can accelerate the initiation of multicenter clinical trials, boosting global healthcare innovation based on increasing the opportunity for patient access to novel therapeutic agents.

This study has several limitations. First, it was a retrospective study that was based on IRB minutes from a single tertiary hospital. This limits the generalizability of our data to the HRPP environment. However, we report our experience as a preliminary study—we provide basic information for the blind review and the monitoring process for IRB reviews, and hope that this will be explored in further studies. In addition, a blind review process to the reviewers can also make the results more scientific and unbiased. Second, only a small number of protocols were included in the blind review. However, as our center has operated HRPP since 2010, and maintains 8 years of experience in IRB review quality assessment, we believe that any degree of bias was minimal. Third, we analyzed the review scores discretionally, and the review points used for consensus among HRPP staff had no scientific criteria. Further studies should investigate and validate this as a multicenter study involving a larger number of protocols.

Conclusion

Blind review is an effective method for overseeing and ensuring the quality of IRB reviews, as well as overall GCP compliance. The usefulness of the blind review process has not been established because of insufficient evidence. However, according to our study, this process seems to be an effective monitoring method in the highly challenging environment of clinical research ethics.

List Of Abbreviations

IRB: Institutional Review Board; GCP: Good Clinical Practice; QA: Quality Assurance, HRPP: Human Research Protection Programs; HPC: Human research Protection Center; SOP: Standard Operating Procedures; ICH: International Council for Harmonization

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets analyzed during the current study are not publicly available due to the confidentiality of IRB review; however, they are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors read and approved the final manuscript to be published. S.Y.R contributed to the research questions and design, as well as to data collection. S.P(1), Y.H.N, C.R.A, S.M.K, M.S.K. J.S.K, and S.Y.R. contributed to the study concept and design. S.P(1), C.M.N, S.P(2), and S.Y.R contributed to the analysis and interpretation of data.

Acknowledgements

The authors would like to thank all HPC staff members for assisting in data collection, and their constant dedication through their efforts on IRB and HRPP.

References

1. Resnik DB. Some reflections on evaluating institutional review board effectiveness. *Contemporary clinical trials*. 2015;45(Pt B):261-4.
2. Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Journal of the Indian Medical Association*. 2009;107(6):403-5.
3. Grady C. Do IRBs protect human research participants? *Jama*. 2010;304(10):1122-3.
4. Greene SM, Geiger AM. A review finds that multicenter studies face substantial challenges but strategies exist to achieve Institutional Review Board approval. *Journal of clinical epidemiology*. 2006;59(8):784-90.

5. Park S, Nam CM, Park S, Noh YH, Ahn CR, Yu WS et al. 'Screening audit' as a quality assurance tool in good clinical practice compliant research environments. *BMC medical ethics*. 2018;19(1):30.
6. Park S, Noh YH, Rha SY, Kim WH, Cheon JH. Institutional Board Review for Clinical Investigations on Inflammatory Bowel Diseases: A Single-Center Study. *Intest Res*. 2015;13(3):274-81.
7. Penn ZJ, Steer PJ. Local research ethics committees: hindrance or help? *British journal of obstetrics and gynaecology*. 1995;102(1):1-2.
8. Veatch RM. Problems with Institutional Review Board inconsistency. *Jama*. 1982;248(2):179-80.
9. Goldman J, Katz MD. Inconsistency and institutional review boards. *Jama*. 1982;248(2):197-202.
10. Angell E, Sutton AJ, Windridge K, Dixon-Woods M. Consistency in decision making by research ethics committees: a controlled comparison. *Journal of medical ethics*. 2006;32(11):662-4.
11. Edwards SJ, Stone T, Swift T. Differences between research ethics committees. *International journal of technology assessment in health care*. 2007;23(1):17-23.
12. Taljaard M, Brehaut JC, Weijer C, Boruch R, Donner A, Eccles MP et al. Variability in research ethics review of cluster randomized trials: a scenario-based survey in three countries. *Trials*. 2014;15:48.
13. Trace S, Kolstoe SE. Measuring inconsistency in research ethics committee review. *BMC medical ethics*. 2017;18(1):65.
14. Hotopf M, Wessely S, Noah N. Are ethical committees reliable? *Journal of the Royal Society of Medicine*. 1995;88(1):31-3.
15. Abbott L, Grady C. A systematic review of the empirical literature evaluating IRBs: what we know and what we still need to learn. *J Empir Res Hum Res Ethics*. 2011;6(1):3-19.
16. Hall DE, Feske U, Hanusa BH, Ling BS, Stone RA, Gao S et al. Prioritizing Initiatives for Institutional Review Board (IRB) Quality Improvement. *AJOB Empir Bioeth*. 2016;7(4):265-74.
17. Lynch HF, Nicholls S, Meyer MN, Taylor HA. Of Parachutes and Participant Protection: Moving Beyond Quality to Advance Effective Research Ethics Oversight. *J Empir Res Hum Res Ethics*. 2019;14(3):190-6.
18. Guideline IHJCS. Integrated addendum to ICH E6 (R1): guideline for good clinical practice E6 (R2). 2015;2:1-60.
19. Fleiss JLJPb. Measuring nominal scale agreement among many raters. 1971;76(5):378.
20. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. John Wiley & Sons; 2013.
21. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*. 1990;43(6):543-9.
22. de Vet HCW, Mullender MG, Eekhout I. Specific agreement on ordinal and multiple nominal outcomes can be calculated for more than two raters. *Journal of clinical epidemiology*. 2018;96:47-53.
23. Koski G. Beyond compliance...is it too much to ask? *Irb*. 2003;25(5):5-6.
24. Anderson EE, DuBois JM. IRB decision-making with imperfect knowledge: a framework for evidence-based research ethics review. *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics*. 2012;40(4):951-69.

25. Taylor HA. Moving beyond compliance: measuring ethical quality to enhance the oversight of human subjects research. *Irb*. 2007;29(5):9-14.
26. Taylor HA, Ehrhardt S, Ervin AM. Public Comments on the Proposed Common Rule Mandate for Single-IRB Review of Multisite Research. *Ethics & human research*. 2019;41(1):15-21.

Supplementary Tables

Supplementary Table 1. Scoring of GCP requirements

		Panel #1	Panel #2	Panel #3	Panel #4	Panel #5	Panel #6	Panel #7
Protocol								
1	Risk and benefit analysis	0	10	10	10			
	Determining the continuing review interval	0	5	10	0			
	Research resources	0	10	5	0			
	Total score	0	25	25	10			
Protocol								
2	Risk and benefit analysis	10	5	10	10			
	Determining the continuing review interval	5	0	10	0			
	Research resources	10	0	0	0			
	Total score	25	5	20	10			
Protocol								
3	Risk and benefit analysis	0	10	10	10			
	Determining the continuing review interval	0	0	0	10			
	Research resources	10	10	10	0			
	Total score	10	20	20	20			
Protocol								
4	Risk and benefit analysis	10	10	0	10	10	10	10
	Determining the continuing review interval	10	10	10	0	10	10	0
	Research resources	5	5	5	0	5	5	5
	Total score	25	25	15	10	25	25	15
Protocol								
5	Risk and benefit analysis	10	5	10	10	10	10	10
	Determining the continuing review interval	10	10	10	10	10	10	10
	Research resources	0	10	10	10	0	0	0
	Total score	20	25	30	30	20	20	20
Protocol								
6	Risk and benefit analysis	10	10	10	0	10	10	0
	Determining the continuing review interval	0	10	0	0	10	10	0
	Research resources	10	0	10	0	0	10	0
	Total score	20	20	20	0	20	30	0
Protocol								
7	Risk and benefit analysis	10	10	10	10	10	10	10
	Determining the continuing review interval	0	10	10	0	0	10	0

	Research resources	10	10	10	5	10	10	5
	Total score	20	30	30	15	20	30	15
Protocol								
8	Risk and benefit analysis	0	0	10	10	10	10	10
	Determining the continuing review interval	10	10	10	10	10	10	10
	Research resources	0	0	0	0	10	0	10
	Total score	10	10	20	20	30	20	30

Supplementary Table 2. Scoring of Protocol review points

		Panel #1	Panel #2	Panel #3	Panel #4	Panel #5	Panel #6	Panel #7
Protocol 1	Protection of private information	10	0	0	0			
	Consent for retained tissue for FISH analysis	10	5	10	10			
	Applied revised Bioethics law in review	10	0	10	10			
	Total Score	30	5	20	20			
Protocol 2	Request information regarding previous main study	0	0	10	10			
	Subjects who experienced clinical relapse but enroll in an extension study with double blind	10	0	10	10			
	Necessity of resubmission as a modification review from a previous study	0	0	0	10			
	Total Score	10	0	20	30			
Protocol 3	Consent process for minors	0	10	0	10			
	Concerns about private information and family members when conducting germline mutation test	10	0	10	10			
	Process of pre-screening prior to acquisition of the verbal consent by telephone	0	10	0	10			
	Total Score	10	20	10	30			
Protocol 4	Proper verbal consent process	10	10	10	10	10	0	10
	Protection of private information	0	0	10	10	10	10	10
	Proper process of pre-screening by telephone	10	10	10	10	10	10	10
	Total Score	20	20	30	30	30	20	30
Protocol 5	Informed consent form containing difficult medical terms	0	10	10	10	0	10	0
	Identify exact intervention type for patient/control group	10	0	10	0	10	10	0
	Random assignment method	10	10	10	0	10	10	0
	Total Score	20	20	30	10	20	30	0

Protocol 6	Proper subject target based on investigator's affiliation	10	0	10	0	0	0	0
	Appropriateness of sham procedure for control group	10	10	10	10	10	10	10
	Discussion on additional hypertension medication	0	0	10	10	0	10	0
	Total Score	20	10	30	20	10	20	10
Protocol 7	Plan for human materials management and consent form	10	5	10	5	10	10	0
	Rationale for clinical trial	0	0	0	0	0	10	0
	Plan for adverse events management	10	10	10	0	10	0	10
	Total Score	20	15	20	5	20	20	10
Protocol 8	Considerations for randomization, especially control group	0	10	10	10	0	10	10
	Monitoring and management plan for algorithmic errors	0	0	10	10	0	10	10
	Discussion on whether the study should follow medical device law	0	0	0	0	10	10	10
	Total Score	0	10	20	20	10	30	30

Figures

Figure 1

Flow chart of the Blind Review.

Figure 1

Changes in review scores (a. GCP requirements, b. Protocol review points).

Figure 1

Blind Review results of the IRB panels according to protocols (a. Pilot study [2010], b. Main study [2014-2018]).