Analysis of the Quality of Academic Papers by the Words in Abstracts

Tetsuya Nakatoh[1(✉)], Kenta Nagatani[2], Toshiro Minami[3], Sachio Hirokawa[1], Takeshi Nanri[1], and Miho Funamori[4]

[1] Research Institute for Information Technology, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
nakatoh@cc.kyushu-u.ac.jp
[2] Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
[3] Kyushu Institute of Information Sciences, Fukuoka, Japan
[4] National Institute of Informatics, Tokyo, Japan

Abstract. The investigation of related research is very important for research activities. However, it is not easy to choose an appropriate and important academic paper from among the huge number of possible papers. The researcher searches by combining keywords and then selects an paper to be checked because it uses an index that can be evaluated. The citation count is commonly used as this index, but information about recently published papers cannot be obtained. This research attempted to identify good papers using only the words included in the abstract. We constructed a classifier by machine learning and evaluated it using cross validation. As a result, it was found that a certain degree of discrimination is possible.

Keywords: Bibliometrics · Research investigation · SVM · Citation

1 Introduction

The investigation of related research is a very important task for researchers. Therefore, databases of academic papers are now indispensable for researchers. Appropriate keywords generate lists of papers related to keywords from these databases. They may be very long, but in general, several scales are provided.

The citation count [1] is the most widely used evaluation scale for an paper. Many databases have a function for sorting the search results of papers by the number of citations. Although the citation count is a useful and objective measure, newly published papers cannot be evaluated. One solution to this problem may be an assessment of the journal in which it was published as a substitute for evaluating the paper directly. The impact factor (IF) is a typical measure used to evaluate academic journals, which reflects the annual average number of citations of papers published in that journal. It is the most frequently used standard. The IF has the ability to imply the relative importance of journals within

S. Yamamoto (Ed.): HIMI 2017, Part II, LNCS 10274, pp. 434–443, 2017.
DOI: 10.1007/978-3-319-58524-6_34

a specific field but is not appropriate for comparison across fields. Therefore, the IF is not suitable for ranking a paper search. Another alternative evaluation method would be to use the researcher's evaluation. A symbolic example of this approach is the h-index [4] proposed by Hirsch. Alternatively, the usefulness of orthometrics [15] has been demonstrated in recent years.

These measures are also very useful for researchers. However, the collection and analysis of such information incurs a large cost. In fact, it is said that the number of citations that are still emphasized is already a mechanism for information gathering because it has already been created [7]. Here, we thought that it would be impossible to determine the quality of a paper more directly from the information contained within the paper. The information in the paper is the following information described in the published paper: the title of the paper, the name of the author, the affiliation of the author, the keywords specified by the author, the abstract, and the text. We have evaluated the quality of papers by using data excluding the text of this bibliographic information [8]. In this research, it became clear that the influence of the journal and year of publication is strong.

In this paper, we attempt to classify papers more purely using only the words included in the abstract of the paper. We define good papers as papers with many citations and construct classifiers based on a support vector machine (SVM), featuring words contained in the abstract. The performance of the classifier is evaluated by 10-fold cross validation. In addition, we conduct a qualitative analysis on the words effectively acting on the classification of the paper.

2 Classification Method

An SVM is a pattern recognition model using supervised learning. In this study, we attempt to classify collected papers into two groups: excellent papers and not excellent papers using an SVM.

It is difficult to define an excellent paper. We decided to use a large number of citations for the definition of an excellent paper only in the evaluation of the method. The purpose of this study is to find excellent papers from only the information included in the paper without external information such as the citation count, so it may seem like a contradiction. However, since Citation Count is not used when actually classifying an paper according to this method, there is no inconsistency.

Paper having a citation count equal to or larger than a given threshold is defined as excellent paper. Based on the threshold, classifiers that organize the papers into two groups are constructed by machine learning. The attribute set used for learning is the frequency vector of words contained in the abstract. We use a 10-fold cross-validation for various thresholds and find the optimal parameters for the threshold.

3 Experiment

3.1 Experimental Data

The papers used for the experiments were collected for two different perspectives. One is the papers extracted from the paper database Scopus including the keyword bibliometrics, and the other is the papers published in 15 international conferences and 5 journals, which we believe are important in the computer science field, also extracted from Scopus. We select only the papers with the abstract. The former paper database is called DB_Bibliometrics, and the latter is called DB_CS. DB_Bibliometrics and DB_CS contain 8,072 and 38,766 papers, respectively. A list of journals selected for DB_CS is presented in Table 1.

All papers are classified by a threshold T for the number of citations and constructed as data input into the SVM, where an paper having a number of citations of T or more is positive; otherwise, it is negative. In this study, the classification performance at multiple thresholds T is analyzed, and the following list is used as the threshold: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100).

3.2 Classification

The SVM has multiple kernels available. In the pilot survey, we used the following four general kernel functions using LIBSVM[1]. The parameter used the default value of LIBSVM.

- linear
- polynomial
- radial basis function
- sigmoid

Four classes of kernels were applied to positive/negative split paper sets, and the classification performance was measured by 10-fold cross validation. As a result, it was almost impossible to classify papers, except with the linear kernel. Therefore, we analyzed the linear kernel in more detail. In the subsequent analysis, SVM^{perf}[2], which can classify the linear kernel at a high speed, was used.

In the linear kernel, there is a normalization parameter C to be set. For the default setting, $C = 0.01$, the following list of C values was used as a candidate determined to be useful in the preliminary analysis: (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100).

In the following, we show the normalization parameter C and performance graphs for the classification performance obtained for the chosen citations.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[2] https://www.cs.cornell.edu/people/tj/svm_light/svm_perf.html.

Table 1. Experimental data1 about computer science

Name of journal	# of papers
Journal of the ACM	1,217
VLDB Journal	617
ACM Transactions on Database Systems	776
IEEE Transactions on Knowledge and Data Engineering	2,436
Name of proceedings	# of papers
International Conference on Information and Knowledge Management	1,563
International Conference on Very Large Data Bases	854
Data Mining and Knowledge Discovery	466
European Conference on Research and Advanced Technology for Digital Libraries	424
International Conference on Asian Digital Libraries	438
International Conference on Data Engineering	2,225
IEEE International Conference on Data Mining	2,184
International Conference on Machine Learning	1,376
International Joint Conference on Artificial Intelligence	3,220
ACM IEEE International Conference on Digital Libraries	652
Pacific-Asia Conference on Knowledge Discovery and Data Mining	1,323
European Conference on Principles and Practice of Knowledge Discovery in Databases	1,027
SIAM International Conference on Data Mining	781
Annual International ACM SIGIR Conference on Research and Development in Information Retrieval	6,647
ACM International Conference on Knowledge Discovery and Data Mining	2,178
ACM International Conference on Management of Data	4,916
International Conference on Theory and Practice of Digital Libraries	201
International Conference on World Wide Web	3,245

Classification for DB_Bibliometrics. Figure 1 shows the classification performance with the regularization parameter C when classifying papers with given CC (Citation Count) in DB_Bibliometrics as positive. Of the many types of CC used for the calculation, 1, 10 and 100 are picked up and shown in the Fig. 1. A larger value of C enables stable classification. However, if the value of CC is large, classification is impossible regardless of the value of C.

On the basis of the above results, it can be seen that while the classification performance increases as the regularization parameter C increases, the

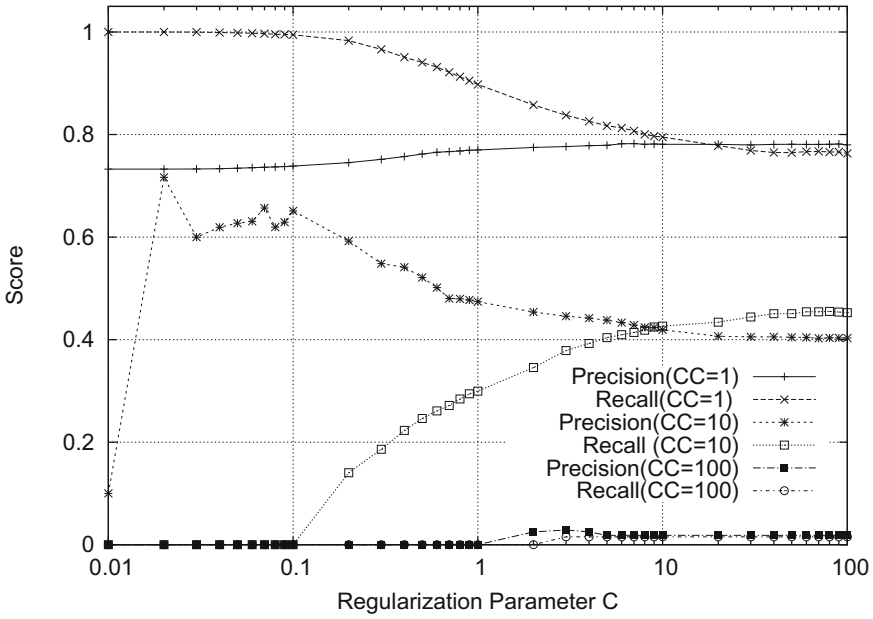

Fig. 1. Search of the regularization parameter for DB_Bibliometrics

classification performance rapidly decreases above a certain level. The relationship between the number of citations and the classification performance when $C = 10$, which seems to result in easier classification performance, is shown in Fig. 2. Although the classification performance is significant compared to the baseline (rate of positive: this is consistent with this line), it may be difficult to extract excellent papers with this classification performance.

Classification for DB_CS. Figure 3 shows the classification performance with the regularization parameter C when classifying papers with given CC (Citation Count) in DB_CS as positive. Of the many types of CC used for the calculation, 1, 10 and 100 are picked up and shown in the Fig. 3. A larger value of C enables stable classification. Regarding $CC = 10$ and $CC = 100$, there is a part where Precision goes down as C increases. However, since Recall is close to zero there, it is not a very good classification. Relatively around $C = 100$ is stable.

On the basis of the above results, we select $C = 100$, which seems to result in easier classification performance, and Fig. 4 shows the relation between the number of citations and the classification performance. Although this is also classified significantly compared with the baseline (rate of positive), it may be difficult to actually extract excellent papers with this classification performance.

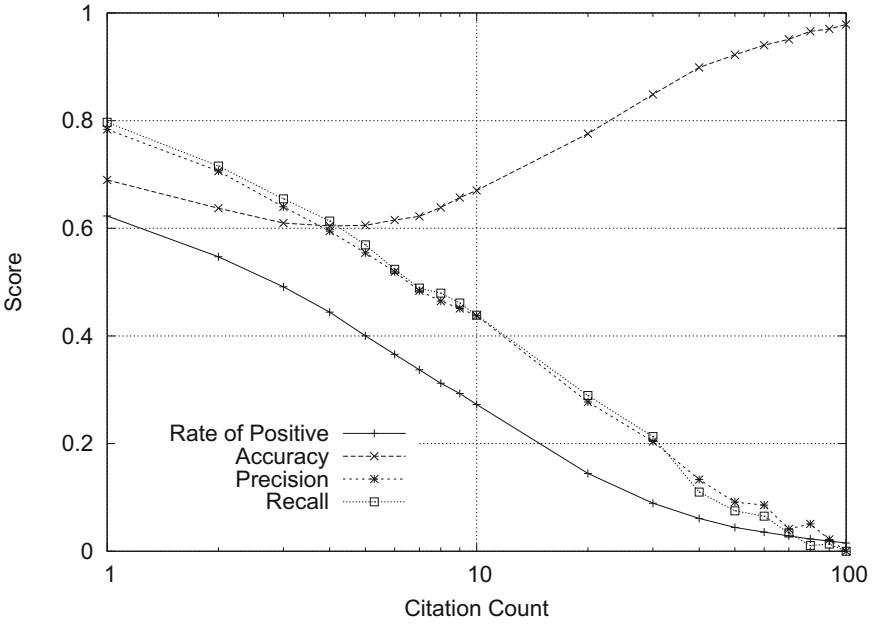

Fig. 2. Classification performance ($C = 10$)

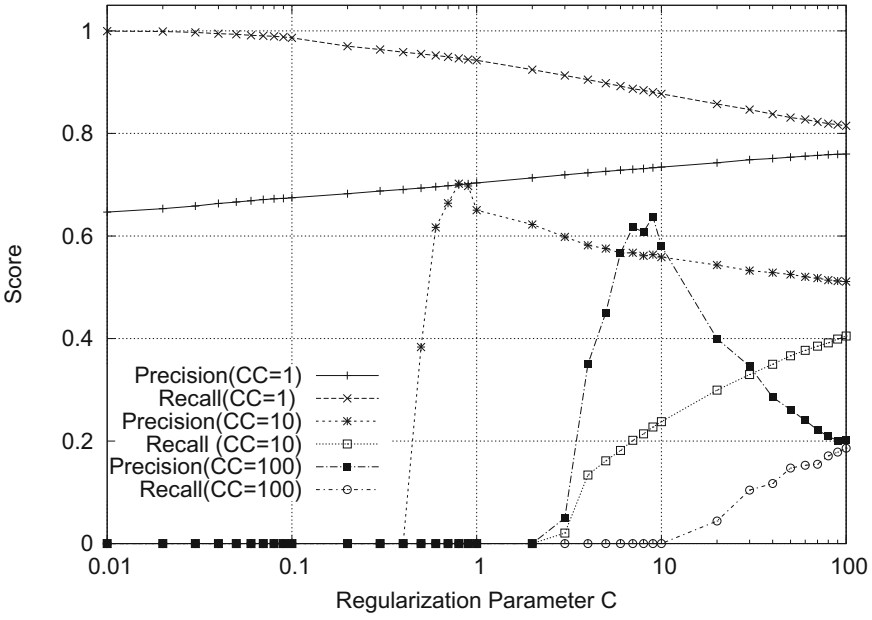

Fig. 3. Search of the regularization parameter for DB_CS

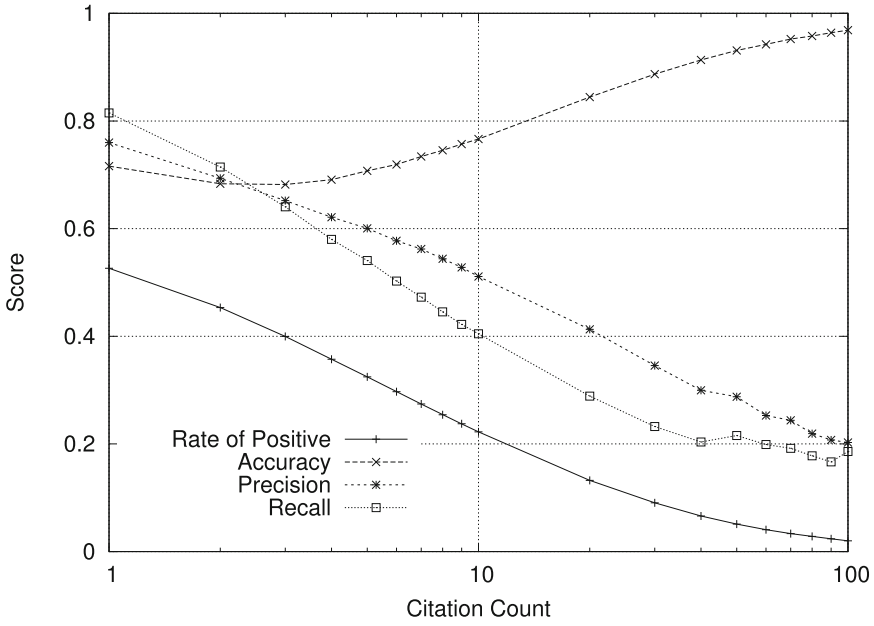

Fig. 4. Classification performance ($C = 100$)

3.3 Qualitative Analysis

In this section, we conduct a qualitative analysis on the words effectively acting on the classification of the paper. In the model generated by SVM, large weight is given to items (words) which greatly influence identification. Therefore, it is possible to judge an important feature word from the magnitude of the weight.

We select data with citation count of 1, 10, 100 as the analysis target. From the results of the classification performance, C with a large value of ROC Area (Receiver Operating Characteristic Area) was selected for each. From the model in the combination, Top 10 of the word weight was obtained for both positive and negative example.

For the characteristic words that classify papers in the bibliometrics field, the top 10 words with large absolute values are shown in the Table 2. Some of the feature words that appeared seem to be technical terms: nanotechnology, sjr (SCImago Journal and Country Rank), hisb (health information-seeking behavior), fret (Forster resonance energy transfer), mis (Minimally-invasive surgery), and wif (Web Impact Factor). All of them appeared on the positive side. It must be shown that a specific concrete research theme attracts many citations. Other words are common words. To judge these meanings, we need more analysis. Among them, it is very interesting that "bibliometrics" which is the field to be analyzed appears on the negative side.

Feature words that classify papers in the Computer Science field are similarly shown in the Table 3. More terms related to specialty appear here: imecho

Table 2. Feature Words for DB_Bibriometrics

CC	C	Polarity	Feature words
1	0.3	+	discuss, spain, terrorism, background, bias, basic, change, attempt, multiple, nature
1	0.3	−	bibliometrics, chinese, right, cooperation, cloud, hypertension, edit, literatures, reserve, imaging
10	0.2	+	nanotechnology, peer, site, percent, bias, illness, background, cocitation, firm, locate
10	0.2	−	bibliometrics, good, aim, education, secondary, mendeley, especially, hospital, big, explore
100	2	+	gs, sjr, hisb, innovative, fret, reconstruction, mis, actor, clear, wif
100	2	−	possible, accord, visibility, category, item, find, bibliometrics, importance, state, important

(a kind of search system), reproduction, yeast, clouddb, congruence, congruence, occf (One-Class Collaborative Filtering), kddcs (K-D tree based Data-Centric Storage), softrank, TAGME (a kind of system), TwitterRank, UMICs (upper-middle-income countries), MWEs (Mulberry water extracts), hilbert, vfdt (variance fractal dimension trajectory), edutella (a P2P network), diffsets, closegraph, sdms (Species distribution models), webml (Web Modeling Language), ordpath (a hierarchical labeling scheme). Many of these appear on the positive side, which seems to indicate that a specific concrete research theme attracts many citations. On the other hand, the negative side seems to contain a lot of sensuous words: noteworthy, metastories, reasonably, reformulate.

Table 3. Feature words for DB_CS

CC	C	Polarity	Feature words
1	60	+	imecho, reproduction, sbook, profitable, typology, yeast, clouddb, cap, holder, redescription
1	60	−	proceeding, copyright, eac, mwes, whilst, sampler, inewsbox, noteworthy, metastories, matrixnet
10	30	+	congruence, epic, illustration, occf, insecure, kddcs, softrank, selector, tagme, twitterrank
10	30	−	copyright, proceeding, γ, ga, reasonably, sampler, reformulate, compactly, bis, umics
100	20	+	acquisitional, hilbert, vfdt, edutella, diffsets, closegraph, splay, sdm, webml, ordpath
100	20	−	copyright, ssl, baseline, denote, historical, recursive, centers, immediate, piece, dnf

4 Related Work

There are many works aiming at research investigation. The citation count is useful for evaluating scientific research. Martin [7] reported that the citation count has gained support as a criterion. Kostoff [5] showed that the citation count as a measure of evaluation has some problems.

It is more appropriate to find related papers if we restrict the papers in a specific research area. Nakatoh et al. [10] proposed the focused citation count (FCC), which restricts the research area of the cited papers by keywords, and showed that more appropriate papers could be extracted.

Even with such examples of the use of the citation count as a measure of an paper's importance, it is not almighty. For example, it is not appropriate to use it as a measure to evaluate a new paper. Therefore, it is common to use an evaluation of the scientific journal in which the paper was published or the researcher who wrote the paper.

A journal's IF [1,3,6] is one of the most popular evaluation measures of scientific journals. Thomson Reuters updates and provides the scores of journals in the Journal Citation Reports every year. Hirsch [4] defined the h-index of a researcher as the largest number h such that the researcher wrote h papers and each of the papers is cited from h papers or more. Scopus provides the h-index score of researchers.

For the detection of appropriate journals, there are also studies that have conducted analyses focused on a research area. Nakatoh et al. [9] proposed a method for selecting appropriate journals by using the citation, which focuses on a specific field to evaluate a journal.

In this study, we attempted to determine the quality of a paper from only the information in the paper. Regarding the judgment of the quality of a paper, there are several studies using checklists [16–18]. However, creating a checklist for each field is a laborious task. Otani et al. [14] considered the important expressions that represent important sentences in scientific papers. Ashok et al. [13] pointed out that it is possible to identify successful novels by their style. These studies support the position of this research.

5 Conclusion

The investigation of related research is very important for research activities. The number of citations is commonly used as an index, but information about recently published papers cannot be obtained. In this study, we attempted to identify good papers using only the words included in the summary. After constructing a classifier utilizing machine learning and evaluating using the cross validation, it became clear that some degree of discrimination is possible.

On the other hand, the discrimination performance is low, and it is difficult to use it to extract papers that are good as it is. It is possible to improve the method using the words used for extraction. In addition, we are planning to evaluate a paper in combination with its other attributes.

Acknowledgement. This work was partially supported by JSPS KAKENHI Grant Number 24500176.

References

1. Garfield, E.: Citation indexes for science: a new dimension in documentation through association of ideas. Science **122**(3159), 108–111 (1955)
2. Garfield, E., Sher, I.H., Torpie, R.J.: The Use of Citation Data in Writing the History of Science. Institute for Scientific Information, Philadelphia (1964)
3. Garfield, E.: The history and meaning of the journal impact factor. J. Am. Med. Assoc. **295**(1), 90–93 (2006)
4. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. USA **102**(46), 16569–16572 (2005)
5. Kostoff, R.N.: Performance measures for government-sponsored research: overview and background. Scientometrics **36**(3), 281–292 (1996)
6. Marshakova-Shaikevich, I.: The standard impact factor as an evaluation tool of science fields and scientific journals. Scientometrics **35**(2), 283–290 (1996)
7. Martin, B.R.: The use of multiple indicators in the assessment of basic research. Scientometrics **36**(3), 343–362 (1996)
8. Nakatoh, T., Hirokawa, S., Minami, T., Nanri, T., Funamori, M.: Assessing the significance of scholarly articles using their attributes. In: 22nd International Symposium on Artificial Life and Robotics (AROB 2017), pp. 742–746 (2017)
9. Nakatoh, T., Nakanishi, H., Hirokawa, S.: Journal impact factor revised with focused view. In: 7th KES International Conference on Intelligent Decision Technologies (KES-IDT 2015), pp. 471–481 (2015)
10. Nakatoh, T., Nakanishi, H., Baba, K., Hirokawa, S.: Focused citation count: a combined measure of relevancy and quality. In: IIAI 4th International Congress on Advanced Applied Informatics (IIAI AAI 2015), pp. 166–170 (2015)
11. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**(2), 404–409 (2001)
12. Wuchty, S., Jones, B.F., Uzzi, B.: The increasing dominance of teams in production of knowledge. Science **316**(5827), 1036–1039 (2007)
13. Ashok, V.G., Feng, S., Choi, Y.: Success with style: using writing style to predict the success of novels. In: 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1753–1764 (2013)
14. Otani, S., Tomiura, Y.: Extraction of key expressions indicating the important sentence from article abstracts. In: IIAI 3rd International Conference on Advanced Applied Informatics, pp. 216–219 (2014)
15. Zahedi, Z., Costas, R., Wouters, P.: How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics **101**(2), 1491–1513 (2014)
16. Schulte, J.: Publications on experimental physical methods to investigate ultra high dilutions – an assessment on quality. Homeopathy **104**(4), 311–315 (2015)
17. Zorin, N.A., Nemtsov, A.V., Kalinin, V.V.: Formalised assessment of publication quality in Russian psychiatry. Scientometrics **52**(2), 315–322 (2001)
18. Dasi, F., Navarro-García, M.M., Jiménez-Heredia, M., Magraner, J., Viña, J.R., Pallardó, F.V., Cervantes, A., Morcillo, E.: Evaluation of the quality of publications on randomized clinical trials using the Consolidated Standards of Reporting Trials (CONSORT) statement guidelines in a Spanish tertiary hospital. J. Clin. Pharmacol. **52**(7), 1106–1114 (2012)