REVIEW ARTICLE

Revisiting Information Technology tools serving authorship and editorship: a case-guided tutorial to statistical analysis and plagiarism detection

Bamidis PD, Lithari C, Konstantinidis ST

Lab of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

With the number of scientific papers published in journals, conference proceedings, and international literature ever increasing, authors and reviewers are not only facilitated with an abundance of information, but unfortunately continuously confronted with risks associated with the erroneous copy of another's material. In parallel, Information Communication Technology (ICT) tools provide to researchers novel and continuously more effective ways to analyze and present their work. Software tools regarding statistical analysis offer scientists the chance to validate their work and enhance the quality of published papers. Moreover, from the reviewers and the editor's perspective, it is now possible to ensure the (text-content) originality of a scientific article with automated software tools for plagiarism detection. In this paper, we provide a step-by-step demonstration of two categories of tools, namely, statistical analysis and plagiarism detection. The aim is not to come up with a specific tool recommendation, but rather to provide useful guidelines on the proper use and efficiency of either category of tools. In the context of this special issue, this paper offers a useful tutorial to specific problems concerned with scientific writing and review discourse. A specific neuroscience experimental case example is utilized to illustrate the young researcher's statistical analysis burden, while a test scenario is purpose-built using open access journal articles to exemplify the use and comparative outputs of seven plagiarism detection software pieces. Hippokratia 2010; 14 (Suppl 1): 38-48

Key words: tutorial, statistical analysis tools, plagiarism detection, guidelines of academic writing, neuroscience case example, emotion statistical analysis, emotion protocol

Corresponding author: Panagiotis D. Bamidis; Lab of Medical Informatics, Medical School, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, tel: +30-2310-999310, email: bamidis@med.auth.gr

I. Introduction

During the last few years, the number of scientific papers published in journals, conference proceedings, and scientific literature in general has significantly increased worldwide. In parallel, Information Communication Technology (ICT) tools provide to researchers novel and continuously more effective ways to present their work and support their arguments. For the purposes of this paper, such tools may be divided into two general categories, namely, tools for the authors, and tools for the reviewers and editors. For instance, software tools facilitating statistics offer an enormous number of possibilities[1], both at the statistical analysis level, as well as, at the presentation and visualization level. Decision support and data mining tools have toped up this arsenal with robust automations and novel visualization paradigms[2,3].

Moreover, bibliography database managers are used to manage information resources by maintaining reference databases and creating bibliographies and reference lists for written works[4]. A number of different citation/reference syntax systems have appeared in the market over the last fifteen years or so, facilitating scientists with versatility and easiness when it comes to one of the more time-consuming and frustrating tasks of research, that of the transformation of references according to specific journal format demands[5]. Reference linking, also reduces the amount of time spent by reviewers for searching and checking references and citations; in fact, most of the journals provide seamless click-through access to abstracts of referenced articles for papers under review. Other reviewing help tools empower the reviewing process by enabling the online (alongside the paper under review) investigation of a new topic and the search for a particular article, the cross-checking of up-to-date publications, and the creation of citation overviews for authors under review[6].

Most of the aforementioned tools provide authors the chance to enhance the quality of their work and increase the possibilities for publication. Moreover, from the editor's perspective, it is important to update paper reviewers on recent developments in peer review, and the available ways to support reviewers in their important task to safeguard the scientific quality of journals. To this end, it is now possible to digitally ensure the originality of a scientific manuscript, by comparing its previous encounters in scientific literature and elsewhere in the web, by means of automated software tools facilitating the detection of different types of plagiarism.

It is, therefore, true, that in the contemporary world of simultaneously growing numbers of article submissions and publication production demands, the aforementioned tools and assets become increasingly important. To this end, this paper places focus on two common aspects of the article writing and review processes, namely, that of statistical analysis supporting any included in the articles

statements or arguments, and second the revisiting of plagiarism, so as to safeguard publication quality.

Thus, the aim of this paper is twofold. First to provide a literature and market review for available tools so that different options are explored, and second to act as a guideline for young authors and fresh reviewers, by demonstrating step-by-step how the use of such tools enables the statistical validity of any claimed results, as well as, the (digital content) originality of the article; both aims are accomplished through specific cases and examples. The envisaged goal is not to come up with specific tool recommendation, however, but rather to provide some information on the efficiency of some tools (especially when these are either widely used or free-to-use. More specifically, with regards to plagiarism, emphasis is not placed at all on the technology and the algorithms that each plagiarism software uses, which are not in the scope of this paper at all, but rather on the comparison of the results from the editor's perspective, in order to form a review process guideline.

So the remainder of this paper is structured as follows. In section II, we review the market and literature of available tools for both statistical analysis and plagiarism. Then in section III, case examples are provided in an effort to demonstrate the different notion identified in the literature - this section is also micro-structured with reference to statistical analysis, and plagiarism. Finally, concerns and issues governing these two pillars of scientific research, as well as, a future outlook are provided in the last section of the paper.

II. Literature and Market review

a. Statistical Analysis Tools

The essence of and need for statistical analysis of any scientific data collections have been the subject of discussion in other papers of this special issue[7], so there is absolutely no further need to engage in any of that discourse in this piece of work. However, there is, nowadays, a plethora of available statistical software pieces which is designed for either analysis and visualization, or meta analysis or the creation of surveys. Some of these are presented in the tables below, but obviously the list is not exhaustive by any means.

Most of the statistic software tools provide basic and advanced analysis on data together with the possibility to create graphs and figures to visualize the results and are widely used by researchers on various fields ranging from life sciences[1] and sociology[8] to engineering[9] and economics[10]. Meta-analysis is defined as the statistic analysis of statistic analyses[11] and special tools and methods[12] are developed to the direction of combining many studies in one. Finally, over the last few years, special software tools, either open source or proprietary, offer the possibility to design questionnaires and surveys. These survey software tools usually offer also some basic statistical analysis and visualization options, along with the ability to export results directly to another statistical tool for analysis.

Table 1: Software tools for basic Statistical analysis.

Statistical Software for Basic Analysis & Visualization	
Short name of Software	**Type of license**
SPSS	proprietary
PSPP	open source http://www.gnu.org/software/pspp/pspp.html
StatView	proprietary
S-plus	proprietary
R	open source http://www.r-project.org/
Excel	proprietary
SalStat	open source http://salstat.sourceforge.net/

Table 2: Software tools for Statistical Meta-analysis.

Statistical Software for Meta-Analysis	
Short name of Software	**Type of license**
MIX (plug-in for Excel)	proprietary (free for developing countries)
MetaWin	Proprietary
Comprehensive Meta Analysis (CMA)	Proprietary
MADAM (toolbox for R)	open source http://cran.r-project.org

Table 3: Software tools for conducting surveys.

Statistical Software for Survey	
Short name of Software	**Type of license**
Lime Survey	open source http://www.limesurvey.org/
Opinio	proprietary
QuestionPro	proprietary
SurveyGizmo	proprietary

b. Plagiarism and its Detection Tools

Plagiarism, is defined as the "use or close imitation of the language and thoughts of another author and the representation of them as one's own original work"[13]. Within academia, plagiarism may take various forms, ranging from that of students to that conducted by professors or researchers themselves. Whatever the form, it has always been considered as an academic dishonesty and fraud and may lead to subsequent forms of expulsion[13]. In academic journal practice, one usually encounters listed codes of ethics followed by some academic journals and to which authors must obey and agree prior to any publication[14].

Plagiarism in scientific articles has been increasing dramatically in the last few years. This is probably an after effect of "content scraping", which is itself an enlargement of the phenomenon of coping and pasting material from Internet websites[15,16]. As a consequence, it is inevitable that journal editors are sometimes faced with concrete dilemmas and the resolution of real puzzles each

time they have to form an article acceptance decision or edit a new journal issue. Despite the authors' written confirmation claiming the refusal of editor's responsibilities regarding plagiarized articles, the fame and the reputation of the journal may seriously be affected, thereby, rendering the process of plagiarism detection a necessary safeguarding procedure.

There exist many different plagiarism categories[17-20]. Maurer et al[21] propose an abstract 4-dimension classification: (i) Accidental (due to lack of plagiarism knowledge, and understanding of citation or referencing style being practiced at an institute), (ii) Unintentional (the vastness of available information influences thoughts and the same ideas may come out via spoken or written expressions as one's own), (iii) Intentional (a deliberate act of copying complete or part of someone else's work without giving proper credit to original creator) and (iv) Self plagiarism (using self published work in some other form without referring to original one). All the above categories are significant. However, accidental plagiarism cannot be accepted for scientific research papers. Unintentional plagiarism can easily happen due to the fact that researchers share their opinions and brainstorming on new ideas. Unintentional plagiarism is very difficult to be proven and putting it simply, the good ethos of each researcher is the mere panacea and the actual limitation of plagiarism. Intentional and self plagiarisms are the two categories that are unacceptable by the academic community. Information and Communication Technologies (ICTs), apart from the facilitation of searching and retrieving huge amounts of research papers, provide also the appropriate tools to discover and evince any scientific misconduct. The latter tools do not always reveal whether the suspicious scientific paper contains plagiarized parts or not with certainty, though, usually provide a starting good indication on the right direction.

The majority of the plagiarism detection tools have been created for students' essays[21-24]. Research papers, case studies and review papers plagiarism have also come to the scene[25-28]. Plagiarism tools have also been created towards this direction, but there is surely more work to be done[29,30].

Tools for plagiarism detection can be divided regarding their comparison policy, into two main categories: (i) tools that search in a document database that the user provides and (ii) tools that conduct an internet-wide searching. Another more technical, but still important aspect for many users, is the classification into two main classes, namely those being web based tools, and those consisting of computer based tools requiring some kind of local installation. "Simple" users (e.g. an author) usually prefer a web based tool that could quickly accomplish their search, while "advanced" or "professional" users may prefer a computer based tool, capable of multi-searching batches of files each time.

Table 4: Free-to-use plagiarism detection tools.

Name	Proprietary /Free	Registration needed	Platform	URL
Plagiarism Detect	Free	Yes	web-based	http://plagiarismdetect.com/
Article Checker	Free	No	web-based	http://www.articlechecker.com/
Dupe Free Pro	Free	Yes	PC Installation	http://www.dupefreepro.com/
DOC Cop	Free	Yes	web-based	http://www.doccop.com/index.html
The Plagiarism Checker	Free	No	web-based	http://www.dustball.com/cs/plagiarism.checker/
Viper - the Anti-plagiarism Scanner 1.5	Free	Yes	PC Installation	http://www.scanmyessay.com
Dupli Checker	Free	No	web-based	http://www.duplichecker.com/
eT Blast (search pubmed, Medline, etc)	Free	No	web-based	http://etest.vbi.vt.edu/etblast3/
plagium	Free	No	web-based	http://www.plagium.com/
SeeSources	Free	No	web-based	http://plagscan.com/seesources/
Chimpsky	Free	No	web-based	http://chimpsky.uwaterloo.ca/
Copytracker	Free	No	web-based	http://copytracker.ec-lille.fr/
crossrefme	Free	No	web-based	http://www.crossrefme.net/
Splat	Free	No	PC Installation	http://splat.cs.arizona.edu/
Wcopyfind	Free	No	PC Installation	http://plagiarism.phys.virginia.edu/Wsoftware.html
Copy Tracker	Free	No	web-based	http://copytracker.ec-lille.fr
Pl@giarism	Free	Yes	PC Installation	http://people.few.eur.nl/span/Plagiarism/index.htm
Pairwise	Free	No	Advanced Server Installation	http://www.pairwise.cits.ucsb.edu/
10 dollar articles	Free	No	web-based	http://www.10dollararticles.com/adc.htm

Moreover, plagiarism detection tools, likewise any other software product, are classified in open source and proprietary. The "simple" user, for example an author who is going to check a few of his/her papers per year, can be fully satisfied by the usability and the effectiveness of open source tools. On the other hand, a more advanced user should use a combination of open source and proprietary tools. Table 4 illustrates some free-to-use plagiarism tools that are accessible through the web, while table 5 lists some licensed plagiarism systems (neither of the lists, though is supposed to be fully complete or exclusive of other (missed) components).

Table 5: Proprietary plagiarism detection tools.

Name	Version comments	Platform	URL
DupeCop	Proprietary	web-based	http://www.dupecop.net/index-online.html
Turnitin	Proprietary	web-based	http://turnitin.com/static/home.html
Glatt Plagiarism Self-Detection Program (GPSD)	Proprietary	web-based/ PC Installation	http://www.plagiarism.com
EVE2	Proprietary	PC Installation	http://www.canexus.com/
Ithenticate	Proprietary	web-based	https://www.ithenticate.com/
SafeAssign	Blackboard plugin	web-based	http://www.mydropbox.com/
Copycatch	Proprietary	PC Installation	http://www.cflsoftware.com/
Plagiarism Detector	Proprietary/ Demo Version	PC Installation	http://plagiarism-detector.com/
PlagiarismDetection.org	Proprietary	web-based	http://www.plagiarismdetection.org/
Plagiarism Finder	Proprietary/ Trial Version in German	PC Installation	http://www.m4-software.com/en-index.htm

III. Case examples

a. Conducting Statistical Analysis with SPSS

An article on neuroscience[31] will be used as a vehicle to demonstrate the proper use of a statistics software package (SPSS) in order to obtain statistically validity of the results analysis.

For purposes of good practice and guideline formation, the specific case is split into discrete steps, namely, research protocol deployment; definition of research parameters; selection of analysis type(s); data input; output and explanation of results; result illustrations and publication preparation.

(i) Research protocol deployment

First of all the research protocol has to be accurately described with detailed reference to participants' selection and the procedure followed. In the example deployed herein, 28 healthy subjects (14 females) participated in the study.

Visual emotion-evocative stimuli were presented to them and Electroencephalographic (EEG) measurements were recorded from 19 channels placed according to the 10-20 system. The International Affective Picture System (IAPS)[32] was used as a pool for the stimuli selection. Each picture from IAPS is rated across two characteristics; valence and arousal. Valence (1-9) indicates how pleasant (9) or unpleasant (1) is the emotion provoked by an image, whereas arousal (1-9) denotes the level of activation, ranging from low (1) to high (9), regardless of the valence[33].

As a result, IAPS stimuli, when plotted in 2-dimensional space (Figure 1) form 4 quadrants; pleasant and high arousing (PHA), pleasant and low arousing (PLA), unpleasant and high arousing (UHA) and unpleasant and low arousing (ULA).

40 pictures from each category were selected for our experiment and presented in a random order to the participants.

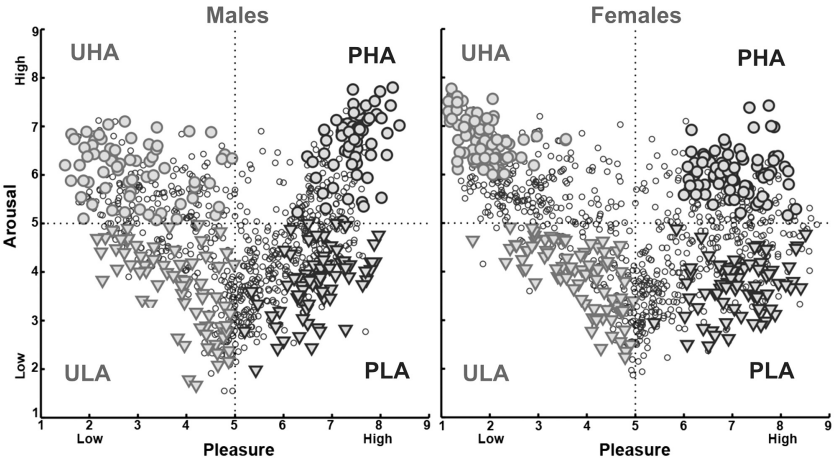

Figure 1: Valence and arousal ratings for pictures shown to males and females[31].

(ii) Definition of research parameters

The signal processing part is following, describing the EEG features extracted from the recordings. For each participant, 5 components, known from literature[34], were extracted (a signal positive peak around 100msec named P100, a negative peak following P100 named N100, P200, N200 and P300). Each component has two characteristics that have to be statistically analyzed; the amplitude and the latency. As a result, for each component and for each characteristic, there is a 28x4 matrix, whose columns represent the 4 stimulus categories, and lines represent the 28 participants. In general, when working with SPSS, the lines represent the participants and the columns represent the questions we want to answer. For example, we want to know if N100 component is affected by arousal or valence and if it is different between males and females.

(iii) Selection of analysis type(s)

Because of the bi-dimensional protocol design, both valence and arousal may affect the amplitude and/or the latency of each component. Valence and arousal are the within subjects factors and gender of the participants is the between subject factor. A 2-way ANOVA will be used to reveal any statistically significant differences raised by valence, arousal or gender. Significance of differences is determined by a p value lower than 0.05. A significant main effect of arousal, for example, means that arousal affected both PHA and UHA in the same way, making them to differ from PLA and ULA. A significant interaction between arousal and valence means that, for example, arousal affects PHA in a way, whereas UHA in another way, denoting a dependence between the two parameters. On the other hand, a gender main effect means that females differ from males across all 4 stimulus categories. A gender by valence interaction would indicate that females differ from males only on pleasant or only on unpleasant stimuli.

(iv) Data input

In the SPSS environment, the variables-columns should be defined in the 'Variable View' tab. The variables are a) PHA, b) PLA, c) UHA, d) ULA and e) gender. Gender is an additional column with zeros in the lines of males and ones in the lines for females or vice versa. From the SPSS menu, in the Analyze tab, we choose "General Linear Model" and then "Repeated Measures". "Valence" and "Arousal", each of them with 2 levels, should be defined as Within Subject Factors. In the window appearing then, the within-subjects variables should be matched with the corresponding combinations of levels produced by SPSS. PHA should correspond to (1,1), PLA to (1,2), UHA to (2,1) and ULA to (2,2) (Figure 2). Finally, "gender" should be defined as the Between Subject Variable. All these definitions can be done by dragging and dropping among tables. On the window appearing by pressing the 'Options' button, OVERALL Factors and Factor Interactions should be chosen, along with the Descriptive statistics in the checkboxes on the lower part of the window. Moreover, the significance level can be changed.

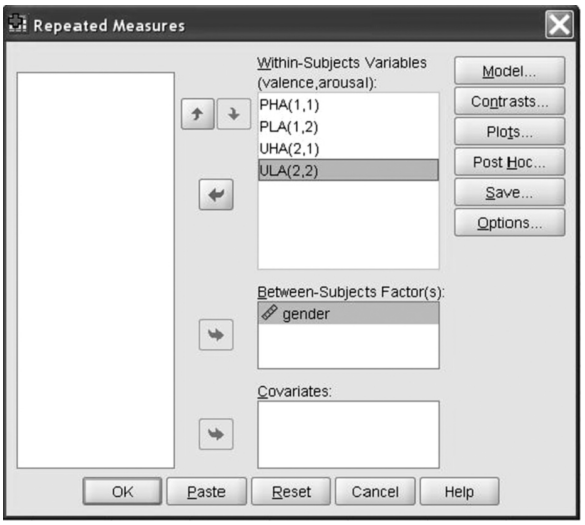

Figure 2: correspondence of combined levels produced by SPSS to variable names defined by the user.

(v) Output and explanation of results

The output file is a separate file in SPSS. All analyses appear in the output file, which is also described in tree structure on the left part of the screen. Under the caption 'Tests of Within-Subjects Effects' the significance of each main effect and interaction is presented. Suppose a significant (p=0.004) main effect of valence on the amplitude of N200 component. The question to be answered next is: In which way valence affects the N200 amplitude? The estimated averages for valence are -6.346μV for the pleasant (level 1) and -7.18μV for the unpleasant (level 2), which means that unpleasant stimuli elicited greater N200 EEG responses. Moreover, there was a significant (p=0.031) gender by valence interaction on N100, with females (1) showing greater responses than males (0) especially to unpleasant stimuli (Figure 3).

(vi) Result illustrations and publication preparation

In order to visualize an interaction, two (2) more figures are needed; one to show the absence of significant difference in males responses for pleasant and unpleasant stimuli, and a second one showing the existence of significant difference in females responses for pleasant and unpleasant stimuli. On the contrary, for the visualization of a main effect only one figure is needed. The final step is to save the output file with the ending ".spv" and the data file with the ending ".sav". The same analysis should be done for amplitude and latency for all components, for all EEG channels.

Further analysis of all electrodes can result in visualization of results on a scalp map/topography. *F values* are calculated for exploring the main effects of valence, arousal and gender, for each component amplitude, and for each channel; in this case, a visualization software, namely, Matlab 6.1, was used for projecting/mapping the F values onto a uniform (reference) head (Figure 4). The F-topographies show gender differences on central and left brain regions for N100 and N200 components.

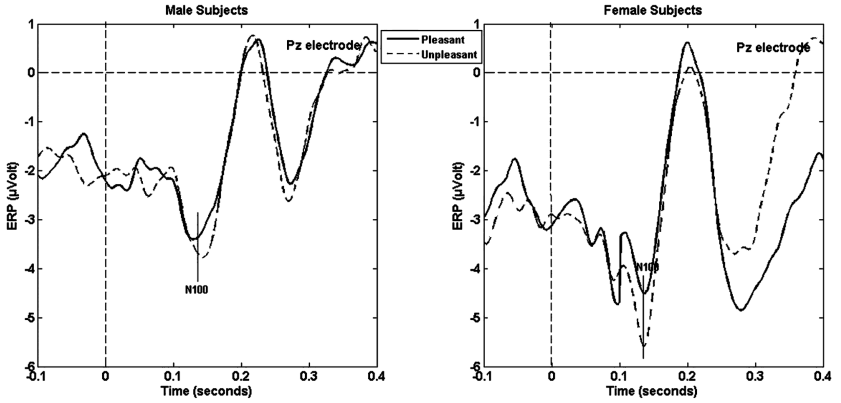

Figure 3: Gender by valence interaction is evident as only females exhibited difference between pleasant and unpleasant stimuli on N100 on Pz electrode[31].

b. A case demonstrating the use of Plagiarism Detection Tools

For the purposes of continuing on the notion and logic flow of the previous case/example, we refer to its last step, that is, publication preparation (from the author's point of view) and/or reviewing (from the reviewer's editor's point of view). Authors wishing to ascertain the integrity of their publication may follow the approach of using plagiarism detection tools. In addition, the same (or more enhanced) set of tools may become an invaluable asset in the hands of a journal (or book) editor, attempting to certify the uniqueness (or at least text-content originality) of the paper in question, thereby safeguarding the journal/books' quality. In order to test how effectively the different tools work, we have constructed an experiment consisting of the following steps:

(i) Test-case scenario, material and tool selection

Open access journals were used to randomly select ten (10) papers/articles, five (5) properly published ones and another five (5) unpublished articles, ready for publication. To this extend, for each article we created an image article by removing its references, due to the fact that the references already exist in the internet and may affect the result of any plagiarism detection tool.

Seven (7) plagiarism software tools were tested for the purpose of this paper; all of them appear in Tables 6 and 7.

(ii) Tool input, use and output

The use of such plagiarism detection tools is relatively easy and is not much altered between web-based applications and PC installations. Uploading the article in question (suspicious file) on the tool and clicking on a button are usually enough for the initialization of the tool. "Copy and Paste of the text" in a web form of the tool is an alternative method to accomplish input and initiate the run.

Each tool uses a different algorithm to detect plagiarism in the article, while some of them use the search Application Programming Interface (API) from known search engines (e.g. Google, Yahoo, etc.). Consequently, the end results and their representations might differ. In some cases, the result is an estimation percentage of the plagiarized text content in the article, while in other cases, the end result is a number representing the amount of web pages containing matched phrases. A combination of the above can also be encountered (e.g. cases of plagiarism software tools 2, 6 and 7 in Tables 6 and 7).

(iii) Explanation of results

An issue of concern is that differences between papers with references and without references are noteworthy. References exist already in the web in the form of titles or references in other papers. In both groups of papers used in this work (published and unpublished), the percentage or "internet hits" is smaller than in image papers (without references), regardless of the level of plagiarism. As a result, the detected plagiarized text was, in some cases, only within the references. For example, tool "Article Checker" detected 5% plagiarized text in paper 2 (Table 6), while if we consider paper 2 without references, the plagiarism percentage is 0% (Table 7).

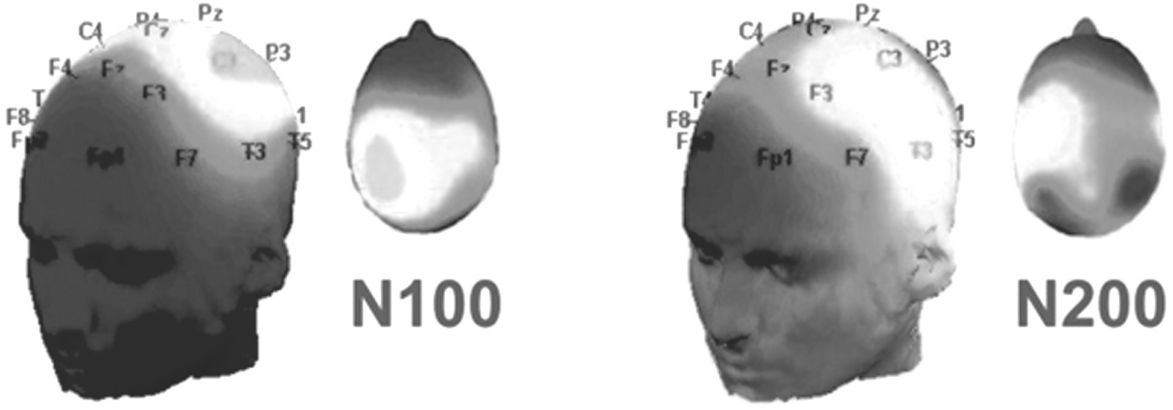

Figure 4: Topographies of component amplitude differences extracted from statistical analysis of all electrodes. F-values are calculated and mapped onto a uniform scalp.

Table 6: Five (5) published articles were selected from open access journal and were tested against seven (7) plagiarism detection tools.

Plagiarism Detection Tools			paper 1		paper 2		paper 3		paper 4		paper 5	
			whole	no ref	whole	no ref	whole	no ref	whole	no ref	whole	no ref
	Plagiarism Detect		3%	2.4%	24%	5.9%	1.6%	0%	65.2%	66.1%	25%	54.9%
	Article Checker	**Google search API**	2% (10/509)	2% (10/423)	5% (9/185)	0% (0/118)	1% (1/85)	0% (0/61)	5% (8/160)	6% (7/110)	5% (9/184)	5% (6/129)
		Yahoo search API	1% (4/509)	1% (4/423)	5% (9/185)	2% (2/118)	0% (0/85)	2% (1/61)	5% (8/160)	6% (7/110)	7% (12/184)	6% (8/129)
	The Plagiarism Checker		Possibly plagiarized - use links above to check (7 links)	Unknown - investigate with links above (4 links)	Unknown - investigate with links above (4 links)	Unknown - investigate with links above (3 links)	No plagiarism suspected	No plagiarism suspected	Possibly plagiarized - use links above to check (6 links)	Possibly plagiarized - use links above to check (6 links)	Possibly plagiarized - use links above to check (6 links)	Possibly plagiarized - use links above to check (5 links)
	Viper		18%	5%	19%	4%	2%	1%	24%	23%	18%	13%
	plagium		Plagium did not find documents making use of the text that you entered	Plagium did not find documents making use of the text that you entered	17% and 41% (matching with online version af the article)	38% and 30% (matching with online version af the article)	15% (with a different web page)	17% (matching with online version af the article)	83% (matching with online version af the article)	89% (matching with online version af the article) 15% (match with other paper of authors)	76% (matching with online version af the article)	75% (matching with online version af the article)
	SeeSources		*	*	16hits	no hits on the Internet	23 hits	12 hits	22hits	46hits	165 hits	6 hits
	crossrefme		2% (7-12% with 4 sources)	2% (7-12% with 4 sources)	2% (7-12% with 4 sources)	2% (7-12% with 4 sources)	3% (7-15% with 4 sources)	4% (8-17% with 4 sources)	26% (10-51% with 4 resources)	32% (12-56% with 4 resources)	5% (7-22% with 4 sources)	5% (7-22% with 4 sources)

* Allowed memory size of the system exhausted.

Each pair of same-colored columns represents a single already published paper. In each pair of columns the left one (titled as "whole") is the paper as it is published, while the right one (titled as "no ref") refers to the same paper without containing the references. Lines correspond to the results of a plagiarism detection tool. The percentage appeared in the cells depict the similarity with other resources existing in the web according to each plagiarism detection tool algorithm. The term "hit" refers to the number of web resources that were found similar to the paper under testing.

Table 7: Five (5) unpublished articles were selected and were tested against seven (7) plagiarism detection tools.

			paper 6		paper 7		paper 8		paper 9		paper 10	
			whole	no ref	whole	no ref	whole	no ref	whole	no ref	whole	no ref
Plagiarism Tools	**Plagiarism Detect**		5%	1%	3.2%	2%	11.3%	0.9%	4%	0.6%	0.6%	0%
	Article Checker	**Google search API**	0% (0/129)	0% (0/100)	0% (0/141)	0% (0/116)	2% (3/144)	0% (0/93)	0% (0/161)	0% (0/131)	0% (0/99)	0% (0/80)
		Yahoo search API	0% (0/129)	0% (0/100)	0% (0/141)	0% (0/116)	1% (2/144)	0% (0/93)	0% (0/161)	0% (0/131)	1% (1/99)	1% (1/80)
	The Plagiarism Checker		No plagiarism suspected	No plagiarism suspected	Unknown - investigate with links above (3 links)	No plagiarism suspected	Possibly plagiarized - use links above to check (6 links)	Unknown - investigate with links above (1 links)	Unknown - investigate with links above (3 links)	Unknown - investigate with links above (1 link)	No plagiarism suspected	No plagiarism suspected
	Viper		10%	7%	9%	1%	11%	0%	4%	2%	4%	4%
	plagium		Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered	Plagium did not find documents making use of the text that you entered	Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered.	Plagium did not find documents making use of the text that you entered.
	SeeSources		7 hits	no hits on the Internet	11 hits	no hits on the Internet	13 hits	2 hits	4 hits	no hits on the Internet	2 hits	2 hits
	crossrefme		8% (18-23% with four sources)	8% (18-23% with four sources)	*	*	*	*	25% (6-46% with 4 sources)	26% (5-47% with 4 sources)	49% (44-67% with 4 sources)	49% (44-67% with 4 sources)

* Allowed memory size of the system exhausted.

Each pair of same-colored columns represents a single unpublished paper. In each pair of columns the left one (titled as "whole") is the paper as it is published, while the right one (titled as "no ref") refers to the same paper without containing the references. Lines correspond to the results of a plagiarism detection tool. The percentage appeared in the cells depict the similarity with other resources existing in the web according to each plagiarism detection tool algorithm. The term "hit" refers to the number of web resources that were found similar to the paper under testing.

Furthermore, most of the papers show a percentage or "internet hits" of plagiarism, but the differences between published and unpublished papers are significant. As Tables 6 & 7 depict, the plagiarism percentage of the published papers (in plagiarism tools that provide percentage of plagiarized text) is considerably higher than the percentage of the unpublished ones.

(iv) Issues of concern

Despite the high percentage of plagiarism in some published papers (no 2, 4 and 5), a closer look indicates that these papers are matching with their online versions in percentages of 38%, 89% and 75% respectively. Plagiarism Percentage revealed by a plagiarism detection tool for an unpublished article is just a warning for a more detailed examination of the article and not a proof. For example, paper 6 with 10% (7% without references) of plagiarized text (using Viper as a plagiarism tool) has similarities in phrases like:

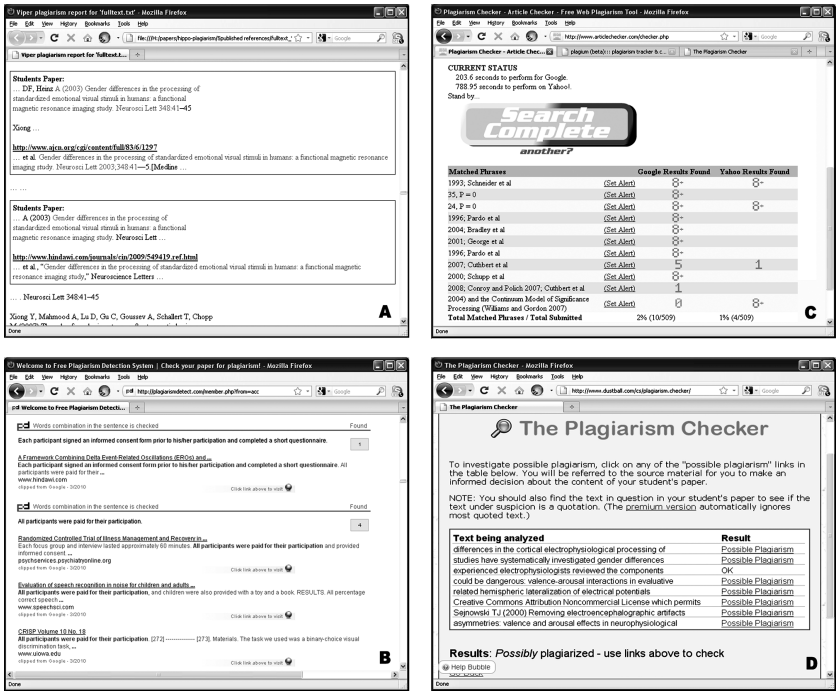

Figure 5: Screenshots of plagiarism detection tools. A and B: copy of the exact phrase with link to the resource of plagiarized text. C and D: link to the source of plagiarized text.

• "been approved by the Research Ethics Committee of …"

• "After complete and detailed description of the study to the"

• "or other serious physical illness, drug or alcohol abuse"

• "stimuli selected from the International Affective Picture System (IAPS)"

• "high resolution structural magnetic resonance imaging (MRI) scans"

• "By registration of the head position at these"

• "each sphere (one per MEG sensor) is fitted to a small patch of the head model (directly under the sensor)"

Acronyms, affiliations, acknowledgments and in general phrases that are commonly used in scientific papers, unveil a false plagiarism alarm. In addition, some of the free plagiarism detection tools provide a copy of the exact phrase that is plagiarized (Figure 5 A and B) being more illustrative, while others provide only the source (Figure 5 C and D).

It is especially worth men-

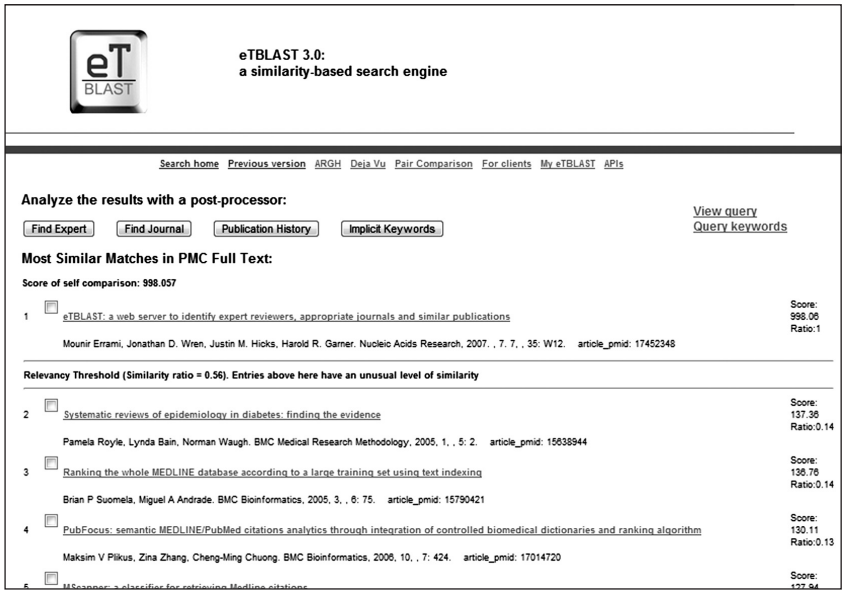

Figure 6: eTBlast: results for a paper existing in PUBMED central. eTBlast separates the results in two groups by a horizontal line; above the line, there are articles with a similarity percentage above the threshold (Similarity ratio = 0.56), while below the line articles with a lower similarity than eTBlast threshold appear. Information regarding the ratio of similarity is provided on the right side of each article. As it is evident, only one article with similarity above the threshold was found, which also happens to be the same article under testing (100% similarity).

tioning eTBlast[29], an online plagiarism tool that searches in MEDLINE, CRISP, NASA, Medical Cases, Pubmed Central (sections), PMC Full Text, PMC METHODS, PMC INTRODUCTION, PMC RESULTS, PMC (paragraphs) and PMC Medical Cases. eTBlast weighs keywords contained in the submitted article to identify a subset of literature in medicine. It then computes a quantitative score based on sentence alignment to justify similarity and relevancy between the submitted article and the selected database (Figure 6). eTblast provides also lists of relevant experts in the field of the submitted article and journals that publish topics relative to the submitted article. eTBlast in conjunction with the Déjà vu database provide identification of highly similar citations[30].

IV. Discussion and outlook

The goal of this paper was to briefly review the current status of ICT tools for writing scientific papers with respect to statistic analysis and plagiarism detection. By use of a specific case/example stemming from the neuroscience field, the paper demonstrated in a step-by-step tutorial fashion, the use of the SPSS package in furnishing the analysis of experimental results with statistical validity, thereby leading to paper preparation for publication. Following that, we revisited the use of plagiarism detection tools, by demonstrating, likewise, step-by-step, how such different tools may be utilized in order to avoid plagiarism. In so doing, the output/efficiency of seven (7) tools was compared in a set scenario exploiting ten (10) articles from open access journals and outlining issues of concern.

It is imperative that both of the above categories of tools are useful for either prospective authors, but also for reviewers and editors seeking a publication decision for an article in question. The combination of these two categories in the current piece of work, does not only consist a unique, and to the best of our knowledge, original approach, but envisages to become a guideline tutorial accommodating the needs of (starting and not only) authors and fresh reviewers or editors. The use of the statistical case, aimed to exemplify the use of a common and widely used statistical package (SPSS), but unlike the case of a user manual, the approach taken herein, used a contemporary neuroscience protocol to reveal some of the secrets behind the analysis, which are nevertheless mandated before any publication attempt.

Moreover, and as mentioned in the introduction section, the abundance of scientific information on the web is both a blessing and a curse. From one hand, researchers are equipped with a plethora of resource finding mechanisms that accomplish the task of literature search "at a click of a button". On the other hand, many of the current researchers and authors may easily indulge into the temptation of "copy-and-paste" as a first attempt, which unless carefully examined and revisited overall at the end, underlies many inherent risks, which may in turn deploy the plagiarism accusation and dishonesty stamps on one's image. The latter situation has been further exploded recently due to the ever increased availability of dynamic information and the catalytic easiness of publishing such information and commentaries by means of artifacts enabled by the social web (Web2.0) e.g. blogs, wikis, etc[35-37]. It is certain, that such information content will be continuously enriched and exploded due to blogs, micro-blogging, social networking, web mashups and aggregators, which inevitably provide insights into people's scientific endeavors. Much work is still needed of course in order to facilitate current plagiarism detection tools with a capacity of efficiently and effectively exploring the above and outputting a valid result, perhaps through the incorporation of predictive analytics and related methods for Web "data mining" where users' posts and queries are garnered from Social Web[38].

To conclude, however, we believe and hope, that the current paper, in the context of this special issue, offers useful guidelines to specific problems concerned with scientific writing and review discourse, and will become a functional and handy tutorial in the hands of many researchers in the future.

References

1. Romualdi C, Lafranchi G. Statistical Tools for Gene Expression Analysis and Systems Biology and Related Web Resources. In: S Krawetz (ed). Bioinformatics for Systems Biology, Humana Press. 2009; 181-205.
2. Bamidis PD, Psarouli E, Stilou S. Using Modern IT Tools to assess the awareness of MDs on radiation issues and plan a continuous education programme. Health Informatics Journal. 2001; 7: 146-151.
3. Stilou S, Bamidis PD, Maglaveras N, Pappas C. Mining Association Rules from Clinical Databases: an Intelligent Diagnostic Process in Healthcare. Proceedings of MEDINFO 2001. 2001; 782-786.
4. Reiss M, Reiss G. Selected aspects of computer-assisted literature management, Wien Med Wochenschr. 1998; 148: 183-186.
5. Smith CM, Baker B. Technology in nursing scholarship: use of citation reference managers, Int J Ment Health Nurs. 2007; 16: 156-160.
6. Elsevier. Supporting Reviewers. http://www.elsevier.com/wps/find/reviewershome.reviewers/supportrev, last access, March 30, 2010.
7. Editorial, Hippokratia. 2010; 14 (Supp 1).
8. Lodha S, Gunawardane P, Moddleton E, Crow B. Understanding relationships between global health indicators via visualization and statistical analysis. Journal of International Development. 2009; 27: 1152-1166.
9. Li S, Duan ZZ, Peng ZS, Chen JY. Survey and Analysis on the Present Situation of the Civil Building Area and Energy Consumption in Anhui Province. Construction Conserves Energy. 2009; 37: 69-71.
10. Hellwig MF. Systemic Risk in the Financial Sector: An analysis of the Subprime-Mortgage Financial Crisis. De Economist. 2009; 157: 129-207.
11. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press INC. 1985.
12. Kugler KG, Mueller LAJ, Graber A. MADAM-An open source meta-analysis toolbox for R and Bioconductor. Source Code for Biology and Medicine, 2010: 5.
13. Wikipedia. www.wikipedia.org, last access, March 30, 2010.
14. ACM, Association of Computing Machinery. ACM Policy and Procedures on Plagiarism. http://www.acm.org/publications/policies/plagiarism_policy, last updated, October 2006.
15. Jones D. Authorship gets lost on Web. http://www.usatoday.com/

tech/news/2006-07-31-net-plagiarism_x.htm?POE=TECISVA, retrieved March 30, 2010.
16. Kock N, Davison, R. Dealing with plagiarism in the IS research community: A look at factors that drive plagiarism and ways to address them. MIS Quarterly. 2003; 27: 511-532.
17. Clough P. Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Internal Report CS-00-05, University of Sheffield. 2000. Retrieved on 21/2/2010 from http://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf
18. Eysenbach G. Report of a case of cyberplagiarism - and reflections on detecting and preventing academic misconduct using the Internet . J Med Internet Res. 2000; 2: e4.
19. Bouville M. Plagiarism: Words and Ideas. Sci Eng Ethics. 2008; 14: 311–322.
20. Ryan G, Bonanno H, Krass I, Scouller K, Smith L. Undergraduate and postgraduate pharmacy students' perceptions of plagiarism and academic honesty. Am J Pharm Educ. 2009; 73: 105.
21. Maurer H, Kappe F, Zaka B. Plagiarism - A Survey. Journal of Universal Computer Science. 2006; 12: 1050-1084.
22. Barrett R, Malcolm J, Lyon C. Are we ready for large scale use of plagiarism detection tools? 4th Annual LTSN-ICS Conference, NUI Galway. 2003; 79-84.
23. White DR, Joy MS. Sentence-based natural language plagiarism detection. Journal on Educational Resources in Computing (JERIC). 2004; 4: Article No.: 2.
24. Bilic-Zulle L, Azman J, Frkovic V, Petrovecki M. Is there an effective approach to deterring students from plagiarizing? Sci Eng Ethics. 2008; 14: 139-147.
25. Weed DL. Preventing scientific misconduct. Am J Public Health 1998; 88: 125–129.
26. Reeves DS, Wise R, Drummond CWE. Duplicate publication: a cautionary tale. Journal of Antimicrobial Chemotherapy. 2004; 53: 411-412.
27. Lippi G, Favaloro EJ. Detection of duplicates and redundancies. A major responsibility of peer-reviewers? Clin Chem Lab Med. 2008; 46: 1796-1797.
28. Scanes CG. Duplicate publication-An unacceptable practice. Poult Sci. 2009; 88: 45.
29. Errami M, Wren JD, Hicks JM, Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Res. 2007; 35: W12-W35.
30. Errami M, Sun Z, Long TC, George AC, Garner HR. Déjà vu: a database of highly similar citations in the scientific literature. Nucleic Acids Res. 2009; 37: D921–D924.
31. Lithari C, Frantzidis CA, Papadelis C, Vivas AB, Klados MA, Kourtidou-Papadeli C, et al. Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. Brain Topogr. 2010; 23: 27-40.
32. Lang PJ, Bradley MM, Cuthbert BN. Motivated attention: affect, activation, and action. In: PJ Lang, RF Simons,M Balaban (eds). Attention and orienting: sensory and motivational processes. Hillsdale, NJ: Erlbaum Associates. 1997.
33. Barrett LF, Russell JA. The structure of current affect: controversies and emerging consensus. Am Psychol Soc Bull. 1999; 8: 10–14.
34. Cuthbert BN, Schupp HT, Bradley MM, Birbaumer N, Lang PJ. Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. Biol Psychol. 2000; 52: 95–111.
35. Eysenbach G. From intermediation to disintermediation and apomediation: new models for consumers to access and assess the credibility of health information in the age of Web2.0. Stud in Health Technol Inform. 2007; 129 (Pt 1): 162-166.
36. Kaldoudi E, Konstantinidis S, Bamidis P. Web 2.0 Approaches for Active, Collaborative Learning in Medicine and Health. Peer-Reviewed Chapter in: S. Mohammed and J. Fiaidhi (eds.). Ubiquitous Health and Medical Informatics: Advancements in Web 2.0, Health 2.0 and Medicine 2.0. IGI Global, Hershey, PA, USA. 2010, ISBN: 978-1-61520-777-0.
37. Hatzipanagos S, Warburton S. Handbook of Research on Social Software and Developing Community Ontologies. Information science reference Series. IGI Global, Hershey, New York. 2009.
38. Kamel Boulos MN, Sanfilippo AP, Corley CD, Wheeler S. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. Comput Methods Programs Biomed. 2010; 100: 16-23.