STATISTICS, BIOMEDICINE AND SCIENTIFIC FRAUD

Claudio De Felice,[1] Alessio Cortelazzo,[2,3] Silvia Leoncini,[3,4] Cinzia Signorini,[4] Joussef Hayek,[3] Lucia Ciccoli[4]

[1]Neonatal Intensive Care Unit, University Hospital, Azienda Ospedaliera Universitaria Senese (AOUS), Policlinico "S. M. alle Scotte", Siena, Italy; [2]Department of Medical Biotechnologies, University of Siena, Italy; [3]Child Neuropsychiatry Unit, University Hospital, (AOUS), Policlinico "S.M. alle Scotte", Siena, Italy; [4]Department of Molecular and Developmental Medicine, University of Siena, Italy

Abstract. A consistent fraction of published data on scientific journals is not reproducible mainly due to insufficient knowledge of statistical methods. Here, we discuss on the use of proper statistical tools in biomedical research and statistical pitfalls potentially undermining the scientific validity of published data. Apart from unaware errors, a growing concern exists regarding data fabrication and scientific misconduct. Indeed, the social impact of false scientific data can be largely unpredictable and devastating, as shown by the worldwide dramatic effects on vaccinations coverage following a retracted paper published on a highly authoritative medical journal. Unfortunately, statistics shows a quite limited power in detecting false science, although a few statistical tools, such as the Benford's law, are known. Taken together, statistics in biomedical sciences i) is a powerful tool to interpret experimental data; ii) has limited power in detecting false science; and iii) first and foremost, is not the result of a simple "click of a mouse", but should be the result of accurate research planning by experienced and knowledgeable users.

Key words: Biomedical sciences; scientific fraud; scientific misconduct; statistics; statistical errors; statistical inference.

BACKGROUND

Statistics can be defined as a "battle against variability" (Prof. Claudio Scala). A consistent fraction of published data on scientific journals is not reproducible [1]. This could be mainly due to insufficient knowledge of statistical methods. On the other hand, there is a concept, largely attributed to the physicist Ernest Rutherford, supposedly saying, "if your experiment needs statistics, you ought to have done a better experiment." There is a lot of truth to this statement when working in a field with high signal-to-noise ratios. Nevertheless, statistical analyses are needed in all fields with a lower signal-to-noise ratio to properly quantify confidence in the study conclusions [2]. Indeed, in the absence of variability there would be little need for data analysis. Variability is avoidable in experiments due to both biological and technical effects [3]. Although biological variability needs to be maintained in order to allow generalization of the results to the population of interest, the tools that allow replicable results to be obtained despite the biological variability are experimental control, randomization, blocking and replication [3]. It is important to distinguish between sources of variation that are merely nuisance factors from those required in order to assess the variability of the effects in the population. The goal for every researcher should be to minimize the confounding factors of the experiment, as well as to sample and quantify the real biological variability in order to generalize conclusions and robustly determine uncertainty in estimates.

Of course, the concept that "there is no statistics without data" cannot be overstated. However, scientific intuition may start from a single case, a context where statistics is, by definition, not applicable. Nevertheless, a single anecdotal evidence can be a good starting point for interesting scientific discovery. For instance, the observation of a sudden drop in the pulse oximeter perfusion index during a transfer flight of a former premature male infant in a condition of severe clear air turbulence led one of the present authors (C.D.F.) to consider pulse oximeter perfusion index as an early marker of subclinical hypoxia. This finding was subsequently replicated on a large cohort study confirming early data (Figure 1) [4]. In a different context, fetal heart rate hypovariability in a singleton term pregnancy led the same author to interpret the finding as a potential marker of a prenatal inflammatory process [5].

On the opposite, the emerging "omics" sciences (*i.e.*,

Correspondence to:
Claudio De Felice
Neonatal Intensive Care Unit, University Hospital, Azienda Ospedaliera Universitaria Senese (AOUS), Policlinico "Le Scotte"
Viale Bracci 16, 53100 Siena, Italy.
Tel.: +39.577.586585; Fax: +39.577.586150
E-mail: geniente@gmail.com; c.defelice@ao-siena.toscana.it

genomics, transcriptomics, proteomics, metabolomics) are facing researchers with important challenges to one of the main principles of statistics *i.e.*, the assumption of data independence (Figure 1).

In the present lecture, we discuss on the emerging need for using adequate statistical tools in biomedical research and the chance to unveil unaware statistical errors possibly in undermining the scientific validity of published data, as well as a mention to the, quite limited, statistical tools to unveil possible scientific frauds.

UNAWARE STATISTICAL ERRORS

"Experience has taught statisticians that data can be misleading and, even worse, wrongly give the semblance of objectivity" [6]. This sentence from Prof. David Rossell (Department of Statistics. University of Warwick, United Kingdom) efficaciously expresses all the concerns of statisticians towards current statistical standards. An important wake-up call in the literature is the alarming rate of non reproducible scientific published findings. In a study of 2011, only 20-25% of preclinical studies were found to be reproducible [7], as well as only 11.3% of basic cancer biology papers [8]. The problem of the poor reproducibility of scientific studies has attracted the attention of the National Institutes of Health [9], as well as non-experts [10]. One of the reasons behind this lack of reproducibility certainly lies in a poor understanding of statistical tools and concepts.

In our own experience, one of the most common mistakes is involving statisticians at the end of research instead that at its beginning. A certainly non-exhaustive list of common statistical errors from our personal experience is reported in Table 1.

Indeed, accurate experimental design, randomization, bias control should intervene much before the experimental procedures are carried out. It has even been suggested that, in genetic association studies, there is a positive relationship between individual study bias and journal impact factor [11]. Therefore, journal prestige and influence are not mandatorily indicators of high quality in research [12].

Unexperienced investigators are likely to make common mistakes, such as "*P*-hacking" [13], overuse of statistical hypothesis testing, and overreliance on the standard error of the mean (S.E.M.) (Table 1) [2]. In particular, there is abundance of confusion and criticism about the meaning of *P* value [14], and understanding of the word "significant", which is often misunderstood. Significant has two distinctive meanings in science: one is that a *P* value is less than a preset threshold (usually 0.05); the other is that an effect is large enough to have

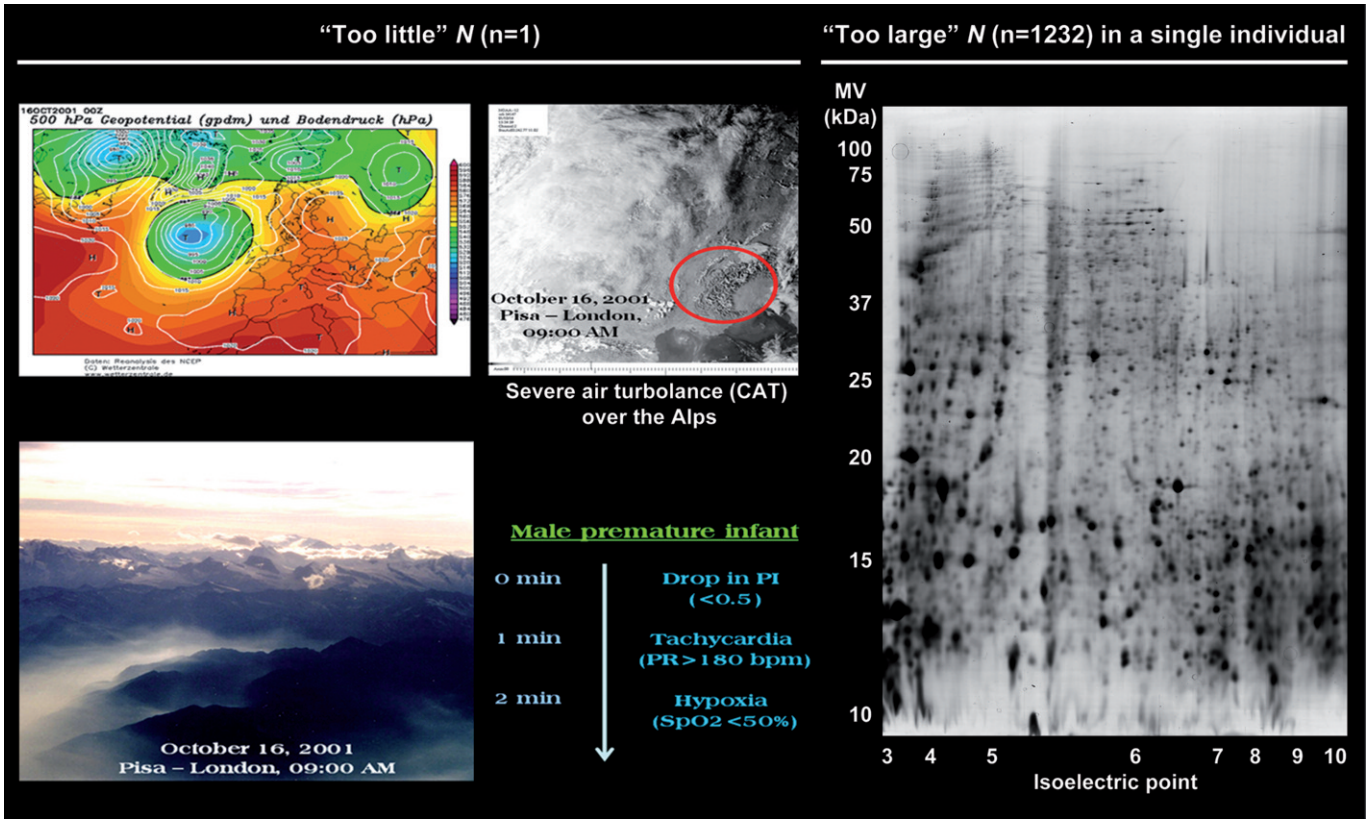


Figure 1. Common challenges in biomedical statistics.

a substantial biological physiologic or clinical impact. These meanings, completely different, are much too often confused. Indeed, the *P* value was never meant to be used the way it is used today [15]. The *P* value could be defined as the probability of seeing an effect as large as or larger that observed in the current experiment if the null hypothesis is true. Nevertheless, the *P* value gives no information about how large the difference (or effect) is. Historically, when the United Kingdom statistician Ronald Fisher introduced the *P* value in the 1920s, he did not mean it to be a definitive test. Fisher intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense of "worthy of a second look" [15]. Even at that time, famous mathematicians and statisticians Jerzy Neyman and Egon Pearson heavily criticized the *P* value as "worse than useless" [15]. In the meantime, other researchers, mostly non-statisticians, have created a hybrid system that squeezed Fisher's *P* value into Neyman and Pearson's reassuringly rigorous rule-based system, thus giving birth to the famous (or "infamous", according to several statisticians) preset value of 0.05 considered as "statistically significant". Actually, the only thing a *P* value could do is to summarize the data assuming a specific null hypothesis. A *P* value of 0.01 actually corresponds to a false-positive rate (type I error) of at least 11%, depending on the underlying probability that there is a true effect, while a *P* value of 0.05 increases this chance to at least 29% [16]. The underlying concept is that significance is no indicator of practical or biological relevance. The term *P*-hacking has been popularized by psychologist Uri Simonsohn and colleagues, intending data-dredging, snooping, fishing, significance-chasing and double-dipping [15]. In different words, *P*-hacking is a

Table 1. Most common errors encountered in statistics applied to life sciences.

Critical issues (personal experience)
Poor research design
Involving statisticians at the beginning of research, not at its end
Lack of a priori sample size calculation/effect-size estimation (statistical power)
Use of wrong statistical tests
Tested null-hypothesis (H_0) not rigorously stated
Study aims and primary outcome measures not clearly stated or unclear
Numerical information given to an unrealistic level of precision
Use of mean ± S.D. to describe non-normal data
Giving S.E.M. instead of S.D.
Lack of reporting on confidence intervals
Failure to prove test assumptions (*i.e.*, normal distribution)
Poor understanding of *P* values ("*P*-hacking" effect)
Significance unsupported by data analysis
"Non-significant" ≠ "no effect" ("non-significant" ≠ "negative")
Statistical significance ≠ biological or clinical relevance
Disregard for Type II error for non-significant results and multiple testing problem
Missing data issue
Inappropriate control group
Failure to use and report randomization
Too many variables involved
Association ≠ cause-effect relationship
Failure to discuss sources of potential bias/confounding factors
Understanding that statistics is not the simple result of a "click of a mouse"

way "to torture the data until it appears to support some pre-conceived idea" [15].

While standard deviation (S.D.) quantifies variation among a set of values, (*i.e.*, S.E.M., computed by dividing the S.D. by the square root of the sample size) does not. Indeed, the range mean ± S.E.M. is a confidence interval, depending on the sample size. This means that range is a 68% confidence interval of the mean when applied to large samples, but can become a 58% confidence interval with N=3 [2].

Doubtless, one of the most common statistic pitfalls in medical research is a failure to prove test assumptions (Figure 2). "Association is not causation" is a quite critical concept that must be kept in mind when drawing conclusions from statistical inference. Besides that (and prior to that), it is critical to test assumptions for any correlations. For instance, homoscedasticity is one of the critical assumptions in linear regression analysis. Figure 2 exemplifies a possible pitfall in the relationship between neonatal birth weight and gestational age at birth. As it can be seen from the scattergram plot (Figure 2), the principle of homoscedasticity is clearly not matched. Further analyses show a non-normal residual errors distribution (Figure 2). The other fundamental requirements for linear regression are the following: continuous variables; a linear relationship between variables; lack of outliers far removed from the mass of data; and independence of the observations. Therefore, the common linear regression analysis cannot be applied in this case. Incidentally, alike the incidence of coronary heart disease and cerebrovascular disease [17], birth weight does not follow a gaussian (*i.e.*, normal) distribution, but is more likely to follow a Weibull hazard model, as birth weight could be considered the final outcome of the "battle for survival" of the fetus.

For any study, the relevance of the research design cannot be overstated. In particular, it should be kept in mind that no statistics - and no statisticians - could ever remedy a poor study design. The example in Figure 3 tries to illustrate this concept by a statistical "divertissement" based on a recent news regarding the lack of natural hibernation in hedgehogs supposedly related to climate change [18].

If we simply test the difference between average temperature in earlier periods (*i.e.*, from 1943 to 1957 in Siena, Italy) *vs.* current data (years 2007-2015), a difference could hardly be detected (Figure 3). However, a more accurate research on the biological signal for the hedgehog natural hibernation appears to be the winter to spring (October to March) temperature. When this information is applied to the same database, the difference between historical and current temperature becomes highly evident (Figure 3).

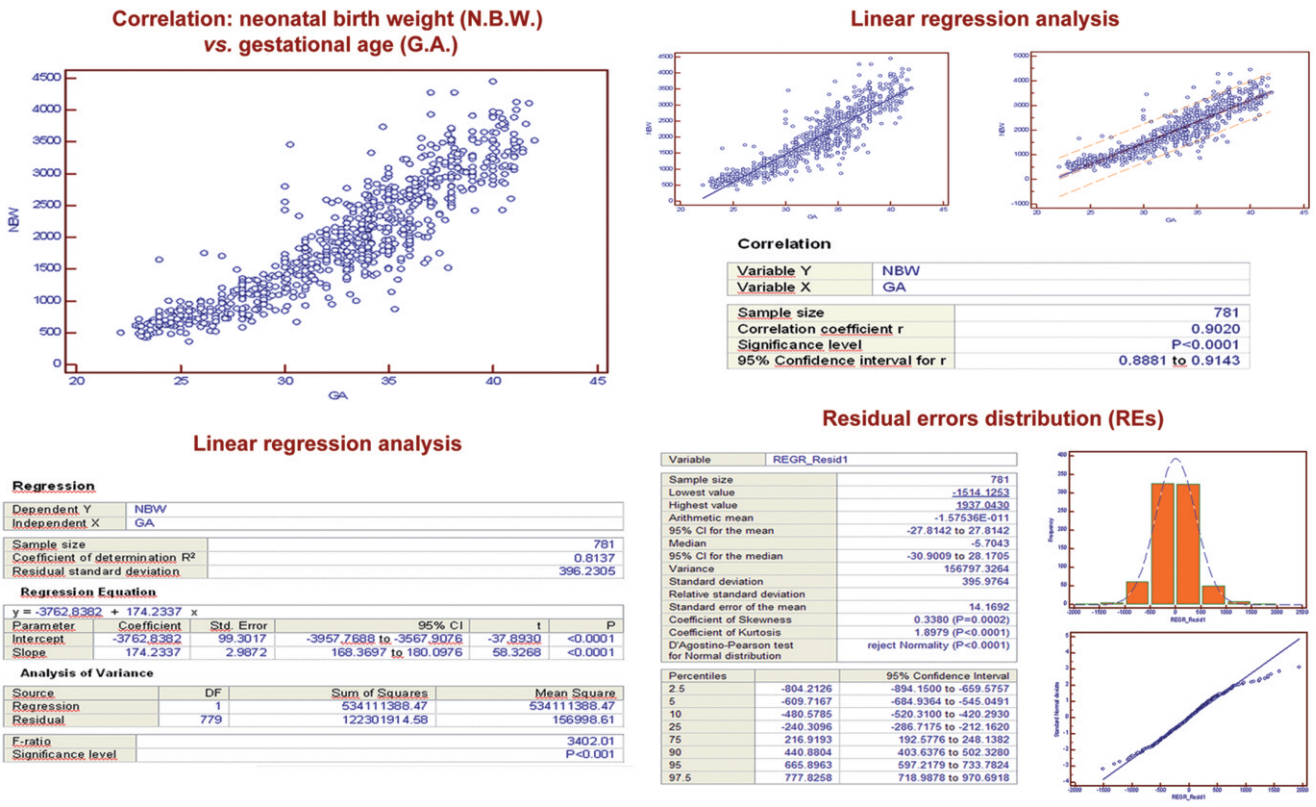


Figure 2. Pitfalls in linear regression analysis.

Another important, often misunderstood, concept is how large N (*i.e.,* sample size) should be in order to allow a meaningful statistical analysis. Of course, this is an ill-posed question. Actually, sample size critically depends on data variance. Nevertheless, Figure 4 illustrates the variation of statistical precision as a function of sample size. The data set refers to N=1457 temperature data from the Siena meteorological archive. A random selection from the whole data set was made in order to have a progressively increasing N (eight groups with sample size ranging from N=5 to N=1457) (Figure 4). Subsequently, deviations from the whole data set were calculated in terms of variance, variance ratio, error mean percentage and error median percentage. Surprisingly, deviation from the real median or mean values largely and unpredictably fluctuates with N ranging from 5 to 100 (mean error: from 2.11% to 35.84%; median error: from 8.2% to 68.6%). Only about a third of the total sample size is evaluated, deviations became negligible (mean error: 0.28%; median error: 2.98%; variance ratio: 0.99). A current challenge to present time statistics is represented by the so called "big data". Indeed, the world's capacity to collect, store and share data has hugely raised in recent times if it is true that 90% of the data in the world has been generated in just a couple of years [19]. Matching with such extremely large data sets will require active methodological research, as well as training a new generation of scientists to develop and deploy the resulting strategies [6].

Moreover, the fractal nature of life [20], represents a further challenge to statistics [21]. Although, it is difficult to understand the real underlying reasons of fractality of nature, the current explanations include the following: i) likely an evolutionary imperative; ii) critical for optimal substrate distribution and metabolic efficiency; iii) robustness and resistance to random errors [22-24].

SCIENTIFIC FRAUD AND SCIENTIFIC MISCONDUCT

The misuse of statistics in medical research can be considered both unethical and having serious clinical consequences [25, 26]. As a result, valuable efforts have been made to enhance the quality of statistics in medical journals [27, 28]. Despite several efforts, little evidence exists that statistical standards have improved over time [29].

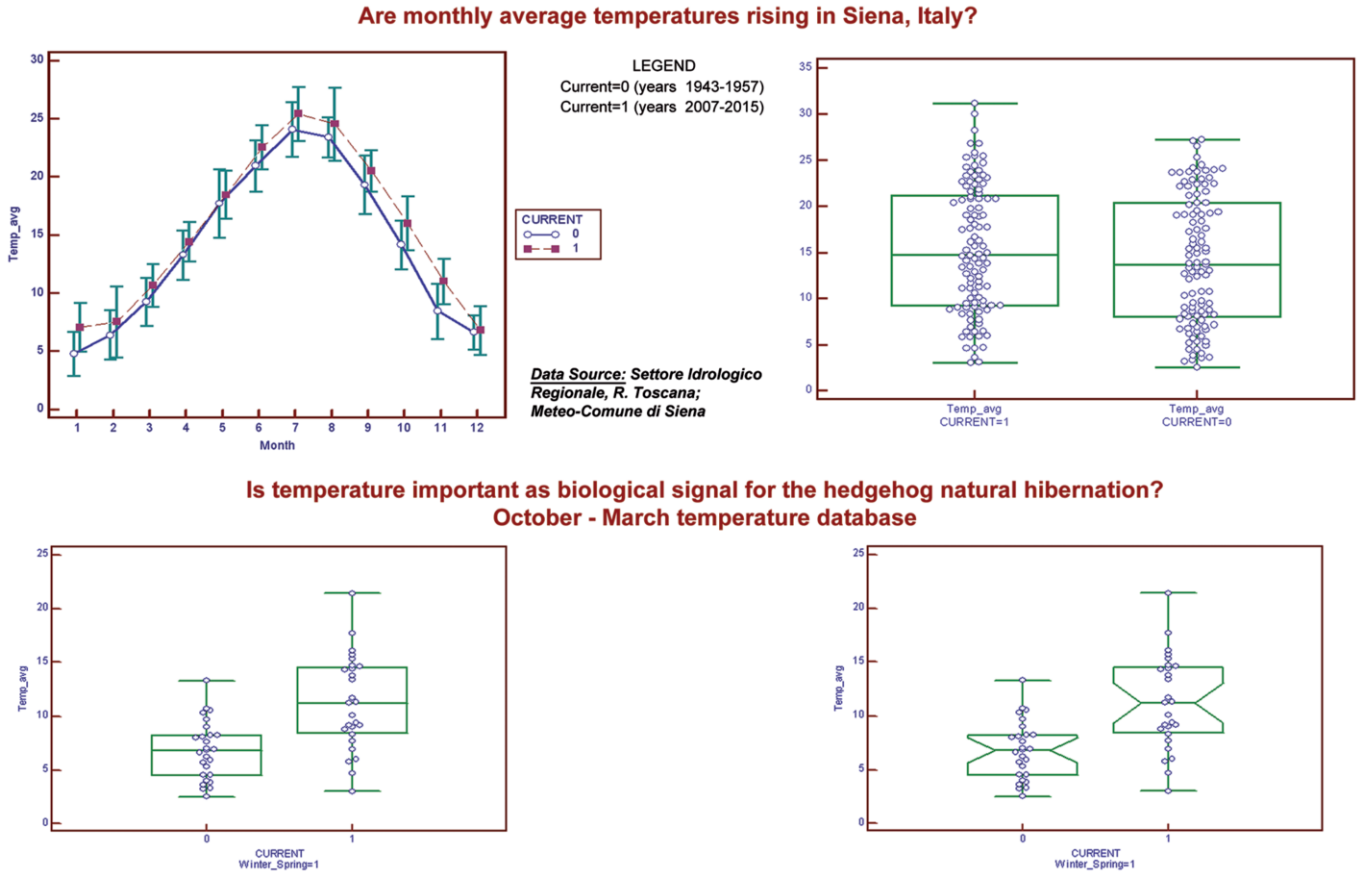


Figure 3. Average monthly temperature and missing of natural hibernation in hedgehogs (Siena, Italy, historical archives).

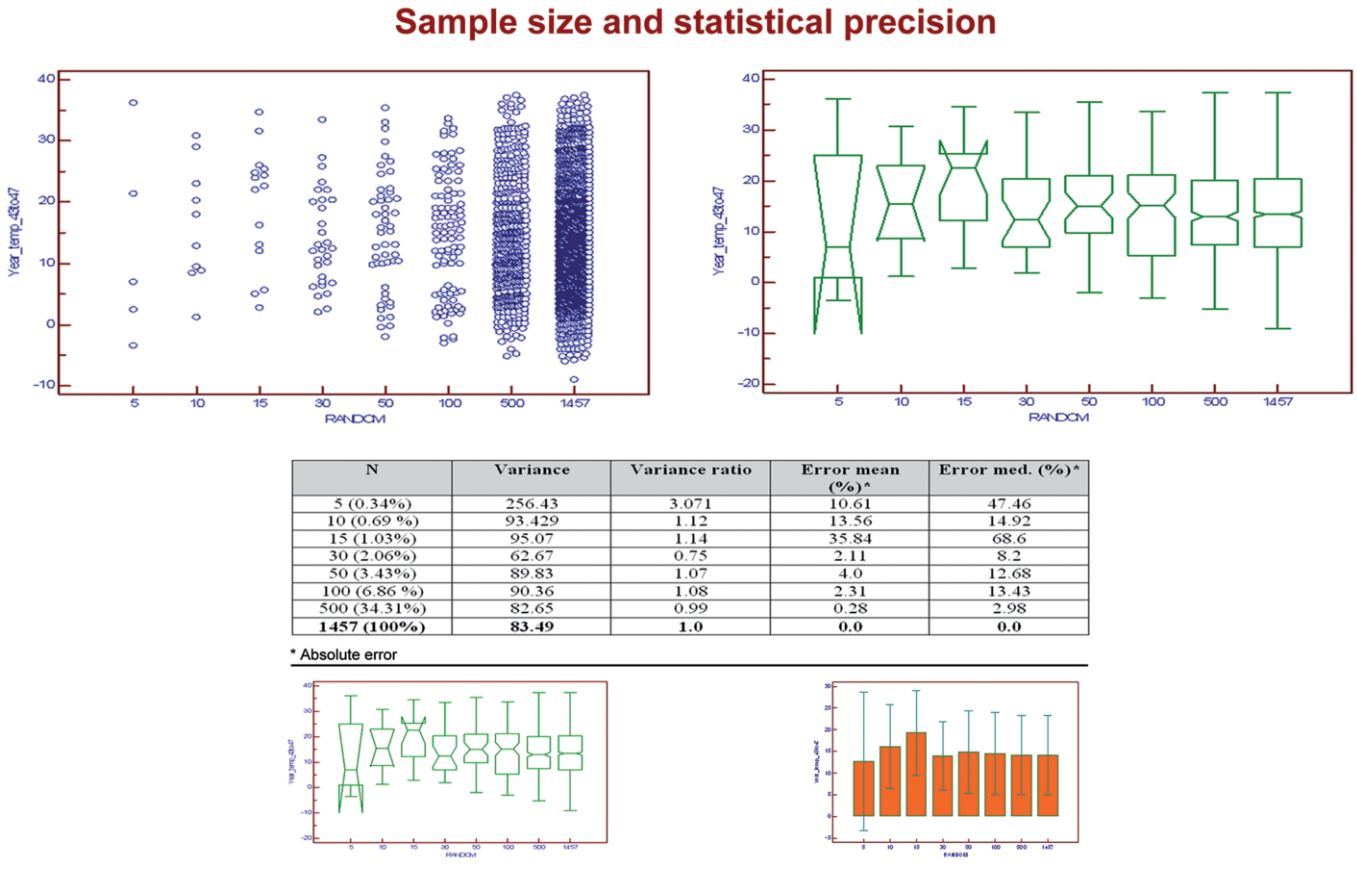
Data fabrication is another very relevant issue in scientific publishing. A meta-analysis of surveys asking scientists about their experiences of misconduct found that, on average, about 2% of scientists admitted to have fabricated, falsified or modified data or results at least [30]. Considering that these surveys ask sensitive questions and have other limitations, it appears likely that this is a conservative estimate of the true prevalence of scientific misconduct [30].

Despite the discovery and development of immunization has been a singular improvement in the health of mankind [31], confidence in vaccination has been declining in recent years. The current global anti-vaccine movement can be linked to a single retracted paper [32] published on February 28, 1998 in the highly respected Lancet journal. Sample size on the original paper was N=12. Further investigations on the leading author uncovered dishonest and unethical medical practices, resulting in losing his medical license. Although a careful review of publicly available information makes it clear that Wakefield's claims regarding vaccine safety are wrong [33, 34], vaccination rates plummeted in the United Kingdom from 92% in 1996/1997 to 80% in 2003/2004 [35], and outbreaks of vaccine preventable diseases followed [36, 37]. Measles remains of high clinical importance given that: i) infection leads to long-lasting immune suppression; ii) complications are of high frequency and severity; iii) there is no specific antiviral treatment; iv) vaccination is effective, cost-effective, and safe, with no demonstrated link between the measles vaccination and autism; v) can be eliminated from a population requiring a coverage with 2 doses of vaccine at rates of 93% to 95%; and vi) endemic transmission can be reestablished if rates of vaccination fall below the elimination threshold [34].

Although the Wakefield's controversy is a very good example of the negative impact of false science on real life, unfortunately, statistical review could do very little against publication of fabricated data. Nevertheless, some hope may originate from the so called Benford's law, *i.e.*, the form of logarithmic distribution of digits in statistical data when are produced by natural or social processes [38]. Indeed, Benford's law has been successfully applied to detect fabricated or falsified data [39] in tax or other financial reports [40-42].

CONCLUSIONS

Statistics is widely accepted as a powerful tool in the scientific research process, with a huge increase in the



N	Variance	Variance ratio	Error mean (%)*	Error med. (%)*
5 (0.34%)	256.43	3.071	10.61	47.46
10 (0.69 %)	93.429	1.12	13.56	14.92
15 (1.03%)	95.07	1.14	35.84	68.6
30 (2.06%)	62.67	0.75	2.11	8.2
50 (3.43%)	89.83	1.07	4.0	12.68
100 (6.86 %)	90.36	1.08	2.31	13.43
500 (34.31%)	82.65	0.99	0.28	2.98
1457 (100%)	**83.49**	**1.0**	**0.0**	**0.0**

* Absolute error

Figure 4. Data numbers and statistical precision.

use of statistical methods for a wide range of medical journals over the past four decades [43-45]. On the other hand, there is also wide consensus on generally low standards resulting in a large proportion of published medical research containing statistical errors [46].

Overall, statistics in biomedical sciences: i) is a powerful tool to interpret experimental data; ii) has little efficacy in detecting false science; and iii) is not the result of a simple "click of a mouse", but should rather be the final result of accurate research planning by experienced and knowledgeable users.

ACKNOWLEDGEMENTS

This *Lectio Magistralis* is dedicated *in memoriam* of Prof. Mario Comporti (1935-2014), an international pioneer in the exploration of oxidative stress in disease, co-editor-in-chief of the Journal of the Siena Academy of Sciences since 2009. He strongly believed in the key importance of scientific data and the critical importance of unveiling methodological influences that could negatively affect data reproducibility.

REFERENCES

1. *Ioannidis JPA. Why most published research findings are false. Plos Med 2005;2:e124.*
2. *Motulsky HJ. Common misconceptions about data analysis and statistics. J Pharmacol Exp Ther 2014;351:200-5.*
3. *Altman N, Krzywinski M. Points of significance: sources of variation. Nat Methods 2015;12:5-6.*
4. *De Felice C, Latini G, Vacca P, Kopotic RJ. The pulse oximeter perfusion index as a predictor for high illness severity in neonates. Eur J Pediatr 2002;161:561-2.*
5. *De Felice C, Dileo L, Parrini S, Latini G. Persistent fetal heart rate hypovariability: a presenting clinical sign of histologic chorioamnionitis at term gestation. J Matern Fetal Neonatal Med 2004;16:363-5.*
6. *Rossell D. Big data and statistics: a statistician's perspective. Metode Sci Stud J 2015;5:143-9.*
7. *Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011;10:712-3.*
8. *Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature 2012;483:531-3.*
9. *Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature 2014;505:612-3.*
10. *Anonymous. Trouble at the lab. The Economist 2013;409:23-7.*
11. *Munafò MR, Stothart G, Flint J. Bias in genetic association studies and impact factor. Mol Psychiatry 2009;14:119-20.*
12. *Brischoux F, Cook TR. Juniors seek an end to the impact factor race. BioScience 2009;59:638-9.*
13. *Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359-66.*
14. *Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995-1004.*
15. *Nuzzo R. Scientific method: statistical errors. Nature 2014; 506:150-2.*
16. *Goodman SN. Of P-values and Bayes: a modest proposal. Epidemiology 2001;12:295-7.*
17. *Schreier NK, Moltchanova EV, Blomstedt PA, et al. Prenatal exposure to wartime stress: long-term effect on coronary heart disease in later life. Ann Med 2011;43:555-61.*
18. *Brancolini R. "Inverno caldo, i ricci non vanno in letargo. E il Centro matildico nel Reggiano li salva". La Repubblica Bologna.it 23 gennaio 2016. Available from: http://bologna.repubblica.it/cronaca/2016/01/23/foto/inverno_caldo_i_ricci_non_sono_andati_in_letargo_per_tempo_e_il_centro_matildico_nel_reggiano_li_ha_salvati-131861537/#1.*
19. *International Business Machines Corporation. IBM Big Data Success Stories. Armonk, NY: International Business Machines Corporation; 2011. Available at: http://public.dhe.ibm.com/software/data/sw-library/big-data/ibm-big-data-success.pdf*
20. *Mandelbrot B. The fractal geometry of nature. San Francisco: W. H. Freeman and Co; 1982.*
21. *De Felice C, Barducci A, Latini G, et al. Node degree distribution in complex microvascular networks: a potential new diagnostic tool for extracellular matrix-related diseases. Fractals 2006;14:251-8.*
22. *Mutch WAC, Lefevre GR. Health, "small-worlds", fractals and complex networks: an emerging field. Med Sci Monit 2003;9:MT19-MT23.*
23. *West GB, Brown JH, Enquist BJ. The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 1999;284:1677-9.*
24. *Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature 2000;406:378-82.*
25. *Altman DG. Statistics and ethics in medical research. Improving the quality of statistics in medical journals. BMJ 1981;282:44-7.*
26. *Gardenier JS, Resnik DB. The misuse of statistics: concepts, tools, and a research agenda. Account Res 2002;9:65-74.*
27. *Gore SM, Jones G, Thompson SG. The Lancet's statistical review process: areas for improvement by authors. Lancet 1992;340:100-2.*
28. *Altman DG. Statistical reviewing for medical journals. Stat Med 1998;17:2661-74.*
29. *García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Med Res Method 2004;4:13-7.*
30. *Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One 2009;4:e5738.*
31. *Nambiar PH, Daza AD, Livornese LL Jr. Clinical impact of vaccine development. Methods Mol Biol 2016;1403:3-39.*
32. *Wakefield AJ, Murch SH, Anthony A, et al. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 1998;351:637-41.*
33. *Harrison JA. Wrong about vaccine safety: a review of Andrew Wakefield's "Callous Disregard". Open Vaccine J 2013;6:9-25.*
34. *Bester JC. Measles and measles vaccination: a review. JAMA Pediatr 2016;doi:10.1001/jamapediatrics.2016.1787.*
35. *Health Protection Agency, United Kingdom. Completed Primary Courses at Two Years of Age: England and Wales, 1966 - 1977, England only 1978 onwards. 2011. Available from:*
http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/VaccineCoverageAndCOVER/EpidemiologicalData/coverVaccine UptakeData/
36. *Jansen VA, Stollenwerk N, Jensen H, et al. Measles outbreaks in a population with declining vaccine uptake. Science 2003;301:804.*
37 *Ramsay ME. Measles: the legacy of low vaccine coverage. Arch Dis Child 2013;98:752-4.*
38. *Benford F. The law of anomalous numbers. Proc Am Philos Soc 1938;78:551-72.*
39. *Diekmann A. Not the first digit! Using Benford's law to detect fraudulent scientific data. J Appl Stat 2007;34:321-9.*

40. *Carslow C. Anomalies in income numbers. Evidence of goal oriented behaviour. Acc Rev 1988;63:321-27.*
41. *Berton L. He's got their number: scholar uses math to foil financial fraud. The Wall Street Journal 1995;B1.*
42. *Nigrini MJ. A taxpayer compliance application of Benford's Law. J Am Taxpay Assoc 1996;18:72-91.*
43 *Altman DG. Statistics in medical journals. Stat Med 1982;1:59-71.*
44. *Altman DG. Statistics in medical journals: developments in the 1980s. Stat Med 1991;10:1897-913.*
45. *Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000;19:3275-89.*
46. *Strasak AM, Zaman Q, Pfeiffer KP, et al. Statistical errors in medical research-a review of common pitfalls. Swiss Med Wkly 2007; 137:44-9.*