Should methods of correction for multiple comparisons be applied in pharmacovigilance? Reasoning around an investigation on safety of oral antidiabetic drugs

Lorenza Scotti[1], Silvana Romio[1,2], Arianna Ghirardi[1], Andrea Arfè[1], Manuela Casula[3], Lorna Hazell[4,5], Francesco Lapi[6], Alberico Catapano[3], Miriam Sturkenboom[2], Giovanni Corrao[1] on behalf of safeguard consortium

ABSTRACT

BACKGROUND: In pharmacovigilance, spontaneous reporting databases are devoted to the early detection of adverse event ‘signals’ of marketed drugs. A common limitation of these systems is the wide number of concurrently investigated associations, implying a high probability of generating positive signals simply by chance. However it is not clear if the application of methods aimed to adjust for the multiple testing problems are needed when at least some of the drug-outcome relationship under study are known. To this aim we applied a robust estimation method for the FDR (rFDR) particularly suitable in the pharmacovigilance context.

METHODS: We exploited the data available for the SAFEGUARD project to apply the r*FDR* estimation methods to detect potential false positive signals of adverse reactions attributable to the use of non-insulin blood glucose lowering drugs. Specifically, the number of signals generated from the conventional disproportionality measures and after the application of the r*FDR* adjustment method was compared.

RESULTS: Among the 311 evaluable pairs (i.e., drug-event pairs with at least one adverse event report), 106 (34%) signals were considered as significant from the conventional analysis. Among them 1 resulted in false positive signals according to r*FDR* method.

CONCLUSION: The results of this study seem to suggest that when a restricted number of drug-outcome pairs is considered and warnings about some of them are known, multiple comparisons methods for recognizing false positive signals are not so useful as suggested by theoretical considerations.

Key words: Spontaneous report database, Bladder safety, Cardiovascular safety, False Discovery Rate, Non-insulin blood glucose lowering drugs, Pancreatic safety, Pharmacovigilance,

(1) Department of Statistics and Quantitative Methods, Division of Biostatistics, Epidemiology and Public Health, Laboratory of Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy

(2)Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands

(3)Department of Pharmacological and Biomolecular Sciences, Centre of Epidemiology and Preventive Pharmacology (SEFAP), University of Milano, Milan, Italy - IRCCS MultiMedica, Sesto S. Giovanni, Milan, Italy

(4)Drug Safety Research Unit, Bursledon Hall, Blundell Lane, Bursledon, Southampton SO31 1AA, UK

(5)School of Pharmacy and Biomedical Sciences, University of Portsmouth, Portsmouth, UK

(6)Italian College of General Practitioners, Florence, Italy.

CORRESPONDING AUTHOR: Dr. Lorenza Scotti, Dipartimento di Statistica e Metodi Quantitativi, Sezione di Biostatistica, Epidemiologia e Sanità Pubblica, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, Edificio U7, 20126 Milano, Italy. Tel.: +39.02.64485846; Fax: +39.02.64485899; E-mail: lorenza.scotti@unimib.it

*DOI: **10.2427/11654***

Accepted on December 17, 2015

KEY POINTS

- *Spontaneous reporting (SR) databases are useful tools to generate signals in a timely and cost-effective fashion.*
- *In pharmacovigilance where a large number of tests are performed simultaneously and one-sided hypothesis are of interest, the robust estimation method for the FDR can be useful to take into account the multiple testing problem.*
- *When a restricted number of drug-outcome pairs is considered and warnings about some of them are known, multiple comparisons methods for recognizing false positive signals are not so useful as suggested by theoretical considerations.*

SPONSOR

This study was funded by a research grant from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 282521 the SAFEGUARD project.

CONFLICT OF INTEREST

The authors declare no conflict of interest.
The results of this study were never presented before

INTRODUCTION

Spontaneous reporting (SR) databases are useful tools to generate signals, i.e. abnormal or unusual reporting patterns suggestive of increased health risks associated with the use of a given drug [1-3]. Although they provide answers in a timely and cost-effective fashion, it should be considered that a wide number of possible associations are concurrently investigated by such approach. This implies a high probability of generating positive signals (i.e. *statistically significant* drug-outcome associations) simply by chance. False positive signals make interpretation of the entire panel of results difficult. It would then be helpful to minimize this source of error to clarify the focus for further research [4,5].

Different approaches addressing massive hypothesis testing have been developed. A conservative approach is to control the Family Wise Error Rate (*FWER*) that is the probability to reject at least one true null hypothesis among all tested; the Bonferroni method is one of the most used to account for this error [6]. A less conservative approach is to control the False Discovery Rate (*FDR*) i.e., the expected proportion of false positive findings among all the rejected hypotheses [7]. It should be considered, however, that a major assumption of FDR is that then p-values have to be uniformly distributed under the null hypothesis. Pharmacovigilance generally aims to detect signals, thus one-sided hypothesis tests are of interest. However, when one-sided hypothesis tests are performed, the uniformity assumption of p-values is systematically violated making the classical FDR approach inapplicable. Recently, Pounds and Cheng proposed a robust method for the estimation of FDR (rFDR) that overcome this assumption [8].

We exploited the data available for the Safety Evaluation of Adverse Reactions in Diabetes (SAFEGUARD) EU project, an international consortium aimed to assess the safety of non-insulin blood glucose lowering (NIBGL) drugs, to evaluate the need to apply multiple testing correction, through the rFDR in pharmacovigilance when a restricted number of hypotheses is tested [9-13].

METHODS

Data sources

We used the data retrieved from two SR databases namely FDA-AERS and EudraVigilance. The FDA-AERS database was set up from 2004 in the United States and receives adverse drug reaction reports from healthcare professionals, patients and drug manufacturers worldwide. A public, anonymized version of the FDA-AERs database is readily accessible by downloading data files from the FDA website. The EudraVigilance database was set up from 2001 to collect adverse drug reaction reports from national regulatory agencies of the European Union and drug manufacturers. Access to a subset of data from the EudraVigilance database for NIBGL drugs was permitted for the SAFEGUARD project. All records recorded from January 1, 2004 until December 31, 2012 were selected since this time-window was available for both databases.

Drug assessment

In the public version of the FDA-AERS database the coding of drug names is highly variable and only partially standardized to the FDA's drug dictionary. This limitation makes difficult the identification of all the NIBGL drugs difficult. Thus, the strategy adopted for the current study was to identify as many NIBGL drugs as possible by mapping reported drug names with a reference list of generic and trade names for the NIBGL drugs. This reference list was compiled manually using the on-line version of Martindale: the Complete drug Reference (www.medicinescomplete.com; accessed 30/11/12). All other drugs were regarded as non-NIBGL agents. NIBGL drugs identified by this process were recoded with their generic name and subsequently standardized using the ATC classification for the purposes of analysis. In the EudraVigilance dataset, drug names were coded to generic drug name at source.

Outcome assessment

The outcomes of interest for the current study were the following: ventricular arrhythmia, heart failure, myocardial infarction, haemorrhagic stroke, ischemic stroke, sudden cardiac death, acute pancreatitis, pancreatic cancer and bladder cancer.

The MedDRA terminology dictionary was used to code reactions in both the FDA-AERS and EudraVigilance databases [http://www.meddra.org/]. In order to extract cases from these databases, each outcome of interest was defined by two pre-specified lists of preferred term.

Raw signal generation

Following Evans et al., signal generation was based on the disproportionality approach [14]. The disproportionality measure used for signal generation in this study was the proportional reporting ratio (*PRR*). The PRR is the ratio between the proportion of outcomes of interest among all reported for a considered drug and the proportion of outcomes of interest among all reported for all other drugs. To evaluate if a drug was significantly associated to a specific outcome, the z-test based on a large-sample normal approximation was performed on the logarithmic transformation of the PRR. A signal was generated whenever the null hypothesis of proportionality for the natural logarithm of *PRR* (i.e., H_0: $ln(PRR)\leq 0$) was rejected ($p\text{-}value \leq 0.05$), favouring the alternative one-sided hypothesis of disproportionality (i.e., *H1:* $ln(PRR)>0$) as more convincing.

Multiple testing

Consider the situation of testing simultaneously m (null) hypotheses of which m_0 are true (Table 1). R, U, V, S and T are unobservable random variables, R representing the number of rejected hypotheses, U and S the number of correctly classified hypotheses and T and V the numbers of erroneously classified hypotheses.

Benjamini & Hochberg originally defined the *FDR* as the expected proportion of false positive findings among all rejected hypotheses, given that at least one null hypothesis is rejected, multiplied by the probability of making at least one rejection $FDR=E(V/R|R>0)Pr(R>0)$ [7]. The estimation of the *q-values*, the natural FDR analogues of p-values, corresponding to a given set of raw ordered p-values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ is based on the local *FDRs*. These are defined to be $lFDR_i=\tilde{v}(p_{(i)})/F\ (p_{(i)})$ for i=1,…,m, where $\tilde{v}(p_{(i)})$ is the estimated expected proportion of false positives when $p_{(i)}$ is used as threshold for evaluating the significance of each test, and $F\ (p_{(i)})$ the proportion of p-values less or equal to , $p_{(i)}$ i.e. $Pr(p\leq p_{(i)})$.

Pound & Cheng proposed a robust estimation procedure for the FDR (*rFDR*) [8], where $\tilde{v}(p_{(i)})$ is estimated as

$$\hat{v}(p_{(i)}) = \begin{cases} \hat{\pi} p_{(i)} & for\ p_{(i)} \leq 1/2 \\ \dfrac{\hat{\pi}}{2} + F(p_{(i)}) - F\left(\dfrac{1}{2}\right) & for\ p_{(i)} > 1/2 \end{cases}$$

where π is the cross-validation estimate of the proportion of true null hypotheses based on the distribution of observed raw p-values. This estimator is modified for $p_{(i)}> 1/2$ to avoid producing exceedingly large $lFDR_i$ values for large $p_{(i)}$s if observed raw p-values follow

TABLE 1

NUMBER OF ERRORS COMMITTED WHEN TESTING M NULL HYPOTHESES

NULL HYPOTHESIS	ACCEPT	REJECT	TOTAL
True	U	V	m_o
False	T	S	$m - m_o$
	m – R	R	m

TABLE 2

CONCORDANT-DISCORDANT MATCHING PAIRS FOR THE COMPARISON OF SIGNAL GENERATED BY THE ROBUST FALSE DISCOVERY RATE (RFDR) ESTIMATION WITH RESPECT TO RAW SIGNALS, ACCORDING TO DATA SOURCE

FDA-AERS DATABASE						EUDRAVIGILANCE DATABASE				
		RFDR Q-VALUES						RFDR Q-VALUES		
		>0.05	≤0.05					>0.05	≤0.05	
RAW	>0.05	85	0	85		RAW	>0.05	120	0	120
P-VALUES	≤0.05	0	55	55		P-VALUES	≤0.05	1	50	51
		85	55	140				121	50	171

a *U-shaped* distribution (as might typically happen in real applied settings) [8]. For each raw p-value $p_{(i)}$, the corresponding *q-value* is $q_i=min_{j \geq i} lFDR_i$. For example, if q_i is less than 0.05, then all hypotheses associated with the *p-values* from $p_{(1)}$ *to* $p_{(i)}$ can be rejected ensuring that *FDR* does not exceed 0.05.

The R (v3.0.2) package *"robust.fdr"* developed by Pounds & Cheng was used for estimating the *rFDR q-values* [http://www.stjuderesearch.org/depts/biostats/documents/robust-fdr.R]

Concordance between raw and rFDR methods

Two *I x J* matrices crossing the I NIBGL drugs and the J adverse reactions of interest (identified using the narrow definition) were separately built from each database. Collapsing these data into as much 2 x 2 contingency tables as NIBGL drug-outcomes pair with at least one report were observed, PRR point estimates and corresponding raw p-values, as well as the rFDR q-values, were calculated.

Concordant-discordant matching pairs comparing raw and rFDR signals were counted. A drug-outcome pair was considered concordant if raw and rFDR method tied the same classification in term of significance of a given signal (i.e., the p- and q-value were either both ≤ 0.05 or both > 0.05), discordant otherwise.

RESULTS

In total, 261 pairs (i.e., 29 drugs x 9 reactions) were evaluable. Table 2 reports the number of concordant and discordant drug-outcome pairs identified by raw and *rFDR* respectively within FDA-AERS and EudraVigilance databases. From the FDA-AERS database, 140 (54%) non-empty cells were obtained of which 55 (39%) concerned raw signals; perfect agreement was observed between raw and *rFDR*. From the EudraVigilance database, 171 (66%) non-empty cells were observed of which 51 (30%) concerned raw signals. Only one false positive signal, i.e., the effect of gliquidone on acute pancreatitis, resulted from the *rFDR*. A total of 71 individual concordant drug-outcome signals were confirmed by the *rFDR* method considering both databases (data not shown).

DISCUSSION

The current study investigated the possible association between 29 antidiabetic agents and 9 outcomes, that is 261 potential drug-outcome signals. Among these, 140 and 171 concerned nonempty cells generated from the EudraVigilance and FDA-AERS databases, respectively. From 30% to 40% of these drug-outcome pairs were detected as significant signals by the conventional analysis, so generating concern about the safety of certain NIBGL agents. It should be remarked that, under the null hypothesis of lack of any association among the investigated pairs, 5% of drug-outcome pairs is expected to be a significant signals by chance. Then, our raw findings suggest that the use of some NIBGL drug may cause one or more side effects, but the open question is whether all these signals are due to the drug effect or if some of these (how many? which ones?) have been generated by chance. The approach followed in our paper was to control for false positive signals through a robust estimation method of the false discovery rate.

The FDR-type methods has been described as particularly suitable for screening purpose, as it is the case of signal generation in the setting of pharmacovigilance [15]. However, theoretical considerations and simulation studies, have shown that the classical FDR approach may be too conservative [16]. This occurs mainly when the assumption of uniformity of the p-values under the null hypothesis required by the classical FDR method is violated, as in the case of one-sided hypotheses. The *rFDR* estimation method was developed to overcome the assumption. However, FDR-type methods, including its robust version, are typically applied in genetics, a setting where thousands of tests are simultaneously performed in almost total absence of a priori knowledge. In the pharmacovigilance setting, however, and particularly in the application presented in the current study, a more restricted number of tests (drug-outcome pairs) is of interest. Our findings showed that among the 96 signals generated from the conventional (raw) approach, almost all (95) were confirmed after the application of the rFDR method. The only signal detected as false positive by the *rFDR* concerns the effect of gliquidone on acute pancreatitis. This evidence was confirmed by another study based on FDA-AERS database [17]. However, to our best knowledge no other studies were published on the effect of gliquidone on the risk of acute pancreatitis.

The empirical evidence that *rFDR* detected a negligible proportion of false positive signals in our application, may have several explanations. First, the number of hypotheses simultaneously tested is restricted in our setting, so that it is possible that the probability of generating a false positive signal is not as much inflated as if the number of tests would be much larger. Secondly, since some of the drug-outcome associations of interest are already known, the number of related reports is expected to be high. For example, the relationship between rosiglitazone and cardiovascular and cerebrovascular outcomes is well known and largely documented in the scientific literature [18-22]. The number of reports regarding the relationship between rosiglitazone and three cardio-cerebro-vascular outcomes, namely, myocardial infarction, heart failure and ischemic stroke was respectively 15,040, 10,018 and 3,004 in the FDA-AERS database and 8,086, 7,270 and 5,967 in the Eudravigilance database. Similar results were found for the pancreatic safety of exenatide. Some evidence suggested a possible role of incretin mimetic drugs in the onset of pancreatic outcomes even if this topic is still debated [23-24]. The number of reports regarding the relationship between exenatide and acute pancreatitis and pancreatic cancer are 2,235 and 222 in the FDA-AERS database and 1,742 and 221 in the Eudravigilance database. All these evidence should lead to highly significant signals that would be unlikely detected as false positive by the *rFDR* approach. Thus, the number of potential false positive signals is limited by design in this setting.

It should be noticed that the analysis of spontaneous report databases is subject to several type of biases that are related to the spontaneous character of the reports. In particular, it is well known that the information reported in these databases are uncontrolled and thus may be affected by a number of reporting related biases. These biases includes the length of time a product has been on the market, country, reporting environment, detailing time and quality of the data [25]. Additionally, reported cases may differs from unreported ones in terms of disease severity or other clinical characteristics. Moreover the ability to assess, analyse and act on safety issues based on spontaneous reporting depends on the quality of the report [25]. Finally, the disproportionality measures calculated using these data may be affected by confounding and cannot take into account difference in patients' clinical profiles and presence of co-medications. Given these limitations, it is possible that some of the signals may be generated erroneously as a results of the combination of different uncontrolled causes. This fact may explain, for example, the signals associate to metformin use. The use of this drug, in fact, was associated to myocardial infarction and heart failure detected in both databases, but it is also well known that metformin has a acceptable cardiovascular safety profile [26]. However, these false positive signals might be due to a systematic error, such as confounding, and cannot be discarded using methods, like the *rFDR*, that act on random error.

CONCLUSIONS

These considerations taken together seem to suggest that when a restricted number of drug-outcome pairs is considered and warnings about some of them are known, multiple comparisons methods for recognizing false positive signals are not so useful as suggested by theoretical considerations.

ACKNOWLEDGEMENTS: *This study was funded by a research grant from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 282521 the SAFEGUARD project. Data analyses were performed at the Laboratory of Healthcare Research and Pharmacoepidemiology, Dept. of Statistics and Quantitative Methods, University of Milano-Bicocca with grants from the Italian Minister for University and Research ('Fondo d'Ateneo per la Ricerca' portion, year 2014).*

References.

[1] Ahmad SR, Goetsch RA, Marks NS. Spontaneous reporting in the United States. In Phamarcoepidemiology, 4th ed. Strom BL (Ed). Chichester, UK: Wiley, 2006:135-59

[2] Edwards IR, Olsson S, Lindquist M, et al. Global Drug Surveillance: The WHO programme for International drug monitoring. In Phamarcoepidemiology, 4th ed. Strom BL (Ed). Chichester, UK: Wiley, 2006:161-83

[3] Waller PC, Coulson RA, Wood SM. Regulatory pharmacovigilance in the United Kingdom: current principles and practice. Pharmacoepidemiol Drug Saf 1996;5:363-75

[4] Thomas DC, Siemiatycki J, Dewar R, et al. The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol 1985;122:1080-95

[5] De Roos AJ, Poole C, Teschke K, et al. An application of hierarchical regression in the investigation of multiple paternal occupational exposures and neuroblastoma in offspring. Am J Ind Med 2001;39:477-86

[6] Hochberg Y, Tamhane AC. Multiple comparison procedure. New York: Wiley, 1987

[7] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 1995;85:289-300

[8] Pounds S, Cheng C. Robust estimation of the false discover rate. Bioinformatics 2006;22:1979-87

[9] Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007;356:2457-71

[10] Psaty BM, Furberg CD. The record on rosiglitazone and the risk of myocardial infarction. N Engl J Med 2007;357:67-9

[11] Singh S, Loke YK, Furberg CD. Long-term risk of cardiovascular events with rosiglitazone: a meta-analysis. JAMA 2007;298:1189-95

[12] Jørgensen CH, Gislason GH, Andersson C, et al. Effects of oral glucose-lowering drugs on long term outcomes in patients with diabetes mellitus following myocardial infarction not treated with emergent percutaneous coronary intervention–a retrospective nationwide cohort study. Cardiovasc Diabetol 2010;9:54

[13] Egan AG, Blind E, Dunder K, et al. Pancreatic safety of incretin-based drugs–FDA and EMA sssessment. N Engl J Med 2014;370:794-7

[14] Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001;10:483-6

[15] Ahmed I, Dalmasso C, Haramburu F, et al. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. Biometrics 2010;66:301-9

[16] Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf 2009;18:427-36

[17] Raschi E, Piccinni C, Poluzzi E, et al. The association of pancreatitis with antidiabetic drug use: gaining insight through the FDA pharmacovigilance database. Acta Diabetol 2013;50:569-77

[18] Nissen SE, Wolski K. Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. Arch Intern Med 2010 ;170:1191-1201

[19] Lipscombe LL, Gomes T, Lévesque LE, et al. Thiazolidinediones and cardiovascular outcomes in older patients with diabetes. JAMA 2007;298:2634-43

[20] Home PD, Pocock SJ, Beck-Nielsen H, et al; RECORD Study Group. Rosiglitazone evaluated for cardiovascularoutcomes–an interim analysis. N Engl J Med 2007;357:28-38

[21] Delea TE, Edelsberg JS, Hagiwara M, et al. Use of thiazolidinediones and risk of heart failure in people with type 2 diabetes: a retrospective cohort study. Diabetes Care 2003;26:2983-9

[22] Lu CJ, Sun Y, Muo CH, et al. Risk of stroke with thiazolidinediones: a ten-year nationwide population-based cohort study. Cerebrovasc Dis 2013;36:145-51

[23] Singh S, Chang HY, Richards TM, et al. Glucagonlike peptide 1-based therapies and risk of hospitalization for acute pancreatitis in type 2 diabetes mellitus: a population-based matched case-control study. JAMA Intern Med 2013;173:534-9

[24] Elashoff M, Matveyenko AV, Gier B, et al. Pancreatitis, pancreatic, and thyroid cancer with glucagon-like peptide-1-based therapies. Gastroenterology 2011;141:150-6

[25] Goldman SA. Limitations and strengths of spontaneous reports data. Clin Ther 1998;20:C40-4

[26] Rojas LB, Gomes MB. Metformin: an old but still the best treatment for type 2 diabetes. Diabetol Metab Syndr 2013:15;5:6

*