An Effective Way to Quantify the Safety Profile of a Drug or Device When Background Toxicity Rates Are Low

by Eric S. Wei, Brian Claggett, Paul E. Greenberg, L.J. Wei

In 2007, Nissen and Wolski (NW) published an article in the *New England Journal of Medicine* [1] regarding potential cardiovascular toxicity from Rosiglitazone (Avandia), a drug that was commonly used to treat Type II diabetes. With the data from 42 comparative clinical trials posted by GlaxoSmithKline (GSK) on its website, NW performed a meta analysis and claimed that Avandia usage significantly increases the patient's risk of having myocardial infarction and likely increases the risk of cardiovascular (CV) related mortality. Their claim prompted vigorous debate within the medical community about the appropriateness of Avandia's continued availability on the market. As a result, the drug was banned in Europe, and its use was severely limited in the United States.

Their claims also triggered numerous personal injury law-

Mr. Wei
is a Partner at DembaWei, LLP,
in New York, NY.

Mr. Greenberg
is Director of the Health Economics Practice at Analysis Group Inc.
in Boston, MA.

Mr. Claggett
is a student at Harvard University,
in Boston, MA.

Dr. Wei
is a Professor at the Department of Biostatistics at Harvard University,
in Boston, MA.

suits against GSK. However, serious concerns have been raised by quantitative scientists about the statistical methods utilized in the analysis by NW. The approach taken by NW in their paper is not unique. Recently, other personal injury and security cases similar to Avandia's have also brought close public scrutiny to the choice of appropriate statistical procedures for quantifying drug safety profiles.

In the Appendix, we present the publicly accessible data for CV-related deaths from these 42 studies. It is important to note that for each study, the number of CV deaths is rather small. In fact, there are 19 studies that do not have any such events. In this article, we show, via this set of data, that the statistical method utilized by NW may not be appropriate, especially when the background toxicity event rates are low. We also suggest a simple, well-known procedure in statistics to quantify the treatment difference effectively with respect to toxicity.

Let us look at the data from a relatively large study (DREAM study in Appendix) to illustrate how NW summarized the contrast of the toxicity rates between Rosiglitazone and its control group. We present the data in the following 2 x 2 table (Table A).

Table A. DREAM Study: CV death data

Number of:	Avandia	Control
deaths	12	10
survivors	2623	2624
patients	2635	2634

NW employed the *relative risk* (RR)—sometimes referred to as the risk ratio—to quantify the group difference [1]. The RR is defined as a ratio of two mortality event rates (the Avandia rate divided by the Control rate) [2]. From Table A, the observed rates for Avandia and Control are 12/2635 = 0.00455 (0.455%) and 10/2634=0.00380 (0.380%), respectively. Therefore, the observed risk ratio is 0.00455/0.00380=1.20. Note that if the true ratio is one, there is no difference between two treatment groups. Since the observed rates are obtained from a sample of the potential patient population, it is important to know what the true RR would be for the entire population. With the data from Table A, the conventional 95 percent confidence interval for the true RR is between 0.52 and 2.77. That is, with a rather high probability, the possible values for the true RR could be as low as 0.52 but as high as 2.77. This set of possible values is quite large. The reason is that the size of the confidence interval for RR depends primarily on the numbers of events occurring in the two groups to be compared. In this example, we only have 22 total events. A large interval indicates a lack of certainty with regard to the true RR.

A similar quantity that can be used to measure the group contrast is called the *odds ratio* (OR) [2]. When the event rates are low, the observed OR is almost identical to its RR counterpart. Note that, for the DREAM study, even with its relatively large sample sizes, the resulting confidence interval for the RR is quite large and may not be used to claim that there is no difference between Avandia and its control group with respect to CV mortality, even though the interval contains the null value of one. In fact, based on this data, it is plausible that Avandia could reduce the rate of CV deaths by nearly 50 percent (as seen by the lower limit of the confidence interval). In this situation, quantitative scientists would conclude that there is a lack of statistical power due to the small number of observed events.

Now, if we triple the sample sizes for the DREAM study but keep the numbers of events the same, the resulting confidence interval would be almost identical to that based on the original data, indicating that there is still a lack of information. This seems counterintuitive. That is, the precision of the observed RR has nothing to do with the sample sizes. The core of the problem is not the inadequacy of the statistical inference methods for RR but the way in which we quantify the group contrast (or difference) using the RR or OR.

To solve this dilemma, one may use the *risk difference* (RD) (the Avandia rate minus the Control rate) to quantify the group contrast. For the data in Table A, the observed RD is 0.076 percent. The conventional 95 percent confidence interval for the true RD is between -0.272 percent and 0.424 percent. Unlike the interval for the RR, this estimate is quite informative for making inferences about the treatment difference. Compared to the observed background CV mortality rates, one may claim that there is no clinically meaningful difference between the two groups. Note that the size of the confidence interval for RD depends on the number of patients in each treatment group. The interval gets smaller as the sample size increases. For example, if we triple the sample sizes in Table A, but keep the event numbers unchanged, the resulting confidence interval for the RD is between -0.091 percent and 0.141 percent, a much tighter interval than the previous one. This is the first drawback of using RR to quantify the group contrast, especially when the event rates are low.

Now consider the following two studies ID-49653/128 (a small trial) and ID-49653/024 (a relatively large trial). Neither study has any CV-related deaths. For convenience, we present the data from these studies in Tables B and C, respectively.

Table B. Study 49653/128: CV death data

Number of:	Avandia	Control
deaths	0	0
survivors	39	38
patients	39	38

Table C. Study 49653/024: CV death data

Number of:	Avandia	Control
deaths	0	0
survivors	774	185
patients	774	185

For either study, because the ratio of event rates, 0.00%/0.00%, is undefined, NW claimed that there was no information about the CV-mortality difference between Avandia and its control group using the RR approach. As a result, they deleted 19 studies (with no events) from their meta analysis. Meanwhile, the exact 95 percent confidence interval for the RD using the data in Table B is between -7.3 percent and 7.1 percent, a relatively large interval due to small sample sizes. The exact 95 percent confidence interval for the RD using the data from Table C is between -1.75 percent and 0.61 percent, a relatively tight interval as a result of the large sample sizes.

These interval estimates for RD are quite informative for the contrast between Avandia and its control with respect to CV-related mortality. With the data shown in Table B, one may claim that there is not enough information to make an inference about the RD, but with the data shown in Table C, we may conclude that these two groups are similar. Unfortunately, we cannot make inferences about the RR with the data shown in Table B or C. In practice, one may replace 0 in these tables with 0.5 and apply the standard inference procedure for RR. Different imputed values for 0, however, may result in different conclusions [3]. This is the second drawback of using RR to quantify the group difference when the event rates are low.

Next, we show the third drawback of using the RR to quantify the difference between two groups when the event rates are low. Specifically, the RR can be quite unstable. To this end, consider the following two studies, ID-49653/085 and ID-AVM100264, whose data are summarized in Tables D and E, respectively.

Table D. Study 49653/085: CV-death data

Number of:	Avandia	Control
deaths	1	0
survivors	137	139
patients	138	139

Table E. Study AVM100264: CV death data

Number of:	Avandia	Control
deaths	2	1
survivors	292	301
patients	294	302

For the first study, there is one event in the Avandia group and none in the control group; for the second study, the sample sizes are more than double those in the first, with two events in the Avandia group and one in the control group. Clinically, these two studies appear to have similar toxicity profiles. However, for the first study, the observed RR is infinitely large (1/138 divided by 0) with a confidence interval from 0.011 to infinity, which is not useful for making an inference about the treatment difference. For the second study, the observed RR is 2.05 (which is drastically reduced from the point estimate of the previous study), with an interval estimate between 0.17 and 31.05, which is still quite large due to the fact that there are only three CV deaths in the study. In any event, the procedure for RR gives us quite different pictures about the relative toxicity profiles for Avandia in these two studies. On the other hand, the estimate for RD with the data from Table D is 0.72 percent, and the exact confidence interval is from -2.56 percent to 3.45 percent; meanwhile, the estimate for RD from Table E is 0.35 percent, with an exact confidence interval from -1.48 percent to 1.92 percent. The two estimates of RD from these two studies are similar while the two estimates of RR are quite unstable (a doubling of risk and an infinite increase in risk are very different). Both RD confidence intervals are informative with respect to the potential CV toxicity. This is the third drawback of using RR to quantify the treatment difference when event rates are low.

Based on the above observations, it is more appropriate to use RD to quantify the group difference than RR or OR, especially when the background toxicity event rates are low. This recommendation is generalized to situation where we deal with incidence rate comparisons. The observed incidence rate is the number of events divided by the exposure time. The *incidence rate ratio* (IRR) between two groups and the *incidence rate difference* (IRD) are commonly used in medical studies. However, when the incidence rate is low, the IRD is a much better summary to quantify the group difference than the IRR.

Despite the fact that it is more appropriate to use the RD to quantify the group difference than relative RR, the RR often finds its way into the public debate about drug safety. Since most cases involving drug or device safety issues have relatively low adverse event rates, we generally use meta analysis to combine information across a set of related studies. It is important to note that the above

three drawbacks associated with the RR metric are not diminished by conducting a meta analysis. If the confidence interval estimates for each individual study are not informative or reliable, the resulting interval estimate via meta analysis will have the same problems. By deleting 19 studies, the 95 percent confidence interval for RR in the meta analysis by NW is between 0.98 and 2.74, which is almost significant (p-value = 0.06) and potentially disconcerting from the perspective of Avandia's drug safety profile. On the other hand, including all 42 study data sets and using the technique proposed by Tian et al. [4] with the RD as the treatment contrast, the resulting 95 percent exact confidence interval for RD is between -0.13 percent and 0.23 percent, which has a rather tight range and indicates no evidence of harm with respect to cardiovascular-related mortality. Note that this finding was recently confirmed in an updated report by NW [5].

In summary, when evaluating the safety of a drug or device, the simple approach of using the RD discussed in this article provides effective and unbiased inferences, with none of the drawbacks associated with the use of RR. For combining information across various studies, recently Tian et al. [4] and Wang et al. [6] proposed novel quantitative methods of meta analysis for making inferences about the RD. These new methods have been utilized successfully in several large-scale legal cases and New Drug Applications to the U.S. Food and Drug Administration. The software for implementing these exact inference procedures can be obtained viathrough the Blue Null Consulting Group at www.bluenull.com.

Appendix

	Avandia Group		Control Group	
Study ID	# of pts.	CV Death	# of pts.	CV Death
49653/011	357	1	176	0
49653/020	391	0	207	0
49653/023	774	0	185	0
49653/093	213	0	109	0
49653/094	232	1	116	0
100684	43	0	47	0
49653/143	121	0	124	0
49653/211	110	3	114	2
49653/284	382	0	384	0
712753/008	284	0	135	0
AVM100264	294	2	302	1
BRL 49653C/185	563	0	142	0
BRL 49653/334	278	0	279	1
BRL 49653/347	418	0	212	0
49653/015	395	2	198	0
49653/079	203	1	106	1
49653/080	104	0	99	0
49653/082	212	1	107	0
49653/085	138	1	139	0
49653/095	196	1	96	0
49653/097	122	0	120	0
49653/125	175	0	173	0
49653/127	56	0	58	0
49653/128	39	0	38	0
49653/134	561	1	276	0
49653/135	116	2	111	1
49653/136	148	2	143	0
49653/145	231	1	242	0
49653/147	89	0	88	0
49653/162	168	1	172	0
49653/234	116	0	61	0
49653/330	1172	1	377	0
49653/331	706	1	325	0
49653/137	204	0	185	1
SB-712753/002	288	1	280	0
SB-712753/003	254	0	272	0
SB-712753/007	314	0	154	0
SB-712753/009	162	0	160	0
49653/132	442	1	112	0
AVA100193	394	1	124	0
DREAM	2635	12	2634	10
ADOPT	1456	2	2895	5

[1] Steven E. Nissen, M.D., and Kathy Wolski, M.P.H., "Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes", New England Journal of Medicine, June 14, 2007, Vol 356 no. 24.

[2]Reference Manual On Scientific Evidence, Second Edition, published by LEXIS for Federal Judicial Center 2000. Washington, D.C.

[3] Sweeting, MJ., Sutton, AJ., and Lambert, PC., "What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data" Statistics in Medicine, 23 (2004), 1351-75.

[4]Lu Tian, Tianxi Cai, Marc A. Pfeffer, Nikita Piankov, Pierre-Yves Cremieux, and L. J. Wei, "Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction", Biostat (2009) 10(2): 275-281 first published online October 14, 2008

[5] Steven E. Nissen, M.D., and Kathy Wolski, M.P.H., "Rosiglitazone Revisited", Archives of Internal Medicine, Vol 170 no. 14 (2010), 1191-1201.

[6] Rui Wang, Lu Tian, Tianxi Cai, and L. J. Wei, "Nonparametric inference procedure for percentiles of the random effects distribution in meta-analysis", Ann. Appl. Stat. Volume 4, Number 1 (2010), 520-532.