ANNEX

Errors and issues with Kaul and Wolf's two working papers on tobacco plain packaging in Australia

OxyRomandie - 29 January 2015

1. Summary

Error #1	Erroneous and misleading reporting of study results
Error #2	Power is obtained by sacrificing significance
Error #3	Inadequate model for calculating power which introduces a bias towards exceedingly large power values
Error #4	Ignorance of the fact that disjunctive grouping of two tests results in a significance level higher than the significance level of the individual tests
Error #5	Failure to take into account the difference between pointwise and uniform confidence intervals
Error #6	Invalid significance level due to confusion about one-tail vs. two-tail test
Error #7	Invalid assumption of long term linearity
Issue #1	Avoiding evidence by post-hoc change to the method
Issue #2	Unnecessary technicality of the method, hiding the methodological flaws of the papers
Issue #3	Very ineffective and crude analytic method
Issue #4	Non standard, ad-hoc method
Issue #5	Conflict of interest not fully declared
Issue #6	Lack of peer review

2. Introduction

We document here the methodological flaws, errors and violation of research standards which we found in the studies by Prof. A. Kaul and M. Wolf. In the sequel, we will refer to the documents given in the references, which are summarized in Table 1 below. The documents are presented in chronological order.

Table 1. Summary of documents referenced in this paper

Ref. Nb.	*Publication date*	*Short description*
[1]	22 May 2013	Kaul and Wolf's project proposal to PMI
[2]	20 March 2014	Notes from meeting in London
[3]	March 2014 (between 20 March and 26 March)	Original version of Kaul and Wolf's first paper (N° 149)
[4]	March 2014 (same day as [3])	PMI media release
[5]	26 March 2014	Comments on Kaul and Wolf's first paper by Cancer Council Victoria
[6]	26 March 2014	Comments on Kaul and Wolf's first paper by UK NHS
[7]	28 March 2014	Reply by Kaul and Wolf to the NHS comments
[8]	10 April 2014	Critique of Kaul and Wolf's first paper by Laverty et al. in The Lancet
[9]	May 2014	Second version of Kaul and Wolf first paper
[10]	30 June 2014	Kaul and Wolf's second paper (N° 165)
[11]	1 July 2014	Media release by IPE Institut für Politikevaluation
[12]	7 July 2014	Critique by Diethelm and McKee in Tobacco Control
[13]	19 July 2014	Kaul and Wolf's response to Laverty et al. critique in The Lancet
[14]	6 August 2014	Response of Japan Tobacco International to the UK Consultation on Standardised Packaging
[15]	7 August 2014	Response of British American Tobacco Ltd to the UK Consultation on Standardised Packaging
[16]	7 August 2014	Response of Philip Morris Limited to the UK Consultation on Standardised Packaging
[17]	31 October 2014	Article in Beobachter: Zürcher Professor forscht für Big Tobacco
[18]	24 December 2014	Article in Beobachter: Tabakmulti «überprüft» Studie
[19]	27 December 2014	T. Angeli (journalist): Universität Zürich lässt «Review» einer Studie durch Philip Morris zu
[20]	2015 (in print)	Critique by Laverty et al. in Tobacco Control

Many of the errors and issues we document here have already been publicly raised, notably in [5], [6], [8], [12] and [20].

The two papers by Kaul and Wolf were published - without peer review - on the website of the Department of Economics of the University of Zürich as part of its Working Paper series. The first study, Working Paper No. 149, was initially published in March 2014 [3] and was replaced in May 2014 with a revised version [9]. The second study is dated June 2014 and was published as Working Paper No. 165 [10].

Kaul and Wolf's studies were the product of a contract apparently concluded between the two professors, their respective universities (Zürich University of University of Saarland), the marketing consulting firm *IPE Institut für Politikevaluation GmbH* and Philip Morris International. In their project proposal to the multinational tobacco company [1], Wolf and Kaul indicate that:

> *The main goal of this project is to analyze whether a causal link between the Plain Packaing [sic] Act 2011 and smoking behavior (smoking prevalence, initiation, and intensity) in Australia can be established. To do so we apply statistical and economic methods to real-world data.*[1]

The University of Zürich authorized a Swiss journalist, Thomas Angeli, to *read* the confidential contract (but not make a copy) [19]. He was able to write down some of its key clauses, including the following:

> *If at any time either Party or either Party's Personnel is contacted by a third party, including any news organization, concerning the Services provided under this Agreement, such Party and/or such Party's Personnel shall make no comment, notify the other Party of the third party contact and refer the third party to such other Party and/or coordinate the information provided to the third party with such other Party.* [19]

Based on this contract, we take it for granted that Kaul and Wolf have given their consent to the communication on the studies issued by the other parties, notably Philip Morris International and *IPE Institut für Politikevaluation GmbH*, the marketing consulting firm of which Kaul is the Director [1].

3. Errors

We have identified the following seven errors. Each one of them applies to both papers of Kaul and Wolf and is sufficient to invalidate their findings.

Error #1 - "*The absence of evidence for an effect should not be misconstrued as evidence for no effect*." – **Erroneous and misleading reporting of study results**

In [7], Kaul and Wolf, objecting to the title used by the NHS to describe their study ("*Plain cigarette packaging doesn't work, says industry funded study*"), state that "*This title is an **incorrect** summary of our results and therefore is **misleading***." (emphasis ours) However, on the day of publication of their first paper on the website of Zürich University, PMI issued a media release [4] which starts as follows: "*The plain packaging experiment in Australia **has not deterred young smokers**, professors from the Department of Economics at Zurich University and the University of Saarland found in a report released today*" (emphasis ours). If the NHS title is incorrect and misleading, so then is this statement in PMI's media release, which was issued in coordination with Kaul and Wolf, as per their contractual arrangement. This means **the two professors have accepted that Philip Morris issue an incorrect and misleading media release**, either explicitly or implicitly by not publicly objecting to the contents of PMI's release, which quotes them extensively.

Still in [7], Kaul and Wolf make the following statement: "*Being experienced empirical researchers, we took care to point out that we "fail to find any evidence for an actual plain packaging effect", which is not the same as claiming we find evidence for no plain packaging effect.*" We see that the PMI's media release show they were not as careful as they claim, or that their attention to the matter was selective. The way they are quoted in the release [4] is conducive to being interpreted as "evidence of no effect": "*We used statistical methodology that gave every possible leeway for detecting a possible plain packaging effect. Nevertheless, the data does not support any evidence of an actual effect of the Australian Plain Packaging Act on smoking prevalence of minors."* On page 13 of the notes on their London meeting [2], Prof. Wolf says: "*I can say from upfront the methodology that we have employed is the one that gives the most leeway to finding an effect, **if there had been any,***" (emphasis ours) implying "but there was none."

Kaul and Wolf's first paper [9] ends with the following conclusion: "*Altogether, we have applied quite liberal inference techniques, that is, our analysis, if anything, is slightly biased in favor of finding a statistically significant (negative) effect of plain packaging on smoking prevalence of Australians aged 14 to 17 years. Nevertheless, no such evidence has been*

discovered. More conservative statistical inference methods would only reinforce this conclusion." Again the message here is "our method would have detected an effect if there had been any."

This message has been received enthusiastically by the tobacco industry. Kaul and Wolf's findings are key pieces of "evidence" used by the tobacco multinationals in their denial of the effectiveness of plain packaging [14-17].

For instance, in its response to the UK consultation on the introduction of regulations for the standardised packaging of tobacco products, PMI presented the result of Kaul and Wolf's studies as follows:

> "*In both studies, using standard techniques for statistical analysis and applying the standard statistical significance level of 5%, the experts found no evidence that "standardised packaging" had had an effect on smoking prevalence among Australians aged 14 to 17 years old (in the case of the March study) or Australians aged 14 and above (in the case of the June study).* ***Kaul and Wolf confirmed that if there had been an effect in reality*** *(including of the magnitude predicted by Pechey and the DH),* ***it would have been reflected in the data****. According to the study, however, no effect was found.*" (emphasis ours) [16]

It can again be presumed that the two professors gave their consent to this misleading and erroneous presentation of their results, in conformity with the contract with PMI stating that each party coordinate with the other parties the information provided to a third party. As far as we know, they have not published a disclaimer.

In their submission to the UK consultation, JTI makes use of Kaul and Wolf' studies as follows:

> *After 18 months, the evidence actually emerging from Australia reinforces the fact that plain packaging does not work:*
>
> • *studies by the Universities of Zurich and Saarland have found that plain packaging has had* ***no effect on smoking prevalence****, either among minors or adults (…)*
>
> (emphasis in original text) [14]

Again an erroneous and misleading statement, with no disclaimer from Kaul and Wolf. One note that the Japanese tobacco company refers to the professors' work for PMI as *"studies by the Universities of Zurich and Saarland"*, using the prestige and authority of the two academic institutions to give greater credit to these studies.

In the comments of British American Tobacco Ltd to the UK consultation [15], the results of Kaul and Wolf's two studies are summarized as follows: "*The Roy Morgan population survey data, which* ***shows that there has been no change*** *in the pre-existing trend in youth or adult smoking since the introduction of Plain Packaging.*" (emphasis ours). Again an erroneous and misleading statement, with no disclaimer from Kaul and Wolf.

Error #2 - "*A formal power analysis demonstrates that the power of our inference methods is remarkably high.*" – **Yes, but this power is obtained by sacrificing significance, thus rendering the inference method defective and Kaul and Wolf's conclusions invalid**

Kaul and Wolf's response to the critique by Laverty et al. in The Lancet [13] was almost entirely based on the power they attribute to their results:

> "*...any actual reduction will only turn out to be statistically significant with a certain probability, and this probability is known as the power of the test. Therefore, the authors [e.g. Laverty et al.] need to attach a power (number) to the specific effect of 1·25 percentage points (unless they have a power of 1 in mind, which is unrealistic).*
>
> *Second, an effect as large as 1·25 percentage points is not needed to be detected with any reasonable power. For example, power against a reduction of 0·5 percentage points is about 0·65; power against a reduction of 1·0 percentage point is about 0·80; and power against a reduction of 1·25 percentage points about 0·85.1 Power of 0·8 is a commonly accepted industry standard, so even the power against a reduction of only 0·5 percentage points is not unreasonably low.*" [13]

In the revised version of their first paper [9], Kaul and Wolf added a section entitled "*Power Analysis*" to address the critique by Laverty et al.

> "*...monthly observed prevalence is rather unstable over time and the deviations from the fitted trend line are typically quite large. This might raise the concern of whether our trend analysis has any reasonable power at all against a possible plain packaging effect. (...) We address this concern by carrying out a formal power analysis. In particular, we consider the following inference method to test for a plain packaging effect which is consistent with our previous analyses. (...) The resulting numbers are presented in Table 2. One can see that power is generally quite high instead of unreasonably low.*" [9]

The power of a test is the probability that it rejects the null hypothesis (commonly designed by the symbol H_0) when it is false. In the case of the studies under consideration, the power of the tests is (equivalently) the probability of concluding that there is a plain packaging effect when such an effect indeed exists. However, effects can be small or large. Kaul and Wolf have used the symbol **Δ** to represent the magnitude of the effect, i.e. "***Δ** is the (fraction of)*

percentage points by which plain packaging has lowered prevalence beyond the time trend". They have determined the power of their algorithm for various values of Δ. Below, we reproduce Table 2 in [9], to which we have added the first row.

Effect Δ	Power
0.00	0.50
0.25	0.56
0.50	0.64
0.75	0.72
1.00	0.79
1.25	0.85
1.50	0.90

Table 2: (Table 2 in [9] augmented with first line) Power of the inference method detailed in Algorithm 3.2 against a plain packaging effect of size. All numbers are based on B = 50, 000 Monte Carlo repetitions in Algorithm 3.4.

The following graph will help visualizing these results:

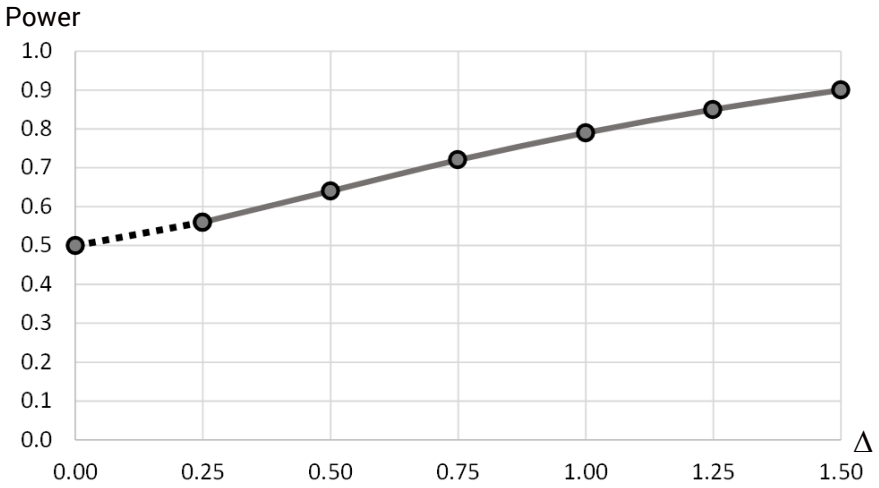

Figure 1: Plot of data in Table 2

The addition of the power value associated with a zero effect (**Δ** = 0), which we have computed using a Monte Carlo simulations with 50,000 iterations (giving a value of 0.5027, which we have rounded to 0.50) reveals the major flaw with Kaul and Wolf's method when we observe that the power associated with a zero effect is simply the *significance level* of the test, i.e. the probability of rejecting H_0 when H_0 is true. This means that Kaul and Wolf's power values are irrelevant because they are associated with an "inference method" whose level of significance is 0.5 (50%). Statisticians know that here is a trade-off between the

significance level and power and that arbitrarily high power (i.e. arbitrarily close to 1.0) can be artificially achieved if one relaxes requirements on the level of significance.

This fatal defect plagues both of Kaul and Wolf's papers. In the abstract of their second paper [10], they conclude by saying: "*A formal power analysis demonstrates that the power of our inference methods is remarkably high.*" Again taking the power figures in their Table 2 – last column showing the most complete test (IM3) - and completing them with the figure corresponding to no plain packaging effect (**Δ** = 0), we get the following table:

Effect Δ	Power (IM3)
0.00	0.48
0.25	0.67
0.50	0.85
0.75	0.96
1.00	0.99

Table 3. (Table 2 in [10] augmented with first line) Power against a permanent plain packaging effect over the period 01/2013–12/2013. The inference method IM-3 is detailed in Algorithm 3.2. All numbers are based on B = 50,000 Monte Carlo repetitions in Algorithm 3.4.

See also Figure 2 below which represents graphically the results in the table:

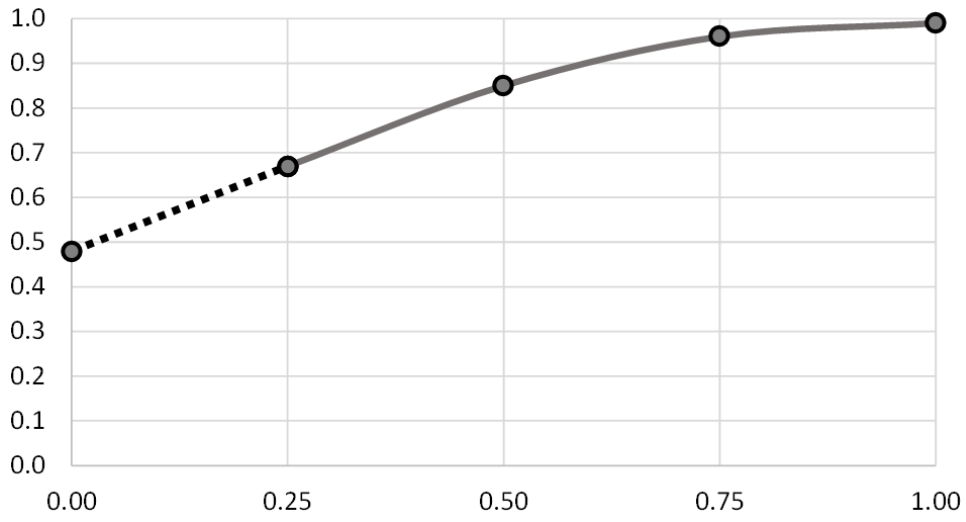

Figure 2: Plot of data in Table 3

One could see that Kaul and Wolf's most powerful "algorithm" (IM3) indeed achieves high power, but by sacrificing significance: its significance level is 0.48 (48%), against the 0.05 (5%) which is the established standard. This is sufficient to invalidate both papers, since no conclusion can be derived from such a defective method.

In [20] it is shown that the real power of Kaul and Wolf method is low (0.55 for $\mathbf{\Delta} = 1$), i.e. its ability to detect a prevalence decline of 1% below the trend line is almost equivalent to flipping a coin.

Error #3 – **Inadequate model for calculating power which introduces a bias towards exceedingly large power values**

In the power analysis of both papers [9] and [10], Kaul and Wolf use a model of plain packaging effect which greatly increases the power of their test. Indeed, instead of assuming a gradual reduction of prevalence from the trend, they assume a sudden change in December 2012, of magnitude **Δ**, and no further change in the subsequent months (step 3. in Algorithm 3.3). This is unrealistic, as plain packaging is mostly expected to reduce smoking *uptake*, which is spread over time: new smokers in a given year do not all take up smoking in December; furthermore, the effect on current smokers will also be gradual owing to the addictiveness of tobacco.

The graph below (Figure 3) illustrates the difference between the sudden effect postulated by Kaul and Wolf and the more realistic gradual model. One can see that the sudden effect model results in the doubling of the "e*ffect area*," replacing triangular area A under the trend line associated with the gradual effect with parallelogram A + B. The power associated with a **Δ** value in Kaul and Wolf's model will therefore correspond to the power associated with 2**Δ** in the more realistic gradual model: the bias in favour of power is thus very large in Kaul and Wolf's model and is not justifiable. The same observation is made in [20].

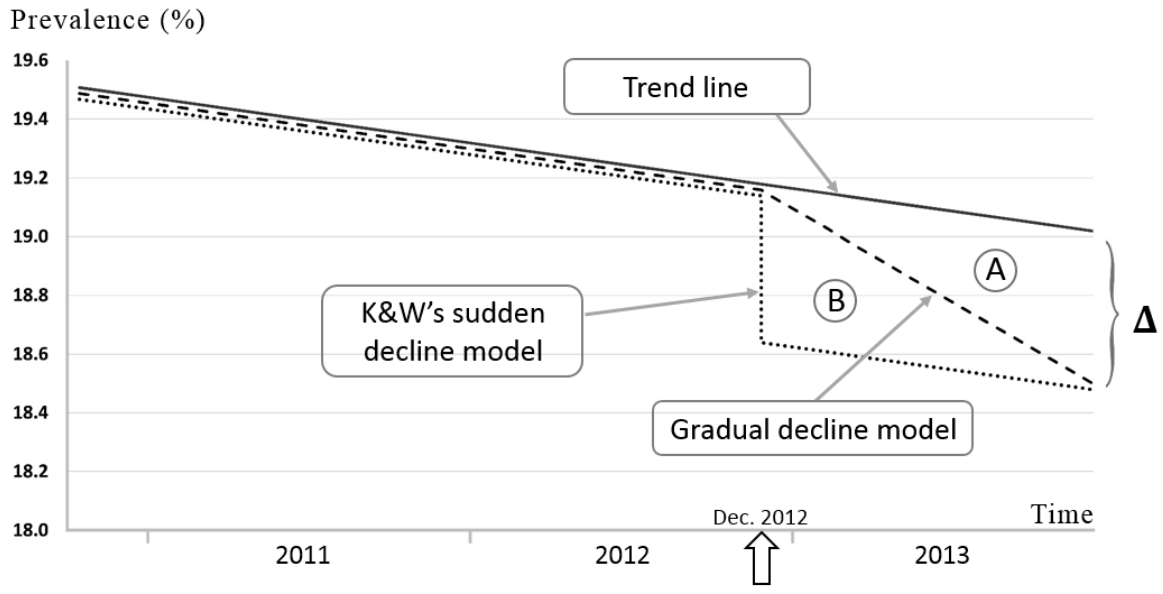

Figure 3: Gradual decline model vs. K&W sudden decline model

Error #4 – **Ignorance of the fact that disjunctive grouping of two tests results in a significance level higher than the significance level of the individual tests**

Suppose that for testing assumption H_1 two tests, T_a and T_b are available, with respective levels of significance α_a and α_b. If one groups the two tests into a single test T by saying that T is successful if either T_a or T_b or both are, then the significance level of T is $\alpha = \alpha_a + \alpha_b - \alpha_a\alpha_b$. In [9], Algorithm 3.1 and in [10], Algorithm 3.2, Kaul and Wolf carry out a two-sample *t*-test (T_a) for the null hypothesis of no treatment effect (IM-1 in [10]) and their confidence interval method (IM-2 in [10]). Then they group the two tests as follows: "*Overall, evidence for a plain packaging effect is established if at least one of these two approaches, IM-1 or IM-2, finds evidence. We call this 'combined' approach inference method 3 (IM-3).*" While they claim that the significance levels associated with the t-test and their confidence interval method are 0.05 (5%), they fail to say that the resulting combined test is then 0.05 + 0.05 – 0.5x0.5, i.e.0.0975, i.e. close to 10%. Admitting that the 5% significance levels of the individual tests were correct (which is not the case for the confidence interval method), the claim that their overall result satisfies the 5% significance level is therefore erroneous.

Error #5 – **Failure to take into account the difference between pointwise and uniform confidence intervals**

Kaul and Wolf are well aware of the difference between a pointwise confidence interval which applies to a pre-selected month and uniform confidence intervals which apply to a whole period of several consecutives months. In [9], they say:

> "*We have computed pointwise confidence intervals. That is, the confidence of 90% holds for any given month. Doing so is appropriate if one is interested in whether there is a plain packaging effect in any specific month, say in December 2012. But if one is interested in whether there is any plain packaging effect at all over the 13 months under consideration, it is more appropriate to compute uniform confidence intervals, where the 90% confidence holds over all 13 months together. Again, this would result in wider intervals.*"

In [10], they repeat the same statement almost verbatim and finish the paper with the following conclusion:

> "*…if the guiding research question is whether there is a plain packaging effect at all, one must adjust the confidence intervals to take the possibility of "cherry picking" into account (that is, the possibility of searching for a statistically significant effect over the entire period). Such an adjustment requires the use of uniform confidence intervals, in which case there is again no evidence for a plain packaging effect on smoking prevalence.*"

They unfortunately see only one side of the issue. They are right to observe that using the wider uniform confidence intervals would make it less likely to have an entirely negative confidence interval, in other words less likely that evidence be found. In their first study, Kaul and Wolf found no pointwise confidence interval that was entirely negative during the plain packaging period. So they claimed that basing their test on the pointwise (and narrower) confidence intervals, they had applied "*quite liberal inference techniques*," adding that "*our analysis, if anything, is slightly biased in favor of finding a statistically significant (negative) effect of plain packaging on smoking prevalence*." What they fail to see however is that this was done at a considerable cost: the use of the narrower pointwise intervals leads to *excessively* large power (see tables and graphs in Error #2 above) and extremely low significance level ($1 - 0.95^{13} = 0.49$, i.e. 49%, for the 13-month period used in [9] and $1 – 0.95^{12} = 0.46$, i.e. 46%. for the 12-month period used in [10]). This explains the results we found in our simulations for $\boldsymbol{\Delta} = 0$ which are presented under Error #2 above. What Kaul and Wolf did is simple: they took the best – for their purpose - of both alternatives: they took the

high power associated with the narrow pointwise confidence intervals and they took the desired (medium) significance level associated with the wider uniform confidence intervals. With as final outcome, a flawed method and false results.

Error #6 - **Invalid significance level due to confusion about one-tail vs. two-tail test**

In the abstract of their second study [10], Kaul and Wolf summarize their findings as follows:

> "*Our main results can be summarized as follows. First, if a statistical significance level of 5% is required, then there is no evidence at all for a plain packaging effect on smoking prevalence. Second, if one is willing to accept a relatively low level of statistical significance (that is, 10%), then there is evidence for a very short-lived plain packaging effect on smoking prevalence, namely in December 2012 only (after which smoking prevalence is statistically indistinguishable from its pre-existing trend)…*"

This is an important result of the study, as it is also covered in the media release issued by *IPE Institut für Politikevaluation GmbH,* which was relayed by the Reuters news agency [11]:

> "*The experts found no evidence for a plain packaging effect on smoking prevalence using standard techniques for statistical analysis, in particular requiring a statistical significance level of 5%, which is the standard in applied research. Only when the experts structured their analysis in a way that favoured finding an effect, in particular, by requiring a statistical significance level of 10% only, could they detect* 'evidence for a very short-lived plain packaging effect on smoking prevalence, namely in December 2012 (…)*.'"*

We are told that, while something happens in December 2012 when we take a low significance level (10%), at the standard statistical significance level of 5%, the evidence disappears. This is based on the following text in the [10]:

> *"The confidence level could be changed from 90% to 95%. The latter is more standard in applied research and would result in wider confidence intervals. If the confidence level is changed to 95%, then there is no evidence for a plain packaging effect whatsoever, since even the confidence interval for 12/2012 contains zero."*

However, this is incorrect. The criterion for evidence is explained as follows in [10]; "*If the entire interval lies below zero, then there is evidence (at the 90% confidence level) that plain packaging has lead to a reduction in prevalence."* This criterion is clearly a one-tailed use of the confidence interval – it has to be entirely *below* zero – and in such a case a 90% confidence interval leads to a 5% significance level, as is shown in Figure 4 below.

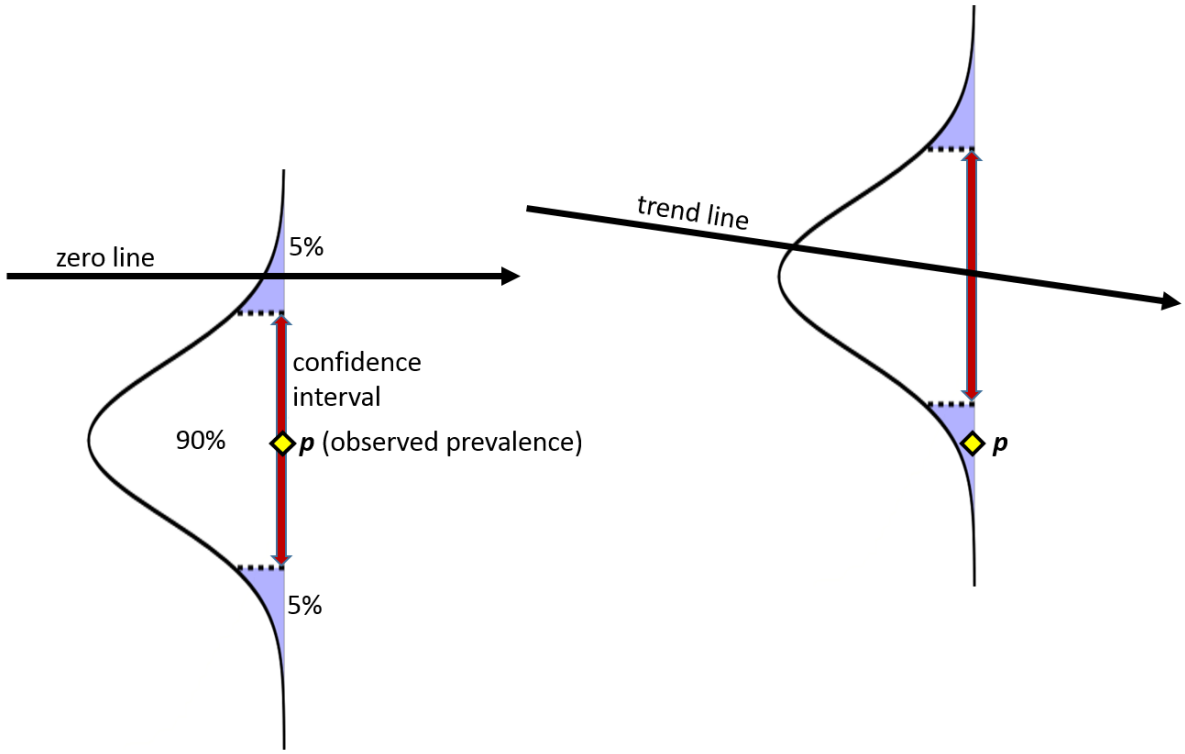

Figure 4: Example of a confidence interval that is considered evidence of a plain packaging effect by Kaul and Wolf. On the left, the interval is shifted, on the right it is centred on the trend line.

Notwithstanding the other methodological flaws of the paper, the statement "*if a statistical significance level of 5% is required, then there is no evidence at all for a plain packaging effect on smoking prevalence"* is therefore **false**. Based on their methodology, the authors found a plain packaging effect when a 5% statistical significance level is required. They simply reported the results associated with a 5% level of significance as having a 10% level of significance.

Error #7 – **Invalid assumption of long term linearity**

In both papers [9] and [10], Kaul and Wolf base their analysis on the assumption that smoking prevalence in Australia follows "*a simple linear time trend*." In [2] they explain that in Australia, like in "*all the OECD countries*," there is a continuous downward trend in smoking prevalence which is well modelled by a declining straight line. "*We see essentially the same line in all countries*" regardless of whether they have "*heavy anti-smoking measures*" with a "*minus 0.4 percentage point effect per year.*"[2] In other words, it should come as no surprise that plain packaging is ineffective, as, in Kaul and Wolf's view, all tobacco control measures produce no effect whatsoever: the decline in prevalence observed over the past 15 years across OECD countries, including Australia, is the result of a "*pre-existing*" continuous and uniform trend, best modelled by a straight line.

However, the assumption of a pre-existing linear trend is contradicted by the marketing survey data used by Kaul and Wolf. In [12] Diethelm and McKee observed that the lack of linearity of the data used in the first study is so severe that this alone suffices to invalidate the authors' findings. See Figure 5 below, which is extracted from [12]:

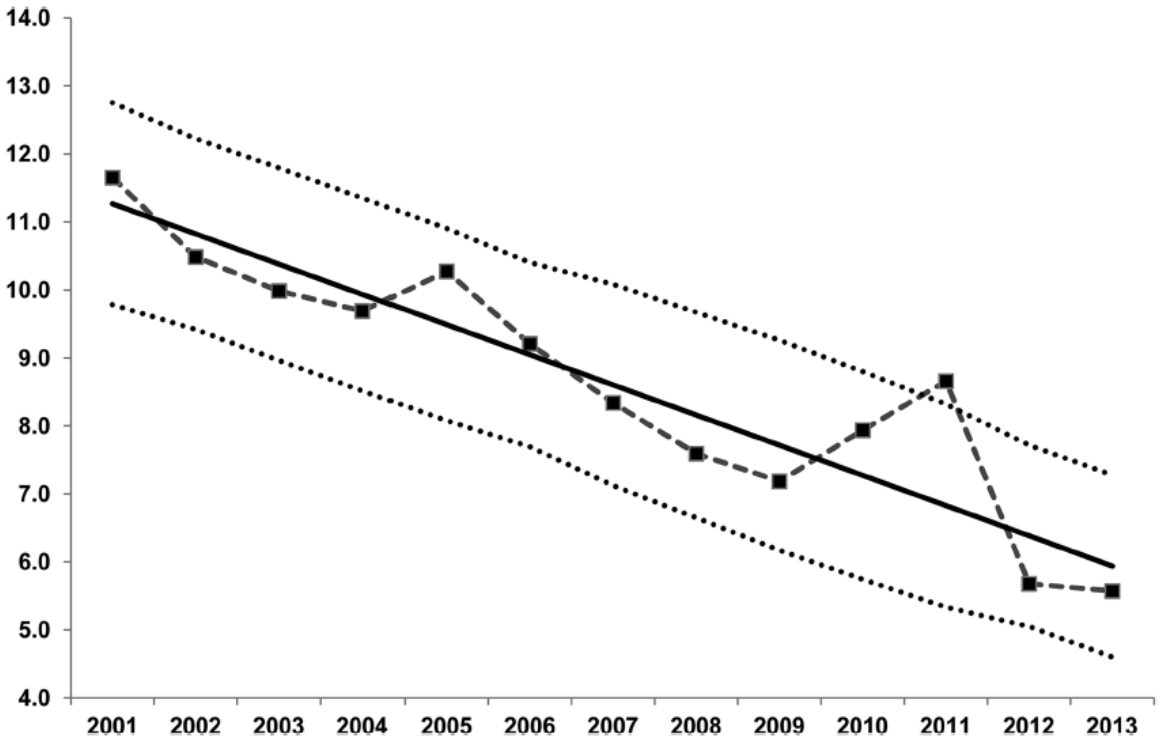

Figure 5. Observed average annual prevalence (squared dots linked by dashed line) vs expected annual prevalence based on the regression line (continuous line) and associated 95% uniform confidence interval (dotted line).

One see that in 2011, the average annual prevalence is largely outside the 95% uniform confidence interval. The data in [10] also suffers from the same weakness. Observing that "*the nonparametric fit resembles a straight line in the second two thirds of the observation period,*" the authors arbitrarily truncated out the first 42 months of observation, explaining unconvincingly that this was "*[f]or simplicity and for ease of reproducibility of our results by other researchers*". [10]

4. Issues

In addition to the errors documented above, the two papers suffer from a number of issues which are not errors properly speaking, but features which seriously affect the credibility and integrity of the studies.

Issue #1 – **Avoiding evidence by post-hoc change to the method**

In Kaul and Wolf's first paper ([3] and [9]), the inference method based on confidence intervals is applied in step 3 of their Algorithm 3.2, which reads as follows:

> *"Compute individual 90% confidence intervals for plain packaging effects from December 2012 until December 2013, as detailed in Section 3.2.3. If at least one of the resulting 13 confidence intervals is entirely negative, this is considered evidence for a plain packaging effect."*

In their second paper [10], Algorithm 3.2 is virtually the same as in [9], with the exception of a slight change:

> *"Compute individual 90% confidence intervals for plain packaging effects from 01/2013 until 12/2013, as detailed in Section 3.2.3. If at least one of the resulting 12 confidence intervals is entirely negative, this is considered evidence for a plain packaging effect."*

The authors take now only 12 confidence intervals, instead of 13 in the previous paper, excluding December 2012. No explanation is provided for such a change in the inference criterion. But we see that if criterion used in the first paper had been applied, the authors would have been forced to conclude that there is evidence for a plain packaging effect. They dealt with this situation by extracting December 2012 from the period of analysis and treating it as an isolated month that constitutes an exception, switching from a uniform approach to a pointwise approach. This is clearly *post-hoc* and not methodologically sound.

Issue #2 – **Unnecessary technicality of the method, hiding the methodological flaws of the papers**

In [10], Kaul and Wolf describe their "analysis based on confidence intervals" using the following "algorithmic" approach:

> "*Algorithm 3.1 (Computation of Confidence Intervals for Plain Packaging Effects)*
>
> 1. *Compute a 90% prediction interval for the observed prevalence based on the fitted time trend (that is, assuming no plain packaging effect). This means if another random sample (with the same sample size) had been chosen instead for this month, then the resulting observed prevalence would have fallen in this interval with 90% confidence (assuming no plain packaging effect). Or, alternatively, 90% of all possible random samples (with the same sample size) would have resulted in observed prevalence numbers falling in this interval (assuming no plain packaging effect). By construction, this interval is centered at the linear time trend.*
> 2. *Subtract the observed prevalence based on the original data from the upper and the lower interval end points.*
> 3. *The thus shifted resulting interval can be interpreted as a 90% confidence interval for the actual (treatment) effect of plain packaging. By construction, this interval is centered at the deviation from the linear time trend. If the entire interval lies below zero, then there is evidence (at the 90% confidence level) that plain packaging has lead to a reduction in prevalence.*"

Most readers will not realize that the criterion established with this "algorithm" is strictly equivalent to the following:

> If any observed monthly prevalence during the plain packaging period is below the 90% confidence interval around the trend line, then there is evidence of a plain package effect.

No need for an algorithm to formulate it. Kaul and Wolf's so called "*more informative analysis*" is actually this very crude and naïve test.

More precisely, Kaul and Wolf test, for each monthly prevalence, whether the entire shifted confidence interval is negative, i.e. its upper limit is less than zero:

$$(p - p^*) + (c^{sup} - p^*) < 0 \qquad (1)$$

where $(p - p^*)$ is the prevalence deviation from the trend line, p being the observed prevalence and p^* the expected prevalence on the trend line and $(c^{sup} - p^*)$ is the shifted upper limit of the 90% confidence interval, where c^{sup} designates the upper limit of the 90% confidence interval around the trend line.

As the distribution around the trend line is normal, therefore symmetrical, the confidence interval extends equally on both side of the trend line, i.e. $(c^{sup} - p^*) = (p^* - c_{inf})$. Thus equation (1) can be rewritten as $(p - p^*) + (p^* - c_{inf}) < 0$, or, as the p^* cancel each other, $p - c_{inf} < 0$, or, finally, $p < c_{inf}$.

Equation (1) is strictly identical to saying that the observed monthly prevalence is below the lower limit of the 90% confidence interval around the trend line.

Issue #3 – **Very ineffective and crude analytic method**

Once the real and crude nature of the so called "analysis based on confidence intervals" is revealed (see Issue #2 above), one realizes that the method is ineffective and has low power. When we repeated Kaul and Wolf's power analysis by Monte Carlo simulation, we noticed that a simple *t*-test, at equal level of significance, outperforms Kaul and Wolf's confidence interval methods for all **Δ** values (which quantify the effect), as is shown in Figure 6 below:

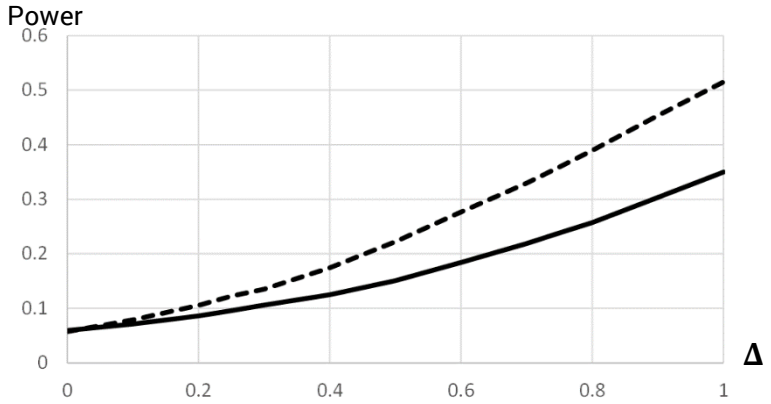

Figure 6. Comparison of the power of Kaul and Wolf's confidence interval method (continuous line) with a t-test (dashed line), at equal levels of significance, obtained by Monte Carlo simulation with 50,000 iterations.

It is easily understood that the confidence interval method makes poor overall use of the information as it is very sensitive to outliers. It would for example fail to detect as evidence a situation in which monthly prevalence figures were systematically and largely below the trend line for all 12 or 13 months if none of them was below the lower limit of the confidence interval. On the other hand, it would take as "evidence" a pathological situation in which all monthly prevalence figures would be *above* the higher limit of the confidence interval except one that would be below the lower limit of the confidence interval.

Issue #4 – **Non standard, ad-hoc method**

Although Kaul and Wolf claim in [10] that they use "*standard analytic techniques that are easy for other researchers to replicate"* their analytic techniques are far from standard, nor are they easy to replicate. They use an idiosyncratic algorithmic approach, resorting to no less than *four* algorithms, which are not described in any textbook, unsurprisingly as they are flawed. In order for us to verify the results presented in their power analysis, which were obtained by Monte Carlo simulation, we had to write a computer program, which was not an easy task.

Issue #5 – **Contradiction and lack of transparency about the way data was obtained**

Kaul and Wolf are not very clear about the way they obtained the marketing data they use in their studies. In [2], to the question by C. Cox asking whether the samples were balanced across ages, Wolf replies: "*From a statistical point, it would not be desirable that in one month they are all 14 and next month they are all 17. I don't know.* ***I was given the data to analyse***." (emphasis ours). As the London meeting was dedicated to discussing the findings which Kaul and Wolf published in the first paper, the data under consideration was thus the data for 14-17 minors. A bit further down in the same document, Wolf repeats: "*I was given the data at the end of January*." The passive voice suggests that Wolf was the passive recipient of the data, which someone else had obtained and gave to him. The presumption here must surely be that the data was obtained by Philip Morris and given to the two professors. The contradiction in the way Kaul and Wolf commented on this question reinforces that presumption.

In their reply to the critique by Laverty et al. in The Lancet, Kaul and Wolf say:

> "*The data we have worked with are* ***publicly available****, and our analyses are described in detail and can be replicated.*" (emphasis ours)

This is simply false. The data are not *publicly available*, but need to be bought from marketing survey firm Roy Morgan Research. In [2] Kaul describes how the Roy Morgan data was chosen by the two professors:

> *"We have reviewed several data-sets because it was not the case that Philip Morris asked us to do the research with the Roy Morgan data-set but they gave us the option to have look at all the data-sets that were available and this was the best data-set we could find."*

So, through Philip Morris, Kaul and Wolf could review all available data sets. One could presume that if PMI gave the two professors such an option, they provided them with the data sets. This remains unclear and needs to be clarified.

In their second paper [10], the footnote to the title of the paper reads as follows:

> "*Philip Morris International provided the funding for this research. At no time did we provide Philip Morris International with access to the underlying data for minors (14–17 years old). The data for adults were provided to us by Philip Morris International.*"

Here, the two authors admit that the data for adults were given to them by Philip Morris. They do not say, however, who gave them the data for minors: they merely imply that they obtained the data by other undisclosed means and did not grant PMI access to them. These explanations are confusing and unconvincing. If Philip Morris bought the marketing survey data from Roy Morgan for adults, they can as easily buy the data for minors.

Issue #6 – **Conflict of interest not fully declared**

In both of their papers, Kaul and Wolf acknowledge that their studies were funded by Philip Morris International. However, they failed to reveal the full extent of the involvement of Philip Morris in the project, which goes beyond funding. Indeed, the contract between the University of Zürich and Philip Morris International contains the following clause:

> *University agrees that, prior to submission to publisher of a manuscript describing the results for publication, University shall forward to PMIM 30 days prior to planned publication a copy of the manuscript to be submitted to PMIM for review and comments and University will take into account in good faith the said comments.* [19]

The ability granted to Philip Morris to supervise the study, review the manuscript before publication and propose changes [18] makes Philip Morris a *stakeholder* in the research work. The study is no longer a project of the University of Zürich with funding from a tobacco company, it is a joint project between the University and the tobacco multinational. For the tobacco companies, such types of projects belong to the category of "*extra-muros* research". A complete declaration of conflict of interest requires that such an arrangement be fully disclosed. (Of course, the recommended policy is for the university to prohibit its members from concluding such an arrangement with a tobacco company).

Furthermore, the role played by *IPE Institut für Politikevaluation GmbH* in the project is not disclosed. The fact that IPE has issued a press release which was coordinated with the publication of the second paper suggests that this marketing consulting company is also a player in the project. Both Kaul and Wolf seem to have a vested interest in IPE, Kaul as its director and Wolf as its "*senior researcher*". Again, for a full disclosure of conflict of interest (which should list not only the actual situations of conflict of interest, but also the potential ones), the role of IPE in the project (do they also receive money from Philip Morris for the project?) and the role of the two authors in IPE should have been declared.

Issue #7 – **Lack of peer review**

In the NHS critique of Kaul and Wolf's first paper [6], the lack of peer review was questioned and was also related to the potential conflict of interest as follows:

> "*The research does not appear to have been peer reviewed, meaning it has not been scrutinised by independent experts in the field for methodological rigor, or to check if the conclusions are reliable. This increases the risk of misleading findings, which can reach the public and mainstream media before they have been properly scrutinised.*
>
> *There is a clear potential conflict of interest in receiving funding from a leading tobacco company when attempting to carry out impartial research assessing smoking data. The risk of misleading information being presented is further increased when the research is not peer reviewed. Given that both these factors are present in this particular study, the results should be interpreted with caution.*"

In [7], the authors reply to this critique by saying that

> "(…) *peer-review takes time and findings are typically communicated in working papers in order to allow for a methodological debate and to disseminate findings at an early stage. (…) We will be submitting our study to a peer-reviewed outlet in due time. Given the straightforward nature of the data and the statistical methodology, we do not expect changes to the basic findings during the reviewing process.*"

It seems that the authors and their financial sponsors had no time to wait for the methodological debate to take place: they immediately made political use of the findings. The communication surrounding the papers by Philip Morris was political from the outset, and orchestrated in close collaboration with the authors: a PMI press release about the first study [4] was issued on the same day as the authors posted their working paper on the website of the University of Zürich; it triggered media reactions in different parts of the globe. The authors themselves did not wait for the result of the methodological debate to communicate to the UK Chantler's team (in charge of a preliminary report on plain packaging) their findings on the lack of effectiveness of plain packaging with minors.

The media release published on 1st July 2014 by *IPE Institut für Politikevaluation GmbH* announcing publication (on 30 June) of their second paper was accompanied on the same day (at 8am!) by a news release issued by Reuters. This was well coordinated.

All tobacco multinationals referred to the studies in their submissions to the UK Consultation on standardised packaging. None of them mentioned that they were not peer-reviewed, including Philip Morris, the sponsor of the study.

References

[1] Wolf, M. and Kaul, A. Project proposal: Intervention Analysis: the Effect of Plain Packaging for Tobacco Products on Smoking Behavior in Australia - A Quantitative Evaluation Applying Statistical Methods. Submitted to Philip Morris International (PMI), Lausanne, 22 May 2013. Available from: http://angelisansichten.ch/wp-content/uploads/2014/12/Project-Proposal.pdf

[2] Meeting to discuss "The (Possible) Effect of Plain Packaging on the Smoking Prevalence of Minors in Australia: A Trend Analysis" working paper. Attendees: Kaul A, Wolf M, Cox C, Collis J and Edwards L. King College London, 20 March 2014. Available from: https://www.kcl.ac.uk/health/Packaging-review/packaging-review-docs/meetingsandbriefings/Professors-Kaul--Wolf-%28University-of-Zurich%29-20-March-2014.pdf

[3] Kaul A and Wolf M. The (Possible) Effect of Plain Packaging on the Smoking Prevalence of Minors in Australia: A Trend Analysis. University of Zurich Department of Economics Working Paper Series. March 2014; Available from: http://www.oxyromandie.ch/docs/econwp149-march-2014.pdf

[4] Philip Morris International. Researchers Find No Evidence Plain Packaging 'Experiment' Has Cut Smoking. Media release. March 2014. Available from: http://www.pmi.com/eng/media_center/Pages/plain_packaging_experiment.aspx

[5] Cancer Council Victoria, Comments on Kaul & Wolf "The (possible) effect of plain packaging on the smoking prevalence of minors in Australia: a trend analysis", Melbourne, 26 March 2014. Available from http://www.cancervic.org.au/downloads/tobacco_control/2013/Cancer_Council_Victoria_comments_on_Kaul_Wolf.pdf

[6] UK National Health Services. Plain cigarette packaging doesn't work, says industry funded study. 26 March 2014. Available from: http://www.nhs.uk/news/2014/03March/Pages/Plain-fags-packs-dont-work-says-industry-funded-study.aspx

[7] IPE Institut für Politikevaluation GmbH. Reply by Ashok Kaul and Michael Wolf to the NHS choices comment Reply to "Plain cigarette packaging doesn't work, says industry

funded study“ by Ashok Kaul and Michael Wolf. 28 March 2014. Available from: http://www.ipe-saarland.de/deutsch/news/.

[8] Laverty AA, Watt HC, Arnott D, and Hopkinson NS. Standardised packaging and tobacco-industry-funded research. Lancet, 2014; 383(9926):1384. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24726722

[9] Kaul A and Wolf M. The (Possible) Effect of Plain Packaging on the Smoking Prevalence of Minors in Australia: A Trend Analysis. University of Zurich Department of Economics Working Paper Series. May 2014; Available from: http://www.econ.uzh.ch/static/workingpapers.php?id=828

[10] Kaul A and Wolf M. The (Possible) Effect of Plain Packaging on Smoking Prevalence in Australia: A Trend Analysis. University of Zurich Department of Economics Working Paper, June 2014. Series. Available from: http://www.econ.uzh.ch/static/workingpapers.php?id=844

[11] IPE Institut für Politikevaluation GmbH. IPE Institute for Policy Evaluation: Research Released on Smoking Prevalence in Australia Following Plain Packaging. Media release. 1st July 2014. Available from: http://www.ipe-saarland.de/app/download/8738689194/Media+Release+-+University+of+Zurich+and+Saarland+Report+-+July.pdf?t=1409657935

[12] Diethelm P and McKee M. Tobacco industry-funded research on standardised packaging: there are none so blind as those who will not see! Tobacco Control, 2014. Available from: http://tobaccocontrol.bmj.com/content/early/2014/07/07/tobaccocontrol-2014-051734.short

[13] Kaul A and Wolf M. Sandardised packaging and tobacco-industry-funded research. The Lancet 2014;**384**(9939):233-34. Available from: http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736%2814%2961210-1.pdf

[14] Japan Tobacco International. JTI’s response to the UK Department of Health’s consultation on the introduction of regulations for standardised packaging of tobacco products. 6 August 2014. Available from: http://www.jti.com/download_file/1569/629/

[15] British American Tobacco UK Ltd. Consultation on the introduction of regulations for the standardised packaging of tobacco products. Response of British American Tobacco UK Limited. 7 August 2014. Available from: http://www.bat.com/group/sites/uk__9d9kcy.nsf/vwPagesWebLive/DO9DKJEB/$FILE/medMD9MWB4B.pdf?openelement

[16] Philip Morris Limited. Response to the Consultation on “Standardised Packaging” 7 August 2014. Available from: http://www.pmi.com/eng/tobacco_regulation/submissions/Documents/UK%20-%20%20Standardised%20Packaging%20Submission%20PML.pdf

[17] Angeli T and Hostettler O. Zürcher Professor forscht für Big Tobacco. Beobachter 22/2014, 31 October 2014. Available from: http://www.beobachter.ch/justiz-behoerde/buerger-verwaltung/artikel/rauchen_zuercher-professor-forscht-fuer-big-tobacco/

[18] Angeli T. Uni Zürich: Tabakmulti «überprüft» Studie. Beobachter 26/2014. 24 December 2014. Available from: http://www.beobachter.ch/justiz-behoerde/buerger-verwaltung/artikel/rauchen_uni-zuerich-tabakmulti-ueberprueft-brisante-studie/

[19] Angeli T. Universität Zürich lässt «Review» einer Studie durch Philip Morris zu. 27 December 2014. Available from: http://angelisansichten.ch/universitaet-zuerich-laesst-review-einer-studie-durch-philip-morris-zu/

[20] Laverty A, Diethelm P, Hopkinson N et al. Use and abuse of statistics in tobacco industry funded research on standardised packaging. Tobacco Control, 2015. In print.