A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered

Georgia Salanti\textsuperscript{a,b,*,} Valeria Marinho\textsuperscript{c}, Julian P.T. Higgins\textsuperscript{a}

\textsuperscript{a}MRC Biostatistics Unit, Cambridge, UK
\textsuperscript{b}Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece
\textsuperscript{c}Clinical and Diagnostic Oral Sciences, Barts and The London School of Medicine and Dentistry Queen Mary, University of London, London, UK

Accepted 3 October 2008

Abstract

Objective: To illustrate the potential and challenges of the simultaneous analysis of a network of trials, using as a case study the investigation of the relative effectiveness of four topical fluoride treatments and two control interventions (placebo and no treatment) in preventing dental caries in children.

Study Design and Setting: We performed multiple-treatments meta-analysis within a Bayesian framework by synthesizing six Cochrane reviews. We explored the compatibility between direct and indirect evidence and adjusted the results using a meta-regression model to take into account differences in the year of randomization across studies.

Results: The validity of our conclusions for the superiority of fluoride toothpaste as indicated from the initial network analysis using Bayesian methods was challenged when we adjusted for possible confounders. The network was dominated by studies comparing placebo with toothpaste, which were older and had been carried out in populations with higher baseline risk than studies involving other fluoride modalities.

Conclusion: After adjusting for possible differences across studies, we did not find clear evidence that any topical fluoride modality is more effective than any other. Multiple-treatments meta-analysis methods allow for more detailed investigations than naive methods in the analysis of indirect evidence on treatment effects. © 2009 Elsevier Inc. All rights reserved.

Keywords: Meta-analysis; Topical fluoride therapy; Mixed treatment comparisons; Multiple-treatments meta-analysis; Incoherence; Meta-regression; Dental caries

1. Background

Methods for combining clinical trials making different treatment comparisons were first described explicitly over a decade ago [1]. Only recently, however, have they become more widely implemented, with the increased complexity of analyses that underpin clinical guidelines and health technology appraisals, such as those produced by the UK National Institute for Health and Clinical Excellence. We refer to these joint analyses as multiple-treatments meta-analyses. They are also known as ‘mixed treatment comparisons meta-analyses’ or ‘network meta-analyses,’ reflecting the network of comparisons that arises when collating studies involving different selections of treatments [2]. The benefits of multiple-treatments meta-analysis have been described previously [2–5], and include increased precision and ability to produce a ranking of the efficacy of different treatments included in the analysis.

Concerns have been expressed about the validity of the multiple-treatments meta-analysis methods, because they rely on assumptions that are difficult to test [3,5]. Statistical methods that allow quantification of the ‘coherence’ in a network of treatments have been recently proposed [6–8]. However, a clear conceptual strategy is not yet available for evaluation of a complete network, and most of the multiple-treatments meta-analyses to date do not investigate the assumptions of the underlying model [8–12]. The principal aim of this article is to illustrate the potentials and challenges of a simultaneous analysis of studies involving multiple treatments through the use of a case study of a comparison among four fluoride modalities against each other, against placebo, or against no-treatment controls, as identified in six Cochrane reviews. We, in particular, aim to explore the validity of some of
the underlying assumptions and in examining interstudy variation and its impact on the ranking of treatments.

The assumptions of multiple-treatments meta-analysis and the appropriate models under different scenarios are outlined and discussed in Salanti et al. [2]. First, we describe the fluoride data in more detail.

2. Topical fluoride therapies for preventing dental caries

The use of fluoride has greatly reduced tooth decay in the last few decades [13]. Systemic (ingested) fluoride therapies (e.g., water fluoridation) and topical fluoride therapies (e.g., fluoride toothpaste) are in common use throughout the world, either alone or in combination. The use of topically applied fluoride products, which are much more concentrated than the fluoride in drinking water, has increased over recent decades, and fluoride-containing toothpastes (dentifrices), mouth rinses, gels, and varnishes are the modalities most widely used at present. Toothpastes are, by far, the most widespread form of fluoride in use, and although the reasons for the decline in the prevalence of dental caries in children from different countries continues to be debated, it has been mainly attributed to the gradual increase in, and regular home use of, fluoride in toothpaste. Toothpastes are typically used daily at home, whereas some modalities (in particular, varnishes and gels) are typically applied infrequently in dental offices, schools, or community-based dental services.

It remains to be demonstrated what the most effective topical fluoride modality is. Although evidence is based on a large number of trials carried out since the 1950s for some forms of topical fluorides, there is more uncertainty about the effectiveness of other forms. In addition, as it is usually the case in health care, indirect evidence (topical fluoride comparisons with placebo or no-treatment controls) is much more abundant than head-to-head evidence directly comparing one topical fluoride with another (available from a much smaller number of trials).

A series of six systematic reviews has been undertaken under the auspices of the Cochrane Oral Health Group to summarize a large body of experimental evidence on the effectiveness of the main topical fluoride modalities. The reviews are based on a single comprehensive search for published and unpublished studies and collate the evidence using a common methodology and measures of effect. The first four reviews investigate the efficacy of fluoride toothpastes (T), gels (G), varnishes (V), and rinses (R), respectively, using placebo controls (P) or no-treatment controls (N) [14–17]. A fifth review brings together the results of these four reviews with investigations of differences in effectiveness between fluoride modalities based on classic meta-regression analyses using the treatments as covariates [18]. A further review collates trials of head-to-head comparisons among the four treatments [19].

In total, 130 studies are available, of which, 121 are head-to-head comparisons and nine compare more than two interventions (eight have three arms and one has four), and these are illustrated in Figure 1.

The main outcome is caries increment, as measured by the change in decayed, missing, and filled tooth surfaces (DMFS) in the permanent dentition of children. To compare DMFS increment in two groups, an effect measure commonly used in dental research, including the original Cochrane reviews, is the prevented fraction (the difference in mean caries increments between the treatment and control groups divided by the mean increment in the control group). For this article, however, the effect measure of interest will be the standardized mean difference (SMD). Mean caries increments are closely related to their standard deviations (they are about the same), and meta-analyses using SMDs will yield materially similar results to those using prevented fractions. For two treatments A and B, if $\text{SMD}_{AB} < 0$, then the caries increment on treatment A is lesser than on B, and hence, A is more effective than B. Table 1 shows the available comparisons.

3. The multiple-treatments meta-analysis model

Consider a study $i$ that compares toothpaste (T) with rinse (R). The estimated treatment effect in this study is the SMD (toothpaste - rinse), denoted by $y_{TR,i}$, with estimated variance, $\hat{\sigma}^2_{TR,i}$. The estimates are assumed to be normally distributed around the true SMD, $\delta_{TR,i}$:

$$y_{TR,i} \sim N\left(\delta_{TR,i}, \hat{\sigma}^2_{TR,i}\right).$$

Given multiple studies of toothpaste vs. rinse, classical meta-analysis models assume either $\delta_{TR,i} = \delta_{TR}$ for
a fixed-effect model, or if the differences between the arms (e.g., toothpaste, rinse, and placebo) are not expected to be heterogeneous, we have analyses for each comparison and use a fixed-effects model. The variance parameter $\tau^2_{TR}$ describes the heterogeneity of (true) SMDs across studies, and is specific to treatment comparison TR.

Suppose that, in addition to having $N_{TR}$ studies comparing toothpaste with rinse, we have $N_{RP}$ studies comparing rinse with placebo (P) and $N_{TP}$ studies comparing toothpaste with placebo. We could conduct separate meta-analyses for each comparison and obtain estimates of $\delta_{TR}$, $\delta_{RP}$, and $\delta_{TP}$ (and, in a random-effects model, of heterogeneity variances $\tau^2_{TR}$, $\tau^2_{RP}$, and $\tau^2_{TP}$). The main assumption underlying multiple-treatments meta-analysis is that we can learn about the comparison $\delta_{TR}$ by combining estimates through a third treatment (here, the placebo). This is equivalent to assuming that the trials are ‘exchangeable’ in their populations, outcomes, design, and conduct [3] or, in other words, that they are sufficiently similar regarding all characteristics other than the comparison(s) being made [5]. This assumption will hold in the absence of a certain type of confounding, as we discuss later. Under the assumption, it holds that

$$\delta_{TR} = \delta_{TP} - \delta_{RP}. \quad (1)$$

This equation implies that combination of any two parameters provides information about the third one. Therefore, using Equation (1), the degrees of freedom for the three-treatment network of studies drop from three to two, resulting in an increase in power.

The idea may be extended to any number of treatment comparisons. When many treatments are available and different subsets of them are reported in different studies, combining direct and indirect evidence in a joint analysis of the network may be expected to increase the precision of the estimates.

Information about the three effect sizes in a three-treatment network might not come solely from mutually exclusive studies. A single study may have three treatment arms (e.g., toothpaste, rinse, and placebo), and contributes information on more than one parameter. For such a study, only two treatment effect estimates are required for the analysis, because it holds (without assumptions) that $\gamma_{TR,i} - \gamma_{RP,i}$. However, these two estimates are correlated, because they use a common group of individuals as a comparator. Selecting the two placebo-controlled comparisons, $\gamma_{TP,i}$ and $\gamma_{RP,i}$, we, therefore, assume that they follow a bivariate normal distribution with sample covariance $c$, which can be estimated [1]:

$$\begin{pmatrix} \gamma_{TP,i} \\ \gamma_{RP,i} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \delta_{TP,i} \\ \delta_{RP,i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{TP,i} & c \\ c & \sigma^2_{RP,i} \end{pmatrix} \right),$$

In a random-effects model, the study-specific, comparison-specific effects, $\delta_{TP,i}$, $\delta_{RP,i}$, are assumed to be drawn from a bivariate normal distribution:

$$\begin{pmatrix} \delta_{TP,i} \\ \delta_{RP,i} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \delta_{TP} \\ \delta_{RP} \end{pmatrix}, \begin{pmatrix} \tau^2_{TP} & \rho \\ \rho & \tau^2_{RP} \end{pmatrix} \right), \quad (2)$$

where $\rho$ is the covariance of the random effects and $\tau^2_{TP}$, $\tau^2_{RP}$ are comparison-specific heterogeneity variances. For simplicity, it is often assumed that all comparisons have the same degree of heterogeneity, that is $\tau^2_{TP} = \tau^2_{RP} = \tau^2$ [1,6]. It then follows that $\rho = 0.5\tau^2$. The model may easily be extended to incorporate trials with more than three arms by increasing the dimension; to model a trial with $T$ arms, the distribution (2) will become a ($T-1$)-dimensional multivariate normal distribution.

### 4. Confounding and underlying assumptions of the network analysis

Joint analysis of the data in a multiple-treatments meta-analysis framework allows novel inferences on treatment comparisons that have not been addressed directly in any studies, and it increases precision for comparisons with few data. However, such gains do not come without strong assumptions. The validity of Equation (1) depends critically on there being no substantive differences between the sources of evidence that inform $\delta_{TR}$, $\delta_{TP}$, and $\delta_{RP}$. Consider a multiple-treatments meta-analysis of data from just three ‘clusters’ of trials: trials of toothpaste vs. rinse, trials of rinse vs. placebo, and trials of toothpaste vs. placebo. Although each individual trial may be randomized and internally valid, there will almost always be trial-level characteristics that can alter the relative effectiveness of the treatments being compared. If these characteristics differ systematically across clusters, then there will be confounding. In particular, evidence on toothpaste vs. placebo will be different from evidence on toothpaste vs. rinse in ways other than the choice of treatment comparisons. Therefore, it will not be possible to distinguish the effects of these characteristics from the effect of the difference between placebo and rinse. Equation (1) reflects an assumption of...
transitivity from indirect to direct evidence. In a recent study, several examples have been given where such transitivity does not hold [20].

Also challenging is the assumption of the ‘common comparator’ treatment, and the transitivity in comparison that this entails. For instance, a common treatment (in our example, placebo) may work differently in toothpaste trials than in rinse trials. The ubiquitous use of fluoride toothpaste makes it difficult to distinguish whether the effect of tooth brushing on caries is mainly a measure of fluoride application, because mechanical plaque removal may also play a role [21]. Traditional reviews of the literature have generally concluded that the effect of oral cleanliness per se on caries is equivocal [22,23], and although experimental evidence before the widespread availability of fluoride toothpaste is available, indicating no effect of oral hygiene alone on caries reduction [24], no systematic review has yet been carried out on this topic. Figure 2 shows a graphical representation of this scenario. In such a situation, combining placebo-controlled trials to learn about toothpaste vs. rinse may yield erroneous results.

It is important to consider both quantitatively and qualitatively whether a joint analysis of the trials could be invalid. This may be achieved for any particular chain of treatment comparisons by: (1) estimating the disagreement between direct and indirect evidence, as outlined in the following section; (2) considering whether treatments that play the role of a ‘common comparator’ (such as placebo or a standard therapy) do really work in a similar way across comparisons, irrespective of the treatments with which they are compared; and (3) examining the distribution of study characteristics that are possible confounders (characteristics that affect treatment effects and that differ across clusters of comparisons).

5. Estimation of incoherence

We now discuss some ‘signals’ of the violation of the network assumptions and some ways to address them. We use the term coherence to describe the presence of agreement between direct and indirect evidence, and incoherence for the converse. For example, suppose we have evidence on the direct comparison δ_{TR} from a meta-analysis of all trials of head-to-head comparisons of toothpaste and rinse and evidence on the indirect comparison δ_{TR} = δ_{TR} - δ_{RP} from the difference between meta-analyses of placebo-controlled trials. Then, the quantity \( \Phi = \delta_{TR} - \delta_{TR} \) expresses the incoherence between the two sources of evidence. In the absence of multi-arm trials, we can estimate \( \Phi \) as \( \Phi = \delta_{TR} - \delta_{TR} \), and its variance as \( \text{var}(\Phi) = \text{var}(\delta_{TR}) + \text{var}(\delta_{TR}) \). The value \( \Phi = 0 \) represents coherence between direct and indirect evidence. This idea can be extended for every closed ‘loop’ of three or more treatments formed in the network of treatments in Figure 1.

6. Implementation

We implement the network model using Bayesian methods in WinBUGS [27], mainly because of the natural way in which full uncertainty in all model parameters can be accounted for. A particular advantage of using a Bayesian framework, however, is the straightforward ability to rank the treatments by calculating the probability that each intervention has the largest treatment effect. Because our use of a Bayesian framework is for convenience rather than because we wish to incorporate prior uncertainty, we use flat normal distributions for (approximately noninformative) prior distributions for parameters that represent means, and flat half-normal prior distributions for standard deviation parameters. We programmed a routine for R (R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org), which calculates all the ‘triangles’ formed by three treatments linked with trials for any given network and estimates the incoherence (available from the first author on request). Because estimates of incoherence are straightforward to obtain, we adopted a frequentist approach for this analysis; a Bayesian equivalent would also be possible.

7. Analysis of the fluoride data

7.1. Pairwise comparison and multiple-treatments meta-analysis for the fluoride data

We first analysed each comparison separately. There are 150 possible comparisons: 121 from the two arm trials, 24 from the eight three-arm trials, and five from the four-armed trials (toothpaste and placebo are not to be compared with no treatment). The SMDs and median heterogeneity standard deviations (specific to each comparison) are given in Table 2.

We then analysed jointly all 140 independent comparisons from the 130 trials in a multiple-treatments meta-analysis: 121 from the two arm trials, 16 from the eight three-arm trials, and three from the four-armed trial. We assumed a common heterogeneity standard deviation for each pairwise comparison. The first column of Table 3 shows the SMDs, using as reference the ‘no treatment’ category. The second column shows the probability that each treatment is most
effective. The estimate of the common heterogeneity, median $\tau$, was 0.18 with 95% credible interval (CrI) of 0.15–0.22.

Toothpaste appears to be the most effective among all fluoride interventions, followed by varnish.

7.2. Incoherence in the network

We estimated the incoherence parameter, $\Phi$, between direct and indirect evidence in each of the 16 possible loops of the network in Figure 1, and the results are presented in Figure 3. Note that the direction of incoherence is arbitrary; here, we defined $\Phi$ for each loop so that the point estimate is positive.

Examination of the confidence intervals reveals that the coherence assumption cannot be rejected using standard level of statistical significance ($\alpha=0.05$), although such analyses do not have high power. None of the loops yields a large value for the incoherence statistic, $\Phi$. However, further investigation is needed before concluding that the network is coherent, because absence of statistical evidence of incoherence does not necessarily imply that there is no confounding in the network. For this reason, it has been suggested that acceptance or rejection of the incoherence assumption should be based on epidemiological and clinical issues rather than on statistical tests [25]. In the following sections, we look at two issues that, from an epidemiological point of view, may lead to violation of the coherence assumption.

7.3. Evaluation of the underlying assumptions

7.3.1. Different placebo effects

The results in Table 3 suggest that placebo has a marginal effect compared with no treatment. If placebo has an effect, then this effect may be different for the placebo versions of the four modalities. If this is the case, then the transitivity assumption is violated. To investigate this assumption, we analyse the data under various assumptions about differential effects across placebo modalities. Thus far, we have assumed that placebo forms of toothpaste, gel, rinse, and varnish are equivalent (or exchangeable). Note that the assumption is not that the absolute caries increments are similar across placebo groups, but that, in any particular trial, a placebo form of one modality could be replaced with one of a different modality. The least restrictive assumption is that all placebos are different, and we denote these as P$_{t}$, P$_{g}$, P$_{r}$, and P$_{v}$. Progressively more restrictive models may be obtained by grouping forms that are similar in terms of application and mechanical function (Table 4). An even more restrictive model than the one we have assumed thus far is that all placebos are equivalent to each other and equivalent to no treatment. Table 4 presents some results from network analyses under five specific models, including the deviance information criterion (DIC) [26].

The DIC is similar to the Akaike’s information criterion for frequentist analysis; the lowest DIC corresponds to the best fitting model [27]. As a rule of thumb, a decrease of three units in DIC indicates a significantly better model. Distinguishing between different placebo modalities does not appear to yield better models or decrease the heterogeneity standard deviation. Furthermore, the model that distinguishes between placebo and no treatment has a marginally preferable DIC than the one that treats no treatment and placebo as equivalent. We conclude that our existing model (number 4) is the most appropriate among this selection.

7.3.2. Possible confounders

We now evaluate the assumption of exchangeability of the sources of evidence by comparing study-level

### Table 2
Results from pairwise meta-analyses of the fluoride trials

<table>
<thead>
<tr>
<th>Comparison (no. of studies)</th>
<th>Mean SMD</th>
<th>95% CrI</th>
<th>Median heterogeneity $\tau$</th>
<th>95% CrI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP (69)</td>
<td>-0.34</td>
<td>-0.41 to -0.28</td>
<td>0.25</td>
<td>0.19 to 0.32</td>
</tr>
<tr>
<td>GP (13)</td>
<td>-0.19</td>
<td>-0.30 to -0.10</td>
<td>0.12</td>
<td>0.02 to 0.25</td>
</tr>
<tr>
<td>RP (31)</td>
<td>-0.29</td>
<td>-0.36 to -0.23</td>
<td>0.16</td>
<td>0.11 to 0.22</td>
</tr>
<tr>
<td>VP (3)</td>
<td>-0.29</td>
<td>-0.13 to -0.29</td>
<td>0.23</td>
<td>0.01 to 0.23</td>
</tr>
<tr>
<td>GN (9)</td>
<td>-0.46</td>
<td>-0.68 to -0.24</td>
<td>0.27</td>
<td>0.15 to 0.54</td>
</tr>
<tr>
<td>RN (4)</td>
<td>-0.46</td>
<td>-0.70 to -0.17</td>
<td>0.11</td>
<td>0.00 to 0.64</td>
</tr>
<tr>
<td>VN (4)</td>
<td>-0.67</td>
<td>-1.33 to 0.01</td>
<td>0.49</td>
<td>0.17 to 1.40</td>
</tr>
<tr>
<td>TR (6)</td>
<td>-0.10</td>
<td>-0.46 to -0.25</td>
<td>0.34</td>
<td>0.16 to 0.84</td>
</tr>
<tr>
<td>RV (4)</td>
<td>-0.11</td>
<td>-0.64 to 0.37</td>
<td>0.33</td>
<td>0.07 to 1.12</td>
</tr>
<tr>
<td>TG (3)</td>
<td>0.04</td>
<td>-0.58 to 0.70</td>
<td>0.27</td>
<td>0.01 to 1.30</td>
</tr>
<tr>
<td>TV (1)</td>
<td>0.06</td>
<td>-0.23 to 0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GV (1)</td>
<td>-0.12</td>
<td>-0.37 to 0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GR (1)</td>
<td>0.13</td>
<td>-0.11 to 0.38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** CrI, credible interval; T, toothpaste; G, gel; R, rinse; V, varnish; P, placebo; N, no treatment.

### Table 3
Results from multiple-treatments meta-analysis of the fluoride trials (without covariates, adjustment for year of randomization and baseline risk, using meta-regression for trials with placebo and no-treatment controls) with ranking of the treatments. The third, fifth, and seventh columns show the probability that each treatment is the most effective one.

<table>
<thead>
<tr>
<th>Fluoride treatments</th>
<th>Year of randomization adjusted to 1994 values</th>
<th>Baseline mean caries level</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adjustment</td>
<td>Mean SMD (95% CrI)</td>
<td>$P$ (best (%))</td>
</tr>
<tr>
<td>No treatment</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Placebo</td>
<td>-0.22 (−0.34, −0.09)</td>
<td>0</td>
</tr>
<tr>
<td>Toothpaste</td>
<td>-0.54 (−0.67, −0.40)</td>
<td>57</td>
</tr>
<tr>
<td>Gel</td>
<td>-0.45 (−0.58, −0.34)</td>
<td>4</td>
</tr>
<tr>
<td>Rinse</td>
<td>-0.50 (−0.63, −0.37)</td>
<td>14</td>
</tr>
<tr>
<td>Varnish</td>
<td>-0.50 (−0.65, −0.34)</td>
<td>25</td>
</tr>
</tbody>
</table>
characteristics across treatment comparisons. Four parameters are particularly susceptible to variation across the treatment comparisons: the length of follow-up, the baseline caries level, the year of randomization, and whether or not the local water supply is fluoridated. Figure 4 illustrates the distribution of these possible confounding factors across studies including each of the four fluoride modalities.

Year of randomization varies notably across comparisons. In particular, placebo-controlled trials of toothpaste, which represent more than half of the network data, include many old studies dating back to the establishment of fluoride as an agent for preventing caries. A dependence of treatment effects on the year of randomization may reflect genuine changes in the effectiveness of fluoride over time, although it may also reflect changes in research quality over time (e.g., adequacy of allocation concealment and blinding). There is a moderate correlation between year of randomization and baseline mean caries levels (correlation coefficient $= 0.35$, $p = 0.005$), perhaps not surprising, given the change in dietary habits and oral hygiene practices between 1954 (the oldest study in our data) and 1994 (the newest study).

Within the toothpaste vs. placebo trials, subgrouping according to whether randomization occurred before or after the median randomization’s year (1968) gives pooled estimates of $-0.36$ ($-0.44$, $-0.28$) ($36$ studies) and $-0.29$ ($-0.35$, $-0.23$) ($33$ studies), respectively. There is considerable heterogeneity in both subgroups. Because toothpaste trials tended to be earlier, and earlier trials had more exaggerated observed effects, the use of these trials in the indirect estimation of other treatment comparisons may introduce bias. In particular, the effects of treatments studied more recently, such as varnish, may be diluted. Such potential confounding can be partly overcome using meta-regression techniques.

### 7.3.3. Meta-regression

We implement a meta-regression model that ‘adjusts’ for time trends in placebo-controlled comparisons. The underlying assumption of the analysis, characterized in Figure 5, is that, within any specific population, the relative effect of any fluoride treatment compared with placebo or no treatment is time-dependent.

Head-to-head comparisons of placebo against no treatment, and among the four fluoride modalities are assumed to be time-invariant. Thus, for example, a comparison of gel vs. placebo in 1985 (denoted by a in Figure 5) may give a similar treatment effect as a comparison of rinse with placebo in 1970 (b), producing a biased underestimate of the difference between gel and rinse. However, an indirect estimate that adjusts for year will give the correct result (c).

Specifically, for each comparison of a fluoride arm ($F$) with a control (placebo or no treatment) arm ($C$), a meta-regression model allows the treatment—placebo difference to depend on year as follows:

$$
\delta_{FC,i} = \delta'_{FC,j} + \beta(Year_i - Year_0)
$$

Here, ‘Year,$i$’ is the year of randomization in study $i$; $Year_0$ is some convenient choice of year to aid interpretation; $\delta'_{FC,j}$ is the treatment effect comparing the fluoride with control in year $Year_0$; and $\beta$ describes the dependence of the fluoride-control treatment effect on year, assumed to be the same across all fluoride modalities. This last assumption corresponds to the presence of similar time-related changes on placebo or no-treatment outcomes, and similar time-related changes across the four fluoride modalities. No regression relationship is modeled for the seven head-to-head comparisons. We set as $Year_0$ the year of randomization of the most recent study (1994). Then, the ranking of the treatments is based on the estimates $\delta'_{FC}$, adjusted to 1994. The estimated coefficient $\beta$ is $0.005$ ($0.001$, $0.009$), adjusted to 1994.

### Table 4

<table>
<thead>
<tr>
<th>Model</th>
<th>Placebo assumption</th>
<th>Median $\tau$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All placebos different</td>
<td>$P_T$, $P_G$, $P_R$, $P_V$</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>Gel and varnish placebos equivalent</td>
<td>$P_T$, $P_G$, $P_R = P_V$</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>Gel, varnish, and rinse placebos equivalent</td>
<td>$P_T = P_G = P_R = P_V$</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>All placebos equivalent</td>
<td>$P_T = P_G = P_R = P_V$</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>All placebos equivalent to no treatment</td>
<td>$N = P_T = P_G = P_R = P_V$</td>
<td>0.19</td>
</tr>
</tbody>
</table>
indicating that older studies gave more enthusiastic results for the effectiveness of the fluoride. Table 3 presents the effect sizes adjusted to 1994 values, along with a revised ranking of treatments by effectiveness. The heterogeneity median standard deviation is unchanged ($\tau = 0.18 \ (0.15, 0.22)$). By taking account of the older age of many of the toothpaste trials, the overall effect of toothpaste is reduced along with its probability of being the most effective treatment. Newer treatments, such as varnish, are pushed substantially up the ranking. However, the DIC value was $-80.5$, indicating that the more complex model does not fit the data better than the one without adjustment.

We repeated the analysis to adjust for baseline caries level, excluding the 14 studies with missing data on this covariate (most of them were older studies; median year of randomization = 1973). The estimated regression coefficient is $\beta = 0.020 \ (0.012, 0.024)$, indicating a small association with baseline caries (Table 3). Again, the ranking of newer treatments is improving in detriment to toothpaste. The DIC and the heterogeneity standard deviation are slightly smaller than those from the meta-regression using year of randomization ($\text{DIC} = -81.7, \tau = 0.16 \ (0.13, 0.20)$).

We did not detect any important association between fluoride treatment effects and either length of follow-up or the fluoridation of water.

### 8. Conclusions and limitations

We have analyzed simultaneously a complex network of clinical trials involving four fluoride modalities and two control interventions with the aim of determining the most effective intervention. The validity of our conclusions from an initial network analysis, indicating superiority of toothpaste, was challenged when we adjusted for possible confounders. Studies supporting the effectiveness of toothpaste were older, and have been carried out in populations with higher baseline risk compared with studies comparing other fluoride modalities with placebo or no treatment. We took these differences into account in a meta-regression model, which impacted on both the heterogeneity and the incoherence. Failure to take these differences into account may lead to an overestimation of the effect of toothpaste. The case study serves as a lesson in the dangers of naïve network meta-analyses, and also of the danger of naïve interpretations of indirect comparisons that do not account for potential confounders.

It is unlikely that we have included all important covariates. The studies included in the network differ also in the frequencies, modes of administration, and

---

Fig. 4. Distribution of the possible confounding factors: box plots for year of randomization, length of follow-up, and baseline mean caries; percentage of the studies carried out in populations with fluoridation in the water.

Fig. 5. Assumptions of a meta-regression analysis in which fluoride-control differences change over time (e.g., because of improved oral hygiene; assumed to affect placebo and no-treatment groups). The scenario relates to a specific population under study, and is not assumed to apply across individuals in different studies.
concentrations of fluoride. Such differences may be associated with the effectiveness of the treatment producing heterogeneity within each pairwise comparison, and may also introduce incoherence if they differ broadly across comparisons, challenging the validity of the network analysis. For example, if placebo-controlled studies of varnish use low-concentration fluoride varnishes, and studies comparing varnish with toothpaste use high-concentration fluoride varnishes, then confounding (and hence incoherence) may occur.

To adjust for possible time and baseline-related differences, we conducted a meta-regression analysis under the assumption of a common (fixed) regression coefficient for all fluoride modalities. This assumption, although plausible, may not be valid in practice. An extension of the meta-regression model (3) allowing the coefficient \( \beta_{IC} \) to be comparison-specific may provide a better fit to the data. Furthermore, given that trial-level characteristics differ across comparisons, the assumption of a common heterogeneity parameter \( \tau \) may not be appropriate. Relaxing this assumption would be at the expense of introducing more parameters, for which, there are few studies providing information.

Multiple-treatments meta-analysis offers many opportunities, including the abilities to enhance precision, to estimate treatment effects that have not been observed directly, and to rank treatments while fully exploiting randomization. In addition to the standard assumptions underlying a conventional meta-analysis (e.g., the studies are similar enough to synthesize), the conventional multiple-treatments meta-analysis model is problematic in the presence of incoherence or when important trial characteristics differ across comparisons. A discussion of the use of different models for different circumstances can be found in Salanti et al. [2]. Many obstacles and challenges remain to be tackled, however. The results of such an analysis may not be valid if the underlying assumptions are violated, and adjusting for confounders to eliminate incoherence will not always be possible. There is a need for further case studies and methodological development, both from statistical and epidemiological points of view, before network analyses can be widely advocated.

References