2 January 2014

Health Freaks on Trial

The call of television

In the early spring of 2013, I received a phone call from a researcher. I am often called by researchers, but this one was different: not a health or education researcher, but a researcher from a television production company. He did want to talk about randomised clinical trials, however. His company, Outline Productions (www.outlineproductions.co.uk), were thinking about doing a series of programmes about home remedies, unconventional treatments for everyday conditions. An example that captured my imagination was to treat male baldness by rubbing into the scalp bull's semen. I never found out whether the recipient had to collect this himself. The idea was to carry out a series of small, short duration, randomised control trials of some of these remedies. He was looking for advice on their design and analysis. I am always keen to communicate what we do to a wider audience and this seemed like a good opportunity. Anything which shows clinical trials in a positive light would be welcome. Too often, I think, clinical trials on TV appear as rather murky affairs where some fictional doctor is fiddling the results or trying to conceal terrible things happening to participants. A few short, entertaining trials where nothing bad happens might stimulate recruitment in the future. And I fancied being on television, though this, it turned out, was not on offer. However, I explained how I would do such trials and told him that, if he had 10 participants in each group, if all the people in one group did better than all the people in the other group, that would be statistically significant and no detailed calculation would be needed. A simple test anybody could understand. He told me that this was exactly what the proponents of these remedies claimed would happen. After several conversations, he told me that he was leaving Outline and moving to another job.

A few weeks later a new researcher called. I told Chloe Seddon what I had told her predecessor and sent her what I had sent him. After a couple of conversations with her, I spoke with her producer, Cat Gale, who left after the pilot programme. TV is a high-turnover world. I signed a contract to help with the design and analysis of a trial for a pilot programme.

The first trial: oatmeal baths

To my disappointment, the pilot trial was not of bull's semen but of oatmeal baths for the treatment of the skin condition psoriasis. The production team recruited some volunteer psoriasis sufferers. The team were concerned, as are so many researchers, that if I randomly allocated the participants to groups, the oatmeal and control groups would, by chance, be very different in severity or some other important variable. I decided not to argue with them, but offered to use a minimisation algorithm to avoid this. In due course, a list of participant IDs arrived, with their ages, genders, and severity of psoriasis classified as mild, moderate, or severe. I allocated them to groups using minimisation and sent back the list. Some of the first group pulled out before trying the treatment; a second group were recruited and I allocated them so as to minimise the final differences between the groups.

The actual treatment is to put a cupful of oatmeal into a porous bag and run the bathwater through it, then the psoriasis sufferer soaks in the bath. The control group had a placebo bag containing rolled up tights. Data were collected before and after five days of baths. In due course the data arrived and I carried out a simple analysis. Of the 24 participants allocated, only 18 completed the trial, 9 in each group. The original balance was not preserved, but it did not turn out very badly and the psoriasis severity was exactly balanced. The main

outcome variable was a questionnaire scale, the Dermatology Life Quality Index (DLQI), which is scored from 0, no effect at all on patient's life, to 30, extremely large effect on patient's life. The results are shown in Figure 1. The oatmeal group had slightly higher scores at the start and there was a lot of overlap between the groups at the end, so no obvious advantage to the oatmeal treatment and nothing using my simple test. The analysis of covariance produced the estimated effect that the Oatmeal group have a higher mean DQLI score by 0.9 and we estimate that for a very large group of participants we would get a difference, Oatmeal minus Control, between −2.5 and +4.4 points on the scale (95% confidence interval), the difference not being significant, P = 0.9. We also had daily scores for four things: skin less inflamed, less itchy, condition improved, and negative effect on life, each scored 1 (best) to 5 (worst). I decided that the way to analyse these would be to find the average score for days 2 to 5, day 1 being the baseline. None of these showed any evidence of a difference between groups. There was also an objective measurement using the ANTERA 3D optical device, which gave us measurements of the average skin redness, variability in redness, depth of lesions, and skin texture, and a total measure. None of these gave significant differences. After I sent Outline my initial analyses, they asked me for two more: a comparison of the percentage change from baseline for the ANTERA 3D and a comparison of final versus baseline for everybody. I pointed out that comparisons of changes from baseline are inefficient compared to the analyses I had done, because the difference contains two lots of measurement error, and we had already done a lot of analyses. I did it anyway and to my profound disappoint the last one I did produced a P = 0.04. However, to my delight, the producer said that she accepted my argument that after 14 different tests comparing the groups, one just making significant did not mean anything and they would not seize on this to report success.

It is noticeable in Figure 1 that the DLQI fell from the start to end, most points being to the right of and below the line of equality. The fall in DLQI from start to end was statistically significant (P=0.02). I did not do this as part of my original analysis, because in trials like this symptom scores often fall from the baseline. This may be because of placebo effects or because of regression towards the mean, where we select people who agree to take part because their symptoms are particularly bad and just by chance then tend to fall. This is why a control group or a crossover trial is much better than a before and after study. It is consistent with bathing helping. It just doesn't prove it.

I didn’t see the pilot, but I did see the final programme when broadcast as the third in the series. This is what the programme looked like. It started with a series of short clips of suggested remedies, inverting the body in a shoulder stand to prevent hair loss, toothpaste for love bites, turmeric and full fat yoghurt, rubbing chillies onto feet and toes, both for problems unspecified at that point, and unrelated reactions of three GPs, primary care physicians, who were then introduced as series presenters. These clips were repeated throughout the series. After a title sequence, the series was introduced by GP Dr. Pixie McKenna, along with her fellow GPs Drs. Ellie Cannon and Ayan Panja. We then had the first advocate of a remedy, for oatmeal baths for psoriasis, and a debate between the three GPs, with two in favour of and one against the trial, and a decision to trial. There was also a brief introduction of three trials experts, which I shall come back to. There followed a series of advocates for other remedies: topical vinegar for athletes’ foot, vinegar for sunburn, and eating tree bark for Crohn’s disease. Then viewers were invited to send in their opinions on the oatmeal cure, whether they had heard of it and whether they though it would work. After a commercial break we had an advocate of sexual intercourse for treating migraine, who looked really cheerful. It was nice to see a headache as a reason for having sex rather than for avoiding it. We then had presentations of soya for menopausal flushes and hot pan-fried *Aloe vera* rubbed on the

back for a chest infection. Then we moved to the result of the trial. We had the results of the online poll: 44% of respondents had heard of oatmeal baths for psoriasis and 82% reckoned it would work. I would be amazed if this were a representative population sample. (Similar results were found for every treatment they tested.) Then the oatmeal advocate came in to hear the results of the trial. Dr Pixie McKenna told her, and us, that both groups reported an improvement in their skin. Using the 3D scanner, over the course of treatment the average score for the group that used the oats had lower levels of psoriasis. The following flashed up on the screen:

OATS:	8.7% AVERAGE REDUCTION
NO OATS:	NO REDUCTION

Which was more or less true and this was the P=0.04 test I had not wanted to do. The conclusion was that there is something in the oats treatment and that a bigger trial would be justified, as recommended by the three experts. (Not by this one.) Interestingly, there was no mention of the Dermatology Life Quality Index. The programme ended with a comment that one study had revealed that having sex could relieve migraines, that although some medicines had been extracted from bark, you should not eat it direct from the tree, and trails for the next programme in the series.

The pilot programme was approved by Channel 4. The name was changed to *Health Freaks*. A series of six programmes was commissioned and I signed another contract.

Can duct tape shrink a verruca?

The first broadcast programme featured what had proved to be an amazing trial. This was a trial of duct tape used to treat verrucae, or warts. The advocate of this claimed to have completely cured a verruca on his foot by sticking a small piece of duct tape, opaque plastic adhesive tape also known as gaffer tape or by the brand name Duck Tape, over it for six days, then changing it and repeating. The remedy was tested in a randomised trial over a month. Two groups of 11 participants took part, using surgical tape as a control treatment and 10 duct tape and 7 control participants completed the trial.

The primary outcome variable was the diameter of the verruca. The measurements at baseline and after month are shown in Figure 2. It is very clear that participants who were on active Treatment all had a reduction in the size of the wart, while participants on Control who provided measurement data had very little change in wart size. If we carry out the standard analysis of regression of the end size on the treatment and the start size, the difference between duct tape and control is estimated to be 2.1 mm (95% confidence interval 1.3 to 2.9 mm). This difference was highly significant, $P<0.001$. We can also note that all ten of the Treatment participants had reductions in size greater than any of the Control participants. This would be very unlikely to happen if the Treatment had no effect, less than one in a thousand trials (actually $P = 0.0008$.) Amazingly, my suggestion of the simple test that everybody would understand worked out! All the 10 participants on Treatment had reductions greater than 1 mm and none of the 7 participants on Control did so. All the duct tape participants had a greater reduction than any of the control participants.

Participants were also asked about pain when the wart was squeezed. Results were very variable and, although the Duct Tape group appeared to do better, this was not significant. Two of the Duct Tape patients, and none of the Controls, reported that the verruca had disappeared, though this was not a significant difference. It was not all good news. Significantly more Duct Tape participants than Controls reported difficulty in applying the treatment and significantly more Duct Tape participants reported skin irritation.

I was very excited by these results, as I had never seen a trial where the results were as clear cut. I wanted to share them with the world, but I was disappointed with the way the results were presented on television. Of my wonderful graphs like Figure 2, my quick and easy significance test, my more powerful analysis, and the questionnaire data there was nothing. Only a piece of text appeared at the bottom of the screen, reading:

> DUCT TAPE — 100% showed reduction of 1mm or more
> SURGICAL TAPE — 0% showed reduction of 1mm.

and the lead medic said 'These are really powerful results'.

I did a literature search and found some small trials of duct tape for warts, with varying and debated results, which looked promising overall.[1] I didn't find any published trials for the other remedies we tested, though I'll admit that I didn't look very hard.

Oil pulling: an old remedy for mouth problems

Oil pulling is an old Asian remedy for gingivitis and other problems caused by mouth bacteria, where the user takes some vegetable oil, such as coconut oil, into the mouth and swills it around for a few minutes. Control participants used water instead. Counts of bacteria were recorded for total bacteria, aerobic bacteria, eight different species of bacteria, and plaque index.

Figure 3 shows the results for total bacteria count. Count fell dramatically in both groups, but there is no obvious difference between oil pulling and control. There were seven individual species of bacteria for which measurements were given, divided into four 'bad' bacteria and three 'good' bacteria, and a separate count for aerobic bacteria. The plaque index was an integer score, from 0 = no plaque to 6 = plaque covering 2/3 or more of the crown of the tooth. The bacterial counts all had highly skewed distributions so I analysed them on the logarithmic scale and finished up with a ratio of the mean count in the oil pulling count to the mean count in the control group, adjusted for their counts measured at the start. Table 1 shows the results for the bacterial counts. The individual counts were highly variable, as Figure 3 shows, and differences had to be pretty big to be statistically significant. None of them made it. Some ratios were less than one, some greater, so there was not much to suggest a consistent benefit for oil pulling compared with the control. There was a pretty big and consistent change from beginning to end, however, and most of the ratios of end count to start count were much less than 1.0 and significantly so. The plaque index did not need any logarithmic transformation and showed virtually no difference between the oil pulling and control treatments, difference = 0.01, $P = 1.0$. The change in mean plaque index over time was −0.22, a slight fall, which was not significant, $P = 0.06$.

The results were presented orally as 'In both groups all of our volunteers saw a massive reduction in the amount of bacteria they had in their mouths, down 80% of the levels that they started with, so it isn't showing a preference to the coconut oil or the distilled water.' This was not quite true, three participants showed very little change (Figure 3), but not far out. On screen flashed:

> BACTERIAL COUNT:
> OVERALL 80% REDUCTION

The presenter went on 'However, those who swilled with coconut oil had a slightly greater reduction in the bad bacteria, but not enough to say "That's the thing that made the difference."' 'Right', said the advocate, 'So it wasn't statistically significant'. 'Exactly', confirmed the presenter, going on 'and it also showed a reduction in plaque again showing no

favoured preference to those who were swilling with coconut oil versus those who were swilling with water.' And on screen flashed

> PLAQUE LEVELS:
> NO SIGNIFICANT REDUCTION

I was amused by the advocate bringing statistical analysis into it.

Yogurt and turmeric face masks are spot on

The last trial was of a yogurt and turmeric face mask for the treatment of acne. I was on holiday and as a result the allocation was not minimised or randomised. They did a trial where 11 participants each had the face mask on the right side of the face and the left side was untreated. (I would have flipped a coin for each participant to decide which side would get the mask.) There were three groups of variables: clinical observations were counts of inflamed lesions (spots) and non-inflamed lesions, and a global assessment, made before and after the treatment period. There were also questionnaire scales, made after treatment only, and measurements, before and after, of redness, texture, pore size, and porphyrins, made using a system called Visia. Porphyrins are a group of organic compounds which produce red or purple colours, one of which causes the red colour of blood.

Figure 4 shows the counts of inflamed lesions after treatment, right against left. If the lesions were reduced by the face mask, the points would be to the right of and below the line. Nine of the 11 are to the right, but the difference was not quite big enough for any test of significance that I could think of. We needed ten, not nine. Using the more powerful paired t test I got $P = 0.1$, not significant. For non-inflamed lesion counts, we get $P = 0.07$. I also summed the inflamed and non-inflamed lesion counts to give a total lesion count, which gave $P = 0.08$. Frustrating, isn't it? For the global assessment, there was very little difference, only two participants differing between the sides of the face and both of these being worse on the treated side. This difference was not significant, using the sign test, $P = 0.5$.

The Visia measurement system produced little indication of any difference in redness, texture, or pore size. It did produce the dramatic and highly significant ($P = 0.006$) difference in porphyrins (Figure 5). I had no idea whether a higher porphyrins level is a good or bad thing.

The questionnaire asked participants to rate on a scale of 1 to 5, separately for each side of the face, whether they felt that the skin was less inflamed or had improved in texture, whether they felt their acne had improved, and whether they felt less self-conscious about their acne. All four of these showed a significant benefit on the right, turmeric and yoghurt side compared to the left ($P = 0.02$, 0.004, 0.02, 0.03, respectively).

So in this trial there was one highly significant objective difference, porphyrins, support for the treatment from the questionnaires, and a non-significant difference in the right direction for the main outcome, lesion counts. I reported this as 'We can conclude that there is evidence for a benefit in using yoghurt and turmeric for acne.' The panel on the programme did not agree. They said 'The dermatologist's assessment [spot counts and global] didn't show any difference at all between the treated and the untreated side.' I thought this was a classic confusion between 'not significant' and 'no difference'. They accurately reported the Visia results and interpreted the porphyrins difference as may be making acne worse. They talked about an increase in bacteria, though we did not have any actual bacteria counts. The panel confirmed that the participants had said their skin felt better after the face mask, but the GPs were against it.

Two programmes made without me

This was a series of six programmes and two did not feature trials in which I had a hand. One of these reported a laboratory investigation of the possible antiseptic properties of human breast milk. This did not show any obvious benefits and did not get any statistical analysis. The other reported a study of copper insoles for arthritis, using just three participants and measurements of copper in the blood, gait, and self-reports. They didn't find any changes in blood copper or gate, but the three participants said they felt better. Even I would not have dignified that with statistical analysis. You can judge for yourself and see all the programmes on channel4.com/ healthfreaks.

My ten seconds of fame

The original deal with Outline was that I would be an off-screen advisor, with my name in the credits but my face unseen. Once the series was commissioned, the broadcaster, Channel 4 in the UK, wanted the advisors to be shown on screen. The director wanted me to wear a white lab coat for this, as if I were a wet-lab scientist. I objected strongly, as I did not need a white coat to protect me as I sat at the computer in my loft of an evening, analysing these data. My fellow statisticians would laugh at me. The wonderful Chloe argued my case successfully and it was agreed that I would appear as myself. A trip to London was arranged for me and my two fellow advisors, general practitioner Richard Albardiaz and clinical trial coordinator Rhiannon Pursall. My wife packed a suitcase of carefully ironed shirts and trousers and I set off. I was met at the station and taken to a studio, where I met the other advisors, Chloe, and others with whom I had had phone or email conversations. We three advisors were filmed sitting at a high table, talking about the trials, but warned we would probably be voiced over. I told our University press office, they got a still of this scene (Figure 6), where I look rather like a garden gnome, and in the week of first broadcast put it on the University main webpage. Then in two of the programmes the great scene was broadcast in a voiceover clip lasting 14 seconds.

Would I do it again? Yes! Despite the time pressures, everything being needed for the next day, I really enjoyed doing it. All the people I worked with from Outline were lovely and a pleasure to work with. My ten seconds of fame was seen by my great niece, Lucy, who though it was great to see her Uncle Martin on TV. I was disappointed by how little was made of the statistics and how little time was given to the trials, but not completely surprised. My one previous experience of television had been similar.[2] I thought the programmes did show clinical trials in a positive light, as I had hoped. I also thought that, despite the small sample sizes, short durations, and time pressures, the trials were done to a really good standard. So when Outline calls again, I shall be there.

3823 words.

1. Stubbings A, Wacogne I. (2011) What is the efficacy of duct tape as a treatment for verruca vulgaris? *Arch Dis Child* **96**, 897–899.

2. Bland M. (2005) The *Horizon* homeopathic dilution experiment. *Significance* **2**, 106-109.

Figures and Tables

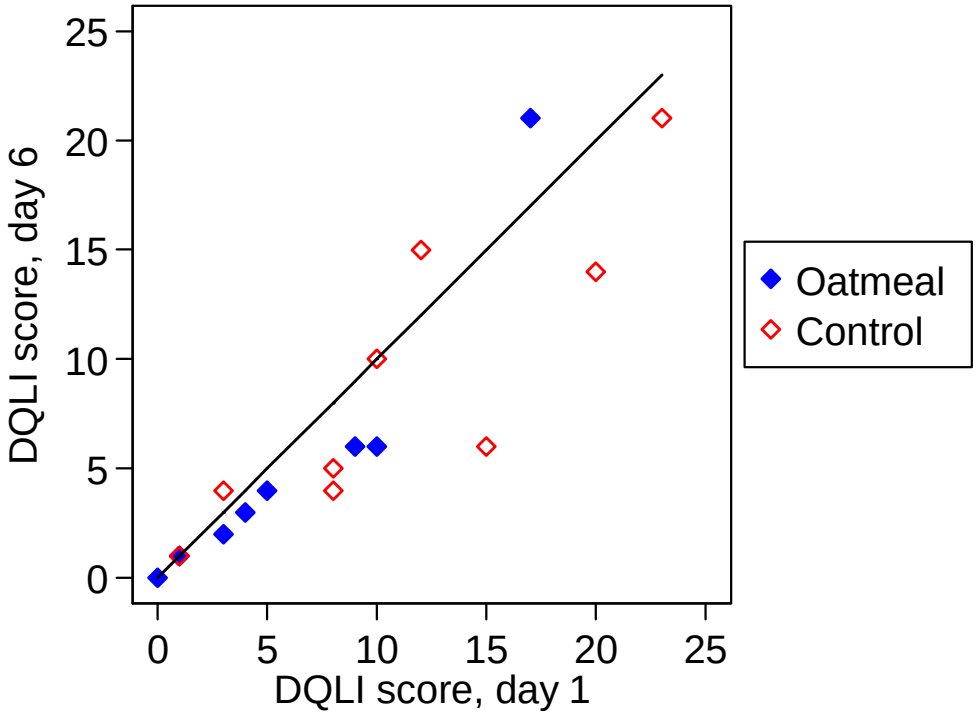

Figure 1. Dermatology Life Quality Index (DLQI) at start and end of the trial, with the line on which the points would lie if they were equal

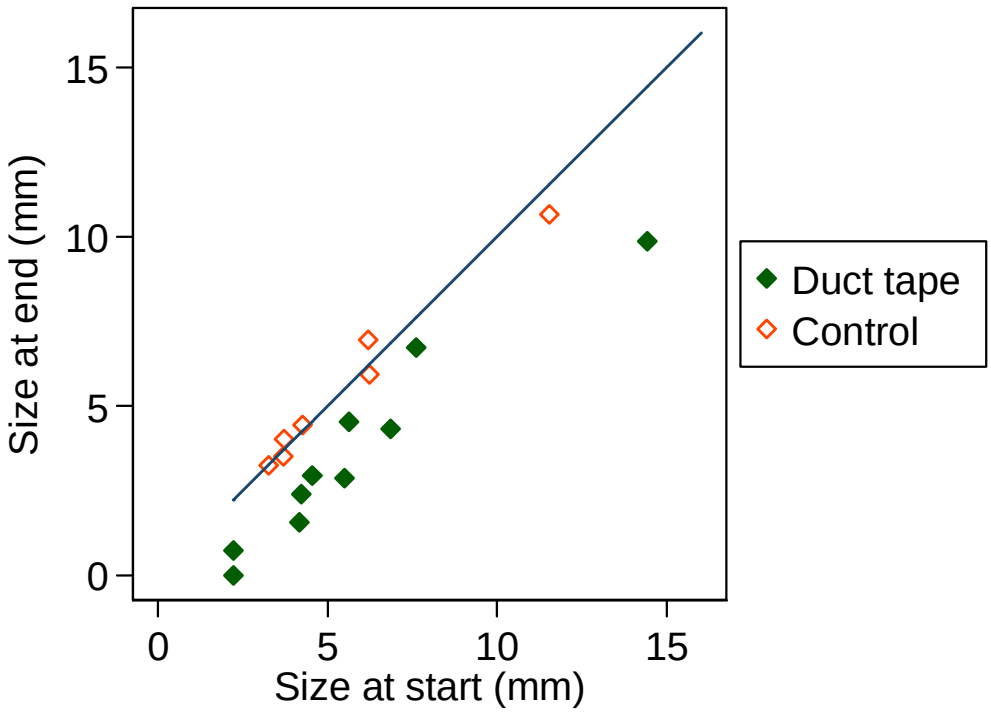

Figure 2. Diameter of the verruca after one month against diameter before treatment, with line of equality.

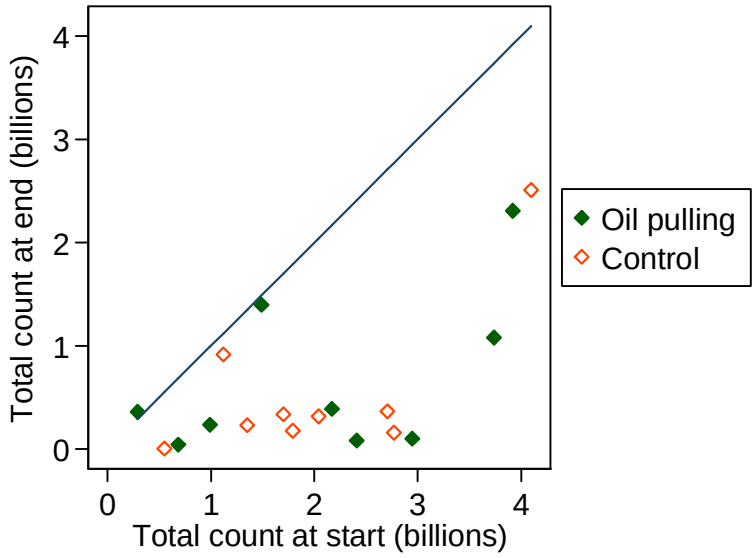

Figure 3. Total bacterial count after and before oil pulling or mouth rinsing with water, with the line of equality

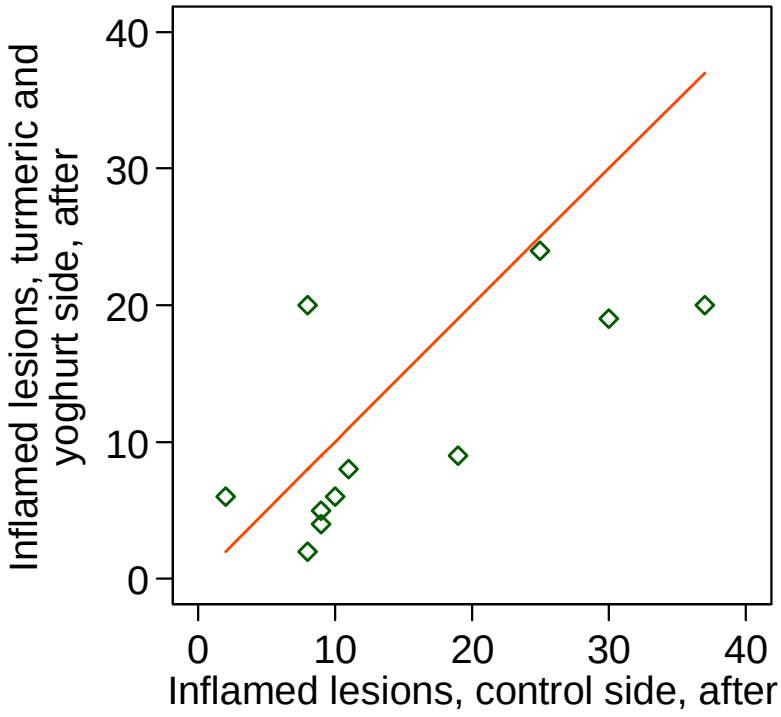

Figure 4. Inflamed lesion counts on either side of the face, after turmeric and yogurt treatment

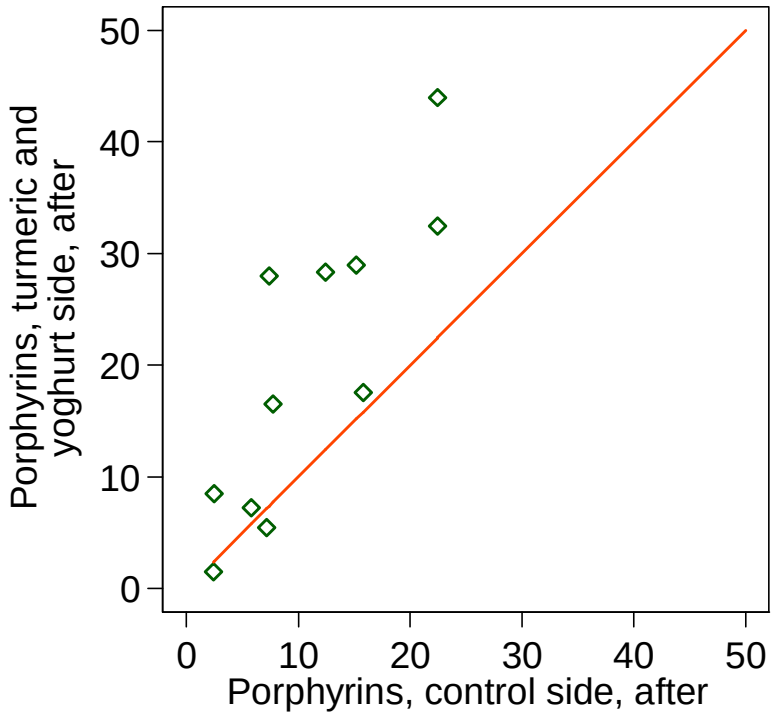

Figure 5. Porphyrins measured on each side of the face after turmeric and yoghurt treatment

Figure 6. Clinical trial advisors Rhiannon Pursall, Richard Albardiaz, and Martin Bland, looking tiny on a high stool

Table 1. Statistical analyses of oil pulling experiment

Bacterial count	Oil pulling/control		End/beginning	
	Ratio	P value	Ratio	P value
Total	1.44	0.6	0.17	<0.0001
Aerobic	0.94	0.5	1.22	0.4
Aerobic/Total ratio	0.67	0.6	7.07	<0.0001
Actinomycetes naeslundii	0.68	0.7	0.34	0.008
Prevotella intermedia	0.86	0.8	0.25	0.0002
Fusobacterium nucleatum	1.03	0.9	0.12	<0.0001
Lactobacillus casei	2.58	0.5	0.002	0.0001
Four ‘bad’ bacteria	0.29	0.3	0.13	0.02
Streptococcus sanguinis	0.48	0.5	1.05	0.8
Neisseria subflava	1.22	0.8	0.23	0.0006
Veillonella dispar	1.06	0.9	0.14	<0.0001
Three ‘good’ bacteria	0.90	0.8	0.18	<0.0001