

Cold Showers and Sickness Absence: A Comment on Buijze et al. (2016)

Joop Adema*

September 22, 2026

Abstract

Buijze et al. (2016) report that ending the daily shower with 30 to 90 seconds of cold water reduces days of sickness absence by 29 percent ($p = 0.003$). Using the trial's own data and specification, I show that this estimate rests on a handful of very long absences that happened to fall in the control arm: six of the eight participants reporting thirty or more days are controls, an arm holding 23 percent of the sample, and most of them report almost no days of feeling ill. Estimators that bound the influence of any single observation put the incidence rate ratio at 0.93 to 0.95, and capping absence at two weeks does the same. The reported standard error is also the smallest available, about a quarter below its bootstrap and robust-variance counterparts, and re-randomising the assignment under the sharp null shows that its t -statistic exceeds 1.96 in 8 percent of draws rather than 5. Underlying this is a measurement problem: the frequency of all four self-reported outcomes collapses above six days, which the fitted negative binomial cannot accommodate. Outcomes equally consistent with the trial's registered aim, including absence over the intervention period itself, show nothing.

Introduction

Not too long ago I came across a study with peculiar findings: 90 days of a short cold shower as part of a daily shower routine reduces the number of sick leave days by 29 percent, and highly significantly so ($p = 0.003$). If something sounds too good to be true, it may be too good to be true. The study's findings initially did not reach much beyond Dutch newspapers, but they have since been picked up by outlets with large international audiences.

The finding comes from Buijze et al. (2016), who ran a randomised trial in the Netherlands in the winter of 2014–2015 and posted the underlying data two years later. My reading of those data is that the headline number does not survive contact with them. This is not because the trial found a small effect that was oversold, but because the estimate is a statement about roughly half a dozen participants rather than about the population the

*Comments welcome. The data analysed here are the trial's posted supporting information, file `pone.0201978.s001.xlsx` (Cool Challenge trial data, 2018). All code and output needed to reproduce every number, table and figure below are in the accompanying replication package.

abstract describes, and because the inference that makes it significant rests on a parametric assumption the data plainly violate. Sections 1 and 2 set out the trial and the commercial movement it sits inside; Sections 3–5 take apart the measurement, the estimate and the registration in turn.

1 The trial

In December 2014 the authors recruited healthy working-age adults who did not usually shower cold, through an open online registration at `coolchallenge.nl`. Participants were randomised into four arms. Three were asked to finish their daily shower with cold water for 30, 60 or 90 seconds respectively, every day for a month; the fourth served as a control and was asked not to adopt cold showers. After the first month participants were free to continue or stop. The authors followed up at 30 and at 90 days, asking among other things for the number of days absent from work and the number of days the respondent felt ill.

The headline specification pools the three cold-shower arms into a single treatment indicator and regresses days of sickness absence over the full 90 days on that indicator and on a binary control for whether the respondent does sport, using a negative binomial model. The estimated incidence rate ratio is 0.71, the 29 percent reduction of the abstract. I reproduce it exactly on the posted data: $IRR = 0.71$ with $p = 0.003$ on 2,220 observations, of which 586, 547 and 570 fall in the 30, 60 and 90 second arms and 517 in the control arm. The reported effect is an intention-to-treat effect, and since 34 percent of the treated stopped and 5 percent of controls started, the implied effect on compliers is larger still. That inflation is applied to whatever the intention-to-treat estimate measures, so it is worth being clear about what that is.

Two features of the design bear on everything below. The intervention was unblinded, which is unavoidable for cold water but means expectations can move the outcome. And both primary outcomes are self-reported by participants who knew their arm and had signed up for a challenge whose promotional material told them what to expect. Where a dice decides who is treated and the outcome is measured administratively, randomisation does most of the work and the analysis reduces to comparing means between arms. Here it does neither.

Participants were also not isolated from one another while the trial ran. One participant, describing the experience in a blog post in December 2015, writes of reading on “the Facebook page” of participants who were enthusiastic about having much more energy throughout the day (Natuurlijk Linda, 2015). The paper does not mention such a forum, and the post does not say who ran the page or whether control participants could read it. A channel on which participants shared how the challenge was making them feel nonetheless means that expectations could spread within and possibly across arms, and that one participant’s report need not be independent of another’s. Both the treatment effect and its standard error rely on the opposite: that each participant’s report depends only on their own assignment, and that reports are independent across participants.

2 Cold exposure as a movement, and the declaration of competing interests

This article should not be read in isolation. It is part of a broader movement that believes in the merits of cold, whose central figure is Wim Hof. Hof gained international attention in 2000 by breaking a Guinness World Record for swimming more than 50 metres under ice. He now promotes the “Wim Hof Method”, a combination of breathing exercises and cold exposure claimed to deliver a range of health benefits, which he sells commercially through online courses. He is a public figure in the Netherlands and has appeared repeatedly on prime-time television, on one occasion in 2014 together with the lead author of the article at hand.

Some of this rests on evidence. Kox et al. (2014) showed that the Hof protocol can attenuate the innate immune response to an endotoxin challenge, and the trial protocol cites that work as its point of departure. But the step from modulating an immune marker in a laboratory to reducing workplace absence over three months is a long one, and the authors do not attempt it. The paper offers no hypothesis about how cold showers would reduce sickness absence, only ad hoc suggestions after the fact.

The published declaration of competing interests names the employment of one coauthor and otherwise states that there are “no patents, products in development or marketed products to declare”. Two names in the trial’s own supporting documents make that difficult to sustain. The study protocol lists Koen de Jong as physiologist on the research team and as one of its authors; de Jong co-wrote a book on Hof’s method with Hof himself, published in February 2015, while the trial was in the field. The same protocol lists Linda Koeman under logistics; she subsequently commercialised cold exposure through the company *ice.club*. Neither is a coauthor, so neither appears in the declaration. PLOS ONE defines a competing interest as “anything that interferes with, or could reasonably be perceived as interfering with, the full and objective presentation” of a submission, and states that such interests may be financial or non-financial, professional or personal, and may arise through a relationship with another person. Under that definition both involvements should have been disclosed, as should the lead author’s prior joint media appearances with Hof.

De Jong’s role was not peripheral. The trial’s recruitment website, of which the Internet Archive holds a copy captured in May 2015 (COOL Challenge, 2015), states that the AMC led the study in collaboration with the magazine *Quest* and with *Sportrusten*, a programme developed by de Jong. Under the heading “Wie zit erachter” (who is behind it), the site presents the study as the work of two people, Buijze and de Jong. It describes de Jong as the author of books on breathing and running who was then writing a book on cold and breathing together with Hof, and states that he helped with the texts, with the website and with bringing together the 2,000 participants. It adds that Hof inspired both of them. The paper’s acknowledgements thank de Jong, Koeman and Hof “for their inspiration and contribution to this study”, and *Quest* for supporting recruitment, so the authors were aware of these contributions. What the paper does not tell its readers is that the person who wrote

the recruitment material and assembled the sample was at the same time writing a book on the method with its best-known promoter.

I do not claim that anyone influenced the result. The point is narrower: enrolment was open to anyone with an internet connection, assignment was unblinded, the outcome was self-reported, and, as Section 4 shows, the published result turns on fewer than ten participants. A design with those four properties has no internal safeguard against interested participation, which is exactly the situation in which a complete declaration of interests does its work.

3 The distribution of reported sick days

Figure 1 plots the distribution of all four self-reported count outcomes, restricted to positive answers below twenty so the mass at zero does not obscure the rest.

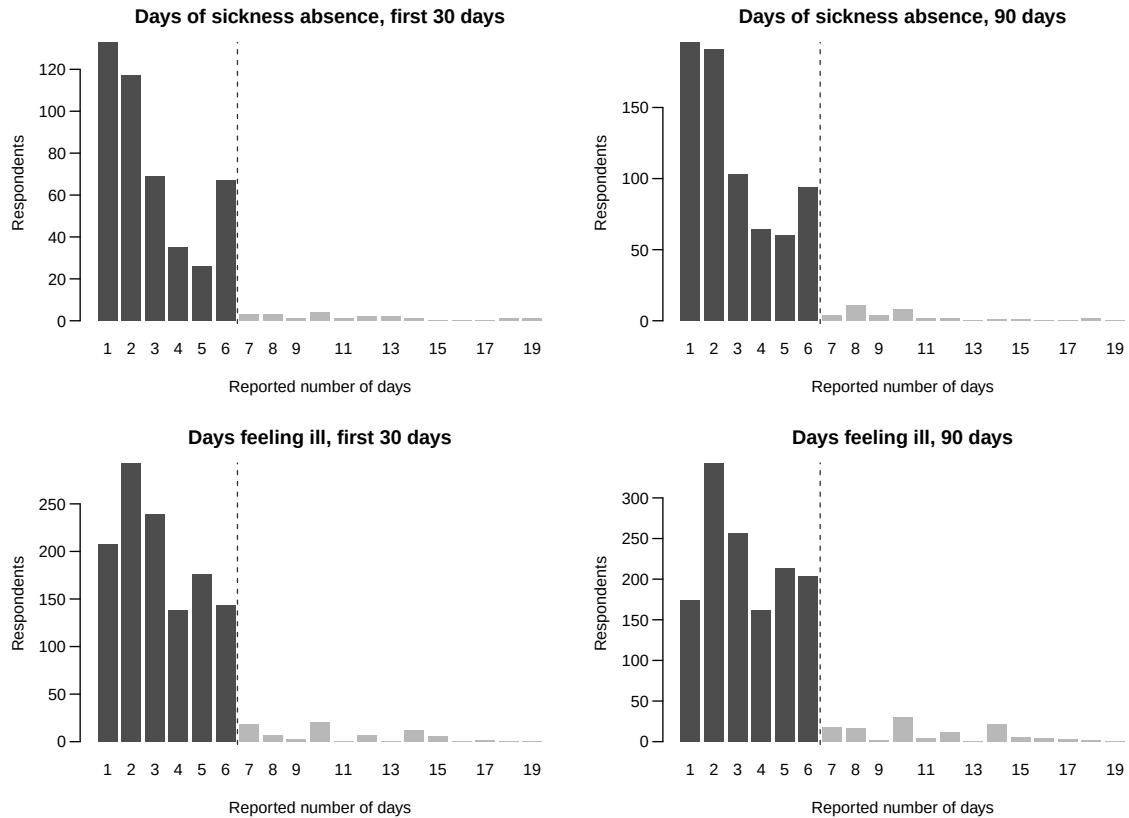


Figure 1: Distribution of the four self-reported count outcomes

Notes: Answers of zero and of twenty or more days are omitted. Bars up to and including six days are shaded darker; the dashed line marks the break between six and seven. All participants with a non-missing answer are included, not only the estimation sample. Source: the trial's posted supporting data.

All four distributions behave in the same way, and none of them behaves like a duration. Answers rise and fall between one and six days, and then essentially stop. Among positive answers, 95 percent of 30-day absence reports, 94 percent of 90-day absence reports, 93 percent of 30-day illness reports and 90 percent of 90-day illness reports are six days or fewer. For 90-day absence, 88 respondents report exactly six days and four report exactly

seven.

No process generating sickness absence produces that. Spells of illness have a roughly constant daily hazard of ending, so the frequency of durations declines smoothly rather than by a factor of twenty between one day and the next. Genuine absence data are also heaped at 5, 7, 10 and 14 days, because people take whole weeks off and round to them; here the heaping is at six and the whole-week values are empty. The likely explanation is the instrument: the questionnaire appears to have offered a bounded set of response options up to six days, with longer absences entered some other way. It is not in the supporting information, so this cannot be verified, but a single respondent reports 8.5 days of feeling ill, which suggests the entry above six was free text. A second anomaly in the self-reports, documented in Appendix B, is that participants who do not work full-time report as many days of absence as those who do.

This alone does not bias the comparison between arms. The instrument was the same for everyone, so the measurement error is non-differential. What it distorts is inference, because the authors do not compare means but fit a fully parametric likelihood.

Table 1: Observed against predicted frequencies of 90-day sickness absence

Reported days	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed	1494	188	183	102	59	59	88	4	11	4	8	2	2
Predicted by the fitted model	1483	270	138	86	59	42	31	23	18	14	11	9	7
Ratio	1.0	0.7	1.3	1.2	1.0	1.4	2.9	0.2	0.6	0.3	0.7	0.2	0.3

Notes: Predicted frequencies are the sum over the estimation sample of the negative binomial density evaluated at each count, using the fitted means and the estimated dispersion parameter from the specification of Buijze et al. (2016). $N = 2,220$.

Table 1 compares the frequencies the fitted model implies with those in the data. The model is asked to reconcile 88 respondents at six days with four at seven, and it cannot: it predicts 31 and 23. Nor is the discrepancy confined to that point, since the model also under-predicts two-day and six-day answers by factors of 1.3 and 2.9 and over-predicts one-day answers by more than 40 percent. The dispersion parameter, which governs the standard errors in this family, is estimated from precisely this mismatch and comes out at $\hat{\theta} = 0.22$, implying overdispersion of about 4.6.

So the reported standard error is produced by an information matrix evaluated at a likelihood that does not describe the data in the body of the distribution, and there is no reason to expect it to be right. It is also, mechanically, a standard error dominated by the fit in the tail: once the spike at six absorbs the body, the shape parameter is pinned down by how far the distribution runs out to the right. That is also where the treatment effect is, and Section 4.1 measures what it does to the reported p -value.

4 What the headline estimate rests on

The effect is easiest to see before any model is fitted. Mean absence over the 90 days is 1.08 days in the 30-second arm, 1.13 in the 60-second arm and 1.11 in the 90-second arm,

against 1.56 days in the control arm. The reported difference is therefore about 0.45 days per participant over three months, and it is the same for all three doses.

That difference comes from a handful of people. Of the ten longest absences in the estimation sample, six are in the control arm, which holds 23 percent of the sample; of the eight participants reporting thirty or more days, six are controls. Their reported illness does not track their reported absence: three of those six report five or fewer days of feeling ill over the same ninety days, and one reports ten, against absences of 30 to 51 days. Whatever kept these participants away from work, it does not look like the kind of illness a cold shower is supposed to prevent, and the paper’s own suggestion that the intervention modulated the intensity rather than the duration of symptoms does not fit them.

Table 2: The headline estimate under alternative samples, estimators and inference

	Incidence rate ratio	Standard error	<i>p</i> -value
Reported specification	0.71	0.116	0.003
<i>Sensitivity to the longest reported absences</i>			
Two longest control absences dropped	0.80	0.116	0.058
Two longest control absences reassigned to treatment	0.85	0.117	0.152
Longest absence in either arm dropped	0.69	0.115	0.001
Absence winsorised at 30 days	0.74	0.115	0.008
Absence winsorised at 14 days	0.83	0.112	0.108
<i>Alternative estimators and inference</i>			
Poisson quasi-ML, robust standard errors	0.71	0.145	0.018
Negative binomial, bootstrapped standard error	0.71	0.146	0.010
Robust negative binomial, T/T bounding	0.95	0.169	0.770
Robust negative binomial, BY/T bounding	0.93	0.130	0.571

Notes: The dependent variable is days of sickness absence over 90 days and every row includes the sport indicator used by Buijze et al. (2016). The first row reproduces their specification. Rows two and three concern the two control participants reporting 51 and 44 days of absence; row four drops the longest absence in the sample, which is in the treatment arm. The bootstrap uses 2,000 resamples and a percentile *p*-value. The robust negative binomial rows apply the bounded-influence estimators of Aeberhard et al. (2014) with Tukey tuning constants of 5, under the two bounding functions those authors propose; results are unchanged for constants between 3 and 8. $N = 2,220$ except where a row drops observations.

Table 2 varies the treatment of the longest absences in the upper panel and, holding the sample fixed, the estimator and inference procedure in the lower one. The upper panel shows how little supports the result. Dropping the two longest control absences, of 51 and 44 days, moves the incidence rate ratio from 0.71 to 0.80 and the *p*-value from 0.003 to 0.058. Reassigning those same two participants to the treatment arm, which randomisation might equally well have done, gives 0.85 and $p = 0.152$. The asymmetry is the informative part: dropping the single longest absence in the whole sample, 62 days, *strengthens* the result to 0.69 with $p = 0.001$, because that participant is in the treatment arm. Which arm the extreme values land in is doing the work. Capping absence at 30 days leaves 0.74 with $p = 0.008$; capping at two weeks leaves 0.83 with $p = 0.108$. It should be said that no single observation overturns the result on its own, since removing any one participant leaves the *p*-value at 0.015 or below. It takes two, which is a low bar for a trial of 2,220 people but not the same thing.

The lower panel separates two problems that are easy to conflate. The first is inference. Poisson quasi-maximum likelihood estimates the same conditional mean without assuming anything about the shape of the count distribution (Gourieroux et al., 1984), and with heteroskedasticity-robust standard errors returns the same 0.71 with $p = 0.018$. A nonparametric bootstrap of the authors' own estimator gives a standard error of 0.146 against the 0.116 they report, with a 95 percent interval for the incidence rate ratio of $[0.53, 0.94]$. The reported precision is therefore overstated; Section 4.1 pins down by how much.

The second problem is the estimate itself. The negative binomial maximum likelihood estimator has an unbounded influence function, so a single large count can move the coefficient by an arbitrary amount, which is why Aeberhard et al. (2014) developed bounded-influence alternatives for this model. Applying them, the incidence rate ratio moves to 0.95 under one bounding function and 0.93 under the other, with p -values of 0.77 and 0.57. Winsorising at two weeks does much the same by cruder means. Every procedure that declines to let a few observations dominate the likelihood returns something close to no effect.

The two panels together support a narrower claim than the paper's. There is a difference in reported absence between the arms; it is concentrated in the extreme right tail of the control distribution, and it is carried by participants whose absences do not correspond to their own reports of feeling ill. The 29 percent is not a statement about the typical participant, and an abstract that reports it as a reduction in sickness absence without qualification invites exactly that reading.

4.1 The null distribution of the t -statistic

That the standard error is unreliable is so far an argument from the shape of the likelihood. It can be checked directly, and without any modelling assumption, because the assignment mechanism is known. Under the sharp null that cold showers changed no participant's absence, each person's reported days are fixed and the only random object in the experiment is who was assigned to treatment. Re-drawing that assignment 2,000 times, holding the realised group sizes at 1,703 and 517, refitting the model on each draw and recording the t -statistic on the treatment coefficient traces out the distribution that statistic actually has under the null. If the reported standard error were right, that distribution would be standard normal and the usual critical values would apply.

Table 3 shows that it is not. For the negative binomial, the null distribution of the t -statistic has a standard deviation of 1.12 rather than 1, and $|t|$ exceeds 1.96 in 8.0 percent of draws rather than 5, and 2.58 in 2.3 percent rather than 1. A test the authors believe is running at the 5 percent level is running at 8 percent, and one they believe is running at 1 percent is running at over 2. The critical values that would actually deliver a 5 percent test are $[-2.11, 2.15]$. Large t -statistics are simply more common in this design than the normal reference admits, for the reason Table 1 gives: the likelihood is wrong where the data are dense, and the dispersion parameter that converts it into a standard error is estimated off that error.

Poisson quasi-maximum likelihood with robust standard errors is the natural repair, since

Table 3: The distribution of the treatment t -statistic under re-randomisation

	Negative binomial MLE	Poisson quasi-ML, robust SE
Observed t -statistic	-3.01	-2.36
p -value from the normal reference	0.003	0.018
Standard deviation of the null distribution	1.12	1.03
Share of draws with $ t > 1.96$	0.080	0.059
Share of draws with $ t > 2.58$	0.023	0.016
Correct 5 percent critical values	[-2.11, 2.15]	[-1.83, 2.22]
Randomization p -value	0.005	0.024

Notes: 2,000 draws of the treatment assignment under the sharp null of no effect for any participant, holding the realised group sizes fixed and refitting the paper’s specification on each draw. The upper block reports the observed statistic and the p -value the estimator’s own normal reference assigns to it. The middle block describes the re-randomisation distribution: if the standard errors were right it would have a standard deviation of one, reject at 1.96 in 5 percent of draws and at 2.58 in 1 percent, and have critical values of ± 1.96 . The randomization p -value is the share of draws with a t -statistic at least as large in absolute value as the observed one.

it is consistent for the conditional mean whatever the shape of the count distribution and does not rely on the negative binomial assumption at all. It behaves much better. Its null distribution has a standard deviation of 1.03 and rejects at 1.96 in 5.9 percent of draws, close to the nominal size. It is not exact, though, and the residual distortion is instructive: the sandwich variance estimator weights each observation by its squared residual, so the same handful of absences of 30 to 62 days that dominate the coefficient also dominate the variance estimate. Being robust to a distributional assumption is not the same as being robust to influential observations. The asymmetry of its critical values, $[-1.83, 2.22]$ against a symmetric ± 1.96 , is the visible trace of that.

Two conclusions follow, and it is worth separating them. The reported p -value is too small, but not enormously so: against its own re-randomisation distribution the observed t of -3.01 gives 0.005 rather than the published 0.003, and the Poisson statistic gives 0.024 rather than 0.018. The sharp null of no effect for anyone can still be rejected, and nothing here says otherwise. What the randomization test repairs is the p -value, not the estimand. It is an exact test of whether assignment is related to the outcome at all, and says nothing about whether that relationship is the 29 percent reduction the abstract reports or the consequence of six participants landing in one arm rather than another. On that question Table 2 is the evidence, and it points the other way.

5 Adherence to the pre-analysis plan

Randomised trials are registered in advance so that the analysis cannot be chosen once the data are in. The registered primary aim of this trial reads: “The primary aim is to determine whether a regular (hot-to-)cold shower can have a beneficial effect on immune function by reducing sickness absence.”

That sentence names a construct, not an estimand. It does not say total days rather than incidence, ninety days rather than thirty, a negative binomial rather than a comparison

of means, the three cold arms pooled rather than separate, or absence conditional on doing sport. Each of those choices was available after the data were collected, and each changes the answer. Table 4 shows what the alternatives give, using the authors’ own estimator and control variable on their own sample. Only the outcome changes, and every outcome in the table is a measure of sickness absence.

Table 4: Outcomes consistent with the registered primary aim

Outcome	Estimate		<i>p</i> -value	<i>N</i>
Days of sickness absence over 90 days	0.71	(IRR)	0.003	2220
Days of sickness absence over the first 30 days	0.80	(IRR)	0.145	2478
Any sickness absence over 90 days	0.88	(OR)	0.242	2220
Sickness absence of more than 4 days over 90 days	0.79	(OR)	0.165	2220
Sickness absence of more than 9 days over 90 days	0.46	(OR)	0.049	2220
<i>Second primary outcome, for comparison</i>				
Days feeling ill over 90 days	0.89	(IRR)	0.077	2218

Notes: IRR denotes an incidence rate ratio from a negative binomial model, OR an odds ratio from a logit. All specifications include the sport indicator used by Buijze et al. (2016). The first row is the outcome the paper reports. Sample sizes differ across rows because the 30-day outcomes are available for participants who did not complete the second follow-up.

The second row deserves attention on its own. The intervention proper ran for thirty days; after that participants were free to stop, and a third of them did. Over that thirty-day window, the window in which the treatment was actually administered, the estimate is 0.80 with $p = 0.145$. The effect appears only once the outcome is extended over sixty further days of voluntary, unmonitored compliance.

The remaining rows fall into an ordering that is hard to read any other way. Arrange the outcomes by how much weight each places on the extreme right tail. Whether a participant was absent at all places no weight there, and returns an odds ratio of 0.88 with $p = 0.242$. Absence of more than four days returns 0.79 with $p = 0.165$, absence of more than nine days 0.46 with $p = 0.049$, and total days, which is unbounded above, $p = 0.003$. The evidence strengthens monotonically in how much of it comes from the longest absences. This is the conclusion of Section 4 again, reached without dropping a single observation.

Two further comparisons belong here. The three doses give the same answer: a test of equal effects across the 30, 60 and 90 second arms returns $p = 0.88$. Thirty seconds of cold water and ninety seconds of cold water are not obviously the same physiological stimulus, and a mechanism that is flat in dose over that range needs saying out loud. And the trial’s second primary outcome, days of feeling ill, returns 0.89 with $p = 0.077$, which the paper reports as a null. The authors reconcile the two by suggesting that the intervention modulated the intensity rather than the duration of symptoms; since the participants driving the absence result are largely not those reporting many days of feeling ill, the reconciliation does not describe the data it is offered for.

None of the outcomes in Table 4 was reported, and no correction for multiple testing was applied. In fairness, a Holm correction across the five absence outcomes still leaves the headline result significant, at an adjusted p of 0.013, so multiple testing is not the problem. The problem is that the authors reported the one outcome in this family that loads most

heavily on six control-group participants, and reported it as the effect of cold showering on sickness absence. Their discussion of mechanisms does not help either. Having found an effect concentrated in the right tail, they do not ask what could make cold showers matter only for people who would otherwise have had very long absences, which is restrictive enough to be worth asking. They suggest instead that a cold shower may work through improved fitness, noting that their 29 percent is “comparable to the effect of regular physical activity” and that Dutch cyclists have 15 percent lower absence than non-cyclists. This supports a claim about fitness and absence with an association, and says nothing about how a cold shower would raise fitness.

6 Conclusion

What the data show is that control participants reported about 0.45 more days of absence over ninety days than treated participants, and that this difference is carried almost entirely by six control participants with absences of a month or more, most of whom reported hardly any days of feeling ill. Bound the influence of those observations, or cap absence at two weeks, or measure absence over the thirty days in which the intervention was actually administered, and the effect is small and statistically indistinguishable from zero. The reported p -value of 0.003 is in any case the most favourable of the available ones, and re-randomising the assignment shows that a nominal 5 percent test on this model rejects 8 percent of the time. Underneath sits a measurement problem, since the instrument appears to stop distinguishing between durations above six days, which leaves the model that produces the p -value misspecified in the body of the distribution and estimating its dispersion off the mismatch.

I do not read any of this as evidence that cold showers do nothing. The trial is simply not informative about the question. Its registration named a construct rather than an estimand and left the analysis open; its outcomes were self-reported by unblinded participants recruited through an open online campaign that told them what to expect; several of the people who organised it had commercial interests in the result that were not declared; and the published estimate rests on fewer than ten of its 2,220 participants.

A replication trial has been registered. If it pre-specifies the estimand rather than the construct, blinds what can be blinded, measures absence administratively where possible, and reports bounded-influence inference alongside the conventional kind, it will tell us considerably more than this one does.

A Internally inconsistent self-reports

Participants were asked for days of absence and days of feeling ill both at 30 and at 90 days. The two windows are nested, so the 90-day count cannot be smaller than the 30-day count. For 7.5 percent of absence reports and 21.6 percent of illness reports, it is.

Two corrections are available and neither is obviously right. Raising the inconsistent 90-day answers to the 30-day value treats the earlier report, given closer to the event, as the more reliable; dropping the inconsistent respondents treats them as uninformative. Table A1

reports both. The first moves the incidence rate ratio from 0.71 to 0.81 and the p -value from 0.003 to 0.038; the second leaves the estimate essentially unchanged. The result is therefore sensitive to how a plainly impossible pattern of answers is handled, and the direction of that sensitivity depends on which defensible correction one picks.

Table A1: The headline estimate under corrections for internally inconsistent reports

Treatment of inconsistent reports	IRR	p -value	N
As reported	0.71	0.003	2220
Inconsistent reports raised to the 30-day value	0.81	0.038	2220
Inconsistent reports dropped	0.72	0.006	2058

Notes: Inconsistent means a 90-day count below the corresponding 30-day count. All rows use the specification of Buijze et al. (2016).

The inconsistency is also not evenly spread across arms, which makes it more than a data-quality problem. Among controls, 4.5 percent of absence reports are internally impossible; in the 30, 60 and 90 second arms the figures are 11.5, 6.7 and 6.8 percent, and the treatment-control difference is significant at $p = 0.005$. Treated participants were thus about twice as likely to report fewer cumulative absence days at the second follow-up than they had already reported at the first. That is differential measurement error on the outcome carrying the paper’s result, and it runs in the direction of the finding, since a treated participant who revises their count downwards makes the treatment look better. It is consistent with what the correction does, because raising those answers back to the 30-day value removes about a third of the estimated effect. Whether the pattern reflects recall, fatigue with the second questionnaire, or participants reconciling their answers with what the challenge had led them to expect, randomisation does not protect against it and an intention-to-treat estimate does not correct for it.

Beyond the bias, the inconsistency is evidence about the quality of the self-reports. On the outcome that carries the paper’s result, roughly one respondent in thirteen gave two answers that cannot both be true, and on the second primary outcome more than one in five did. Combined with the ceiling at six days documented in Section 3, that is a thin basis for a parametric likelihood whose tail behaviour determines the reported standard error.

B Reported sick days by working hours

The baseline questionnaire records whether a participant is a full-time employee, which 38 percent are. The others have fewer working days on which they can be absent: someone working three days a week can miss at most three fifths of the working days a full-time employee can. If the absence question measured missed working days, their counts should therefore be visibly lower. Table B1 shows that they are not. Over the 90 days, full-time employees report a mean of 1.19 days of absence and the others 1.22, and none of the four outcomes differs between the two groups in a t -test. The rank-sum test agrees except for days of feeling ill in the first 30 days, where full-time employees report slightly more ($p = 0.04$), a difference that is gone by 90 days.

Table B1: Self-reported outcomes by full-time employment

Outcome	Full-time employee				Not full-time				<i>p</i> -value	
	<i>N</i>	Mean	Min	Max	<i>N</i>	Mean	Min	Max	<i>t</i> -test	Rank-sum
Absence, first 30 days	928	0.61	0	18	1550	0.55	0	22	0.37	0.15
Absence, 90 days	848	1.19	0	38	1372	1.22	0	62	0.82	0.84
Feeling ill, first 30 days	928	1.98	0	30	1549	1.82	0	30	0.20	0.04
Feeling ill, 90 days	847	3.12	0	90	1371	3.00	0	60	0.62	0.54

Notes: Full-time employment is the baseline indicator reported as “Fulltime employee” in Table 1 of Buijze et al. (2016). All participants with a non-missing answer and a non-missing employment indicator are included, not only the estimation sample. The *t*-test allows unequal variances; the rank-sum test is the Wilcoxon–Mann–Whitney test.

If anything, the illness reports point the other way. Full-time employees report as many days of feeling ill as the others, or slightly more, and have more working days on which to be absent, yet they report no more absence. Either participants counted days of illness rather than working days missed, or the employment indicator does not measure what its label says. In both cases the absence counts are harder to read as days of work lost than the paper suggests.

References

- Aeberhard, W. H., E. Cantoni, and S. Heritier (2014). Robust inference in the negative binomial regression model with an application to falls data. *Biometrics* 70(4), 920–931.
- Buijze, G. A., I. N. Siersevelt, B. C. J. M. van der Heijden, M. G. Dijkgraaf, and M. H. W. Frings-Dresen (2016). The effect of cold showering on health and work: A randomized controlled trial. *PLOS ONE* 11(9), e0161749.
- Cool Challenge trial data (2018). Supporting information file `pone.0201978.s001.xlsx`, *PLOS ONE* 13(8), e0201978.
- COOL Challenge (2015). Homepage of www.coolchallenge.nl, captured by the Internet Archive on 5 May 2015. <https://web.archive.org/web/20150505/http://www.coolchallenge.nl/> (accessed 22 September 2026).
- Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* 52(3), 701–720.
- Kox, M., L. T. van Eijk, J. Zwaag, J. van den Wildenberg, F. C. G. J. Sweep, J. G. van der Hoeven, and P. Pickkers (2014). Voluntary activation of the sympathetic nervous system and attenuation of the innate immune response in humans. *Proceedings of the National Academy of Sciences* 111(20), 7379–7384.
- Natuurlijk Linda (2015). Mijn ervaring met de Cool Challenge 2015 van Wim Hof! Blog post, 19 December 2015. <https://www.natuurlijklinda.nl/cool-challenge-2015/> (accessed 22 September 2026).