Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ignoring multiplicity and clinical significance may undermine trial conclusions: Evidence from a re-analysis of the Grintuss® pediatric cough trial

Bruno Arpino

bruno.arpino@gmail.com

University of Padova

Daniela Bellani

Catholic University of Milan

Research Article

Keywords: Multiplicity adjustment, Confidence intervals, Effect size interpretation, Randomized controlled trial reporting, Pediatric cough, Over-the-counter cough remedies

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8910722/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Title page

Title: Ignoring multiplicity and clinical significance may undermine trial conclusions: Evidence from a re-analysis of the Grintuss® pediatric cough trial.

Authors: Bruno Arpino[1] and Daniela Bellani[2]

1. Department of Statistical Sciences, University of Padua (Italy); bruno.arpino@unipd.it – corresponding author. Via Cesare Battisti, 241, 35121 Padova (Italy).

2. Department of Statistical Sciences at the Catholic University of Milan (Italy); daniela.bellani@unicatt.it.

Ignoring multiplicity and clinical significance may undermine trial conclusions: Evidence from a re-analysis of the Grintuss® pediatric cough trial

Abstract

Background. Reporting and analysis standards for randomized trials are well established, yet evidence used to support pediatric cough remedies sometimes relies on incomplete reporting, extensive hypothesis testing, and interpretation dominated by p-values rather than effect sizes and clinical relevance. We use a pediatric cough trial of Grintuss® syrup as a methodological case study to assess how multiplicity and effect-size–focused reporting may alter trial conclusions.

Methods. We conducted an audit of the trial's design, analysis, reporting, and interpretation against established principles for randomized trials, focusing on prespecification, completeness of reporting, use of within-arm versus between-arm inference, multiplicity, baseline imbalance, and appropriateness of parametric tests for bounded ordinal outcomes. Because individual-level data were not available despite direct contact with the authors, we performed a constrained re-analysis using published figures. We reconstructed approximate means and standard deviations from plotted values and used sample sizes reported or implied in the article to derive approximate 95% confidence intervals for group means and day-4 between-arm differences. We enumerated the hypothesis tests reported or implied and applied Holm's step-down procedure to adjust p-values for multiplicity. For a subgroup comparison on the proportion of improved cough, we computed an absolute risk difference with an approximate confidence interval.

Results. The published report implies at least 15 non-independent hypothesis tests across outcomes, time windows, and subgroups, without prespecification or multiplicity adjustment. After multiplicity control, nominally significant findings most directly interpretable as evidence of efficacy did not remain below conventional thresholds. Reconstructed estimates suggested, at most, a small early

advantage for night-time cough at day 4 with a confidence interval spanning no effect, and no clear day-time advantage at day 4; the report provided no clear between-arm signal at day 8. The subgroup analysis showed a large absolute difference but with wide uncertainty.

Conclusions. In this pediatric cough trial, conclusions appear sensitive to selective p-value reporting, extensive multiplicity, and limited attention to effect size and clinical relevance. For pediatric cough trials and similar studies, robust inference requires prespecified estimands and endpoints, baseline-adjusted between-arm analyses with confidence intervals, explicit multiplicity strategies, and transparent reporting aligned with randomized trial standards.

Keywords: Multiplicity adjustment; Confidence intervals; Effect size interpretation; Randomized controlled trial reporting; Pediatric cough; Over-the-counter cough remedies

1. Background

Reporting and analysis standards for clinical research are well established. In randomized trials, guidance such as CONSORT emphasizes prospective specification of outcomes and analyses, transparent reporting of effect estimates with confidence intervals rather than relying on selective p-values [1]. Complementary methodological guidance—including the Cochrane Handbook—discourages analytic practices that can mislead inference, such as emphasizing within-arm pre–post significance tests, failing to model repeated measures appropriately, and overlooking inflated false-positive risk when multiple outcomes and time points are tested [2]. These principles are intended to make results interpretable, reproducible, and resistant to overstatement.

Yet, despite the availability of such guidance, evidence used to justify clinical or consumer-facing claims about medicines and health products often rests on studies with limited transparency, incomplete

reporting of estimands and uncertainty, and analysis strategies that do not align with the causal questions posed. When evidence is communicated mainly through nominal statistical significance—without confidence intervals, clear primary endpoints, or explicit multiplicity control—readers cannot judge the *magnitude* or *precision* of effects, and small or transient differences may be overinterpreted as clinically meaningful [3-5]. These concerns are amplified in small-to-moderate trials and in settings with many correlated tests [6].

Pediatric cough research illustrates these challenges. Cough is common in childhood, and guidance generally discourages routine use of OTC cough medicines in children—especially those under 4 years—recommending that any products be used sparingly for short-term symptom relief rather than as curative therapy [7-10]. Nonetheless, OTC remedies remain widely used despite uncertainty about their effectiveness in young children [8, 11-12]. Plant-based syrups with mucoprotective properties—often honey-containing—have also grown in popularity, with a heterogeneous evidence base that includes trials and meta-analyses reporting modest short-term benefits but overall low certainty [13-15]. Methodologically, this field is demanding: pediatric cough is often self-limiting, outcomes are frequently bounded ordinal parent-reported scales, and studies commonly examine multiple endpoints and time windows—features that make results sensitive to modeling, reporting, and multiplicity decisions.

In this paper we focus on a specific pediatric cough trial as a case study: a two-centre randomized, double-blind, placebo-controlled trial (N≈100; children ~3–6 years) evaluating Grintuss® pediatric syrup by Canciani et al. [16]. The authors report statistically significant within-arm improvements from baseline to day 4 in the intervention arm for both night-time and day-time cough scores, a between-arm difference at day 4 favoring Grintuss for night-time cough, and a responder-style subgroup analysis

among children with more severe baseline symptoms [16]. The trial is also practically relevant because it is used in product-facing materials (product leaflet and website) to support effectiveness claims and appears to be the only published evidence presented as supporting the product's efficacy. The product is also commercially succesfull: in 2023 Aboca sold more than 12 million units in the Grintuss line overall, which has consolidated leadership in major European markets [17].

Our contribution is therefore not only a re-evaluation of the efficacy of the Grintuss syrup, but as a methodological demonstration of how inference can be distorted when trials deviate from long-established principles of transparent design, analysis, and reporting. We (i) audit Canciani et al.'s trial against established guidelines, (ii) enumerate the battery of tests reported or implied, and (iii) because the original microdata were not available despite direct contact with the authors, reconstruct key descriptive quantities from the published figures to derive approximate effect estimates with 95% confidence intervals and multiplicity-adjusted p-values. Finally, informed by this case study and comparative evidence on honey-based and similar plant-derived syrups [13, 14, 18-19], we propose a checklist for future pediatric cough trials that operationalizes core principles: prespecified estimands and endpoints, explicit multiplicity strategies, complete reporting of uncertainty, and interpretation that distinguishes statistical from clinical significance, baseline-adjusted longitudinal modeling appropriate for ordinal outcomes [1-6].

2. Critical appraisal of Canciani et al.'s trial and alternative analyses

2.1. Main issues: lack of transparency and over-reliance on p-values

As mentioned above, a first issue in this trial is lack of transparency. This is manifested in two ways. First, there is no mention of any pre-registration of the study protocol and hypotheses. The full list of

tested hypotheses is also not explicitly provided and has to be reconstructed from the figures, main text and some assumptions.

The authors mention two broad categories of hypotheses tests: *within-arm* pre–post comparisons (testing whether cough scores decreased from one time point (t_0 or t_4) to a following one (t_4 or t_8) within each treatment arm) and *between-arm* comparisons at follow-up (testing whether the study product group had a lower mean cough score than the placebo group at t_4 or t_8). In addition, the authors considered a subgroup analysis on children with intense cough (cough scores ≥3) at baseline and assessed the difference in proportions of children that improved (cough scores < 3) at follow-up between the two arms.

Specifically, the authors explicitly report:

- Six within-arm *t*-tests comparing day 0 to day 4, day 0 to day 8, day 4 to day 8 scores separately for night-time and day-time cough in the Grintuss group;
- Six analogous within-arm tests in the placebo group;
- Two between-arm independent *t*-tests at day 4 (for night-time and day-time scores);
- One Fisher exact test comparing proportions of children whose day-time cough score improved to scores below 3 at day 4 among those with baseline scores ≥3.

Other two between-arm independent *t*-tests at day 8 (for night-time and day-time scores) were likely conducted but not reported due to statistical insignificance. The authors do not mention whether the Fisher exact test has been conducted also for night-time cough and at time 8. This amounts to a battery of at least fifteen (see Table 1) and likely twenty hypothesis tests, none of which is prespecified in a trial protocol, and all conducted without adjustment for multiple comparisons.

A lack of transparency in Canciani et al. is also manifested for the limits in the details on the conducted analyses. As highlighted in Table 1, some of the results are presented graphically. In this case, the mean score and standard deviation are represented. In addition, when effects are significant, p-values are reported in the Figures and/or in the text. This limits the replicability and assessment of the results because measures of statistical uncertainty (e.g., standard errors) are not reported and p-values are not reported for all tests.

Running multiple hypothesis tests on the same dataset—especially when outcomes are correlated (e.g., day vs night cough, repeated measures at $t_0/t_4/t_8$, and subgroup analyses)—inflates the probability of obtaining at least one "statistically significant" result purely by chance, even if the treatment has no true effect. This is known as the classic multiplicity / familywise error problem: with each additional test, the overall Type I error rate rises unless a prespecified testing hierarchy or multiplicity adjustment is applied. The issue is amplified when tests are non-independent, because correlated endpoints/timepoints tend to move together, creating multiple opportunities to find significance from essentially the same underlying signal (and noise). As a result, selectively highlighting a subset of significant p-values can overstate evidence, whereas a more defensible strategy is to prespecify a primary endpoint, model repeated measures jointly, and report effect sizes with confidence intervals while explicitly addressing multiplicity.

Table 1. Inventory of hypothesis tests reported or implied in Canciani et al.

Type of test	Outcome	Time	Null hypothesis	Reported result	Holm-adjusted p-values
Within tests 1-2	Night-time	$t_0 \rightarrow t_4$	$\mu(A,t_4) - \mu(A,t_0) = 0$; A=1, 0	p = 0.003 for A=1 (Figure 1); p not reported for A=0	0.042
Within tests 3-4	Day-time	$t_0 \rightarrow t_4$	$\mu(A,t_4) - \mu(A,t_0) = 0$; A=1, 0	p = 0.04 for A=1 (Figure 2); p not reported for A=0	0.360
Within tests 5-6	Night-time	$t_0 \rightarrow t_8$	$\mu(A,t_8) - \mu(A,t_0) = 0$; A=1, 0	p = 0.005 for A=1 (in text only); p not reported for A=0	0.065
Within tests 7-8	Day-time	$t_0 \rightarrow t_8$	$\mu(A,t_8) - \mu(A,t_0) = 0$; A=1, 0	p = 0.001 for A=1 (in text only); p not reported for A=0	0.015
Within tests 9-10	Night-time	$t_4 \rightarrow t_8$	$\mu(A,t_8) - \mu(A,t_4) = 0$; A=1, 0	p = 0.05 for A=1 (in text only); p not reported for A=0	0.400
Within tests 11-12	Day-time	$t_4 \rightarrow t_8$	$\mu(A,t_8) - \mu(A,t_4) = 0$; A=1, 0	p = 0.01 for A=1 (in text only); p not reported for A=0	0.120
Between test 1	Night-time	t_4	$\mu(A=1,t_4) - \mu(A=0,t_4) = 0$	p = 0.03 (Figure 1)	0.330
Between test 2	Day-time	t_4	$\mu(A=1,t_4) - \mu(A=0,t_4) = 0$	Not statistically significant (p not reported)	
Between test 3 (subgroup)	Day-time (baseline ≥3)	t_4	prop. improved (A=1) = prop. improved (A=0)	Fisher p = 0.03 (Figure 3)	0.330

Note: A refers to the arms (A= 1, Grintuss treatment; A = 0, control or placebo group).

Another crucial issue is that the interpretation of the effectiveness of Grintuss relies on statistical significance and not on practical/clinical significance (e.g., confidence intervals are not reported and mean differences are not emphasized). This testing strategy introduces a high risk of overinterpretation—particularly given the relatively small sample size and the high number of non-independent tests.

2.2. Adjusting p-values for multiplicity

Canciani et al. [16] conduct a battery of (at least) fifteen hypothesis tests across outcomes (night/day), time windows ($t_0 \rightarrow t_4$, $t_0 \rightarrow t_8$, $t_4 \rightarrow t_8$), study arms (within Grintuss and within placebo), plus between-arm comparisons at t_4 and a subgroup analysis (Table 1). Such a testing strategy raises the probability of at least one "significant" result occurring by chance, unless multiplicity is explicitly addressed. To illustrate how multiplicity affects inference, we applied the Holm step-down procedure to control the familywise error rate (FWER) at α=0.05 [20].

Holm's method is a sequential improvement over Bonferroni. First, the m p-values are sorted from smallest to largest, $p_{(1)} \ldots \leq \ldots p_{(j)} \ldots \leq \ldots p_{(m)}$. In the step-down testing version, the smallest p-value is compared to α/m; if it is significant, the next is compared to $\alpha/(m-1)$, and so on, stopping at the first non-rejection of the null. An equivalent and often more convenient way to present results is via Holm-adjusted p-values, which can be computed as

$$\tilde{p}_{(k)} = max_{j \leq k} \left\{ (m - j + 1) * p_{(j)} \right\}.$$

Intuitively, each ordered p-value is multiplied by the number of hypotheses "still in play" at that step (e.g., m for the smallest, m−1 for the second smallest), and the running maximum is taken to ensure the adjusted p-values are non-decreasing with rank.

The Holm-adjusted p-values (last column of Table 1) change the interpretation of the "significant" findings. Under the full 15-test family, only the smallest within-arm signals remain statistically significant after adjustment (e.g., the Grintuss within-arm day-time change to t_8 retains $\tilde{p}$ =0.015, and the within-arm night-time change to t_4 becomes $\tilde{p}$=0.042). In contrast, the p-values that are most appropriately read as evidence of treatment efficacy—the between-arm night-time difference at t_4 and the subgroup responder comparison (both originally reported as p=0.03)—inflate to $\tilde{p}$=0.33 when the full battery is considered, and to $\tilde{p}$=0.09 even under the most favorable "between-arm only" correction in which only the three between-arm tests included in Table 1 are considered.. Overall, after multiplicity control, the only results that remain below 0.05 are within-arm improvements, which by themselves do not support causal claims about the intervention.

2.3. Approximate interval estimation of between-arm contrasts at day 4

Even ignoring multiplicity, another substantial limitation remains: over-reliance on statistical significance. To evaluate the magnitude and precision of effects that were not reported in Canciani et al., and the robustness of their interpretations, we reconstructed approximate numerical values for group means and standard deviations using visual extraction from Figures 1–3. This process was assisted by digital figure analysis tools and AI-supported image quantification, which allowed us to estimate plotted values with reasonable accuracy. Based on visual inspection and scaling of the error bars, we inferred (approximate) means and SDs of cough scores at each time point for both groups. We focused on day 4 estimates as most of the emphasized significant results in the study refer to this endpoint.

Assuming sample sizes derived from the text (baseline $n = 51$ per group; completers = 47 Grintuss, 44 placebo), we then calculated approximate 95% confidence intervals for both within-group means and between-group differences at day 4. These reconstructed statistics, while inherently approximate due to digitization (likely small) error, nonetheless allow for a more transparent interpretation of the trial's findings.

Tables 2 and 3 report the reconstructed mean and SD of cough scores with approximate 95% CIs for the mean at baseline and day 4 for night-time and day-time, respectively. In both arms, and for both nigh-time and day-time cough mean scores decrease between t_0 and t_4, consistent with natural symptom resolution. Importantly, the standard deviations are large relative to the means, indicating substantial heterogeneity in symptoms and suggesting that mean-based summaries may mask the ordinal and bounded nature of the underlying scale.

Table 2. Night-time cough scores means, SDs, and approximate 95% CIs for the mean

Arm	Time	Mean (≈)	SD (≈)	n	95% CI for the mean (≈)
Grintuss	t_0	1.34	1.47	51	0.93–1.74
Grintuss	t_4	0.85	1.01	47	0.56–1.14
Placebo	t_0	1.56	1.61	51	1.12–2.01
Placebo	T_5	1.35	1.43	44	0.93–1.77

Table 3. Day-time cough scores means, SDs, and approximate 95% CIs for the mean

Arm	Time	Mean (≈)	SD (≈)	n	95% CI for mean (≈)

Grintuss	t_0	1.60	1.25	51	1.26–1.94
Grintuss	t_4	1.20	1.03	47	0.91–1.49
Placebo	t_0	1.60	1.15	51	1.28–1.92
Placebo	T_4	1.45	1.21	44	1.09–1.81

Table 4 reframes the key trial claims in terms of between-arm effect estimates with uncertainty, rather than p-values. For night-time cough at day 4, the estimated mean difference (Grintuss − placebo) is −0.50 points, but the 95% CI (−1.01; +0.01) is wide and includes the null (0), implying that the evidence for a between-group advantage is borderline. For day-time cough at day 4, the estimated difference is smaller (−0.25) and the 95% CI (−0.71; +0.21) is wide and clearly compatible with no effect, reinforcing that any daytime benefit—if present—is uncertain. Presenting results this way makes clear that the trial's precision is limited and that conclusions based primarily on statistical significance levels risk overstating evidence.

Table 4. Day-4 between-arm contrasts (Grintuss − Placebo), approximate

Outcome	**Mean diff (≈)**	**95% CI for diff (≈)**
Night-time (t_4)	−0.50	−1.01 to +0.01
Day-time (t_4)	−0.25	−0.71 to +0.21

Table 5 summarizes the only explicitly reported binary subgroup comparison (children with baseline day-time cough score ≥3). The proportions in improvements (cough score from ≥3 to < 3) differ substantially between arms (92.9% vs 53.8%), corresponding to an absolute risk difference of 39.1 percentage points with an approximate 95% CI of 8.7 to 69.3. While this CI excludes zero and is directionally consistent with benefit, this CI is rather wide and inference remains fragile because the subgroup is small (n = 27), improvement is defined using a dichotomization of an ordinal measure, and the analysis is one of multiple comparisons performed without a prespecified hierarchy or multiplicity

control. Moreover, dichotomizing the 0–5 scale discards information and can magnify apparent effects depending on the chosen threshold. For these reasons, Table 5 should be interpreted as exploratory and hypothesis-generating rather than confirmatory, and it underscores the need for a baseline-adjusted ordinal or repeated-measures model applied to the full sample.

Table 5. Subgroup analysis (day-time cough score ≥3 at day 0): proportion improved at day 4

Arm	**n. improved to < 3 at t_4**	**n. not improved at t_4**	**Total**	**Proportion improved (as %) (95% CI)**
Grintuss	13	1	14	92.9% (66.1–99.8%)
Placebo	7	6	13	53.8% (25.1–80.8%)
Grintuss-Placebo				39.1 (8.7-69.3)*

Note: * Absolute risk difference reported as difference in percentage points.

2.4. Other issues

Two additional statistical issues in Canciani et al. are worth noting: (1) no adjustment for baseline imbalance and (2) treating ordinal outcomes as continuous. The authors report descriptive baseline characteristics for the two arms, and some differences appear non-negligible. In particular, the placebo group has a higher baseline night-time cough score (see also Table 2): the mean at baseline is 1.56 in the placebo arm versus 1.34 in the treatment arm. This baseline gap (Grintuss − placebo = −0.22) matters because the day-4 between-arm contrast is interpreted without baseline adjustment. A simple difference-in-differences intuition illustrates the point: if one subtracts the baseline gap from the unadjusted day-4 mean difference (−0.50), the implied effect shrinks to approximately −0.28. This simple adjustment already shows that unadjusted comparisons at follow-up can overstate (or understate) effects when baseline imbalance is present. A more appropriate approach would be an

ANCOVA model (day-4 outcome adjusted for baseline score), or a mixed-effects/repeated-measures model including baseline and follow-up measurements.

As for the nature of the outcomes, cough is measured on a 0–5 parent-reported ordinal scale. For day-time cough, the categories are: 0 = absent; 1 = for a short period (a few minutes); 2 = for two short periods (≈10 minutes); 3 = frequent cough that does not interfere with normal activities; 4 = frequent cough that interferes with normal activities; 5 = disturbing cough for most of the day. The authors analyze these scores using t-tests, which implicitly treat the scale as interval and approximately normally distributed. While t-tests can be reasonably robust to deviations from normality in large samples (via the central limit theorem), sample sizes here are modest (about 50 per arm at baseline and fewer at follow-up), and the underlying distributions are likely to be seriously skewed because the scale is bounded and discrete, with probable floor effects as symptoms improve. The available information is consistent with skewness: for example, for day-time cough at baseline the paper indicates that only 13 children have scores ≥4–5, while the remainder fall in the 0–3 range, and the mean is relatively close to the lower bound (1.34). At follow-up, skewness is likely even more pronounced as more children shift to low scores (actually only 1 has a score of 4 or 5). Given these features—and the inherently subjective and ordinal nature of the outcome—a preferable strategy would have been to use ordinal approaches (e.g., rank-based methods such as the Mann–Whitney test), or at minimum to complement mean-based analyses with distribution-robust sensitivity analyses (e.g., permutation tests for mean differences).

3. Conclusions and broader methodological implications

Using Canciani et al.'s trial on the effectiveness of the Grintuss syrup [16] as a case study, we combined (i) a structured methodological audit against well-established reporting/analysis standards for

randomized trials and (ii) a constrained re-analysis based on the information that is actually available in the published article. Because the original microdata were not made available despite direct contact with the authors, we reconstructed key descriptive quantities (means and standard deviations) from the trial figures and used the reported/implicit sample sizes to derive approximate effect estimates, 95% confidence intervals (CIs), and multiplicity-adjusted p-values. While approximate by construction, these analyses serve the central methodological purpose of this paper: to show how conclusions can shift when the evidential focus moves from a selective set of nominal p-values to estimands presented with uncertainty and accounting for multiplicity.

Three substantive messages follow. First, the evidential weight in Canciani et al. rests heavily on within-arm (pre–post) significance tests. In randomized trials, however, within-arm change does not identify a treatment effect because it conflates intervention effects with natural recovery, regression to the mean, and measurement reactivity; causal inference depends on between-arm contrasts [21]. Second, when the trial's headline results are reframed as between-arm effect estimates with confidence intervals, the picture is more cautious than the original narrative suggests. The day-4 night-time contrast is small and borderline by conventional thresholds, the day-4 day-time contrast is compatible with no effect, and the paper reports no clear between-arm signal at day 8—against a background of improvement in both groups. Third, the multiplicity implied by the paper's "battery" of tests matters. After controlling the familywise error rate using Holm's procedure [20], the findings that are most plausibly interpreted as evidence of efficacy (the between-arm day-4 night-time comparison and the responder-style subgroup analysis) no longer meet a 0.05 threshold when the full testing family is considered; what remains *statistically* significant is primarily within-arm improvement, which is not probative for efficacy [6, 21]. Even relying on this within-arm effect only, the effect size (mean change) remain small compared to the standard deviation and range of the scale.

These issues are not certainly newly discovered. They reflect long-standing methodological guidance that was already available when Canciani et al. was published and is reinforced in current standards. CONSORT requires that randomized trials report prespecified outcomes and present effect sizes with CIs rather than relying on p-values alone [1, 22]. This focus on estimation is precisely what enables readers to judge practical (clinical) importance—something that cannot be inferred from statistical significance by itself [3]. Indeed, major statistical guidance explicitly cautions against dichotomous "significant/non-significant" thinking and emphasizes that p-values do not quantify effect size, importance, or decision relevance [3-4].

The broader implications extend beyond one product and beyond pediatric cough. Trials of symptomatic, self-limiting conditions—especially those using subjective, bounded, ordinal outcomes—are particularly vulnerable to overinterpretation if they (i) test many endpoints/time windows without a prespecified hierarchy, (ii) emphasize (within-arm) p-values, (iii) omit CIs and baseline-adjusted between-arm estimands, and (iv) apply parametric methods without checking distributional adequacy. For ordinal scales, analytic strategies such as baseline-adjusted ANCOVA (when a mean-based estimand is defensible) and, more naturally, ordinal or distribution-robust longitudinal models (e.g., proportional-odds mixed models) provide estimands that align better with the measurement scale and can be communicated with marginal effects and CIs [21, 23]. Likewise, multiplicity should be treated as a design feature—addressed via a prespecified primary endpoint and structured secondary testing, or via transparent adjustment—rather than handled implicitly by selective reporting [6].

In sum, our re-analysis of the Grintuss cough syrup supports—at most—a small early advantage for night-time cough that is uncertain in magnitude and not clearly durable, and it does not sustain strong claims of efficacy once multiplicity and estimation uncertainty are foregrounded. More importantly for a methods readership, this case underscores a generalizable lesson: transparent, estimand-focused reporting with baseline-adjusted between-arm contrasts, appropriate models for outcome scale/distribution, and an explicit multiplicity strategy is the minimum needed to prevent routine overstatement of evidence in small-to-moderate trials [22]. Our checklist is intended to make these requirements operational for future pediatric cough research and for analogous trials in other symptom-driven conditions where commercial and clinical narratives can otherwise outpace the strength of the underlying evidence.

Declarations

Ethics, Consent to Participate, and Consent to Publish

Not applicable.

Availability of data and materials

The work involves no new data collection and relies solely on re-analyses of results in Canciani et al. (2014)

Competing interests

The authors have no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Funding

This research did not receive specific funding.

Authors' contributions

B.A. and D.B. jointly conceived the study and its methodological focus. B.A. and D.B. designed the appraisal framework, extracted and reconstructed numerical values from the published figures and

tables, and performed the statistical re-analyses . D.B. drafted the background and conclusions, while B.A. drafted the methods. Both authors interpreted the findings, revised the manuscript, and approved the final version.

References

1. Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., ... & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomized trials. BMJ, 340, c869. https://doi.org/10.1136/bmj.c869
2. Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., ... & Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ, 343.
3. Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, *292*(6522), 746-750.
4. Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305-307.
5. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.
6. Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how?. *Journal of clinical epidemiology*, *54*(4), 343-349.
7. Fashner, J., Ericson, K., & Werner, S. (2012). Treatment of the common cold in children and adults. *American family physician*, *86*(2), 153-159.
8. Smith, S. M., Schroeder, K., & Fahey, T. (2014). Over‐the‐counter (OTC) medications for acute cough in children and adults in community settings. *Cochrane Database of Systematic Reviews*, (11).

9. Summerlin, J., & Eiland, L. S. (2025). The Use and Safety of Cough and Cold Medications in the Pediatric Population. *The Journal of Pediatric Pharmacology and Therapeutics*, *30*(1), 17-26.
10. Vogelberg, C., Cuevas Schacht, F., Watling, C. P., Upstone, L., & Seifert, G. (2023). Therapeutic principles and unmet needs in the treatment of cough in pediatric patients: review and expert survey. *BMC pediatrics*, *23*(1), 34.
11. Clark, G., Fitzgerald, D. A., & Rubin, B. K. (2023). Cough medicines for children-time for a reality check. *Paediatric Respiratory Reviews*, *48*, 30-38.
12. Manti, S., Tosca, M. A., Licari, A., Brambilla, I., Foiadelli, T., Ciprandi, G., & Marseglia, G. L. (2020). Cough remedies for children and adolescents: current and future perspectives. *Pediatric Drugs*, *22*(6), 617-634.
13. Cohen, H. A., Rozen, J., Kristal, H., Laks, Y., Berkovitch, M., Uziel, Y., ... & Efrat, H. (2012). Effect of honey on nocturnal cough and sleep quality: a double-blind, randomized, placebo-controlled study. *Pediatrics*, *130*(3), 465-471.
14. Oduwole, O., Udoh, E. E., Oyo‐Ita, A., & Meremikwu, M. M. (2018). Honey for acute cough in children. *Cochrane Database of Systematic Reviews*, (4).
15. Kuitunen, I., & Renko, M. (2023). Honey for acute cough in children—a systematic review. *European Journal of Pediatrics*, *182*(9), 3949-3956.
16. Canciani, M., Murgia, V., Caimmi, D., Anapurapu, S., Licari, A., & Marseglia, G. L. (2014). Efficacy of Grintuss® pediatric syrup in treating cough in children: a randomized, multicenter, double blind, placebo-controlled clinical trial. *Italian journal of pediatrics*, *40*(1), 56.
17. La Repubblica. (2024, June 5). *Prodotti naturali, ricavi in forte crescita per Aboca.* https://www.repubblica.it/economia/rapporti/osserva-italia/le-storie/2024/06/05/news/prodotti_naturali_ricavi_in_forte_crescita_per_aboca-423172279/

18. Carnevali, I., La Paglia, R., Pauletto, L., Raso, F., Testa, M., Mannucci, C., ... & Calapai, G. (2021). Efficacy and safety of the syrup “KalobaTUSS®” as a treatment for cough in children: A randomized, double blind, placebo-controlled clinical trial. *BMC pediatrics*, *21*(1), 29.
19. Cohen, H. A., Hoshen, M., Gur, S., Bahir, A., Laks, Y., & Blau, H. (2017). Efficacy and tolerability of a polysaccharide-resin-honey based cough syrup as compared to carbocysteine syrup for children with colds: a randomized, single-blinded, multicenter study. *World Journal of Pediatrics*, *13*(1), 27-33.
20. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.
21. Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *Bmj*, *323*(7321), 1123-1124.
22. Hopewell, S., Chan, A. W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., ... & Boutron, I. (2025). CONSORT 2025 statement: updated guideline for reporting randomised trials. *The Lancet*, *405*(10489), 1633-1640.
23. Agresti, A. (2010). *Analysis of ordinal categorical data*. John Wiley & Sons.