James Heathers, PhD (j@medicalevidenceproject.org)
David Robert Grimes, PhD (david@retractionwatch.com)

Medical
Evidence
Project

GRIM-*U*

A GRIM-like observation to establish impossible *p* values from ranked tests

Report GRIMU

medicalevidenceproject.org

Medical
Evidence
Project

GRIM-*U*: A GRIM-like observation to establish impossible *p* values from ranked tests

Citation: Heathers, James, and David Robert Grimes. 2026. “GRIM-U: A GRIM-like observation to establish impossible p values from ranked tests.” Available at medicalevidenceproject.org/GRIMU.

More information on this report is available at medicalevidenceproject.org/GRIMU.

This report is a product of the Medical Evidence Project, which aims to reduce harm to patients and improve outcomes by employing forensic scientific techniques to find errors in influential medical articles and guidelines. This work is made possible by a grant from Open Philanthropy and support from donors to the Medical Evidence Project. Learn more about our work and how to support it by going to medicalevidenceproject.org/about.

The Medical Evidence Project is an endeavor of The Center for Scientific Integrity, whose mission is to promote transparency and integrity in science and scientific publishing and to disseminate best practices and increase efficiency in science. Information about the center’s leadership and finances is available at centerforscientificintegrity.org.

Published by The Center for Scientific Integrity, 121 W. 36th St., Suite 209, New York, NY 10018.

© 2026 The Center for Scientific Integrity

All rights reserved. For reprint permission, contact info@medicalevidenceproject.org.

Sections

1 REPORT SUMMARY

Forensic metascience has shown that summary and test statistics (means, standard deviations, *z* scores, *t* and beta values, etc.) can sometimes be mathematically impossible, and they can therefore serve to investigate the accuracy of published scientific papers without access to their underlying data. However, these observations were not made with assessing medical literature in mind. In the small samples of non-normal data commonly found in medical and surgical papers, it is common to compare groups with rank-sum tests. Building on the same principle as the Granularity-Related Inconsistency of Means (GRIM) test, we introduce Granularity-Related Inconsistency of Mann-Whitney *U* (GRIM-*U*), a GRIM-like observation that exploits the discrete nature of this rank-based non-parametric test to identify impossible *p* values in medical papers. There are only so many possible rank sums for any given sample, and many combinations of two datasets produce identical rank sums. For instance, in a small sample (e.g., two samples of $N = 6$) there are only 17 possible *p* values out of 100 (implied at 2 decimal places) or 1000 (implied at 3 decimal places). Applying GRIM-U to a convenience sample of recent medical articles that employed the Mann-Whitney *U* test, we reconstruct the underlying U-statistics (with or without continuity corrections) and compare the reported *p* values against the possible set. Most checked results were found to be consistent, but several display minor discrepancies (e.g., $p =$ 0.171 vs. 0.172) that cannot arise from any permissible rank sum, usually suggesting rounding or software-related discrepancies. In one case considered here, reported *p* values lie below the absolute minimum implied by the given sample size, indicating a likely reporting error. Our findings highlight that *p* values from rank-based tests are intrinsically granular, making them amenable to forensic scrutiny. The implementation of GRIM-U is not challenging but must be qualified by variations in data structure, software used, reporting standards, calculatory methods, etc. We provide analytic formulas, a simple Excel-spreadsheet-based calculator, a lightweight R implementation (*U*-Bend) for investigating the nature of the test, and heuristics for reviewers and editors to flag implausible *p* values efficiently. Adoption of GRIM-*U* alongside existing forensic tools can improve the reliability of statistical reporting in medical research.

2 BACKGROUND AND METHODS

2.1 Introduction

In the tradition of forensic metascience (Heathers 2025), measures of central tendency, variance, and test statistics can be evaluated for possibility, impossibility, or likelihood of accuracy. These tests have typically been deployed to evaluate research in the social sciences (N. Brown and Heathers 2017; Schumm et al. 2025) and less commonly within biomedical, medical, and surgical research. This is due entirely to circumstance and not because there is an absence of problems within medicine to find; in fact, there have been several high-profile episodes within these fields where problems with presented scientific work have been detected via the analysis of published data (e.g., Carlisle 2012; Bolland et al. 2021). It is likely much more scrutiny can be brought to bear on the quality of medical literature if existing forensic metascientific tools can be further developed to include observations more congruent to medical research.

2.2 Analyzing ranked tests

The Granularity Related Inconsistency of Means (GRIM) test (N. Brown and Heathers 2017) relies on a confluence of numerical features observed when means are calculated from integer data. With the correct preconditions, GRIM can establish the possibility or impossibility of a mean relative to its accompanying cell size and rounding. The core observation behind GRIM is trivial: any mean of any set of N integers must have a decimal tail that can be described by X/N where X is an integer, and only a limited number of rounded tails are possible. For instance, in a hypothetical sample of $N =$ 13 integers, 3.23 is $3 + 3/13$ and a possible value for the mean (i.e., X is 42). Likewise, 3.31 is $3 + 4/13$ and possible (and X is 43), but an intermediate value of 3.27 is impossible because there is no value $X/13$ that returns 3.27 rounded to 2 decimal points. The numerator's value would be approximately 42.5, a non-integer and therefore not a possible value. After the GRIM test was described, other researchers noted that standard deviations presented a similar pattern [allard2018; anaya2016]. A formalization of this principle would be: "Due to implied constraints, the possible values presented for some scientific measurements are not continuous but granular." In a related context, the above observation can be extended into *p* values where only a limited number of discrete rank orders are possible.

For simplicity, this manuscript only considers the nomenclature of the Mann-Whitney *U* test, a non-parametric test for comparing two independent groups which are not normally distributed, although the Wilcoxon rank sum test is identical in this context. Consider two groups of $N = 6$ scores designated A and B, with values which allow them to be assembled into increasing rank order:

$$[A, A, B, A, B, A, A, B, A, B, B, B]$$

While infinite sets of values are possible, there are only $12!/6! \times 6!$ multinomial rank orders (i.e., 924 solutions) possible for any given set of 12 scores. To calculate the Mann-Whitney *U* statistic, these orders are devolved into rank sums, specifically in this example:

Table 1: Example data set

A	1	2		4		6	7		9		
B			3		5			8		10	11

i.e., A $= (1 + 2 + 4 + 6 + 7 + 9) =$ 29, and B $= (3 + 5 + 8 + 10 + 11 + 12) =$ 49.

The number of possible rank sums is significantly less than the number of possible orders. While there is only one order for the lowest and highest rank sums (respectively, $[1 + 2 + 3 + 4 + 5 + 6] = 21$ and $[(]7 + 8 + 9 + 10 + 11 + 12] = 57$), other rank sums can be reached by many unique orderings; for instance, a rank sum of 39 has 58 unique orderings which would all be calculated identically, return an identical *z* score, and finally an identical *p* value. The total number of possible rank sums is therefore 21 through 57 inclusive (i.e., there are 37) - but as the lower of the two rank sums is typically used in the calculation, this only represents 19 total calculable *z* scores and therefore 19 distinct *p* values - sufficiently few that all can be listed here (Table 2). When rounded to 2 decimal points, 17 distinct values are possible from a total of 100 reportable figures (i.e., $p = 0.01$ to $p = 1$, discounting values reported greater than or less than a threshold). (Because 0.01 is represented 3 times, only 17 unique values are represented at 2 decimal points.) At 3 decimal points, all 19 values remain distinct but are now drawn from a pool of 1,000 reportable figures (i.e., from $p = 0.001$ to $p = 1$, with similar constraints). Thus, the smaller *p* values are more densely represented (i.e., in particular, the 2-decimal-point *p* values for "significant" or "suggestive" are well represented, e.g., 0.01, 0.02, 0.03, 0.05, 0.07 and 0.09), while higher *p* values are distributed fairly sparsely. Alternatively, at 3 decimal points, only 4 *p* values are possible out of the 501 over $p = 0.5$. Thus, if a *p* value at 3 decimal points was randomly, mistakenly, or dishonestly selected from the possible values for either method of decimal truncation, it would be quite the stroke of luck if that *p* value were possible. This forms the basis for a consistency test that can be easily deployed on a manuscript reporting the *p* values from Mann-Whitney *U* tests. As with the GRIM test, there is a straightforward relationship between the number of potential *p* values in a relevant test output and the sample size. It is straightforward to expand the above to a general rule, which is also necessary as enumerating all possibilities by computational or simulation methods fails; for $N = 30$ unique ranks, there are 2.7×10^{32} (or 30!) unique rank orders within two groups of $N = 15$ to calculate; by $N = 60$, there are 8.3×1081 unique rank orders. Thus, it is more straightforward to consider the rank sums. For 2 even-numbered groups of equal size (and thus also an even-numbered overall N), the lowest possible rank sum is $1 + 2 + 3 + 4 + 5 + ... + N/2$, and the highest possible rank sum is $N/2 + 1 + N/2 + 2 + N/2 + 3 + ... + N$. Every rank sum between these numbers can be represented by the difference between two triangular numbers thus: $\left(\frac{N}{2} + 1\right) \times \frac{N}{4} - \left(\frac{N}{2} + 1 + N\right) \times \frac{N}{4}$ which simplifies to $N^2/4$. As both highest and lowest rank sums can both be included, the total is $1 + (N^2/4)$.

In the previous example where $N = 12$, this is given by possible sums of 21 through 57 (inclusive) which can now be represented as $(144/4) + 1 = 37$. This represents 19 discrete *p* values inclusive (although, as noted previously, because 0.01 is represented 3 times, there are only 17 unique values). Due to the squared term, the number of possible *p* values rises quickly.

To establish a general case, consider a series of measurements 1 through N of which P are in the first sample and $N - P$ are in the second. The lowest possible rank sum is $1 + 2 + 3 + ... + P$ and the highest possible rank sum is $(P + 1) + (P + 2) + (P + 3) + ... + N$. Thus, the difference is: $(P + 1 + N) \times N - P^2 - (P + 1) \times P^2$ which simplifies to $\frac{N^2+N}{2} - (P^2 + P)$. Again, the previous sums of 21 through 57 which can be represented as $(122 + 12)/2 - (62 + 6) + 1 = 37$ (inclusive).

As there are clear constraints on the number of *p* values possible, back-calculating the *U* values from the presented group sizes and the *p* values should yield very close approximations for whole numbers in cases of real *p* values, and reveal granularity errors if *p* values fall between those possible when granular *U* values are transformed into *z* scores and finally to *p* values.

From this point, we can draw several illustrative examples from the medical literature to develop this observation in the real world. While *t* tests are robust to violations of normality at larger sample sizes, samples in medical journals are often small and expected to be non-normal. The use of the Mann-Whitney-Wilcoxon test was previously surveyed in medical journals, and was found in 30% of relevant papers (Kühnast and Neuhäuser 2008). In fields such as plastic and reconstructive surgery, the test is commonplace; searching in the journal *Plastic and Reconstructive Surgery* for "Mann-Whitney" or "Wilcoxon" in the last 12 months yields 128 individual papers and 2,055 papers for the whole journal. Four studies were chosen opportunistically to highlight different observations relevant to test development.

Table 2: Mann-Whitney *U* *p* values for $N = 6$ vs $N = 6$

Unique *p* values	2 dec. points
0.005074868	0.01
0.008239019	0.01
0.013065227	0.01
0.020240571	0.02
0.030638988	0.03
0.045327562	0.05
0.065552161	0.07
0.092695803	0.09
0.128205275	0.13
0.173485468	0.17
0.22976627	0.23
0.297953062	0.30
0.378477593	0.38
0.471169998	0.47
0.575173532	0.58
0.688920556	0.69
0.810181236	0.81
0.936186293	0.94
1.000000000	1.00

Methodology for construction

For any two groups with n_1 and n_2 subjects, the *U* value can be used to determine the *p* value. The *U* value itself is only occasionally reported, but it is possible to reverse-engineer. As above, *U* values are integers when there are no ties but can be half-integers in the presence of tied values, and they are not unique to individual ranks; any ranking that produces a given *U* value for a set of data will have the same *z* score, and therefore the same *p* value. It is trivial to check *U* values given in isolation by back-calculation in any computational platform (this can be easily managed in, e.g., Excel, R, MATLAB, or Python). However, a plausible dataset which produces that *U* value is also required for (a) testing statistical software which does not permit the above (e.g., SPSS, JASP) and (b) inspecting how the relevant data might be structured, as it may be variously plausible or implausible. We developed a general approach that relies on employing simulation-based methods to generate rank configurations for two independent samples that yield a specific target value of the Mann-Whitney *U* statistic, enabling precise comparisons between exact and approximate *p* values. An implementation for this is the attached R code, *U*-Bend, which exploits the fact that the sum of ranks assigned to group 1 is uniquely determined by the desired Mann-Whitney *U* statistic using the identity $R_1 = U + n_1(n_1 + 1)/2$. Each iteration samples a random subset of ranks from the set $\{1, ..., n_1 + n_2\}$ and checks whether their sum equals the target R_1, returning the configuration that does. For fractional *U* values (which simulate mid-ranks caused by ties), an artificial tie is introduced by manually adjusting two ranks to identical fractional values (e.g., both to 5.5). For each rank configuration generated and *U* value obtained, corresponding *p* values are ascertained using both exact *p* values and also asymptotic approximations of the *p* values, as this latter behavior mimics the default calculation in software packages like SPSS. *U*-Bend code is available on GitHub (github.com/drg85/GRIMU).

3 EXAMPLES

3.1 Wagner et al. (2025)

Fronto-orbital distraction osteogenesis (FODO) is a surgical technique for correcting skull deformities. It is often used to treat craniosynostosis—the premature fusion of the skull bones—and so is typically but not exclusively conducted on children. Wagner et al. (2025) recently a series of outcomes of endoscopically assisted FODO with traditional coronal incision. To compare the perioperative characteristics of these procedures, they used the Mann-Whitney *U* test.

In this example, the researchers included 9 endo-FODO patients and 18 traditional coronal incision patients, and the perioperative characteristics return the following *p* values to 3 decimal points:

- Age at surgery $p = 0.999$
- Operative time $p = 0.607$
- Total anesthesia time $p = 0.029$
- Estimated blood loss (mL/kg) $p = 0.012$
- Blood transfusion (mL/kg) $p = 0.001$
- Length of hospitalization $p = 0.678$

There are several interesting characteristics of these *p* values, but it is most instructive to start with the example given by operative time ($p = 0.607$). Back-calculating using the available information is straightforward using the central definitions of the test.

- $p = 0.607$ can be simply transformed into a *z*-statistic of 0.514 (for instance, using pnorm in R or NORM.S.INV in Excel)
- The expected mean is $(n_1 \times n_2)/2$, which is 81.
- The expected SD is $\sqrt{(n1 \times n2 \times [(n1 + n2 + 1)/12]} = 19.44...$
- The formula for the *z*-statistic is $z = (U - M_U)/SD_U$
- This can be rearranged to $U = z \times SD_U + x$

Combining all of the above, we will find the *U* value calculated at 91.00031025—and, hence, that a *p*-value of 0.607 is possible when $U = 91$ (or 91.5 if using a continuity correction). As the Mann-Whitney test uses the lower of the two *U* values, the converse value (71) is used.

This raises the crucial question of tied ranks, which occur if any of the values in the samples are identical. The likelihood of ties is variable. In continuous data with very high precision, less rounding, and a small sample size, ties are incredibly unlikely. By contrast, in discrete data with low precision, higher truncation, and a large sample size, ties are almost inevitable. Presumably, measuring operative time to the minute is somewhere in between these two scenarios (endoFODO 104 [95-112]; FODO 114 [92-122]; both median [interquartile range]). Ties may extend the number of possible *p* values, as they allow the possibility of *U* values to be given as half units and not as whole units. This is also straightforwardly related to the typical continuity correction given for this test, as above; the traditional continuity correction adds a constant to the value of *U*. Combining these factors, we can see that the data is not only consistent, but there are actually two possible *U* values for the data to be correct: $U = 71$ without a continuity correction, and $U = 70.5$ with a continuity correction. As the statistical procedure is presumably the same within all tests, it is therefore possible to list all of the *U* values under both conditions (see Table 3). In doing so, co-comparing different possibilities between the presence or absence of the continuity correction may give an analyst some insight into what procedure was used, as may the idiosyncrasies of the software used to calculate the Mann-Whitney *U* test.

Table 3: The effect of continuity correction (CC) on identical *p* values.

Analysis	No CC	CC
Age at surgery	$U = 81$, $p = 1.000$	$U = 80.5$, $p = 1.000$
Operative time	$U = 71$, $p = 0.607$	$U = 70.5$, $p = 0.607$
Total anesthesia time	$U = 38.5$, $p = 0.0288$	$U = 38$, $p = 0.0288$
EBL (estimated blood loss; mL/kg)	$U = 32.5$, $p = 0.0126$	$U = 32$, $p = 0.0126$
Blood transfusion (mL/kg)	$U = 13.5$ through 19, $p = 0.001$	$U = 13$ through 18.5, $p = 0.001$
Length of hospitalization	$U = 73$, $p = 0.681$	$U = 72.5$, $p = 0.681$

This raises no problematic inconsistencies but does contain several intriguing features.

1. The age data is given as $p = 0.999$ rather than 1. This is presumably due to an approximation given by the statistical software.
2. Ties are definitely present, because in both the presence or absence of continuity correction, there are half units present. Presumably, the authors use the same continuity correction between all tests, and thus we cannot determine if a correction was used.
3. A value is inconsistent, but by an incredibly small increment: $p = 0.678$ is best approximated at $p = 0.681$. This is a deviation not accounted for by rounding of the stated *p* value; if the *p* value rounding is accounted for, *U* values are returned from $72.914 - 72.941$. This narrow range does not contain a whole or half integer number. However, this is an extremely small discrepancy, and it occurs singly.
4. If back-calculating, values like $p = 0.012$ contain the same principle as the RIVETS test (N. J. L. Brown and Heathers 2019)—that the range of potential underlying values represented by the rounded value reported in a paper is significant. In this case, the analytical solution to $p = 0.012$ is 32.158, which is clearly incorrect. But at this level of truncation, the interval between $p > 0.0115$ and $p < 0.0125$ is valid and should be checked. This returns possible *U* values from 31.867 to 32.439, which contains $U = 32$ and thus returns 0.011725883… which truncates to 0.012.
5. The substantial "bunching" effect expected from the general case can be observed here at low *p* values, where several *U* values give consistent values as the entire interval from $p > 0.0005$ and $p < 0.0015$ will truncate to $p = 0.001$. In fact, there are 12 unique *U* values which return $p = 0.001$.

In all, there are no problematic elements from Mann-Whitney *U* results in this paper. The small inconsistencies found were presented to the corresponding author, who did not reply. (Other data presentations were not checked.)

3.2 Brouwers et al. (2024)

A vascularized composite allograft is the transplantation of a multitissue body part (skin, muscle, bone, etc.) such as a hand or face from a donor to a patient. Brouwers et al. (2024) present a porcine model for comparing muscle injury when transplanting limbs cooled on ice vs. perfused via a machine typically used for solid organ transplantation, in the hope of extending the usable surgical life of donated body parts. To compare the perioperative characteristics of these procedures, they used in part the Mann-Whitney *U* test.

In this example, there are $N = 8$ static cold storage (iced) and $N = 8$ machine-perfused pig limbs, and the baseline characteristics include the following *p* values to decimal places as necessary (Table 4):

Table 4: Sample of *p* values vs *U* values as shown in Table 2 of Brouwers et al. (2024)

Analysis	Listed *p* value	Without continuity correction
Harvest	$p = 0.636$	$U = 27.5$, $p = 0.636502$…
Warm ischemia time before storage	$p = 0.005$	$U = 5$, $p = 0.0045$… or $U = 5.5$, $p = 0.00538$…
Warm ischemia time before reperfusion	$p = 0.172$	$U = 19$, $p = 0.172$
Limb weight before intervention	$p = 0.916$	$U = 31$, $p = 0.916$
Temperature before intervention	$p = 0.171$	$U = 19$, $p = 0.172167$…

Again, some features of the above are interesting.

1. The barrier for where multiple *U* values return identical rounded *p* values obviously has a threshold; in this case, $U = 5$ and $U = 5.5$ just return *p* values which round up and down respectively to 0.005. The amount of granularity is dependent on the sample size present—in this case, at $p = 0.005$ (3 decimal places), at 2 groups of $N = 8$, successive *U* values are just close enough together to contain consistent *p* values. If we assume $p = 0.005$ is correctly reported, but we double the group sizes to a hypothetical $N = 16$ for each, *U* has three consistent values (53, 53.5, 54). If we double it again ($N = 32$), *U* has 10 consistent values, 300.5 to 305 inclusive. As this is extremely dependent on sample size, a heuristic for inspecting papers can be given as: **"No meaningful granularity is present past a low significance threshold."**
2. A very minor inconsistency is observed ($p = 0.636502$… instead of $p = 0.636$) which may be due to rounding.
3. A very unusual inconsistency is observed where "WIT before reperfusion" and "temperature before intervention" are $p = 0.172$ and $p = 0.171$ respectively. $U = 19$ returns $p = 0.172167$, which is inconsistent with $p = 0.171$. Interestingly, these two numbers presented together have a heuristic quality an experienced forensic meta-analyst can use, because they cannot co-exist. $U = [18.5, 19, 19.5]$ return $[p = 0.156, p = 0.172, p = 0.189]$, and thus a reasonable flag to begin performing back-calculation might be **"Check any non-significant *p* values drawn from the same sample which differ minimally."**
4. Not shown in our Table 3 were 4 results which reported $p < 0.001$; some of these are cases where all the ice group transplants underwent a different protocol from the machine perfusion group (e.g., the ex vivo storage time was 4 hours for ice, 24 hours for machine perfusion), and thus there was no overlap between samples. This raises the question of how "smaller than a lowest threshold" values (STALT, see, Heathers and Meyerowitz-Katz 2024) should be understood. STALT values are small *p* values (such as, say, $p = 0.00000000000004$ hidden behind the use of a smaller—than symbol, typically $p < 0.01$ or $p < 0.001$). In the correct context—and almost always in the context of a comparison of two group means in a medical or behavioral study—STALT values should be regarded as suspicious. However, the present data are quite uncontroversial: in the above example and likely others, there is a smallest absolute threshold, at $U = 0$. In our initial back-calculation, considering $U = z \times SD_{\text{expected}} + M_{\text{expected}}$, if $U = 0$, the minimum *z* score is equal to the coefficient of variation (mean/SD). Here, no value can exist below $p = .00078$, which is therefore also the underlying *p* value the paper (accurately) reports as $p < 0.001$.

While there were some incongruities in the above results, they are neither substantial nor suspicious. The small inconsistencies found were presented to the corresponding author, who did not reply (other data presentations were not checked).

3.3 Djohan et al. (2023)

Restoring sensation in the breast after reconstruction improves quality of life. Djohan et al. (2023) sought to compare improvements in sensory parameters in neurotized (via a nerve allograft and conduit) vs. non-neurotized abdominally based free flap breast reconstruction. To compare sensory restoration between these groups, they used the Mann-Whitney *U* test for the relevant variables.

Demographic, surgical, and sensory parameters were evaluated in various groups and subgroups using the software program SPSS. This allows an additional opportunity for our observations: statistical software often obfuscates or fails to report the assumptions behind individual tests which are given as built-in functions. In this case, the exact back-calculations can be compared to the equivalent values returned from SPSS using synthetic data designed to reproduce the same *U* value. As R and MATLAB return the same values (designated below as R/M), and continuity correction may or may not be present, this produces six relevant values for each *p* value. An example of an anomaly would be (Table 5):

Table 5: *p* values vs. *U* values of mean subject age in Djohan et al. (2023). No continuity correction.

Variable	Listed *p* value	Re-calculation of *U*	Re-calculation of *p* values
Mean age	0.798	$U = 518$	SPSS: $p = 0.794$ R/M: $p = 0.7987$
		$U = 518.5$	SPSS: $p = 0.799$ R/M: $p = 0.8033$
		$U = 519$	SPSS: $p = 0.803$ R/M: $p = 0.8079$

Sixteen total Mann-Whitney-derived *p* values were presented, and most were slightly different from those found from precise back-calculation. The first assumption in this case should be that this is an artifact of the software being used. As a consequence, sequences were generated that returned the above *U* values and analyzed as the authors did (i.e., in SPSS) using *U*-Bend, which resolved most of the incongruities. **Given the above, three heuristics are suggested: (1) Calculation methods provided by different software packages can produce meaningfully different *p* values for the same *U* values; (2) The specific software used by authors should be deployed to recreate results; and (3) It is critical that papers list both the software and the software version used for recreation.**

Given the above, the results present a consistent but very low-level anomaly which is likely due to the statistical method used, and so they were judged likely not problematic. The small inconsistencies found were presented to the corresponding author, who did not reply. (Other data presentations were not checked.)

3.4 Khajuria et al. (2024)

When connecting blood vessels via microsurgery, the field of view can be compromised by blood and edematous fluid. Khajuria et al. (2024) developed a surgical microsuction/irrigation device and tested its performance compared to that of conventional procedures in a rat femoral vessel model. To compare the performance characteristics of these procedures, the authors used the Mann-Whitney *U* test.

Table 6: *p* values for Khajuria et al. (2024)

Analysis	Listed *p* value
Time to completion	$p = 0.007$
Structured Assessment of Microsurgery Skills (SAMS) score	$p = 0.001$
Wiping events	$p < 0.001$

The presented data is confusing, as none of the listed values are easily replicable. From the Mann-Whitney *U* test, the following terminal values (starting from two mutually exclusive and non-overlapping groups, i.e., the minimum possible *p* value, as above; see Table 6) are observed (see Table 7).

Table 7: Minimum *p* values for $(N = 6$ vs $N = 6)$

U	*p* value
0	0.003947752
0.5	0.005074868
1	0.006485308
1.5	0.008239019

None of the values can be successfully recreated from the *U* test regardless of continuity correction: the uncorrected terminal value at $U = 0$ is $p = 0.0039$, which makes $p < 0.001$ and $p = 0.001$ impossible; likewise, there is no provided value for $p = 0.007$ (although $U = 1$, $p = 0.006458$ is close). The authors also do not report using any specific statistical package. Given the previous example, it seems good practice to re-create an analysis from non-overlapping groups i.e., [1,2,3,4,5,6] vs [7,8,9,10,11,12] within several different software packages (as suggested above), which results in the following (Table 8):

Table 8: Minimum *p* values for $(N = 6$ vs $N = 6)$

Software	Terminal *p* value (i.e., $U = 0$)	$U = 0.5$
JASP 0.19.3	$p = 0.002$	$p = 0.006$
RStudio 2025.05.01 Build 513 (using Wilcoxon)	$p = 0.005075$	$p = 0.006392$
SocSciStatistics.com	$p = 0.00512$	$p = 00652$
Python (using pythononline.net)	$p = 0.0021645021645021645$	$p = 0.006392268870767702$

The above should always be considered when very small changes to initial conditions, potential minor differences in calculatory methods, floating point calculations, internal truncation or rounding, and different final rounding schema have been observed in past investigations—as these are likely observed above in Djohan et al. (2023). The larger differences seen in the above are almost certainly related to the continuity correction used. However, no combination of factors in any software platform can produce a value lower than 0.002. As none of the included values can be reproduced regardless of the method used, it seems prudent to regard the result as flagged and proceed to further analysis of the presented data.

Additional observations on Khajuria et al. (2024) that constitute flags:

- The ethical statement is given only as "After clearance from the institutional animal ethics committee…", which does not explicitly identify any ethical documentation, the institution where the work took place, or the ethics committee that cleared the work.

- It is unclear how the duplicate measurements of the Structured Assessment of Microsurgery Skills (SAMS) scores were combined, as the study states, "Each procedure was recorded, and videos were independently rated by two blinded experts, who were not involved in the study management team, using the SAMS score." Presumably results for each rater were combined, but it is unclear how.
- The observation around time to completion (741.7 ± 203.1 sec vs. 584 ± 155.9 sec; $p = 0.007$) can be investigated easily via other means. An independent-samples t test in R returns a p value of 0.163551, and the GraphPad online calculator returns a similar result ($p = 0.1623$). A t test may not be appropriate if normality cannot be assumed, so it is prudent to try to recreate the dataset manually. As a consequence, SPRITE (see Heathers et al. 2018) was used to generate plausible samples from the mean/SD/n described, and exhaustively applied to the rounded data to generate plausible data that could produce the required p value from the Mann-Whitney U test. As per the values on the previous page, due to minor differences in methods and continuity correction, this could be resolved approximately (but not exactly) to a U value of 0.5 or 1. However, (a) in ~160 minutes of runtime, the lowest possible solution generated for U is 3, and (b) distributions which allow the mean/SD/n to coexist with a low U value are strange (see below; Figure 1), as the SPRITE procedure is generating distributions with curious and specific ordinal properties that still return the correct summary statistics.
- The other Table 1 values were not investigated with SPRITE, as their overlap is at a minimum ~4 SDs of the pooled deviation; any potential solution for these would be comparing the orders of two non-overlapping groups, produce a U value of 0, and recreate the results given above.

The inconsistencies listed above were sent to the corresponding author's email address but returned a 550 ("unknown email"). Consequently, the email was forwarded to other publicly available emails for the author, but there was no reply.

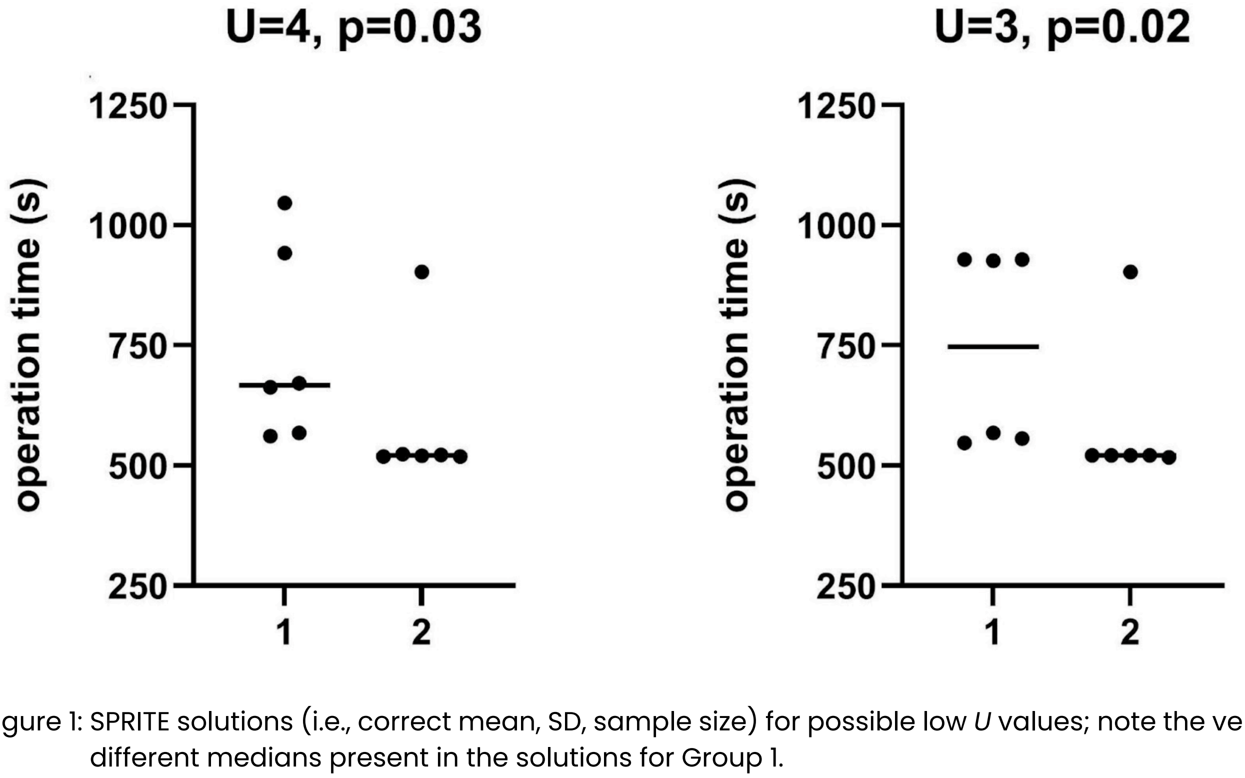

Figure 1: SPRITE solutions (i.e., correct mean, SD, sample size) for possible low U values; note the very different medians present in the solutions for Group 1.

4 DISCUSSION

In addition to those listed in boldface earlier, we offer several heuristics and general observations to keep in mind during the review of any manuscript:

- *p* values to 2 decimal places are sufficiently imprecise that they are less suitable candidates for analysis, and it is distinctly less likely that a valuable version of the above can be performed as if they are present. Happily for the deployment of the test, *p* values in medical research are commonly (a) reported to 3 decimal places and (b) calculated from small sample sizes.
- *U* values can be frequently back-calculated to be given in half units due to ties and/or a continuity correction. The presence or absence of tied ranks usually cannot be determined by simply inspecting the data in a paper.
- Very small inconsistencies in *p* values (e.g., $p = 0.171$ vs $p = 0.172$ with identical group sizes and *U* values) should be regarded as unusual at "non-significant" values, as the granularity of the underlying *p* values likely does not permit them.
- Very small *p* values are normal within medical research. As $U = 0$ has a defined *p* value, the lowest possible *p* value can be calculated. However, the corresponding *p* value in any given paper is unlikely to be reported precisely.
- Group comparisons with small *p* values likely yield indistinguishable ranks; generally, *p* values below a given significance threshold retain multiple possible rank-sums or *U* values.
- The software used to calculate Mann-Whitney *U* values can produce different *p* values (and, presumably, different *z* scores). At different points in our investigations described above, SPSS, R, Python, JASP, Excel, MATLAB, manual calculation, and online calculators were used, and any two may offer different *p* values for identically entered data. The precise continuity correction or calculatory method used may also be unclear from the documentation of these functions or packages. Software may also truncate decimals, and, if truncating, may produce different outputs from identical calculations.
- Finally, and perhaps most importantly here, the role of missing data is not discussed in the above because it is unknowable if not reported. While it feels unlikely from small group comparisons undergoing intensive medical treatment—a situation where every data point may be expensive and valuable—it certainly is possible. Missing data renders the test presented here extremely challenging to interpret without further information.

Given all of the above, some recommendations to journals can be made. In our opinion, authors should explicitly state:

1. the exact *U* (or *W*) statistic for each statistical comparison;
2. the statistical software and version used to produce it;
3. whether or not a continuity correction was applied, if known;
4. how tied values are handled, if known.

(1) and (2) in isolation would often be sufficient to reproduce the exact calculation if potential rank orders can be reconstructed.

Likewise, journal editors and reviewers should treat inconsistencies in ranked-sum *p* values cautiously but seriously. Where *p* values are judged to be impossible, a simple inspection of the data can immediately clarify any presented *p* value without doubt if it is paired with a calculatory method. In addition, one feature of rank-sum tests lends itself very well to post-publication review in medicine: the raw data does not always need to be inspected, because it can be converted into rank sums first, and the rank sums can be communicated instead without the loss of statistical detail. This conversion should be sufficient to preserve privacy or confidentiality in cases where this is a concern.

5 REFERENCES

Software

Allaire, JJ, and Christophe Dervieux. 2024. *Quarto: R Interface to 'Quarto' Markdown Publishing System*. https://CRAN.R-project.org/package=quarto. [↩]

Arel-Bundock, Vincent. 2024. *Tinytable: Simple and Configurable Tables in 'HTML', 'LaTeX', 'Markdown', 'Word', 'PNG', 'PDF', and 'Typst' Formats*. https://CRAN.R-project.org/package=tinytable. [↩]

Ben-Shachar, Mattan S., Daniel Lüdecke, and Dominique Makowski. 2020. "effectsize: Estimation of Effect Size Indices and Standardized Parameters." *Journal of Open Source Software* 5 (56): 2815. https://doi.org/10.21105/joss.02815. [↩]

Boshnakov, Georgi N., and Chris Putman. 2024. *Rbibutils: Read 'Bibtex' Files and Convert Between Bibliography Formats*. https://CRAN.R-project.org/package=rbibutils. [↩]

Chamberlain, Scott, Hao Zhu, Najko Jahn, Carl Boettiger, and Karthik Ram. 2025. *Rcrossref: Client for Various 'CrossRef' 'APIs'*. https://CRAN.R-project.org/package=rcrossref. [↩]

Gagolewski, Marek. 2022. "stringi: Fast and Portable Character String Processing in R." *Journal of Statistical Software* 103 (2): 1–59. https://doi.org/10.18637/jss.v103.i02. [↩]

Jung, Lukas. 2024. *Scrutiny: Error Detection in Science*. https://CRAN.R-project.org/package=scrutiny. [↩]

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here. [↩]

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/. [↩]

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the metafor Package." *Journal of Statistical Software* 36 (3): 1–48. https://doi.org/10.18637/jss.v036.i03. [↩]

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686. [↩]

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook. [↩]

Meta-analysis and primary studies

Allard, Aurélien. 2018. "Analytic-GRIMMER: A New Way of Testing the Possibility of Standard Deviations." Blog post. https://aurelienallard.netlify.app/post/anaytic-grimmer-possibility-standard-deviations/. [↩]

Anaya, Jordan. 2016. "The GRIMMER Test: A Method for Testing the Validity of Reported Measures of Variability." *PeerJ Preprints* e2400v1. https://doi.org/10.7287/peerj.preprints.2400v1. [↩]

Bolland, Mark J., Greg D. Gamble, Alison Avenell, and Andrew Grey. 2021. "Identical Summary Statistics Were Uncommon in Randomized Trials and Cohort Studies." *Journal of Clinical Epidemiology* 136 (August): 180–88. https://doi.org/10.1016/j.jclinepi.2021.05.002. [↩, 3]

Brouwers, Kaj, Anne Sophie Kruit, Dominique van Midden, Her J. H. Zegers, Jonne Doorduin, Erik Koers, Stefan Hummelink, and Dietmar J. O. Ulrich. 2024. "24-Hour Ex Vivo Hypothermic Acellular Perfusion of Porcine Forelimb: A 7-Day Follow-up Study." *Plastic and Reconstructive Surgery* 154 (6): 1138e–1148e. https://doi.org/10.1097/PRS.0000000000011469. [↩, 7]

Brown, Nicholas J. L., and James Heathers. 2019. "Rounded Input Variables, Exact Test Statistics (RIVETS)." PsyArXiv preprint. https://doi.org/10.31234/osf.io/ctu9z. [↩, 7]

Brown, Nicholas, and James Heathers. 2017. "The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology." *Social Psychological and Personality Science* 8 (4): 363–69. https://doi.org/10.1177/1948550616673876. [↩, 3, 3]

Carlisle, J. B. 2012. "A Meta-Analysis of Prevention of Postoperative Nausea and Vomiting: Randomised Controlled Trials by Fujii Et Al. Compared with Other Authors." *Anaesthesia* 67 (10): 1076–90. https://doi.org/10.1111/j.1365-2044.2012.07232.x. [↩, 3]

Djohan, Risal, Isis Scomacao, Eliana F. R. Duraes, Rebecca Knackstedt, Rachel Mangan, and Graham Schwarz. 2023. "Sensory Restoration in Abdominally Based Free Flaps for Breast Reconstruction Using Nerve Allograft." *Plastic and Reconstructive Surgery* 151 (1): 25–33. https://doi.org/10.1097/PRS.0000000000009773. [↩, 9, 10]

Heathers, James. 2025. "An Introduction to Forensic Metascience." https://doi.org/10.5281/ZENODO.14871843. [↩, 3]

Heathers, James, Jordan Anaya, Tim van der Zee, and Nicholas J. L. Brown. 2018. "Recovering Data from Summary Statistics: Sample Parameter Reconstruction via Iterative Techniques (SPRITE)." *PeerJ Preprints*. https://doi.org/10.7287/peerj.preprints.26968v1. [↩]

Heathers, James, and Gideon Meyerowitz-Katz. 2024. "'Yes, but How Much Smaller?' A Simple Observation about p-Values in Academic Error Detection." Center for Open Science. https://doi.org/10.17605/OSF.IO/2SP5B. [↩, 8]

Khajuria, Ankur, Hyung Hwa Jeong, Theodora Papavasiliou, Stelios Chatzimichail, and Joon Pio Hong. 2024. "Application of a Microsuction Background Device for Microanastomosis in a Rat Femoral Vessel Model." *Plastic and Reconstructive Surgery* 153 (1): 91e–94e. https://doi.org/10.1097/PRS.0000000000010512. [↩, 9, 10]

Kühnast, Corinna, and Markus Neuhäuser. 2008. "A Note on the Use of the Non-Parametric Wilcoxon-Mann–Whitney Test in the Analysis of Medical Studies." *German Medical Science* 6: Doc02. https://pmc.ncbi.nlm.nih.gov/articles/PMC2703264/. [↩, 4]

Schumm, Walter R., Duane W. Crawford, Lorenza Lockett, Abdullah AlRashed, and Asma Bin Ateeq. 2025. "Research Anomalies in Criminology: How Serious? How Extensive over Time? And Who Was Responsible?" *Accountability in Research* 32 (1): 22–58. https://doi.org/10.1080/08989621.2023.2241127. [↩, 3]

Wagner, Connor S., Matthew E. Pontell, Carlos E. Barrero, Lauren K. Salinero, Gregory G. Heuer, Jordan W. Swanson, and Jesse A. Taylor. 2025. "A Comparison of Endoscope-Assisted and Open Frontoorbital Distraction for the Treatment of Unicoronal Craniosynostosis." *Plastic and Reconstructive Surgery* 155 (1): 160e–170e. https://doi.org/10.1097/PRS.0000000000011147. [↩, 6]