Title

A simulation study showed that linear regression and Mann-Whitney test can be used to analyse the Days Alive and at Home by day 30 (DAH30) outcome in a randomized controlled trial.

Authorship

Jonathan Alistair Cook

Professor of Clinical Trials & Medical Statistics, Fellow of St Hugh's College

Deputy Director of Oxford Clinical Trials Research Unit

Centre for Statistics in Medicine

Botnar Institute for Musculoskeletal Sciences

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences

University of Oxford, UK

Tel +44(0)1224 223450

@profjacook

What is new?

- The days alive and home at 30 days (DAH30) is a new outcome measure with usual properties which has been proposed for use in RCTs particularly in surgery, its properties are considered using a simulation study.
- It was shown that linear regression and Mann-Whitney U test are valid methods to analyse DAH30 in a RCT in the presence of an additive treatment effect across a range of treatment effects and sample sizes for plausible DAH30 distributions (with lower zero-inflation levels).
- Mann-Whitney had higher rejection rate for small treatment effects, and observable effects are constrained by the properties of the DAH30.

Abstract (260 words – excluding Plain English Summary)

Objective: The aims of the work were to consider the properties of the DAH30 from a statistical perspective, and to conduct a simulation study exploring the use of simple (unadjusted) linear regression and Mann-Whitney test as the method of analysis reflect realised analysis options.

Study Design and Setting:

The days alive and at home by day 30 (DAH30) has been proposed a patient-centric outcome, and clinically relevant outcome suitable for clinical trials. It has unusual statistical properties, and suitability of standard statistical analysis methods is unclear. The properties of DAH30 were reviewed. Simulations based upon 1:1 allocation in a RCT based upon empirical data were conducted reflecting different additive and realised (reflecting the DAH30) treatment effects, sample sizes and distributions with varying and central locations and zero value level. A variety of metrics were used to assess performance (including bias, coverage, rejection rate).

Results:

Linear regression provided a valid estimate of the unadjusted average treatment effect with an additive treatment. This was confirmed in terms of bias, estimation of variance, rejection rate in the absence of an effect, and coverage of the 95% confidence interval for the true realised effect. Mann-Whitney provided greater (power) than linear regression in some situations.

Conclusion:

Simple linear regression is a reasonable analytic option for the DAH30 for estimating the average treatment effect in the RCT cohort (i.e. an intention to treat, or “treatment policy” estimand) where

zero-inflation is relatively low. Mann-Whitney test in some circumstances (small effects and smaller samples sizes) provides better ability (like for like) to detect a difference between the groups.

Plain English Summary:

Simple linear regression can be used to analyse DAH30 outcome in a randomised trial for a range of scenarios which were considered in this study (including relatively proportions of zero values). The DAH30's properties affect the treatment effect than can be estimated. Mann-Whitney test offered better ability to detect a difference of a smaller magnitude for smaller samples sizes.

Keywords

Randomized controlled trial; surgery; days alive and at home; sample size; simulation study

Word count

3441

Article

Background

The days alive and at home by day 30 (DAH30) has recently been proposed a patient-centric outcome, and clinically relevant outcome suitable for clinical trials.[1-3]. However, the DAH30 has unusual statistical properties, and correspondingly which statistical analysis methods are appropriate is not obvious. The DAH30 is typically defined as a simple count of the number of days an individual is at "home" following the receipt of a surgery. Similar names have been used for essential the same outcome (e.g. days alive and out of hospital), and with a variety of related acronyms (DAOH30, DAAH30, and DAAOOH28) sometimes with minor differences (e.g. timing) in definitions as well more substantive and consequential variations (i.e. penalising or not for death).[1-3] A wider set of similar related outcomes have also been defined.[4] The aims of the work were to consider the properties of the DAH30 from a statistical perspective, and to conduct a simulation study exploring the use of simple (unadjusted) linear regression and Mann-Whitney test as the method of analysis reflecting common analysis options.

DAH30

Days Alive and at home timeframes and definitions (e.g. some exclude the first day) have varied though 30 days with a discrete 0 to 30 range appears to be the most common (reflecting emergency and surgery settings); here each unit corresponding to a day in time. However, if a patient dies within the 30-day period, the DAH30 is scored as zero (not as the number of days prior to the death). The DAH30 outcome is therefore a type of bicomponent outcome, where one component is binary (death) and the other is a count (days at home within the 30 day period).

The DAH30 has the potential for very unusual distribution forms. Figure 1 shows what is at least theoretically possible in terms of the extremes and some possible intermediate shapes. Some distribution form while possible are very unlikely in practice (e.g. all with exactly 30 days requiring

everyone to be at home on the first day after surgery). However, distributions with most values close to 30 days[5] or even zero if hospital length of stay was typically over 30 days. Myles and colleagues[1] recently published DAH30 data from 7 studies (including 3 RCTs) for patient undergoing various elective and emergency surgery. Mortality was 3.1% and the distribution had a long tail with a median of 24 days (and mean of approximately 21 days) –Figure 2.

Statistical analysis of the DAH30

A variety of statistical methods have been used to analyse days at home data including Mann-Whitney test, t-test/linear regression, and Poisson regression[6-8] including as a time-to-event outcome.[4, 9, 10] The potential range of distributions (Figures 1 and 2) and what might viewed as a more realistic subset of distributions do not naturally fit any standard distribution. Obviously the data is not normally distributed, as it is discrete and bounded between 0 and 30. The shape (even without zero-inflation) does not readily fit Poisson or even negative binomial distribution, nor a log-normal distribution. This is due to the potential for a large mass over a small range representing a substantial proportion of the observations yet with a long tail (e.g. Figure 2).

A common option where an obvious parametric analysis is not apparent, might be to use Mann-Whiney U test though it does not estimate a treatment effect on the original scale.[11] In principle a mean difference would present an intuitive effect size measure despite concerns about the shape of the distribution. Furthermore, a t-test has been shown to work surprisingly well for discrete ordinal data of various forms.[12, 13] Additionally, recent work has reemphasised the findings noted early in the statistical literature on randomised trials[14-16] that a valid estimate of the average treatment effect (ATE) can be anticipated asymptotically across a wide range of scenarios. The use of a t-test (equivalently) linear regression therefore deserved further consideration for the analysis of the DAH30.

Simulation study

Simulated scenarios

A simulation study of the analysis of the DAH30 using simple linear regression (equivalent to independent t-test (equal variance), and Mann-Whitney in a 2-arm parallel group individually randomised trial with 1:1 allocation. The reference distribution for the simulations mimicking the one reported by Myles et al[1]. The reference distribution was altered in two ways to produce different distributions. First, the mass of the non-zero distribution was shifted up and down in increments of 2 points from to -10 to +6, and also -20 as an extreme shift down (i.e. 9 shifted locations). For example, for the 2-up distribution the proportions for DAH values 1-28 in the reference distribution was used for DAH30 values 3 to 30 in the two-up distribution. The remaining tail of the distribution (DAH30 1 and 2 for this 2-up distribution) was extended as required at similar probability level to the nearest proportion to cover the other values in the DAH30 range (1 to 30). The respective non-zero values were then rescaled to 1 minus the zero-value proportion (i.e. 0.97 to maintain the zero-value proportion at 0.03) to maintain the relative zero and non-zero proportion values in the distribution. The reference distribution and the nine corresponding 3% zero-value distributions with shifted locations are shown in Figure 3. Second, the zero-value proportion was altered for each of these 10 distribution locations while maintaining the shape (and relative density) of the non-zero DAH values. The Zero percentage in the reference distribution was 3%. Plausible range for 30 days post-surgery would seem to be approximately 0 to 10%.[17-20] for a range of settings. Along with the reference distribution level (3%), four other levels were assessed (0,1,5 and 10%). To generate distributions with different proportions of zero values, the proportion of zero values was changed to 0, 0.1, 0.05 or 0.1 and the respective non-zero values were then rescaled to 1 minus the zero-value proportion (i.e. 1.0, 0.99, 0.95, or 0.9) for 0, 1, 5 and 10% zero values). This maintained the relative shape of the non-zero part of the distribution while allowing different zero value proportions.

In total there were 10x5=50 different distributions (table 1). The reference distribution and the corresponding 4 distributions (distributions 1-5 in Table 1) with the same location but varying zero-value proportions are shown in Figure 4. For each of the 50 different distributions a number of simulation scenarios (combination of distribution, N per group and additive treatment effect) were produced. 2N observations were randomly sampled from the respective distribution was generated using a random number between 0 and 1 to select the relative DAH30 score given the corresponding probability distribution for each observation. These were evenly split to mimic (1:1) individually randomised trial into N control group and N treatment group observations. An additive treatment effect (δ_a) applied that ranged from -6 to 6 in increments of one to the observations in the treatment group. This included a 0 value i.e. absence of a treatment effect.

Table 1 – Simulated distributions

No	**Location**[1]	**Median (IQR)**	**Mean (SD)**	**% Zero values**	**N per group (1:1)**	**Additive Treatment effect**[2]
1-5	reference	24 (18,26)	21 (7.2)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
6-10	2 down	22 (17,24)	22 (6.8)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
11-15	4 down	20 (15,22)	18 (6.5)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
16-20	6 down	18 (14,20)	17 (6.1)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
21-25	8 down	16 (12,18)	15 (5.8)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
26-30	10 down	14 (11,16)	14 (5.6)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
31-35	20 down	6 (4,10)	8 (6.4)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6

36-40	2 up	26 (20,28)	23 (7.5)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
41-45	4 up	28 (22,30)	25 (7.6)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6
46-50	6 up	30 (24,30)	26 (7.5)	0,1,3,5,10	50,100,250,500,750,1000	-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6

Notes

1. Location as it differs from the reference distribution. The reference distribution is distribution 3 (3% zero values)
2. The corresponding additive treatment effect was applied universally to all observations in the treatment group.

Values after applying the additive treatment effect to the simulation values in the treatment group were restricted to the 0 to 30 range (31 possible values) i.e. <0 values were scored 0, and >30 values were scored 30. The impact on the DAH30 scale is of a variable treatment effect at the observation (individual) level. As a consequence, the resultant "realised" treatment effect (δ_r) was therefore less than the corresponding additive treatment value, sometimes substantially so. Additionally, the expected zero-value proportion in the treatment group across the scenarios varied (from 0 to 57%) depending upon distribution used and whether the treatment effect pushed the distribution towards or away from the bottom of the distribution. See Supplementary Table S1 for the realised treatment effects. Given the probability distribution function, the probability of $Prob_{MW}¿$ can be calculated exactly. In the absence of a treatment effect it is 0.5. Simulations were carried out for each of the distributions varying the size of the two randomised groups. The combination of control group distribution, treatment effect and sample size lead to 3900 simulation scenarios.

Metrics of interest

The performance of simple linear regression and Mann-Whitney test were evaluated using a range of metrics (Table 2). Linear regression was assessed in terms of bias of the estimated mean difference, 95% confidence interval coverage of the treatment effect, and comparing the empirical standard error, to the model standard error. The rejection rate was also assessed at 5% (two-sided) significance level. Bias and coverage were calculated for the additive treatment effect and the realised treatment effect. The rejection rates between linear regression and Mann-Whitney test for a difference between treatment groups were compared. Additionally, the observed rejection rate was compared to the predicted rejection rate using the non-central t distribution sample size calculation using the realised mean difference and the corresponding group SDs.

Table 2 Simulation study metrics

Name	Definition	Analysis method of interest
Additive effect bias[1]	$\widehat{\beta}_t - \delta_a$	Linear regression
Realised effect bias[1]	$\widehat{\beta}_t - \delta_r$	Linear regression
Rejection rate[2,3]	$Prob\left(pvalue_{\hat{\beta}_t}\right) \leq 0.05$	Linear regression, Mann-Whitney test[4]
Empirical standard error	$\sqrt{Variance\left(\widehat{\beta}_t\right)}$	Linear regression
Model standard error	$\sqrt{E\left(Variance_{Model}\left(\widehat{\beta}_t\right)\right)}$	Linear regression
$Prob_{MW}$ ¿[5] bias	$Prob_{MW}$ ¿- $Prob_{true}$ ¿	Mann-Whitney test[4]
Additive effect coverage[2]	$Prob\left(Upper\,95\,\%confidencelimit_{\hat{\beta}_t} > \delta_a \wedge Low$	Linear regression
Realised effect coverage[2]	$Prob\left(Upper\,95\,\%confidencelimit_{\hat{\beta}_t} > \delta_r \wedge Low$	Linear regression

Notes:

1. the control group is used as the reference group where relevant e.g. $\widehat{\beta}_t$ is the estimated mean difference and a positive number indicates the intervention mean group is higher than the control, and a negative

number that the intervention group mean is correspondingly lower. Where $\delta_a = 0$, the realised treatment effect bias is the same as the additive effect bias metric ($\delta_a = \delta_{r¿}$¿.

2. A 2-sided comparison is made throughout at nominal 5% significance.

3. Rejection rate can be interpreted as the type 1 error/statistical power depending upon presence or not of a treatment effect.

4. Mann-Whitney (U) test is also known as the Wilcoxon sum rank test.

5. Also known as the area under the receiver operating curve.[11, 21, 22] The value for equivalent groups is 0.5.

Implementation

The simulations (see supplementary file Simulation code) were carried out in Stata 17.0 using a user written program that generated a sample of the required size in each group and the desired additive treatment effect in conjunction with the in-built Stata command simulate. All metrics of interest (See Table 2) were calculated in the same number of repetitions (10000) for all simulation scenarios. Summary (mean) values over all repetitions were collected and automatically saved and processed in Excel.

Results

Bias

The simulation results for bias (additive and realised linear regression mean difference estimate, $Prob_{MW}$¿) are provided in Figure 5, Supplementary Figure S1 and S2, and Tables S1-3. The magnitude of the additive treatment effect bias could be as large as 4.1 (distributions 45 and 46 and additive treatment effect of 6). In contrast where the location was centred away from the ends of the distribution, bias was far less across all additive treatment effects. In contrast for the realised

treatment effect, the estimated magnitude of bias was less than 0.05 for simulations scenarios. Similarly, the bias for estimation of the $Prob_{MW}$ ¿ was less than 0.01 across all scenarios. For all three bias measures, the respective sample size per group had negligible impact.

Variance estimation & coverage

The difference in the empirical variance of the mean difference and the model variance of the treatment effect (Supplementary Figure S3) showed excellent agreement with the difference at most magnitude of 0.09 for sample size of 50 per group but typically around 0.05 for this size. For larger sample sizes the difference was within 0.02. The zero-value proportion has very modest influence on the estimates. Coverage of 95% CI for the additive treatment effect varied markedly by the magnitude of effects (Supplementary Figure S4 and Table S4). Difference from the nominal 0.95 level varied from within a reasonable margin of error (0.005) to an almost complete loss of coverage (<0.01). Nominal level was more commonly achieved for smaller sample sizes. Coverage was best when the distribution was located away from the measurement range ends, and when fewer zero % values. For example, for the 4 down and 0% zero values distribution, the absolute loss of coverage never exceeded 0.005 (i.e. 0.5%). In contrast, coverage of the realised treatment effect (Supplementary Figure S5 and Table S5) was markedly better and closely matched with nominal level across all scenarios. The estimated coverage was within +-0.005 (i.e. 0.5%) of the 0.95 target except for a few occurrences with n=50 per group which marginally exceeded this, and never for the larger sample sizes.

Rejection rate

The difference in the rejection rate between Mann-Whitney and linear regression are given in Supplementary Figure S6 and respective rejection rates in Supplementary Tables S6 and S7. The

relative loss could be as large as 0.87 for n=100 (distribution 50 and additive effect of -1) for detecting of a real effect. Substantial losses were mostly observed for smaller additive treatment effects (+-1 or 2) and sample sizes (n=50 or 100). Where the sample size was n=500 or more per group, and the magnitude of additive treatment effect was 3 or more the difference was commonly <0.01. However, a reduction in the rejection rate for simple linear regression of around 0.40 was possible for n=1000. The predicted rejection rate compared well against the empirically observed rate across all scenarios (Supplementary Figure S7). Estimated rejection rate were typically within +-0.02. Larger sample sizes agreed even more closely with the predicted value. Rejection level rate in the absence of a treatment effect ("significance level") was well-controlled across all scenarios (always within 0.006 of the nominal 0.05 level).

Discussion

Simple linear regression is an appropriate analytic option for the DAH30 for estimating an ATE (i.e. an intention to treat, "as randomised" group comparison or "treatment policy" estimand) in the scenarios considered (ordinal variable with restricted range and relatively low zero inflation). The results showed that despite a clearly non-normal distribution, an ordinal count outcome (constrained at both end) with zero inflation, linear regression (unadjusted) provided a valid estimate of the ATE in the presence of an additive treatment effect. This was evident in terms of bias, estimation of variance, and coverage of the 95% confidence interval for the true empirical effect. Furthermore, a standard statistical sample calculation can be used. Though larger sample sizes had relative advantages in terms of rejection rate for small magnitudes of realised treatment effects, the findings regarding bias, estimation of variance and coverage held for the smaller sample sizes considered. This study provides reassurance against concerns about estimation of the treatment effect variance using simple regression when the number of observations is smaller even

though the estimate is known to be asymptotically consistent.[23] The Mann-Whitney test in some circumstances (small effects and smaller samples sizes) provided better ability (like-for-like) to detect a difference between the groups as suggested by others.[12, 24] A unbiased point estimate of the probability that an observation from one treatment group is higher than the other can also be generated with a corresponding confidence interval can be produced[21] (though this was not assessed). Nevertheless, one might not consider it a useful treatment effect to estimate particularly as it is not on the outcome scale, and thus favour a mean difference.[11, 12, 21] An advantage of this simulation study was that it was based upon empirical distribution on the observed distribution in a published study.[1] Furthermore, the location, magnitude of treatment effect and % of zero values varied across what would cover most plausible clinical setting (0 to 10%)[1, 17-20].

The use of linear regression to analyse an ordinal outcome, and one with a zero inflated form, may seem surprising. Certainly currently practice seems to indicate routine application of checks of model parameters which is of questionable value.[25, 26] A few points can be highlighted regarding the estimation of the ITT ("as randomised") effect or a treatment policy estimand as it might more recently be referred to.[27] First the robustness of the application of linear regression (or an ANOVA) in its simplistic form was an early, though insufficiently acknowledged, finding in statistical literature[14, 28, 29]. Second other work has shown that use of regression is valid for analysis for ordinal data even with very skewed original quality of life data.[12, 13]. Third, the limited value of any "normality" checks has been noted by a number of authors[30]. Four, practice perhaps reflects a discontinuity between conducting a RCT as the preferred design but analysing the data from it as if it were any observational study and ignoring the benefits of randomisation in terms of estimation of the ATE.[15, 16, 31]

It was demonstrated that for the scenarios considered that a standard sample size (a t-distribution and a non-central t-distribution under the null and alternative hypotheses respectively with Satterwaite's t test[32]) is adequate for calculating the required sample size. It is important to note this

finding pertains to analysing a RCT. There are some key caveats to add to this. The distribution inputs to a standard sample size need to reflect the properties of the DAH30 in terms of the target difference and the standard deviation of both distributions in the presence of a treatment effect. There are three related points to this that need to be made. First, the target difference needs to be specified as the "realised" effect one; this is the one of interest to an investigator not a simplistic additive effect. Second, the SDs under the alternative need to be correctly specified (applying the target difference for intervention group and following rules of the DAH30. Third, aside from using appropriate inputs in the sample size calculation, one needs to consider whether it is clinically plausible to observe any impact. More rapid and safe discharge is not a modest request.

Limitations

Limitations of the simulation study including the use of one general discrete distribution form, a standard RCT design (two-arm parallel group) with 1:1 simple random allocation, additive treatment effects, and only Mann-Whitney and simple regression were assessed opposed to multiple regression[23] or more complex analysis method such as linear or non-linear mixed longitudinal models. The reference distribution used was based upon published data[1] and the mass of the distribution was shifted and the zero-value percentage varied across a realistic range for many, though not all, settings (e.g. surgical and most emergency). The use of uneven random allocation combined with a more complex sample generation processes (e.g. involving one or two prognostic factors) might produce somewhat differing findings[33, 34] though 1:1 allocation is the most common.

Simple linear regression and a Mann-Whitney test, two simpler analysis methods, were assessed in the simulations study though they reflect credible options for analysing an outcome like the DAH30 particularly for the analysis of a "typical" RCT. The range of simulations, while substantial does not cover all possibilities for example very high level of zero-inflations e.g. >60%). There is no obvious analytic method for an outcome like the DAH30 which does not conform to any standard distribution

form. An alternative approach could have been to estimate the "intention to treat" (i.e. the mean difference in the outcome in the treatment groups without imposing a model form)[23] or another estimator like GEE[15, 16]. Linear regression implies an assumption of homoscedasticity (or put equivalently a simple form of treatment additivity). However, the findings here show that the estimate from simple linear regression is tolerant to a degree of heteroscedasticity in the treatment effect for DAH30 (due to a restricted range). Linear regression will tend to be used due to a desire to adjust for baseline factors (including a baseline outcome measurement though that is not relevant for the outcome of interest here). Another reason for adjustment is control for randomisation factors to ensure the appropriate confidence level (precision). How well the findings of this study translate to multiple regression with adjustment for multiple baseline covariates is uncertain though recent work suggests it may perform similarly well for realistic situations.[15, 16, 35], at least under simple (1:1) randomisation. An altogether different analytic strategy would be to consider days at home as a time-to-event estimation problem though readmissions are problematic[9, 10]. Common analysis methods (e.g. log rank test, Cox regression) are readily available though death is a competing event complicating the analysis. Furthermore, the typical estimand from these readily available analyses (e.g. hazard ratio) is much less intuitive than a mean difference.

The imposition of additivity as the form of treatment effect might be seen as a key limitation. However, as noted by Senn[36] this assumption may be less troublesome in practice as the estimate can be viewed as kind of ATE (across prognostic factors). As demonstrated here, simple linear regression is tolerant in terms of basic inferential statistics (bias and coverage) to a treatment effect that varies according to the baseline level. Here it is perhaps worth noting the restricted distribution has the value of removing the potential for "outliers" to influence estimation.

The underlying properties of the DAH30 raise questions about its value as a clinical trial outcome. Where a substantial proportion of the distribution is close to either zero (died within 30 days or not returned home) or 30 (e.g. discharged home immediately), the detrimental impact on the ability to

observe an effect in the direction of outcome score limit is marked. This is exacerbated where the treatment effect moves towards the same end of the range. Furthermore, if one believes the treatment effect may alter only the DAH30 values for those who do not die, this would lead to a proportion of the results being invariant to the effect of the treatments. A larger timeframe would likely ameliorate the zero inflation though offsetting this may be additional deaths. However, for a distribution like that observed in Myles[1] (with low mortality and the mass of the distribution not close to the extremes), the DAH30 would seem a more reasonable option given relatively low zero-inflation level.

References

1. Myles PS, Shulman MA, Heritier S, et al. Validation of days at home as an outcome measure after surgery: a prospective cohort study in Australia. *BMJ Open*. Aug 18 2017;7(8):e015828.
2. Bell M, Eriksson LI, Svensson T, et al. Days at Home after Surgery: An Integrated and Efficient Outcome Measure for Clinical Trials and Quality Assurance. *EClinicalMedicine*. May-Jun 2019;11:18-26.
3. Jerath A, Austin PC, McCormack D, Wijeysundera DN. Impact of postoperative intensive care unit utilization on postoperative outcomes in adults undergoing major elective noncardiac surgery. *J Clin Anesth*. Jun 2020;62:109707.
4. Granholm A, Anthon CT, Kjær M-BN, et al. Patient-Important Outcomes Other Than Mortality in Contemporary ICU Trials: A Scoping Review. *Critical Care Medicine*. 2022;50(10):e759-e771.
5. Hoffman C, Shah S, Mai M, Miller A, Banki F. Feasibility and Outcomes of Same-Day Surgery in Primary and Reoperative Laparoscopic Hiatal Hernia Repair. *J Gastrointest Surg*. Sep 5 2023.
6. Michael HM, Joel P, Flavia KB, et al. Post-discharge after surgery Virtual Care with Remote Automated Monitoring-1 (PVC-RAM-1) technology versus standard care: randomised controlled trial. *BMJ*. 2021;374:n2209.
7. Asthana V, Sundararajan M, Ackah RL, et al. Heart failure education in the emergency department markedly reduces readmissions in un- and under-insured patients. *Am J Emerg Med*. Dec 2018;36(12):2166-2171.
8. Andersen-Ranberg N, Poulsen LM, Perner A, et al. The Agents Intervening against Delirium in the Intensive Care Unit Trial (AID-ICU trial): A detailed statistical analysis plan. *Acta Anaesthesiol Scand*. Oct 2020;64(9):1357-1364.
9. Ball J, Løchen ML, Carrington MJ, Wiley JF, Stewart S. Mild cognitive impairment impacts health outcomes of patients with atrial fibrillation undergoing a disease management intervention. *Open Heart*. 2018;5(1):e000755.
10. Farmakis D, Parissis JT, Bistola V, et al. Plasma B-type natriuretic peptide reduction predicts long-term response to levosimendan therapy in acutely decompensated chronic heart failure. *Int J Cardiol*. Feb 18 2010;139(1):75-79.
11. Senn S. U is for Unease: Reasons for Mistrusting Overlap Measures for Reporting Clinical Trials. *Statistics in biopharmaceutical research*. 2011;3(2):302-309.
12. Walters SJ. Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. *Health and quality of life outcomes*. 2004;2(1):26-26.
13. Sullivan LM, D'Agostino Sr RB. Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. *Statistics in medicine*. 2003;22(8):1317-1334.
14. Kempthorne O. Why randomize? *Journal of Statistical Planning and Inference*. 1977/02/01/ 1977;1(1):1-25.
15. Wang B, Ogburn EL, Rosenblum M. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics*. Dec 2019;75(4):1391-1400.
16. Yang L, Tsiatis AA. Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial. *The American Statistician*. 2001;55(4):314-321.
17. Liu JK, Braschi C, de Virgilio C, Ozao-Choy J, Kim DY, Moazzez A. Predictors of poor outcomes after cholecystectomy in gallstone pancreatitis: NSQIP analysis of 30-day morbidity and mortality. *Langenbecks Arch Surg*. Dec 31 2022;408(1):5.
18. Coimbra R, Allison-Aipa T, Zachary B, Firek M, Edwards S. A comprehensive analysis of 30-day readmissions after emergency general surgery procedures: Are risk factors modifiable? *J Trauma Acute Care Surg*. Jan 1 2023;94(1):61-67.

19. Rashid A, Gupta A, Adiamah A, West J, Grainge M, Humes DJ. Mortality Following Appendicectomy in Patients with Liver Cirrhosis: A Systematic Review and Meta-Analysis. *World J Surg*. Mar 2022;46(3):531-541.

20. Blanco JF, da Casa C, Pablos-Hernández C, González-Ramírez A, Julián-Enríquez JM, Díaz-Álvarez A. 30-day mortality after hip fracture surgery: Influence of postoperative factors. *PLoS One*. 2021;16(2):e0246963.

21. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: general issues and tail-area-based methods. *Stat Med*. Feb 28 2006;25(4):543-557.

22. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med*. Feb 28 2006;25(4):559-573.

23. Freedman DA. On regression adjustments to experimental data. *Advances in applied mathematics*. 2008;40(2):180-193.

24. Vickers AJ. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Med Res Methodol*. Nov 3 2005;5:35.

25. Nielsen EE, Nørskov AK, Lange T, et al. Assessing assumptions for statistical analyses in randomised clinical trials. *BMJ Evid Based Med*. Oct 2019;24(5):185-189.

26. Nørskov AK, Lange T, Nielsen EE, et al. Assessment of assumptions of statistical analysis methods in randomised clinical trials: the what and how. *BMJ Evid Based Med*. Jun 2021;26(3):121-126.

27. Keene O. Adherence, per-protocol effects, and the estimands framework. *Pharm Stat*. Jul 21 2023.

28. Fisher RA. *The design of experiments*. Edinburgh: Oliver and Boyd; 1935.

29. White RF. Randomization and the analysis of variance. *Biometrics*. Jun 1975;31(2):555-571.

30. Schmidt AF, Finan C. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*. 2018/06/01/ 2018;98:146-151.

31. Rosenberger WF, Uschner D, Wang Y. Randomization: The forgotten component of the randomized clinical trial. 2019.

32. *Stata: Release 15. Statistical Software* [computer program]. Version. College Station, TX:; 2017.

33. Wang B, Ogburn EL, Rosenblum M. Rejoinder to "Robustness of ANCOVA in randomized trials with unequal randomization" by Jonathan W. Bartlett. 2020.

34. Freedman DA. Statistical models for causation: what inferential leverage do they provide? *Eval Rev*. Dec 2006;30(6):691-713.

35. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine*. 2008;27(23):4658-4677.

36. Senn S. Controversies concerning randomization and additivity in clinical trials. *Stat Med*. Dec 30 2004;23(24):3729-3753.