Supplemental Online Content

Widmer AF, Atkinson A, Kuster SP, et al. Povidone iodine vs chlorhexidine gluconate in alcohol for preoperative skin antisepsis. *JAMA.* Published online June 17, 2024. doi:10.1001/jama.2024.8531

This supplemental material has been provided by the authors to give readers additional information about their work.

SUPPLEMENTARY APPENDIX

Appendix A: **Note on sample size considerations**

Design: Multicenter cluster-randomized crossover two arm trial.

Primary endpoint: Difference in proportions of patients with SSIs for cardiothoracic and abdominal surgery following disinfection of the surgical site with povidone-iodine (PI) or chlorhexidine gluconate (CHG).

Null hypothesis: PI is inferior to CHG

Alternative: PI is non-inferior to CHG.

Framework for the trial

Swissnoso is an operational surveillance program, and the trial should fit into the existing operational structures and processed.

Expected total number of cardiac and abdominal surgeries in an 18 month period is approximately between 3000 and 3500.

There is a fixed number of monthly clusters of treatment assignment for each center (Basel, Bern, Zürich), with varying number of patients in each cluster.

We assumed no carryover effect, which is believed to be reasonable for cases in which one product is logistically fully replaced with another. [1]

Individual patient randomization was not feasible for this study, due to operational and cost constraints.

Definition of non-inferiority

- We defined non-inferiority in terms of difference in proportions between the proportion of patients with SSIs following disinfection with CHG (reference), p1, and the proportion of patients with SSI following disinfection with PI, p2.
- We defined the non-inferiority margin to be Δ, so that for non-inferiority we need to achieve (p1-p2)> -Δ, with the lower Wald type 95% CI limit excluding –Δ.

As baseline measurement for the crude SSI infection rate using CHG, as a guideline we used the meta-analysis in the article by "WHO's recommendation for surgical skin antisepsis is premature".[2] Since the rates were in line with those from historical Swissnoso data, the latter rates were used in the sample size calculation.

Primary analysis model

We assume the difference in proportions is assessed using a chi-square test.

We did not dimension the study to explicitly demonstrate the presence of inter-cluster or inter-period differences, nevertheless, these aspects where investigated with supplementary analyses (*eTable 1*).

Sample size calculation

We dimensioned the study based on the crude combined SSI rate for both cardiothoracic and abdominal surgery, assuming the endpoint is a weighted average of the rates from both types of surgery. The weights were estimable from the historical data from Swissnoso Surgical Site Infection surveillance in Switzerland.

Assumptions

Significance level = 5%

Power = 80, 90%

SSI:

Reference treatment	CHG p1 = 0.075
Comparator treatment PI	p2 = 0.075 (under the null hypothesis)
Non-inferiority margin	Δ = 0.025

Sample size calculated using function in package "TrialSize" in R, function "TwoSampleProportion.NIS").

Recommendation

We designed the study to have 80% power at a two-tailed significance level of 0.05 with a non-inferiority margin of -2.5%. We estimated that the study would need to enroll at least 1374 patients in each group (CHG/PI) to achieve the desired power.

Discussion of sample size calculation and assumptions

For the original sample size calculation, we assumed an individual patient randomized design not considering cluster correlations, rather than calculating sample size based on the cluster randomized design. This was the subject of considerable discussion in the trial team, and is a limitation. Nevertheless, alternative cluster-based sample size calculations have limitations of their own, especially when considering crossover designs with many changes (as was the case in our study) - see discussion below. We performed a post-hoc sensitivity analysis on the sample size calculation, taking into account potential cluster correlations.

The posthoc sample size calculation was performed using the newly (at the time the analysis was performed, 2020) available "Shiny RCT Calculator" (https://clusterrcts.shinyapps.io/rshinyapp/ [6]) in which we used the observed SSI rate in A, and estimated the correlation rho= 0.002 from the appropriate mixed effects model. For a binary outcome, and assuming a "multiple-period cluster randomized cross-over" design with two periods (AB/BA), with "exchangeable correlation" structure, variable cluster sizes (coefficient of variation =2), a baseline SSI rate on arm A of 5%, with on average 1000 patients per cluster, we would be able to detect an effect size of 1.2%, with 80% power and 5% significance level. This would imply that our study design, with many more periods/clusters is adequately powered in terms of investigating the primary endpoint.

We considered a number of alternative approaches to calculate the sample size, none of which could model our proposed design.

The above cluster based approach does not explicitly take the non-inferiority margin into account, which was not the case in our original patient level sample size calculation, and makes other assumptions not in line with the study design (e.g. continuous outcome, equal cluster sizes, two period design, no decay in correlation between periods).

There is always a balance to be struck between calculating the sample size in a relatively straightforward manner based on a number of verifiable assumptions, and using a more complex approach that follows the design more closely, but relies on more uncertain assumptions. In our study, this meant either assuming individual patient randomization and using existing methods for non-inferiority trials, or using one of the available methods for dimensioning cluster-randomized crossover trials and making assumptions concerning the period and cluster correlation coefficients.

The gold standard of study design would have been an individual patient randomized trial. Randomization at the patient level aims to achieve balance of patient baseline characteristics. In the end, and paraphrasing Giraudeau et al. [7], we argued that the "negative impact of the clustering effect on the sample size [from cluster randomization]" was balanced out by the "benefit of ….a cross-over trial"; this benefit in our study coming from the number of planned periods (planned 9 crossovers, 18 periods). One of the aims of randomization is to achieve patient balance between the arms in a trial. The analysis shows that baseline covariate balance on arms A and B were mostly achieved (refer to *Table 1* of the manuscript), albeit the numbers on each arm were less balanced than for a typical individual patient randomization of similar size (but may occur similarly in an individually randomized trial).

However, even individually randomized controlled clinical trials suffer from limitations as reviewed by Frieden TR[8]. In our study, all eligible patients were enrolled eliminating a risk of participation bias and providing external validity. Cluster randomized trials allow for a larger sample size.

Appendix B: Supplementary analyses on cluster, period and procedure effects

Methods

Supplementary analyses of the primary outcome included fitting mixed effects logistic models with SSI as endpoint, including center as fixed effect and period as random effect (e.g. following the approaches "M1" and "M2" outlined in Turner et al. [9], *eTable 1*). The estimates from a generalized estimating equation (GEE) population mean model with jack-knifed standard errors were also compared (defined as in Arnup et al. [10]); this latter model was mentioned in the original study protocol.

Results

Univariable logistic models with SSI as dependent variable and center as fixed effect identified a slightly lower infection rate in Center B (OR 0.6, 95% CI 0.4-0.9) compared to Center A (the reference level). Mixed effects model M2 with the intervention and cluster treated as fixed effects and a random effect on the period had similar estimates (OR 0.6, 95% CI 0.5-0.9).

The Breslow-Day test for homogeneity did not indicate any significant differences between SSI rates between CHG and PI after adjusting for the center. Similarly, although the SSI rate for cardiac surgeries was lower than that for abdominal surgeries (OR 0.4, 95% CI 0.3-0.6), the Mantel–Haenszel test did not indicate a difference in the SSI rates between CHG and PI following adjustment for procedure group.

Appendix C: Supplementary analysis on shortened follow-up cardiac surgeries during COVID-19.

Due to resource and time constraints, the second follow-up period was shortened for cardiac surgeries from 1 year to between 90 days and 1 year. The estimated potential effect on the number of "missed" infections was judged to be limited (10% of infections > 90 days for Picasso vs 14% for Swissnoso benchmark, p=0.6), see *eTable 2*.

Appendix D: Missing data

- From the statistical analysis plan (SAP): "If the missingness of any one variable (covariate or endpoint) in the full analysis set is more than 10%, a missing data analysis will be performed."
- There was no endpoint missingness for all analyses in the main manuscript. Variables with more than 10% missingness, AND included in the supplementary analyses (i.e. *eTable 1*) are: Hemoglobin, creatinine, C-reactive protein, smoking.
- From the SAP: Patients with missing records are identified in the data set, and compared to those without missing records to determine if there are systematic patterns of missingness in the data. To this end, a logistic regression model will be fitted with the missingness indicator as dependent variable and the relevant covariates as independent variables (i.e. those included in Table 1 of the analysis). Graphical methods will be used to visualize the missingness.

eTable 4 shows variables in Table 1 stratified according to missingness. Missingness patterns are shown in *eFigure 8.*

In summary, patients with missing data have/are:

- A higher infection rate
- More likely to be from Center A
- More likely to be female
- Younger
- Slightly lower hemoglobin
- Slightly elevated creatinine
- More likely to be diabetic
- More likely to have hypertension
- Higher BMI
- Less likely to have a lower Americal Association of Anesthiologists score
- Receive antibiotic prophylaxis later
- Shorter duration operation
- More likely an endoscopic procedure

- More likely to be clean-contaminated category
- Much more likely abdominal surgery

The adjusted model provides insights into which other auxiliary variables should be included in the imputation model since they may be related to the missingness process.

Multiple imputation

Methods

SAP: “Methods for handling missing covariate and endpoint data are defined in section 7.5: Baseline variables will be multiply imputed if the missingness is more than 10% in the full analysis set. Multiple imputation will be performed assuming missing at random (MAR) using the MICE packaged in R, i.e. using multiple imputation using chained equations.

Missing SSI (yes/no) information will be multiply imputed using the MICE package in R assuming MAR if the missingness is above 10% for the full analysis set.”

Comment: There was no endpoint missingness.

A total of 50 completed data sets will be multiply imputed, and then the appropriate analysis model will be fitted to these complete data sets, with Rubin’s rules used to calculate point estimates and their standard errors.

Once this process is complete the results from the complete case analysis using the per protocol analysis set will be compared to the results from fitting the appropriate analysis model to the multiply imputed data sets. If the point estimates of the covariates in the analysis model from the analysis with the multiply imputed data sets are outside of the 95% confidence intervals of the estimates from the complete case analysis with the analysis model, it will be concluded that further sensitivity analyses will be required to investigate plausible departures from the missing at random assumption. Such sensitivity analyses will be defined in detail post hoc depending on the variable, but will be based on either delta, or reference based, multiple imputation methods, whichever seems most appropriate.”

Results

- We multiply impute missing data for the following variables: Hemoglobin, creatinine, C-reactive protein, smoking
- The substantive model is any analysis which includes the imputed variables in the main document, which is actually limited to just the analyses in *eFigure 1*.
    - We repeat the relative risk calculations on each of the multiply imputed data sets, and compare the point estimates and 95% CIs with those from the complete case analysis. In a slight change to the analyses in Figure S1, we compare the results from fitting a multivariable adjusted logistic regression model to the complete cases, and then to the multiply imputed data.
    - The substantive model has dependent variable SSI (0/1), and independent variables for the arms (CHG/PI) and the respective marker or comorbidity (smoking).
- The imputation model includes the following variables: SSI (indicator variable, arm indicator, procedure type, Center, hemoglobin, creatinine, C-reactive proteine, smoking BMI, timing of surgical antimicrobial prophylaxis, duration of surgery, ASA-score, endoscopic surgery, wound class.
- For hemoglobin we use a linear regression imputation approach, for creatinine/C-reactive protein predictive mean matching, and logistic regression for smoker within the MICE package.
- Following imputation, we recreate the categorical variables (with cutoffs for the markers) and then fit each of the now “full” the imputed data sets to the logistic regression models. The point estimates and variances are then averaged using “Rubin’s rules” in the usual way.

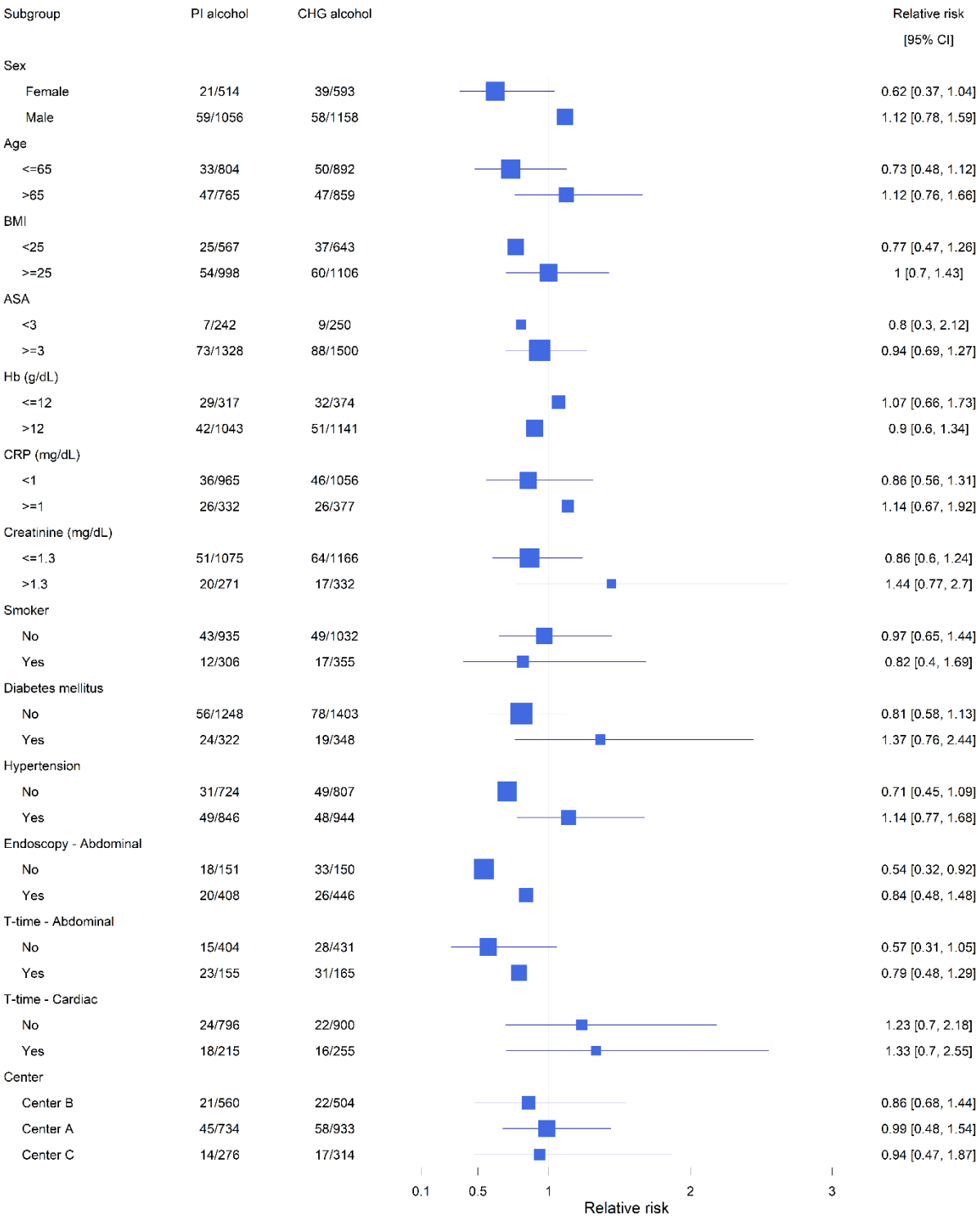
eTable 5 summarizes the results.

Conclusions and recommendations: The results from multiply imputing the missing data under the missing at random assumption were broadly consistent with those from the complete case analysis, with the 95% confidence intervals from the imputed data containing the point estimates from the complete

241 case analysis. Whilst this is not conclusive, it implies that the MAR assumption is plausible, and we
242 decided to not continue with sensitivity analyses.

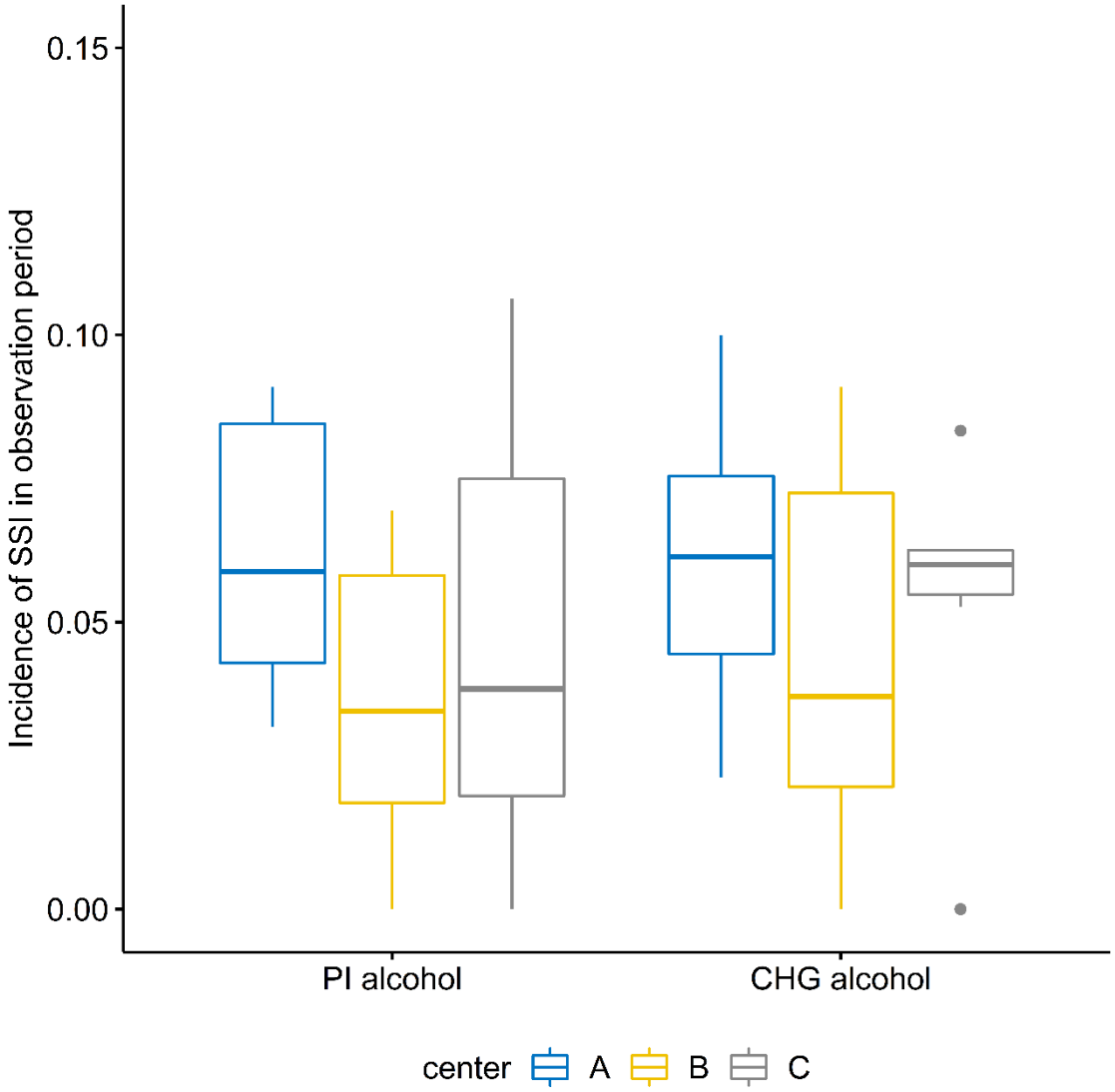
243

eFigure 1: Forest plot of predefined subgroup analyses (and laparoscopy in colon surgery), CHG as reference.



(CHG: Chlorhexidine gluconate, PI: Povidone iodine, CI: Confidence interval, Age: Age in years, CRP: C-reactive protein, BMI: body mass index in kg/m^2, ASA: American Society of Anesthesiologists Score, Hb: Hemoglobin, T-time: Surgery duration T-score over 75th percentile)

253 **eFigure 2:** Infection rates for the periods, stratified by intervention and center.



254

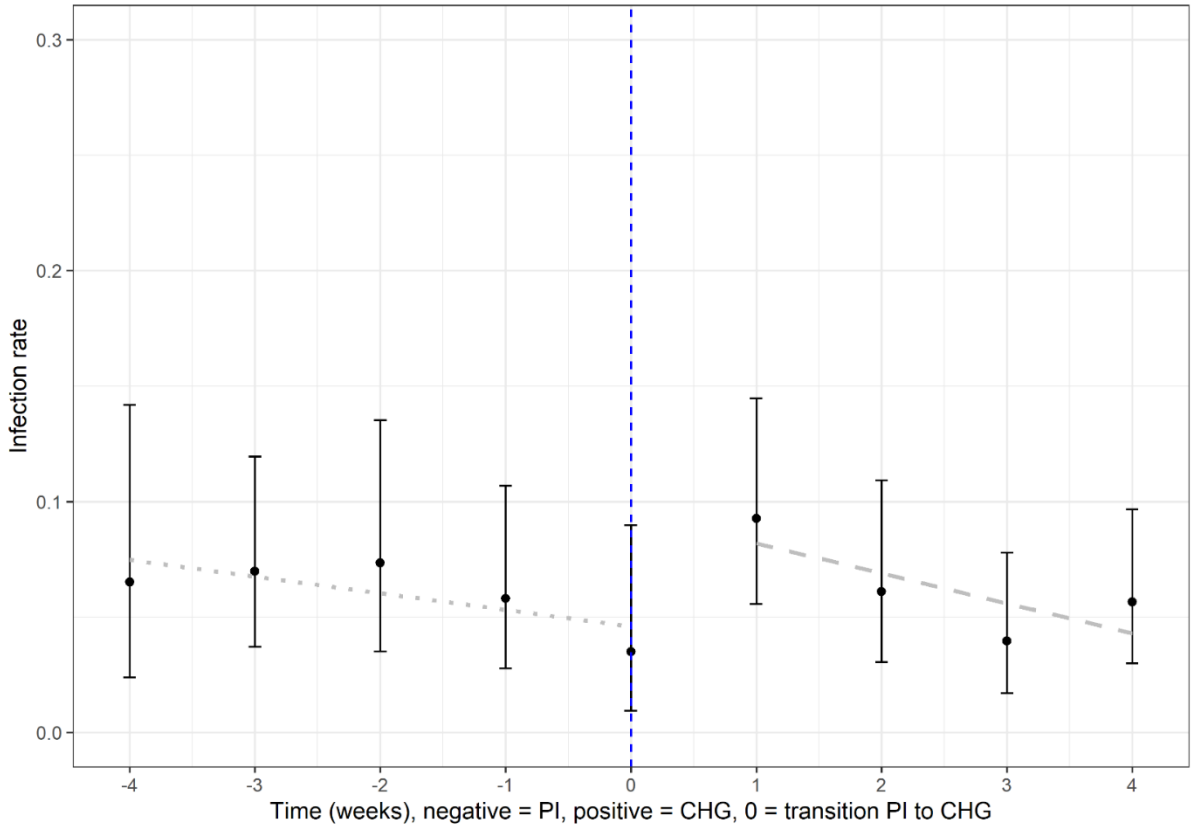
255 *(CHG: Chlorhexidine gluconate, PI: Povidone iodine, SSI: Surgical Site Infection)*

256

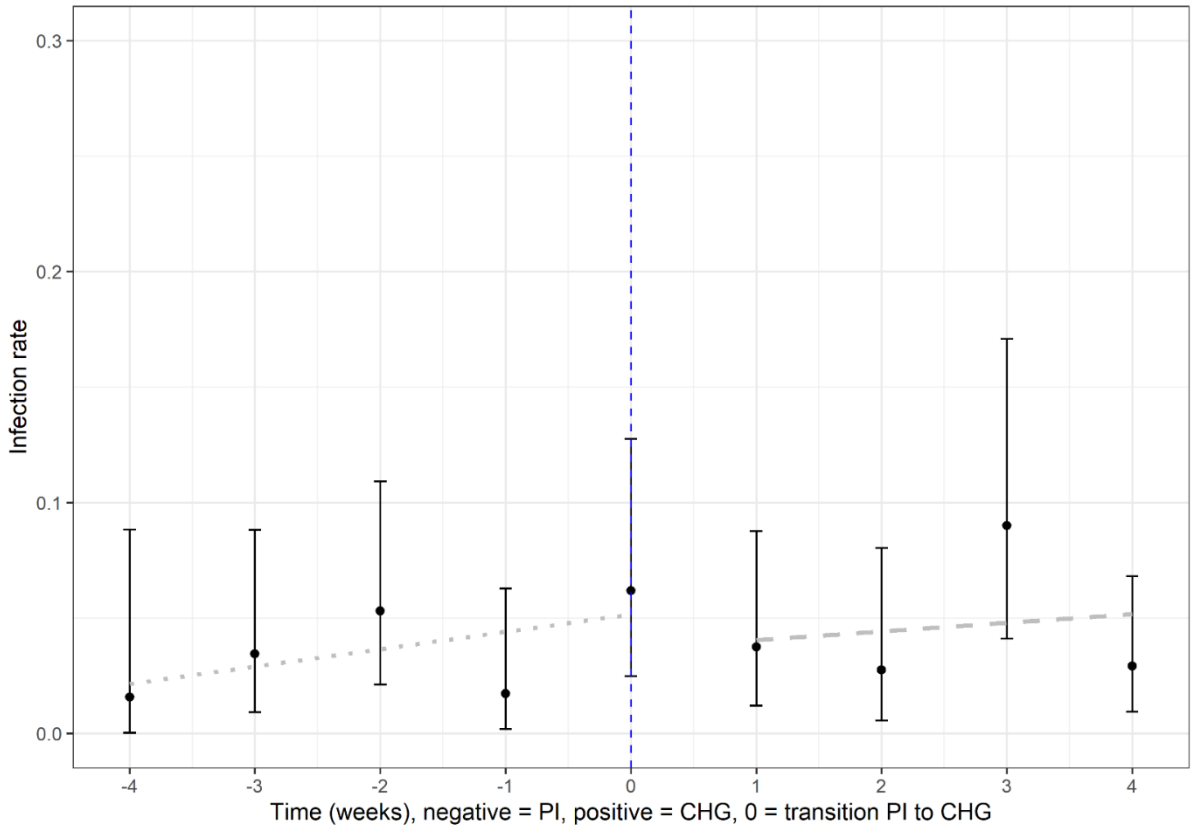
eFigure 3: Interrupted time series analysis of the effect of switching in either direction, stratified by center.

Comments: Time 0 is the switching time from povidone iodine (PI) to chlorhexidine gluconate (CHG), with negative times the 4 weeks on PI, and positive times the 4 weeks following the switch to CHG.

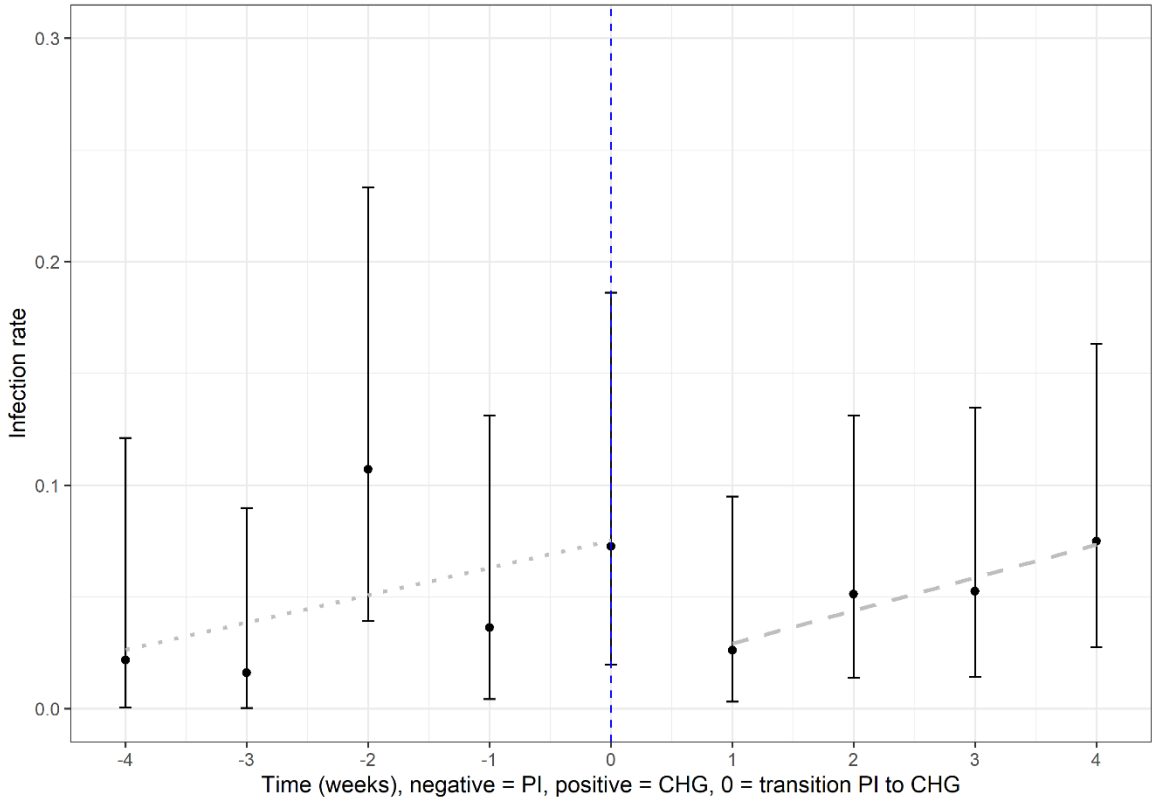
Center A



Center B



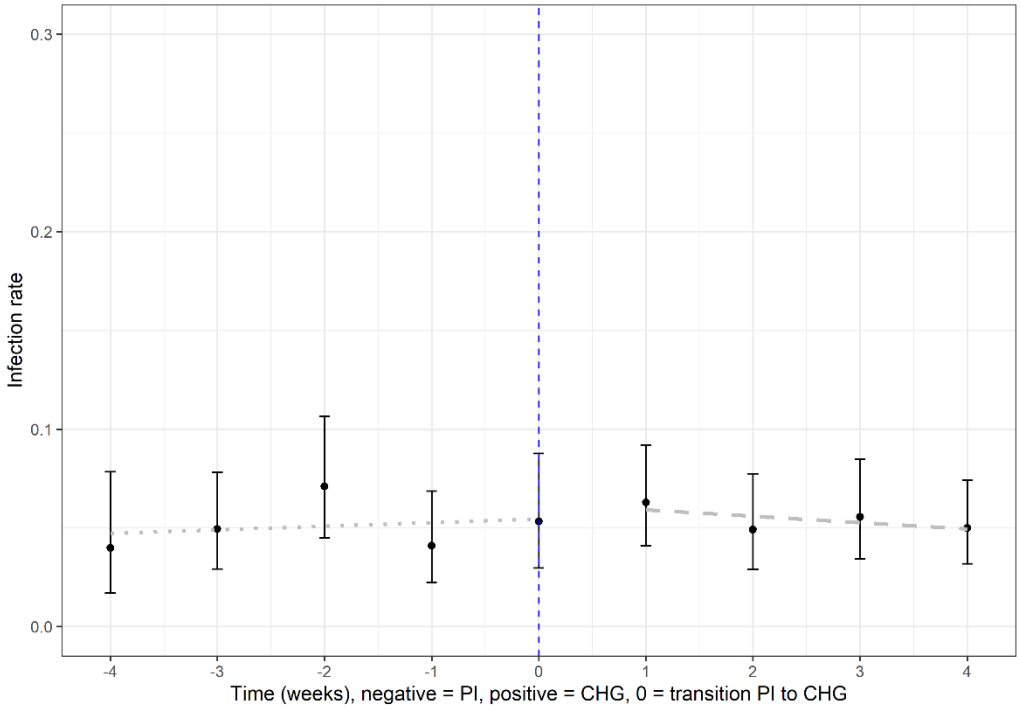
Center C



271

272

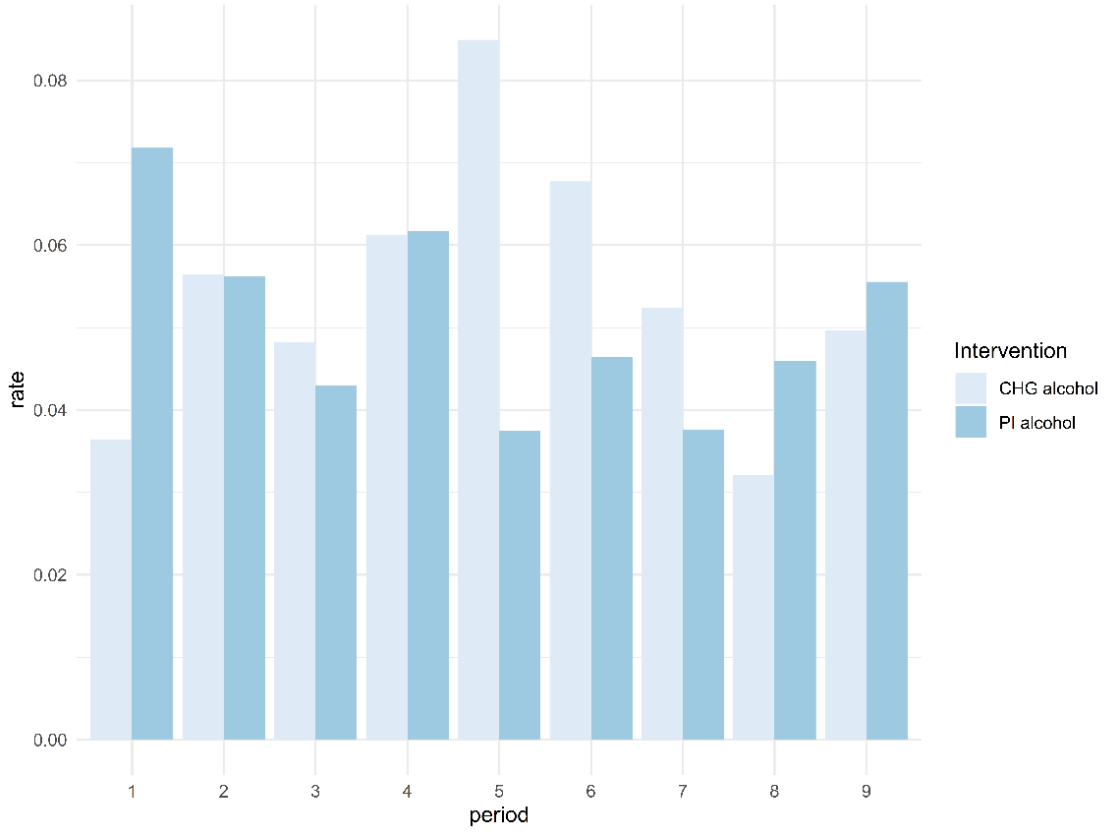
273 All centers



274

275 (*CHG: Chlorhexidine gluconate, PI: Povidone iodine*)

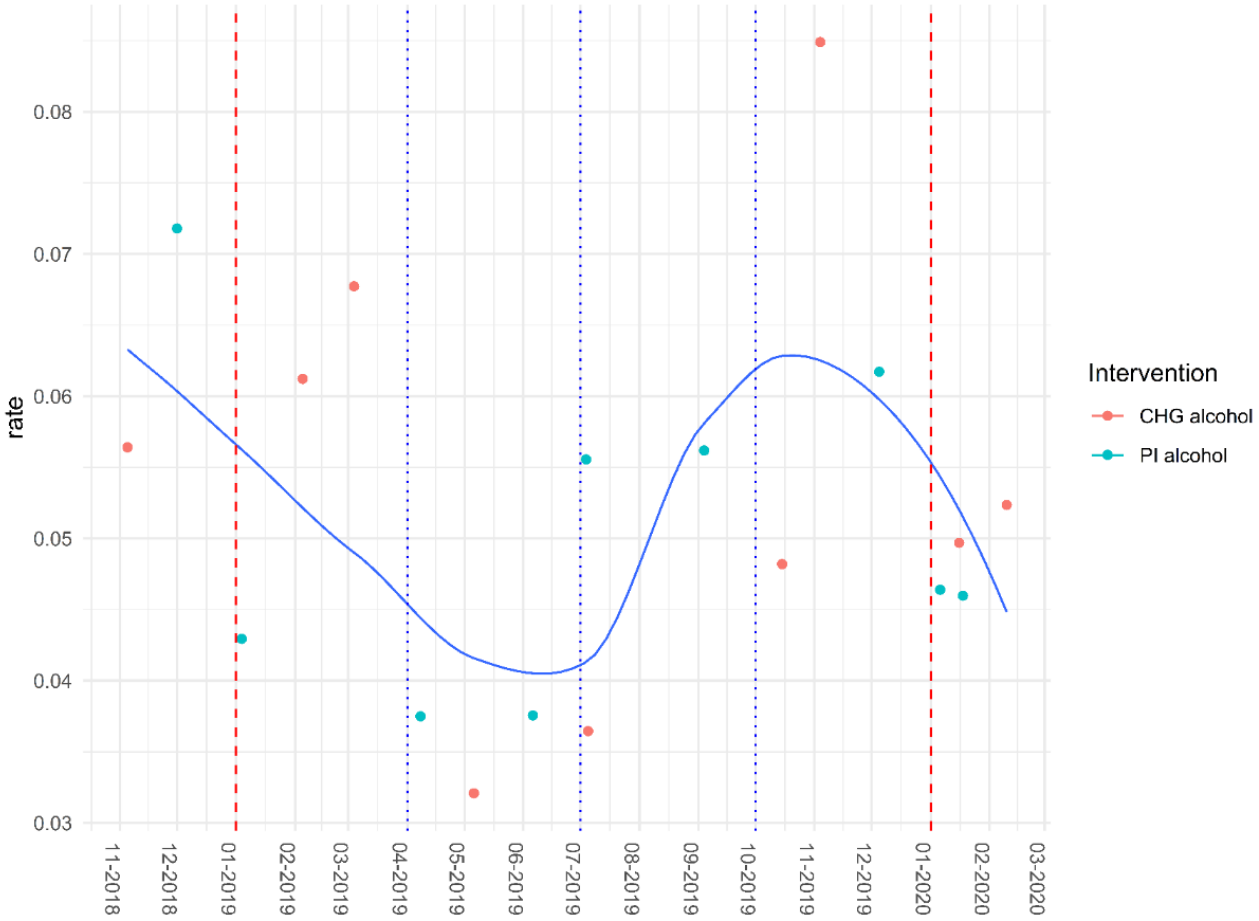
eFigure 4: Infection rates for the periods, stratified by intervention.



(CHG: Chlorhexidine gluconate, PI: Povidone iodine)

eFigure 5: Infection rates for the periods, plotted at the start of the calendar month and year, stratified by intervention; LOESS smoother added (solid, blue).

Comment: There seems to be some degree of seasonality (which is known for SSIs), but the monthly changes of cluster randomization takes this into consideration.



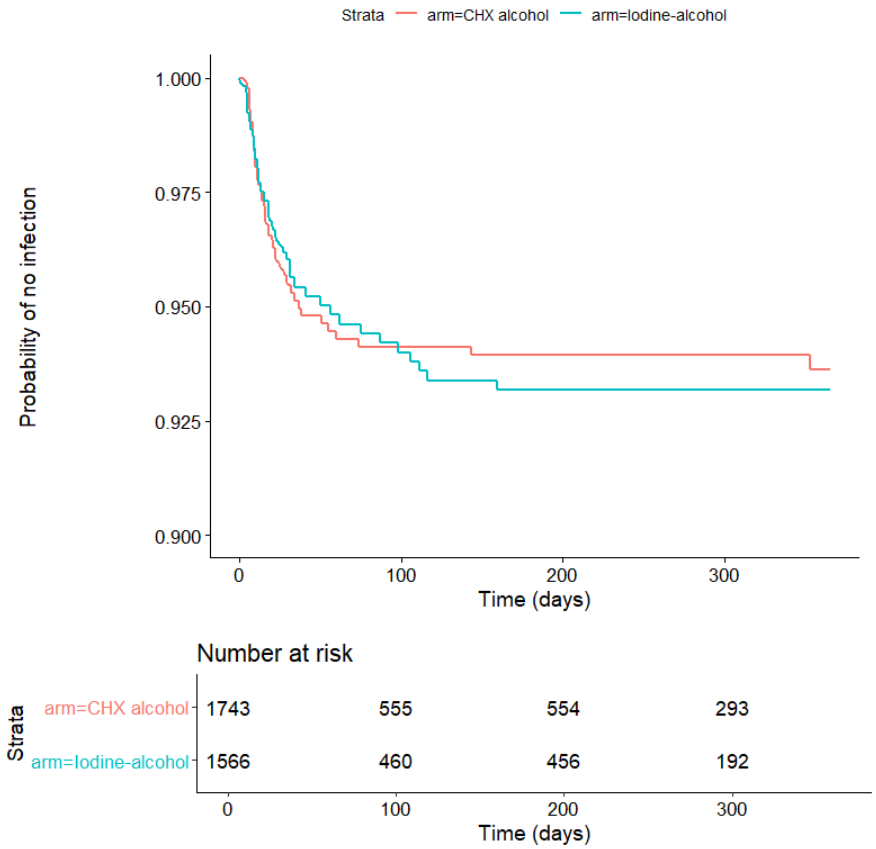
(*CHG: Chlorhexidine gluconate, PI: Povidone iodine)*

290 **eFigure 6:** Kaplan-Meier curve on probability of remaining free of surgical site infection (Logrank test, p

291 =0.8).

292

293



294

295 *(CHX: chlorhexidine)*

296

297

298 **eFigure 7:** Randomized allocation of monthly clusters per center.

Study month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Center A	CHG	PI	CHG	PI	CHG	PI	PI	CHG	PI	CHG	CHG	PI	PI	CHG	CHG	PI	CHG	PI
Center B	PI	CHG	PI	CHG	PI	CHG	CHG	PI	PI	CHG	PI	CHG	CHG	PI	PI	CHG	CHG	PI
Center C	*	*	CHG	PI	PI	CHG	CHG	PI	CHG	PI	CHG	PI	PI	CHG	CHG	PI	PI	CHG

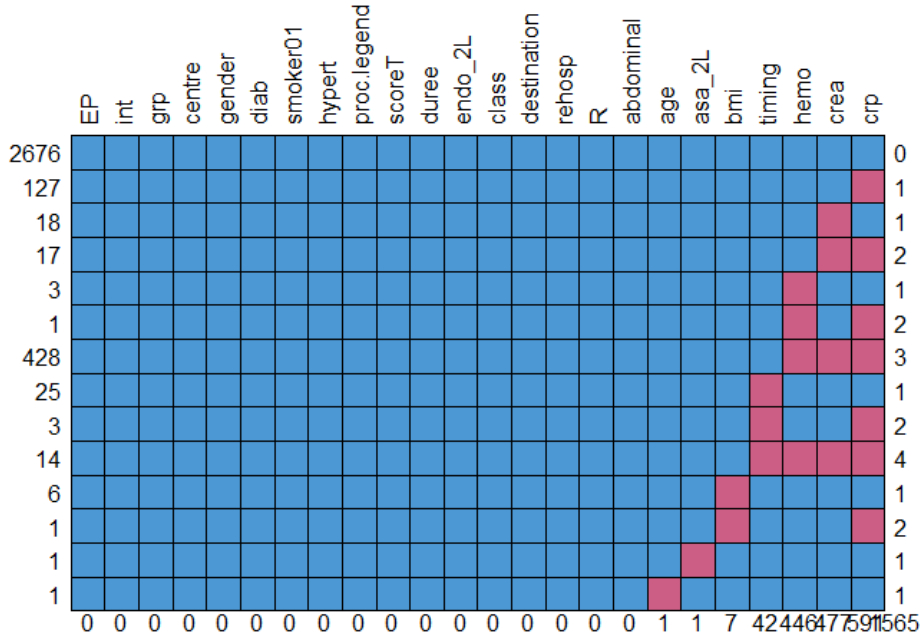
299

300 **Center C: Study started two months after the other centers.*

301 *(CHG: Chlorhexidine gluconate, PI: Povidone iodine)*

302

eFigure 8: Missingness patterns. Variables (top) with missing data (red) and variables without missing data (blue) with the number of patients with the respective missingness pattern (left) along with the number of number of missing variables (bottom); for example the top missingness pattern has 428 patients in which hb (hemoglobin), crea (creatinine) and crp (C-reactive protein) are all missing.



eTable 1: Fitted logistic regression models with SSI as endpoint, including center as fixed effect and period as random effect (*Supplementary Analysis, Appendix B*).

	PI in alcohol	**CHG in alcohol**
Primary outcome, supplementary analysis	Odds ratio [95% CI]	Odds ratio [95% CI]
Model M1: Cluster and cluster-period effects treated as random[a]	0.9 [0.7, 1.3,]	1 (reference)
Model M2: Cluster treated as fixed effect and cluster-period effects as random	0.9 [0.7, 1.3]	1 (reference)
GEE + sandwich or jack-knifed SEs	0.9 [0.7, 1.3]	1 (reference)
Secondary outcomes		
SSI rate (30 days), n/N (%)	27/1570 (1.7%)	27/1724 (1.5%)
Time to infection, median [IQR], days	13 [8, 25]	14 [9, 22]

[a] *The intra-cluster correlation coefficient was estimated to be 0.002*

(GEE: Generalized estimating equation , SE: Standard error, SSI: Surgical site infection, CHG: Chlorhexidine gluconate, PI: Povidone iodine, CI: Confidence interval)

315 **eTable 2**: Comparison of the number of infections from cardiac surgeries for the trial and for the 18

316 months prior to the trial starting from Swissnoso Surveillance (*Supplementary Appendix C*).

% of infections	**In hospital**	**0-30 days**	**31-90 days**	**>90 days**
Trial	19%	46%	25%	10%
Swissnoso benchmark (18 months prior to start of study)	31%	36%	19%	14%

317

318

eTable 3: Preoperative antibiotic prophylaxis stratified by intervention (five most used agents).

	PI in alcohol		**CHG in alcohol**	
1	Cefuroxime	71.0%	Cefuroxime	71.0%
2	Amoxicillin/Clavulanate	11.5%	Amoxicillin/Clavulanate	13.4%
3	Gentamicin	2.4%	Gentamicin	3.2%
4	Metronidazole	2.2%	Metronidazole	2.4%
5	Piperacillin/Tazobactam	2.0%	Clindamycin	1.7%

(CHG: Chlorhexidine gluconate, PI: Povidone iodine)

eTable 4: Variables in *Table 1* stratified according to patients with no missingness (R=0) and patients with one or more missing records (R=1).

	Univariable		Adjusted (logistic regression model) R as dependent variable
Patient characteristic[a]	R=1 (missing)	R=0 (no missing)	Odds Ratio (95% CI)
N	645	2676	-
SSI	7.4%	4.8%	-
Center			
B	26.5%	33.4%	0.2 (0.1, 0.3)
A	69.5%	45.6%	1 (reference)
C	26.5%	33.4%	0.1 (0.1, 0.2)
Male sex	47.3%	71.3%	-
Age, years	53 [39, 65]	67 [58, 74]	-
Hemoglobin, g/dL	13.3 [11.9, 14.3]	13.6 [12.2, 14.7]	-
Creatinine, mg/dL	0.86 [0.71, 1.05]	0.93 [0.78, 1.10]	-
C-reactive protein, mg/dL	0.51 [0.28, 1.37]	0.30 [0.23, 1.00]	-
Diabetes	16%	11.2%	-
Hypertension	65%	41.6%	-
BMI, kg/m^2	29.5 [24.5, 38.9]	26.4 [23.6, 30.0]	1.1 (1.0, 1.1)
Abdominal surgery	94%	21.5%	196 (99, 400)
Timing relative to incision, min	-33 [-45, -21]	-41 [-54, -30]	1.0 (1.0, 1.0)
Surgery duration, min	70 [49, 132]	230 [173, 281]	1.0 (1.0, 1.0)
Endoscopic surgery	76.6%	21.5%	-
ASA-Score 3-5	57.4%	91.9%	0.6 (0.4, 0.8)
Wound class: clean	20.2%	76.6%	1 (reference)
Wound class: clean-contaminated	73.5%	12.6%	0.3 (0.2, 0.5)

Categorical variables as N(%), continuous as median [IQR]

[a] *Only those variables shown with systematic differences (p<0.1); CI confidence interval; not shown – Smoked cigarettes, Type of intervention, T-Score, ASA-Score, reintervention, rehospitilisation, destination; adjusted model identified using forwards selection then backwards deletion using p<0.1 as inclusion criteria.*

(SSI: Surgical site infection, BMI: Body mass index, ASA: American Society of Anesthesiologists)

eTable 5: Comparison of estimates from the complete case analysis and following multiple imputation.

	Complete case analysis	Multiply imputed data (K=50)
	Odds ratio (95% CI)	Odds ratio (95% CI)
Hemoglobin		
<=12 g/dL	1 (reference)	1 (reference)
>12 g/dL	0.89 (0.74, 1.70)	0.89 (0.60, 1.32)
Creatinine		
<=1.3 mg/dL	1 (reference)	1 (reference)
>1.3 mg/dL	1.21 (0.81, 1.75)	1.22 (0.83, 1.80)
C-reactive protein		
<=1.0 mg/dL	1 (reference)	1 (reference)
>1.0 mg/dL	1.87 (1.30, 2.67)	1.66 (1.14, 2.42)
Smoked cigarettes		
No	1 (reference)	1 (reference)
Yes	0.77 (0.51, 1.15)	0.84 (0.55, 1.29)

Supplementary Appendix References

1. Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. *Contemp Clin Trials*. May 2007;28(3):316-23. doi:10.1016/j.cct.2006.10.004

2. Maiwald M, Widmer AF. WHO's recommendation for surgical skin antisepsis is premature. *Lancet Infect Dis*. Oct 2017;17(10):1023-1024. doi:10.1016/S1473-3099(17)30448-6

3. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol*. Dec 1981;114(6):906-14. doi:10.1093/oxfordjournals.aje.a113261

4. Arnup SJ, McKenzie JE, Hemming K, Pilcher D, Forbes AB. Understanding the cluster randomised crossover design: a graphical illustraton of the components of variation and a sample size tutorial. *Trials*. Aug 15 2017;18(1):381. doi:10.1186/s13063-017-2113-2

5. Reich NG, Myers JA, Obeng D, Milstone AM, Perl TM. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One*. 2012;7(4):e35564. doi:10.1371/journal.pone.0035564

6. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. *Int J Epidemiol*. Jun 1 2020;49(3):979-995. doi:10.1093/ije/dyz237

7. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Stat Med*. Nov 29 2008;27(27):5578-85. doi:10.1002/sim.3383

8. Frieden TR. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials. *N Engl J Med*. Aug 3 2017;377(5):465-475. doi:10.1056/NEJMra1614394

9. Turner RM, White IR, Croudace T, Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med*. Jan 30 2007;26(2):274-89. doi:10.1002/sim.2537

10. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. *J Clin Epidemiol*. Jun 2016;74:40-50. doi:10.1016/j.jclinepi.2015.11.013