TITLE: Prospective, Single-Blinded, Randomized-Controlled Trial Comparing the Performance Profiles of Two Non-Cross-Linked Porcine Dermal Matrices in Abdominal Wall Reconstruction

Statistical Analysis
Document Date: February 18, 2025
NCT # NCT02228889

Principal investigator: Jeffrey E. Janis, MD, FACS (Professor of Plastic Surgery)

Comparative Outcomes of Strattice vs. XenMatrix in Complex Abdominal Wall Reconstruction, PI: Jeffrey Janis, MD

Molly A. Olson, MS

Updated: 18 February, 2025

Contents

Statistical Analysis Plan

Overview

This study aims to evaluate the effect of using XenMatrix or Strattice mesh on the occurrence of surgical site occurrences (SSO) within six weeks post-surgery. Patient-level data were drawn from two sources:

- A single-site randomized controlled trial (RCT), and
- A single-site observational study.

These datasets were combined using a random-effects meta-analysis to generate an overall estimate of the treatment effect of mesh type on the risk of SSO.

Statistical Analysis

Descriptive Analysis

An unadjusted analysis comparing mesh was performed.

- Baseline Characteristics: Descriptive statistics (e.g., means or medians for continuous data, counts and proportions for categorical data) were generated for both the RCT and observational cohorts.
- Group Comparisons:
  - Continuous variables were compared using the Kruskal–Wallis test.
  - Categorical variables were compared using Pearson's Chi-square test.

Primary Outcome Analysis

SSO within 6 weeks was analyzed using logistic regression. Separate logistic regression models were fit for the RCT and for the observational cohort, with mesh type (XenMatrix vs. Strattice) as the primary predictor. Simple logistic regression and multivariable logistic regression was used for each study type (RCT and observational). Both models were adjusted by hernia width, CDC wound class, immunosuppression, and recurrent hernia to account for potential confounding in the observational group. Covariates were selected based on evidence in the literature as well as potential imbalance. The final number of covariates was limited to avoid model overfitting.

Confounders:
- Hernia width
- Wound class
- Immunosuppression
- Recurrent hernia

After obtaining separate estimates from the RCT and observational cohort, a random-effects meta-analysis was performed to derive a combined estimate. A combined estimate for the log odds ratio was calculated using inverse variance weighting and a restricted maximum likelihood estimator. Between-study heterogeneity was assessed by the I^2 measure.

Forest plots and summary tables for the log odds ratio and odds ratio were generated to summarize results.

Statistical and Study Design Considerations

Rationale for Combining RCT and Observational Study Data

- RCT was terminated early for difficulties with enrollment, resulting in very imbalanced group sizes.
- While some guidelines caution against combining observational studies with RCTs in a single meta-analysis due to methodological differences [2], other sources emphasize that in many situations, the benefits of incorporating all available evidence can outweigh the drawbacks, provided proper methods are used [2, 3]
- [3] proves a framework for synthesis of non-randomized and randomized studies.
  - In this study, the observational data come from the same single site as the RCT and are not subject to publication bias.

Choice of Statistical Methods

- Frequentist methods were chosen for broader audience familiarity. Although Bayesian multi-level models are an option, they can be sensitive to prior specification, and the intended audience may be less familiar with them.
- There are not enough events to perform analysis for other outcomes.
- Propensity score matching was not used because it can substantially reduce the sample size (i.e., discarding unmatched patients).

- Statistical methods chosen to maximize data (Propensity score matching methods throw away data). Methods were also chosen based on what audiences would be most familiar with (Frequentist approaches). Bayesian multi-level models are also a choice, but audiences are not as familiar with them and they can be sensitive to priors. as well as
- Because the study was stopped due to practice changes, interpreting results from an early stopped trial can be complex due to potential bias.

- The RCT power calculation is not valid when adding observational data to "boost" sample size, because the assumptions behind RCT power calculations (particularly randomization and control over selection) do not apply to non-randomized data.

- Here is an example of a study using the same methods: Varges D, Manthey H, Heinemann U, et alDoxycycline in early CJD: a double-blinded randomised phase II and observational studyJournal of Neurology, Neurosurgery & Psychiatry 2017;88:119-125.

Methodology implementation notes

- Six patients were missing hernia size; mesh size was used as a proxy in those cases. While not ideal, this approach maximizes the use of available data.

Additional comments

- The RCT was stopped due to practice changes, which can introduce bias. Early-stopped trials may overestimate the treatment effect or otherwise affect the interpretability of results.
- With a relatively small sample size and potential imbalance in covariates, randomization alone might not ensure balance (even if p-values are > 0.05). Hence, adjusting for known confounders is critical.
- Be wary of exaggerated findings, as smaller studies are more prone to exaggeration.
  - Known as a Type M (magnitude) error. In other words, a finding must be large enough to exceed the statistical significance threshold despite wide confidence intervals, which systematically inflates the estimated effect size. Consequently, although we observed a strong effect, its true magnitude may be substantially lower than our point estimate suggests.[1]

References

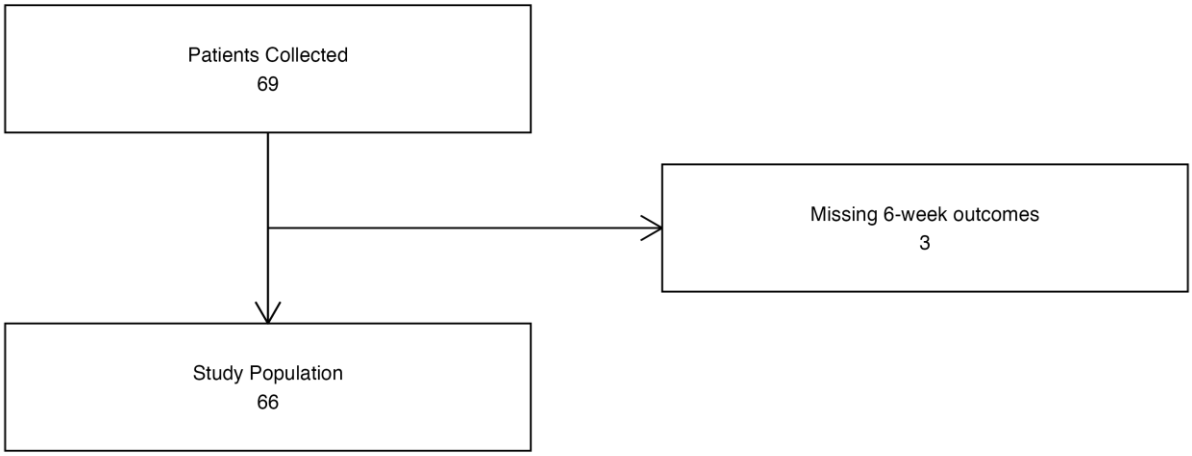
Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science, 9(6), 641-651. https://doi.org/10.1177/1745691614551642

René-Sosata Bun, Jordan Scheer, Sylvie Guillo, Florence Tubach, Agnès Dechartres, Meta-analyses frequently pooled different study types together: a meta-epidemiological study, Journal of Clinical Epidemiology, Volume 118, 2020, Pages 18-28, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2019.10.013.

Sarri G, Patorno E, Yuan H, et alFramework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision makingBMJ Evidence-Based Medicine 2022;27:109-119.

Shrier I, Boivin JF, Steele RJ, Platt RW, Furlan A, Kakuma R, Brophy J, Rossignol M. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. Am J Epidemiol. 2007 Nov 15;166(10):1203-9. doi: 10.1093/aje/kwm189. Epub 2007 Aug 21. PMID: 17712019.

Additional Exclusion Criteria



Tables

Baseline Characteristics

Table 1a: Baseline demographics, randomized patients

Variables	N	Strattice	XenMatrix	Test Statistic
		(N=16)	(N=30)	
Gender : Male	46	6/16 (37.50)	16/30 (53.33)	X^2_1=1.05, P=0.31[2]
Days from surgery to last follow-up	46			$F_{1,44}$=5.51, P=0.02[3]
N		16	30	
Median (interquartile range)		1157.50 (727.67—1672.42)	544.00 (140.92—1226.92)	
Range		76.00—3100.00	0.00—2157.00	
Mean±SD		1191.38±774.56	710.13±632.56	
Age	46			$F_{1.44}$=2.51. P=0.12[3]
N		16	30	
Median (interquartile range)		60.00 (55.83—68.58)	57.00 (43.00—62.08)	
Range		27.00—75.00	20.00—73.00	
Mean±SD		59.25±13.79	53.57±13.50	
Body mass index	46			$F_{1,44}$=1.23, P=0.27[3]
N		16	30	
Median (interquartile range)		29.61 (26.79—33.45)	33.10 (26.46—37.38)	
Range		21.22—40.21	20.91—42.16	
Mean±SD		30.14±5.03	32.21±6.21	
BMI >30kg/m2 : Yes	46	8/16 (50.00)	18/30 (60.00)	X^2_1=0.42. P=0.51[2]
Diabetes : Yes	46	7/16 (43.75)	7/30 (23.33)	X^2_1=2.05, P=0.15[2]
Hypertension : Yes	46	7/16 (43.75)	14/30 (46.67)	X^2_1=0.04, P=0.85[2]
COPD : Yes	46	5/16 (31.25)	5/30 (16.67)	X^2_1=1.30, P=0.25[2]
Immunosuppression : Yes	46	2/16 (12.50)	10/30 (33.33)	X^2_1=2.35, P=0.13[2]
Primary hernia diagnosis : Yes	46	14/16 (87.50)	28/30 (93.33)	X^2_1=0.45, P=0.50[2]
Previous repair : Yes	46	11/16 (68.75)	20/30 (66.67)	X^2_1=0.02, P=0.89[2]
Number of prior hernia repairs	46			$F_{1,44}$=1.93, P=0.17[3]
N		16	30	
Median (interquartile range)		2.50 (0.00—4.58)	1.00 (0.00—2.00)	
Range		0.00—9.00	0.00—6.00	
Mean±SD		2.69±2.73	1.43±1.61	
History of mesh repair : Yes	46	12/16 (75.00)	17/30 (56.67)	X^2_1=1.51, P=0.22[2]
Hernia width	41			$F_{1,39}$=2.14, P=0.15[3]
N		16	25	
Median (interquartile range)		15.80 (9.95—18.71)	12.60 (9.20—14.80)	
Range		2.10—30.30	2.20—22.20	
Mean±SD		14.91±7.41	11.67±5.48	
Hernia width (imputed)	46			$F_{1,44}$=0.68, P=0.41[3]
N		16	30	
Median (interquartile range)		15.80 (9.95—18.71)	12.90 (9.48—18.65)	
Range		2.10—30.30	2.20—30.00	
Mean±SD		14.91±7.41	13.06±6.46	
Hernia length	41			$F_{1,39}$=0.10, P=0.75[3]
N		16	25	
Median (interquartile range)		16.60 (9.84—20.00)	11.80 (8.20—21.60)	
Range		2.30—28.00	3.50—27.70	
Mean±SD		15.30±6.90	14.20±7.89	
Hernia length (imputed)	46			$F_{1,44}$=0.01, P=0.93[3]
N		16	30	
Median (interquartile range)		16.60 (9.84—20.00)	16.25 (9.00—21.41)	
Range		2.30—28.00	3.50—27.70	
Mean±SD		15.30±6.90	15.05±7.49	
Hernia area	41			$F_{1,39}$=1.33, P=0.26[3]
N		16	25	
Median (interquartile range)		275.97 (96.08—354.27)	153.40 (45.61—301.15)	
Range		8.61—700.00	7.70—545.69	
Mean±SD		265.97±200.88	195.61±157.45	
Total OR time (min)	43			$F_{1,41}$=0.29, P=0.59[3]
N		14	29	
Median (interquartile range)		471.50 (443.50—483.58)	417.00 (336.33—547.67)	
Range		292.00—560.00	212.00—838.00	
Mean±SD		462.57±58.22	451.17±142.81	
ASA class	46			X^2_2=2.21, P=0.33[2]
2		3/16 (18.75)	3/30 (10.00)	
3		13/16 (81.25)	24/30 (80.00)	

Table 1a: Baseline demographics, randomized patients *(continued)*

Variables	N	Strattice	XenMatrix	Test Statistic
4		0/16 (0.00)	3/30 (10.00)	

Note:

N is the number of non-missing value. 1 Kruskal-Wallis. 2 Pearson. 3 Wilcoxon.

Stat, P is the test statistic and p-value.

cata = check all that apply

Table 1b: Baseline demographics, observational patients

Variables	N	Strattice	XenMatrix	Test Statistic
		(N=12)	(N=8)	
Gender : Male	20	6/12 (50.00)	3/8 (37.50)	X^2_1=0.30, P=0.58[2]
Days from surgery to last follow-up	20			$F_{1,18}$=1.18, P=0.29[3]
N		12	8	
Median (interquartile range)		311.50 (93.67—469.17)	570.50 (156.75—1293.50)	
Range		54.00—855.00	64.00—1751.00	
Mean±SD		327.08±266.89	715.38±669.49	
Age	20			$F_{1,18}$=0.01, P=0.94[3]
N		12	8	
Median (interquartile range)		55.00 (49.42—60.58)	57.50 (35.33—67.58)	
Range		32.00—70.00	30.00—73.00	
Mean±SD		54.75±10.16	53.12±17.01	
Body mass index	20			$F_{1,18}$=2.30, P=0.15[3]
N		12	8	
Median (interquartile range)		36.48 (32.12—38.26)	32.97 (25.34—35.71)	
Range		24.17—47.60	21.49—39.67	
Mean±SD		35.59±6.02	31.23±6.49	
BMI >30kg/m2 : Yes	20	10/12 (83.33)	5/8 (62.50)	X^2_1=1.11, P=0.29[2]
Diabetes : Yes	20	2/12 (16.67)	0/8 (0.00)	X^2_1=1.48, P=0.22[2]
Hypertension : Yes	20	7/12 (58.33)	2/8 (25.00)	X^2_1=2.15, P=0.14[2]
COPD : Yes	20	1/12 (8.33)	1/8 (12.50)	X^2_1=0.09, P=0.76[2]
Immunosuppression : Yes	20	4/12 (33.33)	0/8 (0.00)	X^2_1=3.33, P=0.07[2]
Primary hernia diagnosis : Yes	20	4/12 (33.33)	7/8 (87.50)	X^2_1=5.69, P=0.02[2]
Previous repair : Yes	20	7/12 (58.33)	3/8 (37.50)	X^2_1=0.83, P=0.36[2]
Number of prior hernia repairs	20			$F_{1,18}$=0.33, P=0.57[3]
N		12	8	
Median (interquartile range)		1.00 (0.00—1.58)	0.00 (0.00—2.17)	
Range		0.00—9.00	0.00—6.00	
Mean±SD		1.50±2.54	1.25±2.19	
History of mesh repair : Yes	20	6/12 (50.00)	3/8 (37.50)	X^2_1=0.30, P=0.58[2]
Hernia width	19			$F_{1,17}$=5.27, P=0.03[3]
N		11	8	
Median (interquartile range)		11.70 (9.75—16.47)	16.50 (15.27—19.93)	
Range		2.80—30.00	15.00—22.50	
Mean±SD		13.23±7.78	17.54±2.94	
Hernia width (imputed)	20			$F_{1,18}$=2.89, P=0.11[3]
N		12	8	
Median (interquartile range)		12.45 (9.82—20.15)	16.50 (15.27—19.93)	
Range		2.80—30.00	15.00—22.50	
Mean±SD		14.62±8.86	17.54±2.94	
Hernia length	19			$F_{1,17}$=0.01, P=0.91[3]
N		11	8	
Median (interquartile range)		22.50 (20.10—25.08)	22.35 (19.17—27.41)	
Range		2.40—32.70	13.50—30.00	
Mean±SD		20.27±9.01	22.61±5.47	
Hernia length (imputed)	20			$F_{1,18}$=0.07, P=0.80[3]
N		12	8	
Median (interquartile range)		22.05 (20.00—24.91)	22.35 (19.17—27.41)	
Range		2.40—32.70	13.50—30.00	
Mean±SD		20.25±8.59	22.61±5.47	
Hernia area	19			$F_{1,17}$=2.69, P=0.12[3]
N		11	8	
Median (interquartile range)		284.28 (231.27—422.23)	362.85 (322.09—467.22)	
Range		6.72—600.00	230.90—657.00	
Mean±SD		298.77±188.98	398.54±130.54	

Table 1b: Baseline demographics, observational patients *(continued)*

Variables	N	Strattice	XenMatrix	Test Statistic
Total OR time (min)	19			$F_{1,17}$=0.34, P=0.57[3]
N		12	7	
Median (interquartile range)		431.00 (383.17—540.58)	551.00 (345.17—575.83)	
Range		310.00—696.00	289.00—1018.00	
Mean±SD		465.08±117.42	538.86±240.86	
ASA class	20			X^2_2=3.59, P=0.17[2]
2		1/12 (8.33)	1/8 (12.50)	
3		11/12 (91.67)	5/8 (62.50)	
4		0/12 (0.00)	2/8 (25.00)	

Note:

N is the number of non-missing value. 1 Kruskal-Wallis. 2 Pearson. 3 Wilcoxon.

Stat, P is the test statistic and p-value.

cata = check all that apply

Surgical Characteristics

Table 2a: Surgical demographics, randomized patients

Variables	N	Strattice	XenMatrix	Test Statistic
		(N=16)	(N=30)	
Types of hernia diagnosis*cata*				
Ventral hernia : Yes	46	15/16 (93.75)	28/30 (93.33)	X^2_1=0.00, P=0.96[2]
Flank hernia : Yes	46	0/16 (0.00)	2/30 (6.67)	X^2_1=1.12, P=0.29[2]
Parastomal hernia : Yes	46	3/16 (18.75)	3/30 (10.00)	X^2_1=0.70, P=0.40[2]
Umbilical hernia : Yes	46	0/16 (0.00)	0/30 (0.00)	NA
Parastomal hernia : Yes	46	3/16 (18.75)	3/30 (10.00)	X^2_1=0.70, P=0.40[2]
CDC wound class	46			X^2_3=3.08, P=0.38[2]
1		7/16 (43.75)	17/30 (56.67)	
2		5/16 (31.25)	4/30 (13.33)	
3		2/16 (12.50)	2/30 (6.67)	
4		2/16 (12.50)	7/30 (23.33)	
VHWG grade	46			X^2_3=3.46, P=0.33[2]
1		1/16 (6.25)	1/30 (3.33)	
2		2/16 (12.50)	9/30 (30.00)	
3		11/16 (68.75)	13/30 (43.33)	
4		2/16 (12.50)	7/30 (23.33)	
Kanters grade	46			X^2_2=1.12, P=0.57[2]
1		1/16 (6.25)	1/30 (3.33)	
2		6/16 (37.50)	16/30 (53.33)	
3		9/16 (56.25)	13/30 (43.33)	
Primary fascial repair : Yes	46	14/16 (87.50)	22/30 (73.33)	X^2_1=1.23, P=0.27[2]
Unilateral component separation : Yes	46	2/16 (12.50)	2/30 (6.67)	X^2_1=0.45, P=0.50[2]
Bilateral component separation : Yes	46	8/16 (50.00)	17/30 (56.67)	X^2_1=0.19, P=0.67[2]
Bridged repair : Yes	46	2/16 (12.50)	6/30 (20.00)	X^2_1=0.41, P=0.52[2]
Mesh position*cata*				
Onlay mesh placement : Yes	46	0/16 (0.00)	3/30 (10.00)	X^2_1=1.71, P=0.19[2]
Underlay mesh placement : Yes	46	12/16 (75.00)	24/30 (80.00)	X^2_1=0.15, P=0.70[2]
Inlay mesh placement : Yes	46	0/16 (0.00)	2/30 (6.67)	X^2_1=1.12, P=0.29[2]
Sublay mesh placement : Yes	46	4/16 (25.00)	8/30 (26.67)	X^2_1=0.02, P=0.90[2]

Note:

N is the number of non-missing value. 1 Kruskal-Wallis. 2 Pearson. 3 Wilcoxon.

Stat, P is the test statistic and p-value.

cata = check all that apply

Table 2b: Surgical demographics, observational patients

Variables	N	Strattice	XenMatrix	Test Statistic
		(N=12)	(N=8)	
Types of hernia diagnosis*cata*				
Ventral hernia : Yes	20	11/12 (91.67)	7/8 (87.50)	X^2_1=0.09, P=0.76[2]
Flank hernia : Yes	20	2/12 (16.67)	1/8 (12.50)	X^2_1=0.07, P=0.80[2]
Parastomal hernia : Yes	20	4/12 (33.33)	0/8 (0.00)	X^2_1=3.33, P=0.07[2]

Table 2b: Surgical demographics, observational patients *(continued)*

Variables	N	Strattice	XenMatrix	Test Statistic
Umbilical hernia : Yes	20	0/12 (0.00)	0/8 (0.00)	NA
Parastomal hernia : Yes	20	4/12 (33.33)	0/8 (0.00)	X^2_1=3.33, P=0.07[2]
CDC wound class	20			X^2_3=3.64, P=0.30[2]
1		6/12 (50.00)	5/8 (62.50)	
2		3/12 (25.00)	2/8 (25.00)	
3		0/12 (0.00)	1/8 (12.50)	
4		3/12 (25.00)	0/8 (0.00)	
VHWG grade	20			X^2_3=4.49, P=0.21[2]
1		0/12 (0.00)	1/8 (12.50)	
2		6/12 (50.00)	2/8 (25.00)	
3		4/12 (33.33)	5/8 (62.50)	
4		2/12 (16.67)	0/8 (0.00)	
Kanters grade	20			X^2_1=0.83, P=0.36[2]
1		0/12 (0.00)	0/8 (0.00)	
2		5/12 (41.67)	5/8 (62.50)	
3		7/12 (58.33)	3/8 (37.50)	
Primary fascial repair : Yes	19	9/12 (75.00)	6/7 (85.71)	X^2_1=0.31, P=0.58[2]
Unilateral component separation : Yes	20	1/12 (8.33)	0/8 (0.00)	X^2_1=0.70, P=0.40[2]
Bilateral component separation : Yes	20	4/12 (33.33)	7/8 (87.50)	X^2_1=5.69, P=0.02[2]
Bridged repair : Yes	19	4/12 (33.33)	1/7 (14.29)	X^2_1=0.83, P=0.36[2]
Mesh position[cata]				
Onlay mesh placement : Yes	20	0/12 (0.00)	0/8 (0.00)	NA
Underlay mesh placement : Yes	20	12/12 (100.00)	6/8 (75.00)	X^2_1=3.33, P=0.07[2]
Inlay mesh placement : Yes	20	1/12 (8.33)	1/8 (12.50)	X^2_1=0.09, P=0.76[2]
Sublay mesh placement : Yes	20	0/12 (0.00)	2/8 (25.00)	X^2_1=3.33, P=0.07[2]

Note:

N is the number of non-missing value. 1 Kruskal-Wallis. 2 Pearson. 3 Wilcoxon.

Stat, P is the test statistic and p-value.

cata = check all that apply

6-week Outcomes

Table 3a: 6-week outcomes, randomized patients

Variables	N	Strattice	XenMatrix	Test Statistic
		(N=16)	(N=30)	
Bulge : Yes	46	0/16 (0.00)	0/30 (0.00)	NA
Recurrence : Yes	46	0/16 (0.00)	0/30 (0.00)	NA
SSO within 6 weeks : Yes	46	1/16 (6.25)	11/30 (36.67)	X^2_1=5.01, P=0.03[2]
SSO Type (cata)				
Skin necrosis : Yes	46	0/16 (0.00)	1/30 (3.33)	X^2_1=0.55, P=0.46[2]
Enterocutaneous fistula : Yes	46	0/16 (0.00)	2/30 (6.67)	X^2_1=1.12, P=0.29[2]
Dehiscence : Yes	46	0/16 (0.00)	1/30 (3.33)	X^2_1=0.55, P=0.46[2]
Seroma : Yes	46	0/16 (0.00)	2/30 (6.67)	X^2_1=1.12, P=0.29[2]
Hematoma : Yes	46	0/16 (0.00)	0/30 (0.00)	NA
Infection : Yes	46	1/16 (6.25)	10/30 (33.33)	X^2_1=4.21, P=0.04[2]
Mesh infection : Yes	46	0/16 (0.00)	3/30 (10.00)	X^2_1=1.71, P=0.19[2]

Note:

N is the number of non-missing value. 1 Kruskal-Wallis. 2 Pearson. 3 Wilcoxon.

Stat, P is the test statistic and p-value.

cata = check all that apply

Table 3b: 6-week outcomes, observational patients

Variables	N	Strattice	XenMatrix	Test Statistic
		(N=12)	(N=8)	
Bulge : Yes	20	0/12 (0.00)	0/8 (0.00)	NA
Recurrence : Yes	20	1/12 (8.33)	0/8 (0.00)	X^2_1=0.70, P=0.40[2]
SSO within 6 weeks : Yes	20	4/12 (33.33)	4/8 (50.00)	X^2_1=0.56, P=0.46[2]
SSO Type (cata)				
Skin necrosis : Yes	20	0/12 (0.00)	1/8 (12.50)	X^2_1=1.58, P=0.21[2]
Enterocutaneous fistula : Yes	20	2/12 (16.67)	0/8 (0.00)	X^2_1=1.48, P=0.22[2]

Table 3b: 6-week outcomes, observational patients *(continued)*

Variables	N	Strattice	XenMatrix	Test Statistic
Dehiscence : Yes	20	2/12 (16.67)	0/8 (0.00)	X^2_1=1.48, P=0.22[2]
Seroma : Yes	20	1/12 (8.33)	0/8 (0.00)	X^2_1=0.70, P=0.40[2]
Hematoma : Yes	20	0/12 (0.00)	0/8 (0.00)	NA
Infection : Yes	20	1/12 (8.33)	3/8 (37.50)	X^2_1=2.55, P=0.11[2]
Mesh infection : Yes	19	0/11 (0.00)	0/8 (0.00)	NA

Note:

N is the number of non-missing value. 1 Kruskal-Wallis. 2 Pearson. 3 Wilcoxon.

Stat, P is the test statistic and p-value.

cata = check all that apply

Results

Executive Summary

Unadjusted Random Effects Meta-Analysis (Simple Logistic Regression):

Effect Estimate (Odds Ratio): 3.718 (0.897, 15.404)
Significance: The pooled estimate was not statistically significant.

Heterogeneity:

I^2 (total heterogeneity / total variability): 3.29%
H^2 (total variability / sampling variability): 1.03

Test for Heterogeneity:
Q(df = 1) = 1.0340, p-val = 0.3092

There appears to be minimal variability between the two studies. The I^2 value of 3.29% indicates that nearly all of the total variation can be explained by chance rather than true heterogeneity, and an H^2 of 1.03 reflects low excess variability. The Q test (Q = 1.034, p = 0.3092) is nonsignificant, confirming that there is no strong evidence of meaningful heterogeneity.

Adjusted Random Effects Meta-Analysis (Multivariable Logistic Regression):

Effect Estimate (Odds Ratio): 12.476 (1.745, 89.199)
Significance: The pooled estimate was statistically significant.

Heterogeneity:

I^2 (total heterogeneity / total variability): 0.00%
H^2 (total variability / sampling variability): 1.00

Test for Heterogeneity: Q(df = 1) = 0.2830, p-val = 0.5947

There is effectively no evidence of meaningful heterogeneity between the RCT and the observational study, provided by 0.00% I^2 value. Similarly, $H^2 = 1.00$ supports the absence of excess variability beyond what would be expected by sampling error alone. Finally, the Q test (Q = 0.2830, p = 0.5947) is non-significant, confirming that the studies are consistent and do not exhibit statistically detectable heterogeneity.

When covariates were included in the model, a previously non-significant association between the exposure and outcome became statistically significant. This shift suggests that the unadjusted analysis may have been confounded—i.e., certain variables were masking or biasing the relationship.

Note

Be wary of exaggerated findings, as smaller studies are more prone to exaggeration. This is known as a Type M (magnitude) error. In other words, a finding must be large enough to exceed the statistical significance threshold despite wide confidence intervals, which systematically inflates the estimated effect size. Consequently, although we observed a strong effect, its true magnitude may be substantially lower than our point estimate suggests.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science, 9(6), 641-651.

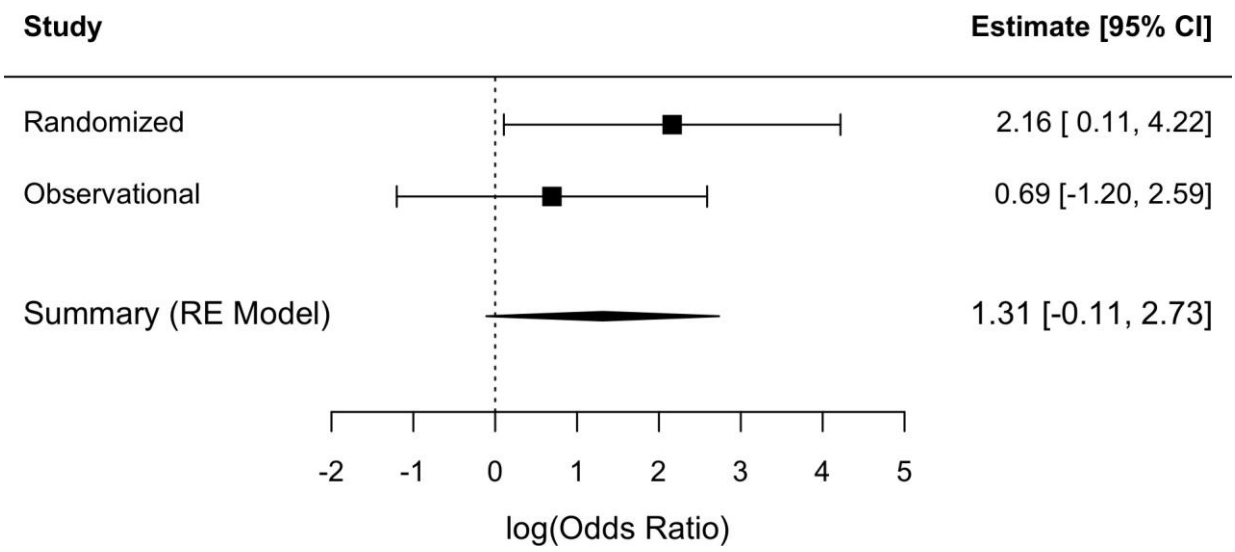
Random Effects Meta-Analysis: Simple Logistic Regression

Simple logistic regression = logistic regression with no adjustment for confounders

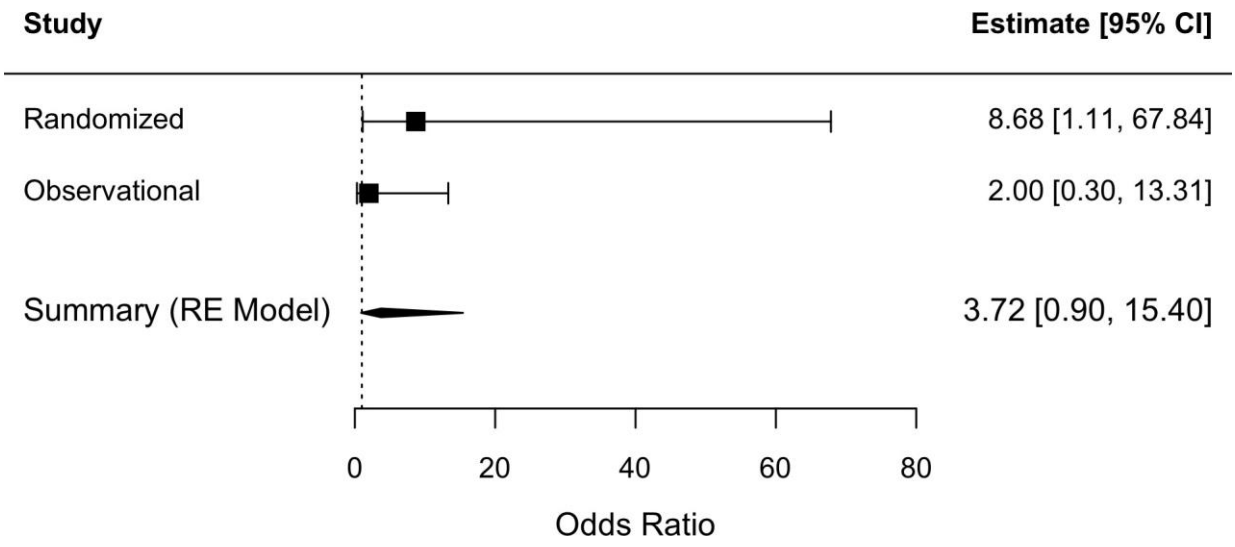
Random Effects Meta-Analysis: Simple Logistic Regression

Study	Estimate (log OR)	95% CI (log OR)	Estimate (OR)	95% CI (OR)	P-value
Randomized	2.162	(0.365, 5.123)	8.684	(1.441, 167.853)	0.049
Observational	0.693	(-1.143, 2.597)	2.000	(0.319, 13.423)	0.459
RE Meta-Analysis	1.313	(-0.108, 2.735)	3.718	(0.897, 15.404)	0.070

Forest Plot: Log Odds Scale



Forest Plot: Odds Scale

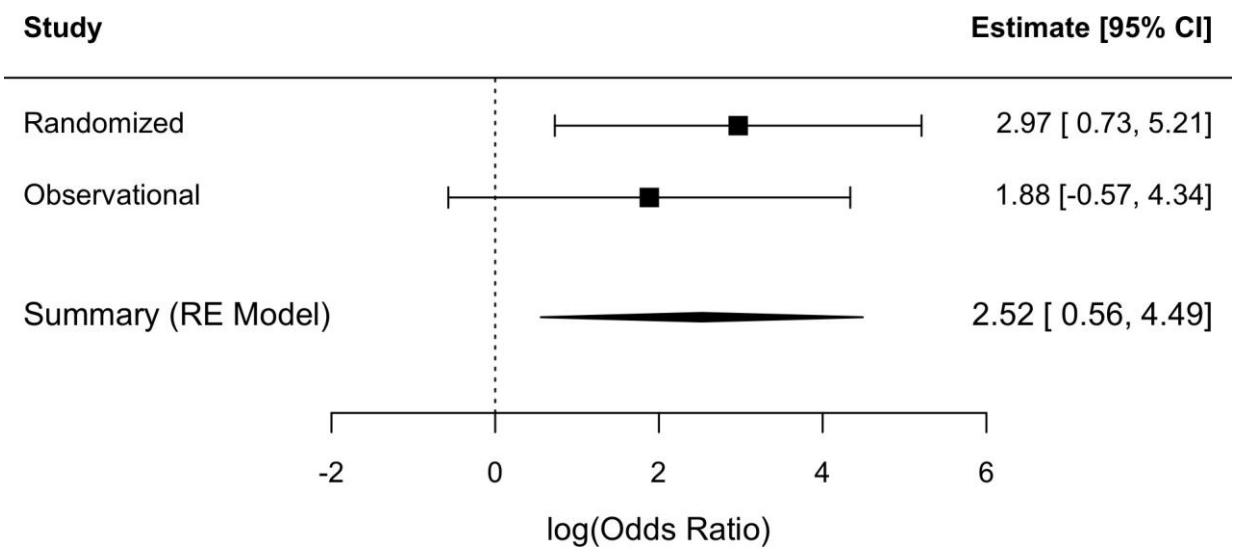


Random Effects Meta-Analysis: Multivariable logistic regression

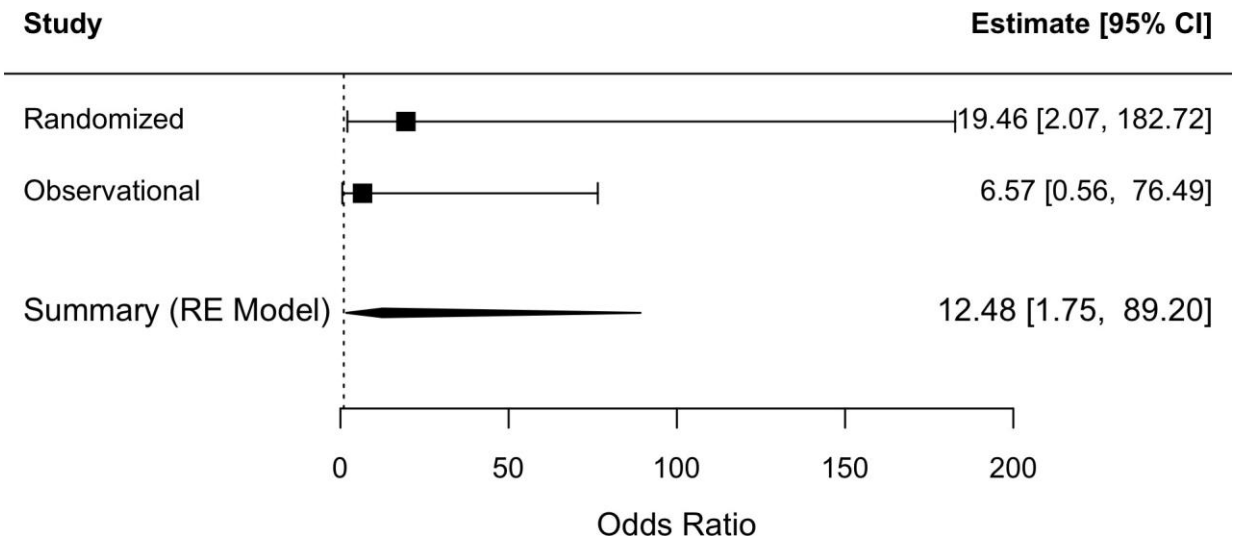
Random Effects Meta-Analysis: Multivariable Logistic Regression

Study	Estimate (log OR)	95% CI (log OR)	Estimate (OR)	95% CI (OR)	P-value
Randomized	2.968	(0.815, 6.261)	19.458	(2.258, 523.717)	0.023
Observational	1.882	(-0.898, 5.783)	6.569	(0.407, 324.629)	0.230
RE Meta-Analysis	2.524	(0.557, 4.491)	12.476	(1.745, 89.199)	0.012

Log Odds Scale



Odds Scale



FAQ

Q: What is overfitting?

A: Overfitting in regression happens when a model is too complex, often due to too many parameters relative to a small sample size, causing it to capture noise rather than the true data pattern. With limited data, the model lacks sufficient information to generalize well, leading to poor performance on new data because it is overly tailored to the specific quirks of the training set.

Q: What is a chunk (or composite) test?

A: A chunk or composite hypothesis test is often used for categorical variables to assess whether the entire variable contributes significantly to the model, rather than just specific categories. By testing all categories together, this approach helps determine if the variable as a whole has an effect, rather than focusing on individual category differences alone.

Q: What can I say when I have a large p-value?

A. This is a friendly reminder that absence of evidence does not mean evidence of absence. A large p-value doesn't prove there is no effect; it simply indicates that we don't have strong evidence to detect a difference. Instead of incorrectly concluding "there were no differences," it's more accurate to say, "there was no evidence of a difference." Additionally, you can use the confidence interval to provide more context. For example, if comparing complication rates, and the confidence interval suggests the difference could be as much as ±5 percentage points, you might say: "If there is a difference between procedure A and procedure B, it is likely no larger than 5 percentage points." Similarly, for odds ratios, if the confidence interval includes a range up to 2, you could say, "There may be at most a 100% increase in odds or up to a 50% decrease in odds." This acknowledges uncertainty without overinterpreting the results.

Answer generously borrowed and adapted from this blog post(link) by Thomas Stewart, PhD.

Also see this article for explanation of interpretation of evidence and suggestions about communicating levels of evidence.

Q. What is validation and optimism-corrected statistics?

A. When a model is fit to a dataset, it tends to perform better on that dataset than it would on new, unseen data, leading to overly optimistic predictions and performance statistics. To adjust for this, techniques like cross-validation or bootstrapping are used to correct for this optimism, producing "optimism-corrected" statistics, which provide a more realistic estimate of the model's performance on future data. These corrections help ensure that the model's results are not overly tailored to the training data and are more generalizable.

Q. What are Brier scores (B)?

A. The Brier score measures how well a model predicts binary or categorical outcomes by calculating the mean squared difference between the predicted probabilities and the actual outcomes. It ranges from 0 to 1, with lower scores indicating better predictive accuracy and calibration. A Brier score closer to 0 suggests that the model's predicted probabilities align well with the observed outcomes, making it a useful metric for evaluating the accuracy of probabilistic predictions.

Q. What is the g-index (g)?

A. The g-index measures the ability of a model to discriminate between outcomes based on the Gini mean difference, reflecting the typical difference between predictions for any two randomly selected individuals. In essence, a higher g-index indicates better discrimination, meaning the model provides a wider range of predictions that can distinguish between individuals more effectively. For example, in a time-to-event analysis, a larger g-index suggests that the model can predict meaningful differences in survival times or hazard rates between patients.

For time-to-event outcomes: A higher g-index indicates larger differences in predicted log relative hazards between patients, which implies better discrimination. For binary/ordinal outcomes: The g-index reflects the typical log odds ratio between predictions for any two patients, meaning how much their odds of an outcome differ. For continuous outcomes: The g-index represents the typical difference in predicted values, showing how far apart predictions are for different patients. Overall, a higher g-index means the model is better at distinguishing between individuals based on their predicted outcomes.

Here are improved explanations for each of your questions:

Q. What is validation and optimism-corrected statistics?

A. When a model is fit to a dataset, it tends to perform better on that dataset than it would on new, unseen data, leading to overly optimistic predictions and performance statistics. To adjust for this, techniques like cross-validation or bootstrapping are used to correct for this optimism, producing "optimism-corrected" statistics, which provide a more realistic estimate of the model's performance on future data. These corrections help ensure that the model's results are not overly tailored to the training data and are more generalizable.

Q. What are Brier scores (B)?

A. The Brier score measures how well a model predicts binary or categorical outcomes by calculating the mean squared difference between the predicted probabilities and the actual outcomes. It ranges from 0 to 1, with lower scores indicating better predictive accuracy and calibration. A Brier score closer to 0 suggests that the model's predicted probabilities align well with the observed outcomes, making it a useful metric for evaluating the accuracy of probabilistic predictions.

Q. What is the g-index (g)?

A. The g-index measures the ability of a model to discriminate between outcomes based on the Gini mean difference, reflecting the typical difference between predictions for any two randomly selected individuals. In essence, a higher g-index indicates better discrimination, meaning the model provides a wider range of predictions that can distinguish between individuals more effectively. For example, in a time-to-event analysis, a larger g-index suggests that the model can predict meaningful differences in survival times or hazard rates between patients.

For time-to-event outcomes: A higher g-index indicates larger differences in predicted log relative hazards between patients, which implies better discrimination. For binary/ordinal outcomes: The g-index reflects the typical log odds ratio between predictions for any two patients, meaning how much their odds of an outcome differ. For continuous outcomes: The g-index represents the typical difference in predicted values, showing how far apart predictions are for different patients. Overall, a higher g-index means the model is better at distinguishing between individuals based on their predicted outcomes.

g is on the original scale gr is on the ratio scale (e.g. odds/hazard) gp is on the probability or risk scale

Q. What is DXY?

A. Dxy, or Somers' Dxy, is a statistic used to measure a model's discrimination ability, which refers to how well the model can distinguish between different outcomes. Dxy ranges from -1 to 1, with 1 indicating perfect discrimination and 0 indicating no better than random guessing. For binary outcomes, Dxy is related to the concordance index (c-index) and can be expressed as Dxy = 2 * (c-index - 0.5). In time-to-event models, Dxy represents the rank correlation between the predicted log relative hazard and observed survival times, similar to a rank-based R2. Higher Dxy values indicate that the model does a better job at ranking individuals according to their predicted outcomes.

For example:

Dxy = 0.5 means 75% of pairs are concordant (the predictions are in the correct order), and 25% are discordant. Dxy = 0.2 means 60% of pairs are concordant, and 40% are discordant. According to Dr. Frank Harrell, there are no universally "acceptable" values for Dxy, as the significance depends on the difficulty of predicting the specific outcome (e.g., lower values may still be informative for difficult-to-predict outcomes like mortality).

Q. What is a restricted cubic spline? What are knots?

A. Restricted cubic splines are a flexible tool used in statistical models to better capture non-linear relationships between a continuous variable and an outcome. Instead of assuming a straight-line relationship, restricted cubic splines break the data into sections at specific points, called "knots," and fit smooth curves between them. The "restricted" part means the curves become straight lines at the ends, ensuring the model doesn't overreact to extreme values. This method helps create a smoother, more realistic fit to the data without making the model too complicated or unstable, making it useful when simple linear relationships aren't sufficient.

Q. What is a tipping point analysis?

A. A tipping point analysis is a type of sensitivity analysis used to assess how robust the results of a study are to potential changes or assumptions, especially in the presence of missing data or uncertain parameters. It identifies the point at which the study's conclusions would change if certain assumptions or values were adjusted. In the context of missing data, for example, tipping point analysis can determine how much the imputed or missing data would need to shift in order to alter the outcome of the analysis. This helps evaluate the stability of the results and the impact of potential biases on the conclusions.

Q. What are the benefits of adding the propensity score as a covariate when there aren't enough events to include multiple covariates directly?

A. When you don't have enough events to directly include multiple covariates in a model, adding the propensity score as a covariate offers several benefits. It reduces dimensionality by summarizing multiple covariates into a single variable, improving model stability and avoiding overfitting. This approach efficiently adjusts for confounders without the need for a large number of parameters, maintaining statistical power and reducing bias in small sample settings. Additionally, it helps avoid multicollinearity and provides flexibility across different types of models, making it a valuable tool when dealing with sparse data.

Reference: Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. doi: 10.1080/00273171.2011.568786. Epub 2011 Jun 8. PMID: 21818162; PMCID: PMC3144483.

R Session info

```
## R version 4.4.1 (2024-06-14)
## Platform: x86_64-apple-darwin20
## Running under: macOS 15.3
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRlapack.dylib;  LAPACK version
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] broom.helpers_1.17.0 broom_1.0.7          rms_7.0-0
##  [4] Hmisc_5.2-2          tangram_0.8.2        magrittr_2.0.3
##  [7] R6_2.5.1             ggplot2_3.5.1        janitor_2.2.1
## [10] dplyr_1.1.4          here_1.0.1           knitr_1.49
## [13] kableExtra_1.4.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6        xfun_0.49           htmlwidgets_1.6.4   lattice_0.22-6
##  [5] vctrs_0.6.5         tools_4.4.1         generics_0.1.3      sandwich_3.1-1
##  [9] tibble_3.2.1        cluster_2.1.6       pkgconfig_2.0.3     Matrix_1.7-0
## [13] data.table_1.16.4   checkmate_2.3.2     lifecycle_1.0.4     compiler_4.4.1
## [17] stringr_1.5.1       MatrixModels_0.5-3  munsell_0.5.1       codetools_0.2-20
## [21] snakecase_0.11.1    SparseM_1.84-2      quantreg_6.00       htmltools_0.5.8.1
## [25] yaml_2.3.10         htmlTable_2.4.3     Formula_1.2-5       tidyr_1.3.1
## [29] pillar_1.10.0       MASS_7.3-60.2       rpart_4.1.23        multcomp_1.4-28
## [33] nlme_3.1-164        tidyselect_1.2.1    digest_0.6.37       mvtnorm_1.3-3
## [37] polspline_1.1.25    stringi_1.8.4       purrr_1.0.2         splines_4.4.1
## [41] rprojroot_2.0.4     fastmap_1.2.0       grid_4.4.1          colorspace_2.1-1
## [45] cli_3.6.3           base64enc_0.1-3     survival_3.6-4      TH.data_1.1-3
## [49] foreign_0.8-86      withr_3.0.2         scales_1.3.0        backports_1.5.0
## [53] lubridate_1.9.4     timechange_0.3.0    rmarkdown_2.29      nnet_7.3-19
## [57] gridExtra_2.3       zoo_1.8-12          evaluate_1.0.1      viridisLite_0.4.2
## [61] rlang_1.1.4         glue_1.8.0          xml2_1.3.6          svglite_2.1.3
## [65] rstudioapi_0.17.1   systemfonts_1.1.0
## To cite R in publications use:
##
##   R Core Team (2024). _R: A Language and Environment for Statistical
##   Computing_. R Foundation for Statistical Computing, Vienna, Austria.
##   <https://www.R-project.org/>.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2024},
##     url = {https://www.R-project.org/},
##   }
##
```

```
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```