European Journal of Surgical Oncology 50 (2024) 108513

Contents lists available at ScienceDirect

European Journal of Surgical Oncology

journal homepage: www.ejso.com

Evaluation of the proportional hazards assumption and covariate adjustment methods in comparative surgical observational studies with time-to-event endpoints

Rui-ming Liang [a,1], Ze-bin Chen [b,1], Qian Zhou [a,c,*]

[a] Department of Medical Statistics, Clinical Trials Unit, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[b] Center of Hepato-Pancreato-Biliary Surgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[c] Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Keywords:
Proportional hazards assumption
Propensity score analysis
Time-to-event endpoint
Surgical oncology
Observational studies

ABSTRACT

Introduction: Comparative studies on surgical treatments with time-to-event endpoints have provided substantial evidence for clinical practice, but the accurate use of survival data analysis and the control of confounding bias remain big challenges.

Methods: This was a survey of surgical studies with survival outcomes published in four general medical journals and five general surgical journals in 2021. The two most concerned statistical issues were evaluated, including confounding control by propensity score analysis (PSA) or multivariable analysis and testing of proportional hazards (PH) assumption in Cox model.

Results: A total of 74 studies were included, comprising 63 observational studies and 11 randomized controlled trials. Among the observational studies, the proportion of studies utilizing PSA in surgical oncology and non-oncology studies was similar (40.9 % versus 36.8 %, P = 0.762). However, the former reported a significantly lower proportion of PH assumption assessments compared to the latter (13.6 % versus 42.1 %, P = 0.020). Twenty-five observational studies (25/63) used PSA methods, but two-thirds of them (17/25) showed unclear balance of baseline data after PSA. And the proportion of PH assumption testing after PSA was slightly lower than that before PSA, but the difference was not statistically significant (24.0 % versus 28.0 %, P = 0.317). Comprehensive suggestions were given on confounding control in survival analysis and alternative resolutions for non-compliance with PH assumption.

Conclusion: This study highlights suboptimal reporting of PH assumption evaluation in observational surgical studies both before and after PSA. Efforts and consensus are needed with respect to the underlying assumptions of statistical methods.

1. Introduction

Surgical procedures for both cancer and non-cancer patients are increasing diverse and sophisticated, which requires physicians to compare the pros and cons of different options and select the best one [1,2]. Randomized controlled trials (RCTs) are typically considered as the gold standard for evaluating efficacy and safety among different treatments [3]. However, the use of RCTs may be limited by methodological and practical challenges, such as recruitment difficulties, poor randomization, high research costs and other issues [4]. Observational studies are found to be more frequently used to evaluate surgical outcomes [5]. In recent years, advancements in real-world data and registry databases have facilitated more studies comparing the effectiveness of different surgical procedures, yielding good results [6,7].

However, the biggest challenge with these observational studies is the confounding bias caused by baseline imbalance [8]. For example, comparing overall survival between surgical and non-surgical groups may introduce bias, as surgically eligible patients often have less severe conditions and longer survival. Multivariable regression analysis and propensity score analysis (PSA) are two commonly used methods to

* Corresponding author. Department of Medical Statistics, Clinical Trials Unit, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China; No.58, Zhongshan Road 2, Guangzhou, 510080, China.
E-mail address: zhouq49@mail.sysu.edu.cn (Q. Zhou).
[1] Contributed equally.

https://doi.org/10.1016/j.ejso.2024.108513
Received 18 March 2024; Received in revised form 3 June 2024; Accepted 26 June 2024
Available online 27 June 2024

address this issue [9]. Multivariable analysis is limited by event numbers, and PSA compensates for this shortcoming [10]. Nevertheless, the quality of PSA application in surgical research needs substantial improvements [5].

On the other hand, more surgical studies focus on long-term endpoints post-surgery, leading to a rise in the application of time-to-event endpoints [2].While Cox regression is the most widely used method for survival endpoints [11], its use in surgical research is far from satisfactory [12,13]. Previous study in surgical RCTs have found that testing and reporting of the proportional hazards (PH) assumption was scarce and violation of the PH assumption was frequently observed [13]. The PH assumption refers to the treatment effects of the intervention and control groups remaining relatively constant throughout the follow-up period [14]. If the standard Cox regression analysis is used ignoring the violation of PH assumption, it may lead to biased results, underpowered trials, or even misleading conclusions [15,16]. Therefore, PH assumption should be assessed before conducting Cox regression and handled if violated. However, in observational surgical studies with survival endpoints, the application and quality reporting of multivariable Cox analysis considering PH assumption and PSA to control bias remains obscure [12].

Therefore, we aimed to assess the testing of the PH assumption in Cox models within comparative surgical RCTs and observational studies with survival endpoints, and to evaluate the application of PSA in observational studies. Comprehensive suggestions on these issues were given according to the application practice. The results of the study are applicable to other observational studies as well.

2. Methods

2.1. Study design and eligibility criteria

This study is a systematic review and methodological evaluation of published surgical papers in five general surgical journals (Annals of Surgery, The British Journal of Surgery, JAMA Surgery, Journal of the American College of Surgeons, Surgery), and four general medical journals (New England Journal of Medicine, British Medical Journal, Lancet, The Journal of the American Medical Association) selected according to their impact factors. We searched Web of Science for reports published between January 1, 2021, and December 31, 2021, in these journals, with document types of clinical trials, reviews, or articles. The main inclusion criteria were observational studies or RCTs comparing the effectiveness and/or safety of different surgical treatments in humans. Detailed search strategy and eligibility criteria were provided in Appendices 1-2. As this was a systematic review of published studies, Institutional Review Board (IRB) approval was not required.

2.2. Data extraction

2.2.1. Baseline characteristics of included articles

For each article, the following information was extracted: journal name, first author's name, study design, disease types, comparison types, number of study groups, sample size, and funding. For RCTs, we also collected the phases of trials, types of hypothesis test, randomization ratio, and use of blinding.

2.2.2. Reporting of PH assumption testing

We assessed the reporting of PH assumption testing regarding survival curves or Cox models by reviewing the Methods and Results sections of an article. We considered authors to have addressed the PH assumption if they reported specific methods for PH assumption testing, or provided methods to address deviations from PH assumption, even without detailed testing methods. Otherwise, we classified it as unassessed. If an article mentioned PH assumption testing, we recorded the detailed methods used and the results whether the PH assumption was met.

2.2.3. Reporting of methods to control for confounding and use of PSA

Through reviewing the methodology section, we identified the methods that used for confounding control, including multivariable analysis, PSA, and others (such as matching based on prognostic factors or coarsened exact matching). For each article that utilized PSA, we first assessed the balance of baseline characteristics before and after PSA. Only when standardized mean difference (SMD) was <0.1, it was considered well-balanced; SMD ≥ 0.1 indicated imbalance; otherwise, it was classified as unclear.

Meanwhile, the following aspects were also evaluated [5]: type of PSA, primary analysis designation, covariates reporting in propensity score (PS) model, PS model and variable selection, number of PS model variables, and reporting and handling methods of missing data. Two authors (QZ and RL) screened all papers independently and any discrepancies were discussed.

2.2.4. Assessment of risk of bias from observational studies

Newcastle-Ottawa Scale (NOS), the most widely used tool to evaluate the quality of observational studies [17], was adopted to assess the risk of bias of all included observational studies. Total score >7: low bias risk; 5–7: moderate; <5: high bias risk [18].

2.2.5. Statistical analysis

Continuous variables were presented as median (interquartile range, IQR) and compared with Wilcoxon rank sum test due to non-normal distribution. Categorical variables were described as numbers and percentages and compared by Chi-square test or Fisher's exact test. For studies employing PSA, McNemar's test was used to compare the difference in the proportion of testing PH assumption. Some studies with non-survival endpoints as primary outcomes but with survival endpoints as secondary outcomes were also included in our research. To confirm the robustness of the results, sensitivity analyses were conducted by including studies with survival endpoints as primary outcomes. All analysis was performed with SAS software 9.4 (SAS Inc., Cary, N.C., USA).

3. Results

3.1. Study characteristics

A total of 7573 articles were screened from five general surgical journals and four general medical journals. Initially, 47 duplicate articles were excluded, followed by the exclusion of 7422 articles based on titles or abstracts. Detailed evaluations of 104 articles were conducted to determine their eligibility for inclusion. Ultimately, 74 articles were included: 63 observational studies and 11 RCTs (Fig. 1). The list of included articles was in Supplementary Table S1. As shown in Table 1, most observational studies compared two groups, while 9 articles (14.3 %) included multiple groups (5 three-group studies, 2 four-group studies, 1 seven-group study, and 1 ten-group study). Among the included observational studies, the principal comparison was "surgery vs surgery", followed by "surgery vs usual care/no surgery ". And the median sample size was 1099 (IQR: 293-6385). All RCTs primarily compared two surgical procedures. And the median sample size in the included RCTs was 530 (IQR: 112-3625). The results for all included studies were also presented by oncology (n = 47) and non-oncology studies (n = 27) in Table 1. Except for participant numbers, no statistically significant differences in characteristics were found between these two types of studies.

3.2. Reporting of PH assumption testing

Table 2 displayed PH assumption assessment results of all included articles. Approximately 18 out of 74 total articles (24.3 %) reported the assessments of PH assumption (22.2 % in observational studies versus 36.4 % in RCTs, P = 0.445). The proportion of PH assumption

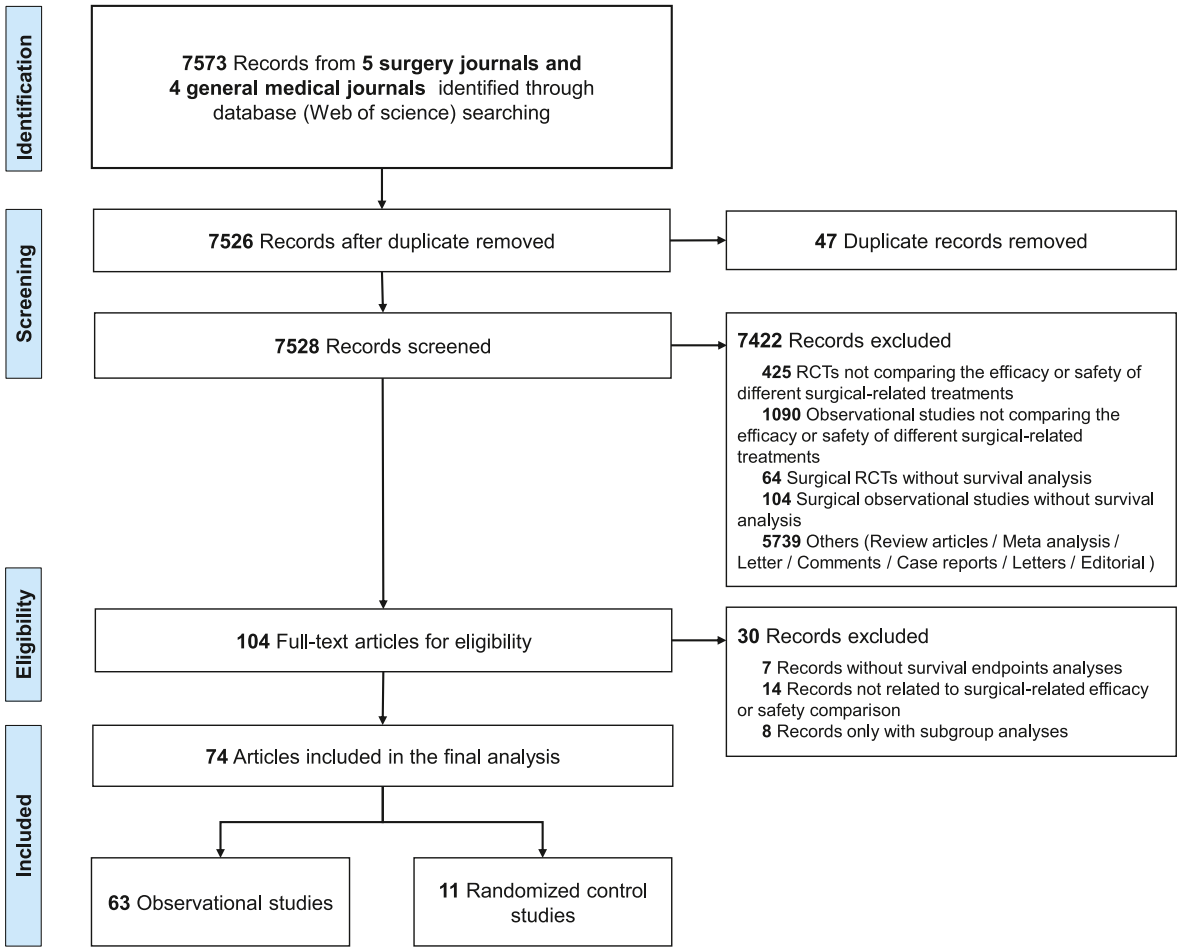

Fig. 1. Flowchart of the study. *RCTs*, randomized controlled trials.

assessment in oncology studies was significantly lower than in non-oncology studies in all included studies (14.9 % versus 40.7 %, P = 0.013). When restricted in observational studies, the difference of proportion of PH assumption assessment between oncology studies and non-oncology studies was also significant (13.6 % versus 42.1 %, P = 0.020, Supplementary Table S6). Methods used for evaluation included graphical assessment of Kaplan-Meier (KM) curves, visual inspection of log-log plots, Schoenfeld residuals, the interaction between treatments and time, Kolmogorov type supremum test, or not reported. Schoenfeld residuals and log-log plots were two most frequently used methods (50.0 % and 22.2 % respectively). No significant differences were found between observational studies and RCTs and between the oncological and non-oncological studies.

3.3. PSA in observational studies

We assessed the methods employed for confounding control among the 63 observational studies. Among these, 73.0 % (46/63) utilized multivariable adjustment, 39.7 % (25/63) employed PSA, and 4.8 % (3/63) used other methods like coarsened exact matching or prognostic factor matching. Twenty studies employed both multivariable adjustment and PSA simultaneously. Among the observational studies, the proportion of studies utilizing PSA in surgical oncology and non-oncology studies was similar (40.9 % versus 36.8 %, P = 0.762, Table 3). Of the observational studies using PSA, around one-third (32.0 %, 8/25) had imbalanced baseline data before adjustment, and 68.0 % (17/25) had unclear balance. After PSA, 40.0 % of articles (10/25) reported good baseline balance, while 60.0 % (15/25) had unclear balance. Matching (76.0 %) was the primary PS method, followed by weighting (24.0 %). Logistic regression model was the most frequently used model to estimate PS (72.0 %, 18/25). Most studies (72 %, 18/25) reported the variables used to construct the PS model, but nearly two thirds (64.0 %, 9/25) still failed to report variable selection methods of model building. Among the 25 observational studies using PSA, 72.0 % of them made conclusions according to results from PSA. Table 3 showed that no significant difference in the employment of PSA was observed between oncology and non-oncology studies.

3.4. PH assumption testing before and after PS

As shown in Table 4, the proportion of testing PH assumption between observational studies with or without PSA was comparable before PSA (28.0 % vs 18.0 %, P = 0.371). Among PSA studies, the proportion of PH assumption testing after PSA was slightly lower than that before PSA, but the difference was not statistically significant (24.0 % versus 28.0 %, P = 0.317).

3.5. Risk of bias assessment for the included observational studies

The overall reporting quality of the 63 included observational studies was moderate, with a mean NOS score of 6.3 ± 0.3 (Supplementary Table S2). Main deficiencies included the absence of clear evidence regarding outcome events before study initiation due to the retrospective cohort nature (92.1 %). Insufficient follow-up data (79.4 %) and inadequate follow-up duration (61.9 %) were two other common issues. Studies were also grouped based on PH assumption assessment and PSA use, but no statistically significant differences in reporting quality were found between these groups (Supplementary Tables S3–S5).

Table 1
Baseline Characteristics of Included articles.

Variables	Total (N = 74)	Observational study (n = 63)	RCT (n = 11)	P value	Oncology study (n = 47)	Non-oncology study (n = 27)	P value
Journal				<0.001[d]			0.005[d]
British Medical Journal	1 (1.4 %)	1 (1.6 %)	0 (0.0 %)		0 (0.0 %)	1 (3.7 %)	
Lancet	2 (2.7 %)	0 (0.0 %)	2 (18.2 %)		0 (0.0 %)	2 (7.4 %)	
New England journal of medicine	4 (5.4 %)	0 (0.0 %)	4 (36.3 %)		0 (0.0 %)	4 (14.8 %)	
The Journal of the American Medical Association	2 (2.7 %)	2 (3.2 %)	0 (0.0 %)		0 (0.0 %)	2 (7.4 %)	
Annals of surgery	33 (44.6 %)	32 (50.8 %)	1 (9.1 %)		22 (46.8 %)	11 (40.8 %)	
JAMA Surgery	5 (6.8 %)	3 (4.8 %)	2 (18.2 %)		4 (8.5 %)	1 (3.7 %)	
Journal of the American College of Surgeons	2 (2.7 %)	2 (3.2 %)	0 (0.0 %)		2 (4.3 %)	0 (0.0 %)	
Surgery	17 (23.0 %)	17 (27.0 %)	0 (0.0 %)		14 (29.8 %)	3 (11.1 %)	
The British journal of surgery	8 (10.7 %)	6 (9.4 %)	2 (18.2 %)		5 (10.6 %)	3 (11.1 %)	
Arms				0.095[d]			0.716[d]
Two	65 (87.8 %)	54 (85.7 %)	11 (100.0 %)		42 (89.4 %)	23 (85.2 %)	
More than 2 groups	9 (12.2 %)	9 (14.3 %)	0 (0.0 %)		5 (10.6 %)	4 (14.8 %)	
Type of comparison				0.105[d]			0.271[d]
Surgery vs Surgery	39 (52.7 %)	31 (49.1 %)	8 (72.7 %)		27 (57.4 %)	12 (44.5 %)	
Surgery vs Usual care/No surgery[a]	18 (24.3 %)	18 (28.6 %)	0 (0.0 %)		10 (21.3 %)	8 (29.6 %)	
Surgery vs Surgery plus other treatments	2 (2.7 %)	1 (1.6 %)	1 (9.1 %)		2 (4.3 %)	0 (0.0 %)	
Surgery vs Drugs[b]	2 (2.7 %)	2 (3.2 %)	0 (0.0 %)		0 (0.0 %)	2 (7.4 %)	
Others[c]	13 (17.6 %)	11 (17.5 %)	2 (18.2 %)		8 (17.0 %)	5 (18.5 %)	
Number of patients analyzed, median (IQR)	1039 (293-3887)	1099 (293-6385)	530 (112-3625)	0.238	676 (196-2059)	3825 (1158–65,416)	<0.001
Funding				0.102[†]			0.281[e]
Yes	35 (47.3 %)	27 (42.9 %)	8 (72.7 %)		20 (42.6 %)	15 (55.6 %)	
No	39 (52.7 %)	36 (57.1 %)	3 (27.3 %)		27 (57.4 %)	12 (44.4 %)	

RCT, randomized controlled trial; *IQR*, interquartile range.
[a] Two studies with three-group parallel comparisons were included.
[b] One study with three-group parallel comparisons was included.
[c] Six studies with more than two groups comparisons were included.
[d] Fisher's exact test.
[e] Chi-square test.

Table 2
Reporting of proportional hazards (PH) assumption assessments.

Characteristics	Total	Observational study	RCT	P value	Oncology study	Non-oncology study	P value
Assessment of PH assumption **(N=74)**				0.445			0.013
No	56 (75.7 %)	49 (77.8 %)	7 (63.6 %)		40 (85.1 %)	16 (59.3 %)	
Yes	18 (24.3 %)	14 (22.2 %)	4 (36.4 %)		7 (14.9 %)	11 (40.7 %)	
Results of PH assumption testing reported **(N=18)**				1.000			0.627
No	6 (33.3 %)	5 (35.7 %)	1 (25.0 %)		3 (42.9 %)	3 (27.3 %)	
Yes	12 (66.7 %)	9 (64.3 %)	3 (75.0 %)		4 (57.1 %)	8 (72.7 %)	
Methods used to assess the PH assumption **(N=18)**[a]							
Graphical evaluation of KM curves	1 (5.6 %)	1 (7.1 %)	0 (0.0 %)	1.000	0(0.0 %)	1 (9.1 %)	1.000
Log-log plots	4 (22.2 %)	3 (21.4 %)	1 (25.0 %)	1.000	1 (14.3 %)	3 (27.3 %)	1.000
Schoenfeld residuals	9 (50.0 %)	6 (42.9 %)	3 (75.0 %)	0.577	3 (42.9 %)	6 (54.5 %)	1.000
Treatment * time interaction	1 (5.6 %)	1 (7.1 %)	0 (0.0 %)	1.000	0 (0.0 %)	1 (9.1 %)	1.000
Kolmogorov type supremum test in the cox model	1 (5.6 %)	1 (7.1 %)	0 (0.0 %)	1.000	0 (0.0 %)	1 (9.1 %)	1.000
Not reported	5 (27.8 %)	5 (35.7 %)	0 (0.0 %)	0.278	3 (42.9 %)	2 (18.2 %)	0.326

RCT, randomized controlled trial; *PH*, proportional hazards; *KM*, Kaplan-Meier.
[a] An article can employ multiple PH assumption testing methods simultaneously.

3.6. Sensitivity analysis

In the sensitivity analysis, 57 observational studies and 8 RCTs with survival endpoints as the primary analysis were included. The baseline characteristics, PH assumption, and PS reporting did not show significant changes (Data not shown).

4. Discussion

Proportional hazards assumption and covariate adjustment are two crucial issues in clinical research with survival data. This study identified 74 comparative surgical studies with time-to-event endpoints, the majority of which were observational studies. Most studies did not evaluate PH assumption in Cox regression in observational studies, with less than one-fourth reporting PH assumption testing. In addition, PH assumption assessments were significantly lower in surgical oncology studies compared to non-oncology studies. Furthermore, the study highlights the suboptimal reporting of PH assumption evaluation in observational surgical studies both before and after PSA. These findings are consistent with the studies by Lin et al. [19] and Handorf et al.'s [20], which highlighted two main challenges in observational studies with survival outcomes: non-proportional hazards and confounding bias.

For survival data analyzed with Cox regression, addressing violations of the PH assumption and accurately choosing statistical methods are crucial [16,21]. However, our study revealed poor reporting of PH

Table 3
Reporting of propensity score (PS) analysis in observational studies.

Characteristics	Total	Oncology study	Non-oncology study	*P* value
Reporting of PS analysis in all observational studies (n=63)				
PS analysis used				
No	38 (60.3 %)	26 (59.1 %)	12 (63.2 %)	0.762
Yes	25 (39.7 %)	18 (40.9 %)	7 (36.8 %)	
Reporting of PS analysis in selected observational studies using this method (n = 25)				
Baseline data balance a prior to PS analysis[a]				0.640
No	8 (32.0 %)	5(27.8 %)	3 (42.9 %)	
Yes	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	
Not clear	17 (68.0 %)	13(72.2 %)	4 (57.1 %)	
Baseline data balance after PS analysis[a]				0.378
No	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	
Yes	10(40.0 %)	6(33.3 %)	4 (57.1 %)	
Not clear	15(60.0 %)	12(66.7 %)	3 (42.9 %)	
PS methods[b]				
Matching	19 (76.0 %)	15(83.3 %)	4 (57.1 %)	0.299
Adjustment	1 (4.0 %)	0(0.0 %)	1 (14.3 %)	0.280
Weighting	6 (24.0 %)	3(16.7 %)	3 (42.9 %)	0.299
Stratification	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	NA
Model used to estimate PS				0.171
Logistic regression model	18 (72.0 %)	11 (61.1 %)	7 (100.0 %)	
Generalized boosted method	1 (4.0 %)	1 (5.6 %)	0 (0.0 %)	
Not reported	6 (24.0 %)	6 (33.3 %)	0 (0.0 %)	
Details of variables included in PS reported	18 (72.0 %)	12(66.7 %)	6 (85.7 %)	0.626
Number of variables included in PS models, median (IQR)	10 (6, 17)	10 (6, 15)	8 (6, 25)	0.832
Variable selection methods used in PS models				0.430
Data-driven	3 (12.0 %)	3 (16.7 %)	0 (0.0 %)	
Pre-defined	6 (24.0 %)	3 (16.7 %)	3 (42.9 %)	
Not reported	16 (64.0 %)	12 (66.7 %)	4 (57.1 %)	
Rate of missing data reported	7 (28.0 %)	4 (22.2 %)	3 (42.9 %)	0.355
Methods to handle missing data reported	8 (32.0 %)	5 (27.8 %)	3 (42.9 %)	0.640
PS analysis considered as the primary analysis	18 (72.0 %)	12(66.7 %)	6 (85.7 %)	0.626

PS, propensity score; *IQR*, interquartile range; *NA*, not available.
[a] Variable with standardized mean difference (SMD) < 0.1 was considered well balanced between comparison groups.
[b] An article can employ multiple PSA methods simultaneously.

Table 4
Evaluation of proportional hazards (PH) assumption before and after propensity score (PS) analysis.

Characteristics	PS used		Total	*P* value
	Yes	No		
Before PS				0.371[a]
PH assumption tested	7 (28.0 %)	7 (18.0 %)	14 (22.0 %)	
PH assumption not tested	18 (72.0 %)	31 (82.0 %)	49 (78.0 %)	
After PS				–
PH assumption tested	6 (24.0 %)	–	–	
PH assumption not tested	19 (76.0 %)	–	–	
***P* value**				
	0.317[b]	–	–	

PS, propensity score; *PH*, proportional hazards.
[a] Chi-square test to compare the proportions of PH assumption testing before PS analysis between PS used or not.
[b] McNemar's test to compare the difference of PH assumption testing in studies using PS.

assumption testing. Typically, the reporting of Cox analysis assumptions may receive less attention when survival endpoints are not the primary focus of a study. To ensure robustness, we conducted sensitivity analyses using survival endpoints as the primary focus, with no alteration to our conclusions. Previous studies on PH assumption assessment in various field reported proportions ranging from 0 to 53 % [12,22,23]. For instance, Kuitunen et al. examined studies on total joint arthroplasty, and found that 40 % mentioned PH assessment, with 36 % reporting PH assumption violation [12]. Our proportion was slightly lower, possibly because we didn't restrict surgery types. Kuemmerli et al. focused on PH assumption reporting in surgical research [13], comparing adherence to CONSORT guidelines and statistical method reporting in surgical RCTs and general medical journals. They found poorer adherence and less PH reporting in surgical RCTs, with only two out of 25 studies reporting formal PHA testing [13]. These findings highlighted the need for improved PH reporting in current literature. In addition, only 11 % of studies used PSA and assessed PH simultaneously. Even after PSA, considering PH assumption testing with Cox regression is essential to avoid biased results [24–26].

Besides multivariable Cox model, PSA, aiming to mimic RCTs [5], has become a prevalent approach for bias control in observational studies. Only 25 articles (40 %) among the included observational studies employed PSA to control for confounding factors. Matching was the most commonly used PSA method, with the logistic model being the most frequently used PS model [27,28]. Similar to other studies [10,29,30], we outlined seven conventional PSA steps in Supplementary Fig. S1. However, we observed a low proportion of studies testing PH assumption of Cox regression after PSA. Prior research primarily concentrated on one aspect, with limited studies addressing both issues simultaneously [19,31]. Recently, researchers have increasingly emphasized the importance of concurrently considering the PH assumption and PS [19,20,31–33], indicating a promising research direction ahead. For example, Conner proposed adjusted restricted mean survival time (RMST) using and integrated KM estimator with inverse probability weighting [33]. This method performs well in both proportional and non-proportional hazards scenarios and has gained acceptance recently. Similarly, Ni introduced the Stratified RMST Model, aiming to reduce confounding effects in observational studies by simultaneously adjusting PS [31]. In summary, if the PH assumption deviates when using Cox after PSA, it's advisable to consider these alternative solutions rather than applying statistical methods blindly.

To aid researchers in choosing suitable statistical methods and enhancing literature reporting, we've outlined key points and analytical steps for using these methods in Table 5 and Supplementary Fig. S1 as a reference. We summarized six methods commonly used for the PH

Table 5
Summary of several methods to analyze survival data with non-proportional.

Methods	Description	Explanation	Advantages (+) and disadvantages (−)	R packages and functions
Stratified Cox regression [38]	Variables not meeting PH are stratified, while those meeting PH are directly included in the Cox model.	Applicable for few non-proportional hazards variables with limited categories and secondary focus. Not applicable for many non-proportional hazards variables or when their effects are essential.	+ General adjustment for confounding variables. - Cannot estimate the effects of stratified variables. - Reduced precision and power with excessive stratified variables.	Package survival, "coxph" function with strata
Separate Cox regression for different time periods [12,38]	Fit separate Cox regressions within different time intervals to obtain corresponding HR, ensuring PH assumption within each interval.	Applicable when distinct short-term and long-term effects exist or when HR abruptly changes, such as a transition from >1 to <1. Segmentation based on prior clinical knowledge to determine the appropriate breakpoints.	+ Easy interpretation. - Selection of time points and segments affects HR estimation. - PH assumption must be satisfied within each interval, sometimes impractical.	Package survival, "coxph" function; Package coxphw, "coxphw" function with template = "PH"
Cox regression with time-dependent covariates [38,44]	To address non-proportional hazards variables, introduce an interaction term between the variable and time in the Cox regression, creating a time-dependent Cox model.	It can assess PH assumptions or capture time-dependent effects of study factors. A time-dependent covariate is the product of a covariate with a predefined function of time. The function choice, like $\gamma(t) = t$ or $\gamma(t) = \log(t)$, affects results, considering model complexity and overfitting.	+ Flexible time function selection based on data. + Obtain effects in multiple time periods with one model. - Requires careful function selection for study factor-time interaction. - Complex modeling risks overfitting. - Limited by sample size and event count.	Package survival, "coxph" function; Package coxphw, "coxphw" function with template = "PH" and a predefined time-by-covariate interaction
Weighted Cox regression [38]	Weighted Cox regression assigns weights to event times based on survivor function and follow-up probability.	A well interpretable average effect (average HR) could be provided. It can be considered when only a single effect size is needed.	+ No need for additional parameters. + Applicable to small sample sizes. - Averaged HR may hide directional changes over time and temporal treatment effects.	Package coxphw, "coxphw" function with template = "AHR"
Restricted mean survival time, RMST [33,45,46]	RMST refers to the area under the KM curve up to a specified time point, which indicates the treatment effect up to that particular time point. A longer RMST indicates a better treatment effect.	Predefined time points significantly influence results, often chosen for clinical relevance or slightly below maximum follow-up times. Initially used in RCTs, RMST is now widely applied in observational studies, incorporating covariate adjustment methods such as ANCOVA, pseudo-values, and IPW.	+ Absolute differences can be provided, more conservative and interpretable. + Independent of PH assumption. Robust test results regardless of survival curve crossing. - Require predefined time points without a unified standard.	"akm.rmst" function. The function is available on GitHub (https://github.com/s-conner/akm-rmst). Package survRM2, "rmst2" function.
Landmark Cox analysis [47,48,49]	Analysis is based on a predefined landmark time, excluding subjects with events or lost follow-up before it. Only subjects' status at the landmark time is considered.	Suitable for investigating post-baseline factor relationships with survival outcomes, especially when curves cross. Landmark time should be chosen based on clinical experience or literature review; if unsure, conduct sensitivity analyses with multiple time points.	+ Easily to visualize. + Provides effect sizes for different time periods and addresses immortal time bias. - Results may vary with different landmark times. - Early landmark time may lead to misclassification, while late landmark time may reduce sample size and power. - Excluding patients with endpoint events before the specified time may disrupt trial randomization.	Package survival, "coxph" function.

PH, proportional hazards; *HR*, hazard ratio; *IPW*, inverse probability weighting; *RMST*, Restricted mean survival time; *KM*, Kaplan-Meier; *ORR*, objective response rates; *OS*, overall survival.

assumption testing, including graphical methods (KM curves, log-log plots, Schoenfeld residuals) [34,35], and three statistical tests (Grambsch-Therneau test, treatment-time interactions, Kolmogorov-type supremum test) [23,24,35,36]. All of these methods were used in the articles we evaluated. As results may vary, it's advisable to utilize both graphical methods for visualizations and statistical tests for a thorough assessment [37]. Meeting Cox regression's basic assumptions, especially the PH assumption, enables standard Cox regression usage. Otherwise, alternative methods must be considered. There were two studies reporting that the PH assumption was violated after assessment, one employed Cox model with time-dependent covariates and another one fitted different cox models in different time periods. However, there is no one size fits all approach currently. Table 5 outlines six methods suitable for survival data analysis when PH assumption is violated. The application scenarios, as well as the advantages and disadvantages of each method, were provided. We want to emphasize that when choosing appropriate method for survival data analysis after the violation of PH assumption, researchers need to consider various aspects, including analysis objectives, evidence of time-dependent effects, event numbers, sample size, and relevance of non-proportional factors [38]. Besides the methods listed in Table 5, parametric models like accelerated failure time models [39], or more flexible spline-based approaches such as the Royston-Parmar spline model [40,41], serve as alternatives to semi-parametric modeling. Random Survival Forest (RSF), a machine learning method, is also gaining popularity for survival data analysis due to its freedom from

proportional hazards assumptions and advantages with high-dimensional and nonlinear data [42,43]. After understanding their analysis needs, researchers can choose methods by weighing their advantages and disadvantages.

There are some limitations in the study. Firstly, we did not conduct a comprehensive search for articles comparing surgical treatments but focused instead on those published in selected nine journals over a period of one year. As the study aimed to assess reporting quality and make methodological evaluations, a snapshot of articles in related top journals could provide representative results. In future studies, we plan to include articles spanning multiple years to analyze potential shifts or patterns. Secondly, our findings regarding the assessment of PH assumption testing were based on authors' reporting, without reconstructing individual patient-level data for secondary confirmation. Some studies may have performed PH assumption testing but not reported it, potentially leading to underestimation of the assessment proportion. However, our findings were consistent with previous research, showing a similar proportion of deviation from the PH assumption, ranging from 24 % to 28 % [15,21,50]. Thirdly, this study did not limit the disease type, potentially affecting reporting across different disease areas.

5. Conclusions

In observational studies comparing different surgical options for survival outcomes, there is a low proportion of studies employing survival analysis and adequately reporting the PH assumption testing, as well as using PSA to control for confounding. Notably, the reporting of PH assumption testing after employing PSA is even lower. Ignoring the assumptions of statistical methods may lead to unreliable results, and it's crucial to explore alternative statistical methods if assumptions are not met. This study provided a methodological paradigm to use statistical methods in surgical comparative research with survival outcomes. In the future, precise method use and reporting will enhance surgical research quality.

Funding

No funding.

CRediT authorship contribution statement

Rui-ming Liang: Formal analysis, Writing – original draft. **Ze-bin Chen:** Writing – original draft. **Qian Zhou:** Study design, formal analysis, Writing – review & editing. All the authors have read and approved the final manuscript.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejso.2024.108513.

References

[1] Are C, Murthy SS, Sullivan R, et al. Global Cancer Surgery: pragmatic solutions to improve cancer surgery outcomes worldwide. Lancet Oncol 2023;24:e472–518. https://doi.org/10.1016/s1470-2045(23)00412-6.
[2] Robinson NB, Fremes S, Hameed I, et al. Characteristics of randomized clinical trials in surgery from 2008 to 2020: a systematic review. JAMA Netw Open 2021;4:e2114494. https://doi.org/10.1001/jamanetworkopen.2021.14494.
[3] Polley MC, Schwartz D, Karrison T, et al. Leveraging external control data in the design and analysis of neuro-oncology trials: pearls and perils. Neuro Oncol 2024. https://doi.org/10.1093/neuonc/noae005.
[4] Pronk AJM, Roelofs A, Flum DR, et al. Two decades of surgical randomized controlled trials: worldwide trends in volume and methodological quality. Br J Surg 2023;110:1300–8. https://doi.org/10.1093/bjs/znad160.
[5] Lonjon G, Porcher R, Ergina P, et al. Potential pitfalls of reporting and bias in observational studies with propensity score analysis assessing a surgical procedure: a methodological systematic review. Ann Surg 2017;265:901–9. https://doi.org/10.1097/SLA.0000000000001797.
[6] Aminian A, Wilson R, Al-Kurd A, et al. Association of bariatric surgery with cancer risk and mortality in adults with obesity. JAMA 2022;327:2423–33. https://doi.org/10.1001/jama.2022.9009.
[7] Che WQ, Li YJ, Tsang CK, et al. How to use the Surveillance, Epidemiology, and End Results (SEER) data: research design and methodology. Mil Med Res 2023;10:50. https://doi.org/10.1186/s40779-023-00488-2.
[8] Courvoisier DS, Lauper K, Kedra J, et al. EULAR points to consider when analysing and reporting comparative effectiveness research using observational data in rheumatology. Ann Rheum Dis 2022;81:780–5. https://doi.org/10.1136/annrheumdis-2021-221307.
[9] Andrew BY, Alan Brookhart M, Pearse R, et al. Propensity score methods in observational research: brief review and guide for authors. British journal of anaesthesia 2023;131:805–9. https://doi.org/10.1016/j.bja.2023.06.054.
[10] Benedetto U, Head SJ, Angelini GD, et al. Statistical primer: propensity score matching and its alternatives. Eur J Cardio Thorac Surg 2018;53:1112–7. https://doi.org/10.1093/ejcts/ezy167.
[11] Kim TO, Kang DY, Ahn JM, et al. Impact of target lesion revascularization on long-term mortality after percutaneous coronary intervention for left main disease. JACC Cardiovasc Interv 2024;17:32–42. https://doi.org/10.1016/j.jcin.2023.10.068.
[12] Kuitunen I, Ponkilainen VT, Uimonen MM, et al. Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. BMC Musculoskelet Disord 2021;22:489. https://doi.org/10.1186/s12891-021-04379-2.
[13] Kuemmerli C, Sparn M, Birrer DL, et al. Prevalence and consequences of non-proportional hazards in surgical randomized controlled trials. Br J Surg 2021;108:e247–8. https://doi.org/10.1093/bjs/znab110.
[14] Freidlin B, Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? J Clin Oncol : official journal of the American Society of Clinical Oncology 2019;37:3455–9. https://doi.org/10.1200/jco.19.01681.
[15] Alexander BM, Schoenfeld JD, Trippa L. Hazards of hazard ratios - deviations from model assumptions in immunotherapy. N Engl J Med 2018;378:1158–9. https://doi.org/10.1056/NEJMc1716612.
[16] Stensrud MJ, Hernan MA. Why test for proportional hazards? JAMA 2020;323:1401–2. https://doi.org/10.1001/jama.2020.1267.
[17] Low CJW, Ling RR, Lau M, et al. Mechanical circulatory support for cardiogenic shock: a network meta-analysis of randomized controlled trials and propensity score-matched studies. Intensive Care Med 2024. https://doi.org/10.1007/s00134-023-07278-3.
[18] Tan NKW, Tang A, MacAlevey N, et al. Risk of suicide and psychiatric disorders among isotretinoin users: a meta-analysis. JAMA dermatology 2024;160:54–62. https://doi.org/10.1001/jamadermatol.2023.4579.
[19] Lin Z, Zhao D, Lin J, et al. Statistical methods of indirect comparison with real-world data for survival endpoint under non-proportional hazards. J Biopharm Stat 2022;32:582–99. https://doi.org/10.1080/10543406.2022.2080696.
[20] Handorf EA, Smaldone MC, Movva S, et al. Analysis of survival data with nonproportional hazards: a comparison of propensity-score-weighted methods. Biometrical journal Biometrische Zeitschrift 2022:e202200099. https://doi.org/10.1002/bimj.202200099.
[21] Trinquart L, Jacot J, Conner SC, et al. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol : official journal of the American Society of Clinical Oncology 2016;34:1813–9. https://doi.org/10.1200/JCO.2015.64.2488.
[22] Chai-Adisaksopha C, Iorio A, Hillis C, et al. A systematic review of using and reporting survival analyses in acute lymphoblastic leukemia literature. BMC Hematol 2016;16:17. https://doi.org/10.1186/s12878-016-0055-7.
[23] Jachno K, Heritier S, Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. BMC Med Res Methodol 2019;19:103. https://doi.org/10.1186/s12874-019-0749-1.
[24] Makkar RR, Yoon SH, Chakravarty T, et al. Association between transcatheter aortic valve replacement for bicuspid vs tricuspid aortic stenosis and mortality or stroke among patients at low surgical risk. JAMA 2021;326:1034–44. https://doi.org/10.1001/jama.2021.13346.
[25] Li RA, Liu L, Arterburn D, et al. Five-year longitudinal cohort study of reinterventions after sleeve gastrectomy and roux-en-Y gastric bypass. Ann Surg 2021;273:758–65. https://doi.org/10.1097/sla.0000000000003401.
[26] Garland SK, Falster MO, Beiles CB, et al. Long-term outcomes following elective repair of intact abdominal aortic aneurysms: a comparison between open surgical and endovascular repair using linked administrative and clinical registry data. Ann Surg 2023;277:e955–62. https://doi.org/10.1097/sla.0000000000005259.
[27] Yao XI, Wang X, Speicher PJ, et al. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. J Natl Cancer Inst 2017;109. https://doi.org/10.1093/jnci/djw323.

[28] Akmal SR, Beier MA, August DA. Propensity score methods in the surgical oncology literature. Surg Oncol 2022;42:101776. https://doi.org/10.1016/j.suronc.2022.101776.
[29] Loke YK, Mattishent K. Propensity score methods in real-world epidemiology: a practical guide for first-time users. Diabetes, obesity & metabolism 2020;22(Suppl 3):13–20. https://doi.org/10.1111/dom.13926.
[30] Narita K, Tena JD, Detotto C. Causal inference with observational data: a tutorial on propensity score analysis. Leader Q 2023;34. https://doi.org/10.1016/j.leaqua.2023.101678.
[31] Ni A, Lin Z, Lu B. Stratified restricted mean survival time model for marginal causal effect in observational survival data. Ann Epidemiol 2021;64:149–54. https://doi.org/10.1016/j.annepidem.2021.09.016.
[32] Lu B, Cai D, Tong X. Testing causal effects in observational survival data using propensity score matching design. Stat Med 2018;37:1846–58. https://doi.org/10.1002/sim.7599.
[33] Conner SC, Sullivan LM, Benjamin EJ, et al. Adjusted restricted mean survival times in observational studies. Stat Med 2019;38:3832–60. https://doi.org/10.1002/sim.8206.
[34] Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med 1995;14:1707–23. https://doi.org/10.1002/sim.4780141510.
[35] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515–26.
[36] Mentias A, Smedira NG, Krishnaswamy A, et al. Survival after septal reduction in patients >65 Years old with obstructive hypertrophic cardiomyopathy. J Am Coll Cardiol 2023;81:105–15. https://doi.org/10.1016/j.jacc.2022.10.027.
[37] Meuli L, Kuemmerli C. The hazard of non-proportional hazards in time to event analysis. Eur J Vasc Endovasc Surg : the official journal of the European Society for Vascular Surgery 2021;62:495–8. https://doi.org/10.1016/j.ejvs.2021.05.036.
[38] Dunkler D, Ploner M, Schemper M, et al. Weighted cox regression using the R package coxphw. J Stat Softw 2018;84:1–26. https://doi.org/10.18637/jss.v084.i02.
[39] Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 1992;11:1871–9. https://doi.org/10.1002/sim.4780111409.
[40] Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med 2002;21:2175–97. https://doi.org/10.1002/sim.1203.
[41] Royston P, Parmar MK. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. Trials 2014;15:314. https://doi.org/10.1186/1745-6215-15-314.
[42] Tian D, Yan HJ, Huang H, et al. Machine learning-based prognostic model for patients after lung transplantation. JAMA Netw Open 2023;6:e2312022. https://doi.org/10.1001/jamanetworkopen.2023.12022.
[43] Ishwaran Hemant, Kogalur Udaya B, Blackstone Eugene H, et al. Random survival forests. Ann Appl Stat 2008:841–60. https://doi.org/10.1214/08-AOAS169.
[44] Zhang Z, Reinikainen J, Adeleke KA, et al. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med 2018;6:121. https://doi.org/10.21037/atm.2018.02.12.
[45] Trinquart L, Bill-Axelson A, Rider JR. Restricted mean survival times to improve communication of evidence from cancer randomized trials and observational studies. Eur Urol 2019;76:137–9. https://doi.org/10.1016/j.eururo.2019.04.002.
[46] Charu V, Tian L, Kurella Tamura M, et al. Using restricted mean survival time to improve interpretability of time-to-event data analysis. Clin J Am Soc Nephrol 2024;19:260–2. https://doi.org/10.2215/cjn.0000000000000323.
[47] Dafni U. Landmark analysis at the 25-year landmark point. Circulation Cardiovascular quality and outcomes 2011;4:363–71. https://doi.org/10.1161/circoutcomes.110.957951.
[48] Sun G, Yafasova A, Baslund B, et al. Long-term risk of heart failure and other adverse cardiovascular outcomes in granulomatosis with polyangiitis: a nationwide cohort study. J Rheumatol 2022;49:291–8. https://doi.org/10.3899/jrheum.210677.
[49] Putter H, van Houwelingen HC. Understanding landmarking and its relation with time-dependent cox regression. Statistics in biosciences 2017;9:489–503. https://doi.org/10.1007/s12561-016-9157-9.
[50] Liang F, Zhang S, Wang Q, et al. Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. Ann Oncol 2018;29:1320–4. https://doi.org/10.1093/annonc/mdy075.