Defining benchmark values for outcomes of comprehensive resection of primary retroperitoneal liposarcoma: a retrospective multicenter study

Fabio Tirotta,[a,*] Marco Fiore,[b] Sylvie Bonvalot,[c] Dirk Strauss,[d] Piotr Rutkowski,[e] David E. Gyorki,[f] Winan J. van Houdt,[g] Dario Callegaro,[b] Markus Albertsmeier,[h] Dimitri Tzanis,[c] Ferdinando Cananzi,[i,j] Jason K. Sicklick,[k] John Mullinax,[l] Michelle Wilkinson,[d] Valerie P. Grignol,[m] Kenneth Cardona,[n] Toufik Bouhadiba,[c] Marko Novak,[o] Sergio Valeri,[p] Mark Fairweather,[q] Samuel J. Ford,[a,r] Jacek Skoczylas,[e] Hayden Snow,[f] Andrew J. Hayes,[d] James Hodson,[s] Chandrajit P. Raut,[q] and Alessandro Gronchi,[b] on behalf of the Transatlantic Australasian Retroperitoneal Sarcoma Working Group

[a]Department of Sarcoma and General Surgery, Midlands Abdominal and Retroperitoneal Sarcoma Unit, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[b]Department of Surgery, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[c]Department of Surgery, Institut Curie, Paris, France
[d]Department of Academic Surgery, Sarcoma Unit, The Royal Marsden Hospital NHS Foundation Trust, London, UK
[e]Department of Soft Tissue/Bone Sarcoma and Melanoma, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland
[f]Division of Cancer Surgery, Peter MacCallum Cancer Centre, and Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3000, Australia
[g]Department of Surgical Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[h]Department of General, Visceral and Transplantation Surgery, LMU University Hospital, Ludwig-Maximilians-Universität Munich, Munich, Germany
[i]Sarcoma, Melanoma and Rare Tumors Surgery Unit, IRCCS Humanitas Research Hospital, Milan, Italy
[j]Department of Biomedical Sciences, Humanitas University, Milan, Italy
[k]Division of Surgical Oncology, Department of Surgery, UC San Diego, San Diego, CA, USA
[l]Sarcoma Department, Moffitt Cancer Center, Tampa, FL, USA
[m]Department of Surgical Oncology, The Ohio State University Wexner Medical Center, Columbus, OH, USA
[n]Winship Cancer Institute, Emory University, Atlanta, GA, USA
[o]Department of Surgical Oncology, Institute of Oncology Ljubljana, Ljubljana, Slovenia
[p]Department of Surgery, Università Campus Bio-Medico, Roma, Italy
[q]Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
[r]Institute of Cancer and Genomic Science, University of Birmingham, Birmingham, UK
[s]Research Development and Innovation, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Summary

Background Comprehensive resection represents the standard of care for patients affected by retroperitoneal well- or dedifferentiated liposarcoma (WDLPS/DDLPS). However, reference values to indicate the best achievable results are currently lacking. As such, the study aimed to define clinically relevant benchmark values for intra- and postoperative outcomes of patients undergoing comprehensive resection for primary retroperitoneal WDLPS/DDLPS.

Methods The international, prospectively maintained Retroperitoneal Sarcoma Registry (RESAR; NCT03838718) was used to calculate benchmark values for 22 outcomes, including intraoperative factors, and rates of complications, recurrence and survival. Only low-risk patients undergoing comprehensive resection for WDLPS/DDLPS at high-volume centers between 1st January 2017 and 31st December 2021 were used to calculate the benchmark values. Specifically, "low risk" was defined as age <75 years, with minimal comorbidities, and undergoing a "standard" comprehensive resection including at least colon and kidney with or without other organs—excluding those associated with significant morbidity (e.g., pancreas). Benchmark values were defined based on the 25th or 75th percentiles of the center-level data. To validate the benchmark values, these were applied to two cohorts expected to have inferior outcomes, which were defined by changing one of the exclusion criteria; namely those treated in low-volume centers, and those with American Society of Anesthesiologists (ASA) score ≥3 ("ASA ≥ 3").

eClinicalMedicine 2025;84: 103280
Published Online xxx
https://doi.org/10.1016/j.eclinm.2025.103280

*Corresponding author. Department of Sarcoma and General Surgery, Midlands Abdominal and Retroperitoneal Sarcoma Unit, University Hospitals Birmingham NHS Foundation Trust, Mindelsohn Way, Birmingham B15 2GW, UK.
E-mail address: Fabio.tirotta@uhb.nhs.uk (F. Tirotta).

Findings Of the 1510 patients undergoing surgery, 147 met the inclusion criteria and were included in the benchmarking analysis. This identified benchmark values including: median duration of surgery ≤278 min, intraoperative packed red cell transfusion rate ≤30%, R0/R1 resection rate ≥89%, median length of hospital stay ≤15 days, reoperation rate ≤13%, major postoperative complication rate ≤21%, and 90-day postoperative mortality/failure-to-rescue rates of 0%. The "low-volume centers" cohort failed to meet 10 of these benchmarks, including duration of surgery (median: 293 vs. ≤278 min), R0/R1 resection rate (82% vs. ≥ 89%), major postoperative complication rate (35% vs. ≤21%), and reoperation rate (35% vs. ≤13%), whilst the "ASA ≥ 3" cohort failed to meet seven benchmarks.

Interpretation These novel benchmark values can act as reference values to which sarcoma centers or individual surgeons can compare, which may help to identify performance gaps and improve the quality of care.

Funding "*5 x mille*" fund for healthcare research (Italian Ministry of Health).

Keywords: Benchmark; Retroperitoneal sarcoma; Surgical oncology; Liposarcoma; Quality of surgery

Research in context

Evidence before this study

We conducted a literature search in PubMed for studies published before 1st December 2023, that investigated the use of benchmark values for evaluating outcomes following comprehensive resection of primary retroperitoneal well-differentiated and dedifferentiated liposarcoma (WDLPS/DDLPS). We kept our search broad, using the terms "retroperitoneal," "sarcoma," "surgery," and "benchmark." At the time of our search, no studies had analyzed benchmark values following surgery for retroperitoneal WDLPS/DDLPS. Previous studies have reported widely varying outcomes, largely due to differences in patient selection, surgical approaches, and institutional expertise, making it difficult to establish standardized performance metrics.

Added value of this study

This study provides, for the first time, benchmark values for comprehensive resection of primary retroperitoneal WDLPS/DDLPS, using data from a large, multicenter cohort reporting on surgical and oncologic outcomes. Unlike previous research, which has primarily focused on individual institutional outcomes including multiple histologies and procedures, we establish objective reference points that can be used to evaluate surgical quality across different centers, surgeons, and over time. These benchmark values set a new standard for assessing surgical performance, helping institutions and surgeons measure their outcomes against the "best in class" rather than the average.

Implications of all the available evidence

Our findings support the need for robust benchmark values to improve surgical quality. These benchmarks can help institutions and individual surgeons monitor their performance. Moving forward, further validation in different settings will be important to ensure their broad applicability.

Introduction

Surgery for primary retroperitoneal sarcoma (RPS) is complex, and results in severe postoperative morbidity in approximately 20% of cases.[1] RPS comprises several different histologies, with liposarcoma (LPS), either well-differentiated (WDLPS) or dedifferentiated (DDLPS), being the most common. These specific RPS subtypes usually present as large tumors and carry a considerable risk of local recurrence (LR).[2] Comprehensive (formally termed "compartmental") resection has been found to improve oncological outcomes[3–5] and is currently recommended by international guidelines.[6] However, the surgical approach can vary, with the "standard" compartmental resection of the tumor with colon, kidney, and psoas fascia/muscle often being extended to include other relevant organs, such as the spleen and pancreas on the left side or duodenum/head of the pancreas or the inferior vena cava on the right side. Considering the variation in surgical approach, patient population, and sarcoma center case volumes, it is therefore difficult to make quality comparisons across centers or between different studies.

A novel method, namely benchmarking, may serve to overcome this issue, by producing reference values to enable such comparisons. The concept of benchmarking was born in the field of industry and has recently been applied to several major surgical procedures, including pancreatoduodenectomy,[7] liver transplantation,[8] major hepatectomy,[9] and many others.[10–20] In surgery, the aim of benchmarking is to select a low-risk population undergoing surgery at high-volume centers, and use this to define reference values for clinically relevant outcome indicators. These values then represent the best achievable outcomes, which can be

used as a reference to enable meaningful comparison across centers, surgeons, different geographical areas, as well as over time. Failure to meet the thresholds defined by these benchmark values may highlight areas of underperformance, which can then be addressed to improve the quality of care.

The aim of this study was to identify the best achievable results (i.e., benchmark values) for intra- and postoperative outcomes in patients undergoing comprehensive resection for primary WDLPS/DDLPS by applying the well-described benchmark methodology.

Methods

Study design

The study followed a standardized methodology based on a ten-step procedure, which was initially devised by a panel of experts,[21] before being validated, and refined using a Delphi approach,[22] and has been used in several previous studies in other complex surgical procedures.[7–20] In brief, this methodology first defines the intervention to be benchmarked and the patient cohort of interest, which should represent those expected to have the best possible outcomes. The outcomes being benchmarked must then be selected, with a focus on those that are commonly used, clinically relevant, and where data would be readily available. Eligible centers are then identified, which should be high volume to ensure that they are sufficiently experienced in the intervention of interest, and to maximize the number of cases available for estimating center-level outcomes. Prospective data collection then commences at each center, from which center-level outcomes are summarized based on the median for continuous variables or the proportion for binary outcomes. Benchmarks are then defined based on percentiles of the center-level data, with the 75th percentile used for outcomes where higher values indicate a worse outcome (e.g., complication rates) or the 25th percentile where lower values indicate a worse outcome (e.g., survival rates). These thresholds were proposed to reflect targets that were achievable for low-risk patients treated in the majority of experienced centers; hence, would be aspirational but not unrealistic when applied to the wider cohort of patients treated in standard centers.

The study protocol was presented to, and approved by, the Transatlantic Australasian Retroperitoneal Sarcoma Working Group (TARPSWG) at the group's biannual meeting on 17th March 2021. Data were extracted from the REtroperitoneal SArcoma Registry (RESAR) database (NCT03838718), the largest international, prospectively maintained database of RPS patients in the world, for all adult (age ≥ 18 years) patients undergoing surgery between 1st January 2017 and 31st December 2021. Fifteen specialist sarcoma centers from three continents (five in North America, nine in Europe, and one in Australia) contributed data to the study, all of which had a relevant academic profile.

Ethics

The original RESAR study received ethical approval from *Istituto Nazionale dei Tumori*, Milan, Italy (approval number INT 201/16), while the study protocol of the benchmark study was registered at University Hospitals Birmingham NHS Foundation Trust (CARMS ID: 17539), and ethical approval was granted by the Health Research Authority/NHS Research Ethics Committee (IRAS ID: 305025). Written informed consent was obtained in compliance with local ethical and regulatory requirements.

Study population

Benchmarking was only performed for the subset of high-volume centers, defined as performing an average of ≥13 primary RPS (any histology) resections per year (i.e., ≥65 in the five-year period)[23–25]; seven of the 15 centers were deemed high-volume based on this definition. The intervention being considered for benchmarking was low-risk comprehensive resection of primary localized WD/DDLPS. As such, patients with other tumor histologies, metastatic or recurrent disease, or with tumors at other sites (e.g., pelvis or mesentery) were excluded from the initial RESAR cohort. Comprehensive resection was defined using a similar approach to Bonvalot et al.[3] as the excision of the tumor along with at least the colon and kidney (±other organs). Resections incorporating the excision of organs associated with a higher risk of postoperative morbidity were then excluded, namely the spleen and pancreas, major vascular resection, pancreatoduodenectomy, or a major liver resection.[26] Patients with high-risk medical features were also excluded, comprising: age ≥75 years, body mass index (BMI) of <20 or ≥35 kg/m^2, diabetes mellitus, coronary artery disease, chronic obstructive pulmonary disease (COPD), albumin <3 g/dl or serum creatinine >1.8 mg/dl, Eastern Cooperative Oncology Group performance status (ECOG) grade ≥2, or an American Society of Anesthesiologists (ASA) score ≥3. Patients who did not meet any of the above exclusion criteria, but for whom data were missing for at least one of the criteria were additionally excluded, as they could not be confirmed to be low risk.

Benchmarking outcomes

A total of 22 clinically relevant intra- and postoperative outcomes were considered in the benchmarking analysis. These were selected based on consensus among sarcoma experts within the TARPSWG network, based on the combination of being commonly used in clinical trials, relevant to patient care and long-term prognosis, and feasible to collect across international centers. Intraoperative outcomes included the duration of surgery, amount of blood loss, and the need for packed red

cell (PRC) transfusion. Postoperative outcomes comprised the surgical margins, classified as either macroscopically complete (R0/R1) or incomplete (R2), the total length of hospital stay, and the need for PRC transfusion. The major complication rate was also assessed, defined as the development of at least one postoperative complication of grade ≥3 according to the Clavien-Dindo classification (CDC).[27,28] Rates of individual postoperative complications were also considered, namely bowel anastomotic leak, bleeding, and sepsis, as well as the need for reoperation for postoperative complications. Postoperative mortality was quantified as the 90-day mortality rate with the failure-to-rescue rate also assessed, defined the rate of 90-day mortality in patients developing a major complication.[29] Since the cohort had ongoing follow-up, with the final update of the data being on the 28th February 2024, long-term survival rates at six months and one and three years were additionally assessed, as were the incidence rates of LR and distant metastases (DM) at these times.

Comparative cohorts

To assess the benchmark values and give an example of how they could be utilized in practice, the values were applied to two additional cohorts of patients that would be expected to have inferior outcomes to the benchmarking cohort, which were defined by subtly changing the exclusion criteria. The first was a "ASA ≥ 3" cohort, defined as patients meeting all of the inclusion criteria, except for having an ASA score of ≥3, and therefore deemed high-risk for postoperative complications. The second was a "low-volume centers" cohort, defined as patients meeting all the inclusion criteria, except for being treated at a low-volume center (i.e., <13 RPS resections per year on average). Analysis of a third "ASA ≥ 3 in low-volume centers" cohort was originally planned in the study protocol; however, this was not performed due to the sample size being insufficient, with only N = 4 patients meeting these criteria.

Statistical methods

For the benchmarking analysis, outcomes were first summarized for each of the high-volume centers, using medians for continuous variables and rates for nominal variables. Survival- and recurrence-related outcomes were quantified using time-to-event analyses, which were performed separately for each center. Specifically, overall survival was assessed using Kaplan–Meier curves, with death as the event, and patients censored at the last follow-up. LR and DM were assessed using a competing risks approach to produce cumulative incidence curves, with three potential events, namely death, LR and DM; patients diagnosed with synchronous LR and DM were classified as DM for analysis. The resulting Kaplan–Meier curves (for survival) and cumulative incidence function curves (for LR and DM) were then evaluated, to extract the estimated outcome rates at six months, and at one and three years for each center.

Benchmark values were then identified based on analysis of the center-level data, and defined using the 75th percentile for outcomes where higher values indicated a worse outcome (e.g., complication rates) or the 25th percentile where lower values indicated a worse outcome (e.g., survival rates); percentiles were calculated using the Tukey's Hinges method. Analyses were primarily performed using IBM SPSS v29 (IBM Corp. Armonk, NY), with competing risks analyses performed using the "cmprsk" package in R v4.3.2. For cohort characteristics, continuous variables were summarized as mean ± standard deviation where approximately normally distributed, with median (interquartile range) used otherwise.

Role of the funding source

The "*5 x mille*" research fund (Italian Ministry of Health) supported the development of the RESAR data collection platform, and the hiring of a dedicated clinical research coordinator for RESAR. The funder had no role in the study design, data collection, analysis, interpretation, or writing of the report.

Results

Study cohort for benchmarking

The 15 RESAR centers performed a total of N = 1510 primary RPS resections during the study period (range: 12–334 cases per center). Based on the histological exclusion criteria, N = 597 cases were subsequently excluded, due to having histology other than WD/DDLPS, non-retroperitoneal tumors, or metastatic disease. Of the remaining N = 913, the resected organs were not reported in N = 7, with comprehensive resections performed in N = 624 (68.9%) of the remainder. Of patients treated in high-volume centers, 70.9% (531/749) underwent comprehensive resections, which was significantly greater than the 59.2% (93/157) of those treated in low-volume centers (p = 0.006). Of the N = 531 undergoing comprehensive resections at high-volume centers, a further N = 162 were excluded due to undergoing resections of organs associated with a higher risk of postoperative morbidity, with N = 222 excluded due to meeting one of the high-risk medical exclusion criteria, see Fig. 1. After exclusions, the benchmarking cohort comprised N = 147 patients from seven high-volume centers, ranging from 7 to 51 cases per center, and making up 10–29% of centers' primary localized retroperitoneal WD/DDLPS case volume (Fig. 2). The benchmarking cohort had a mean age of 57.0 ± 11.0 years, with the majority undergoing resection of DDLPS (69.4%), and a mean tumor size of 286 ± 104 mm; further details of the cohort are reported in Table 1.

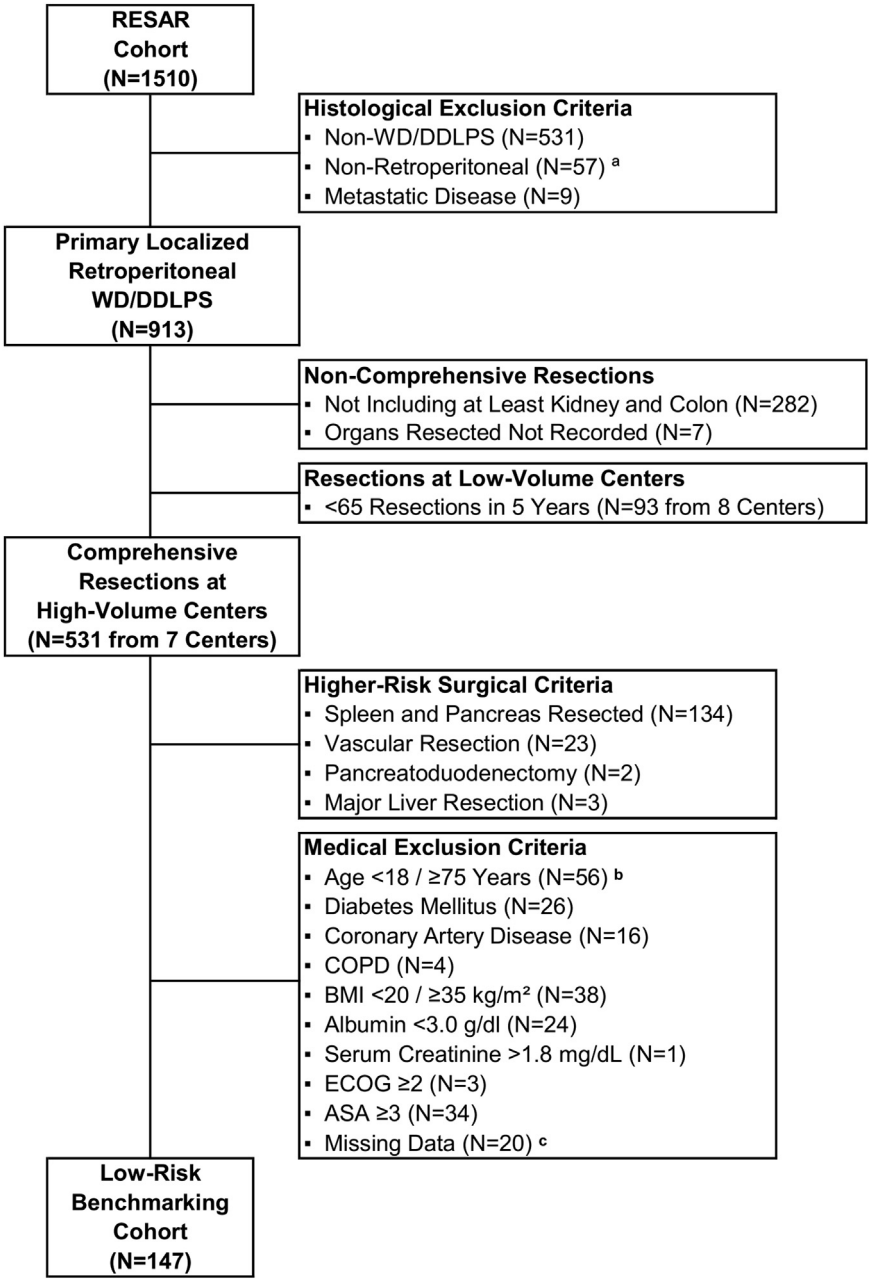

***Fig. 1:* Study flowchart.** Exclusions are applied cumulatively in the order stated. [a]Includes N = 8 with missing data, who were excluded since it was unclear whether they had retroperitoneal tumors [b]Patients aged <18 years were already excluded from the RESAR database at source. [c]Patients who were not reported to meet any of the stated medical exclusion criteria, but who had missing data for at least one of the criteria. Abbreviations: ASA: American Society of Anesthesiologists score, BMI: body mass index, COPD: chronic obstructive pulmonary disease, DD(WD)LPS: dedifferentiated (well-differentiated) liposarcoma, ECOG: Eastern Cooperative Oncology Group performance status.

Benchmark values

The benchmark values derived from the low-risk cohort are reported in Table 2, with selected outcomes visualized in Fig. 3. The benchmark values for the intraoperative outcomes of the median duration of surgery, median intraoperative blood loss, and intraoperative PRC transfusion rates were ≤278 min, ≤588 ml, and ≤30% of cases, respectively. Of the postoperative outcomes, the benchmark values were a R0/R1 resection rate of ≥89%, median length of hospital stay ≤15 days, PRC transfusion rate ≤37% of cases, and a reoperation rate ≤13%. Among the postoperative complications, the

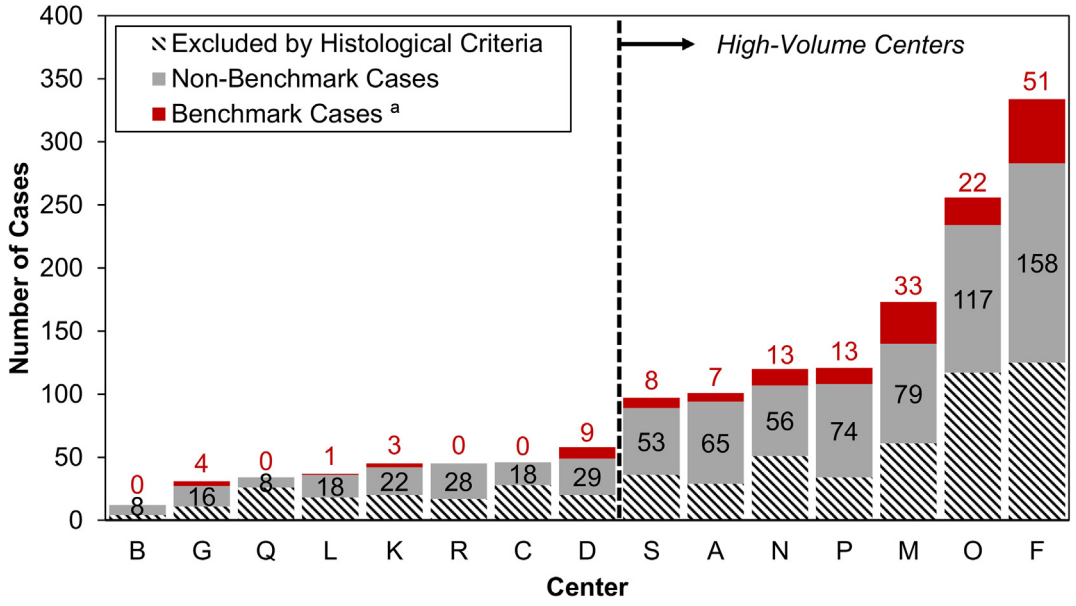

Fig. 2: **Total volume and number of benchmark cases for individual centers.** [a]Represents the number of benchmark cases for the high-volume centers; for low-volume centers, this represents the number of cases meeting all other benchmarking criteria (except for center volume, i.e., the "Low-Volume Centers" cohort).

benchmark value for the major complication rate was ≤21%, with analysis of individual complications returning benchmark values of anastomotic leak ≤3%, bleeding ≤6%, and sepsis ≤4%. Benchmark values for the 90-day postoperative mortality rate and failure-to-rescue rate were both 0%, with six-month, one-year and three-year overall survival of ≥97%, ≥96%, and ≥88%. For the recurrence-related outcomes, benchmark values for LR were ≤1%, ≤14% and ≤22% at six months, one year and three years, respectively, with corresponding rates for DM of ≤8%, ≤11% and ≤15%.

Outcome comparisons

The newly obtained benchmark values were applied to two different patient cohorts to illustrate their potential real-world utility (Table 3). The first patient cohort included N = 32 patients meeting all of the inclusion criteria of the benchmarking cohort, except for the fact that they were considered medically high-risk for surgery (i.e., ASA ≥ 3). This "ASA ≥ 3" cohort exceeded the benchmark values for the median intraoperative blood loss (600 ml vs. ≤588 ml), and the rates of intraoperative PRC transfusion (57% vs. ≤30%) and major postoperative complications (28% vs. ≤21%). Of the individual postoperative complications considered, the cohort exceeded the benchmark values for anastomotic leak (3.1% vs. ≤3%), and sepsis (6% vs. ≤4%). Rates of LR at six months (4% vs. ≤1%) and three years (29% vs. ≤22%) were also inferior to the benchmark values.

The second patient cohort consisted of N = 17 patients meeting the same criteria as the benchmarking cohort, apart from having received surgery in low-volume centers. This "low-volume centers" cohort failed to achieve the benchmark values for several outcomes, including the median duration of surgery (293 vs. ≤278 min) and the rates of R0/R1 resection (82% vs. ≥ 89%), major postoperative complications (35% vs. ≤21%), reoperation (35% vs. ≤13%), anastomotic leak (12% vs. ≤3%), bleeding (24% vs. ≤6%), and sepsis (24% vs. ≤4%), as well as rates of LR at six months (12% vs. ≤1%), overall survival at one year (94% vs. ≥96%) and DM at three years (19% vs. ≤15%).

Discussion

This study utilized a validated[21,22] and commonly used[7–20] methodology to identify benchmark values for 22 intra- and postoperative outcomes in patients undergoing a standard comprehensive resection of primary retroperitoneal WDLPS/DDLPS. Of note, the benchmark value for major postoperative complications was ≤21%, confirming that even in low-risk (benchmarking) cases, RPS surgery carries a significant risk of postoperative morbidity. The benchmark values were then tested in two different cohorts, each of which differed from the benchmarking cohort only for a specific factor (i.e., ASA ≥ 3 and resection at low-volume centers). Such comparisons identified performance gaps where the benchmark values were not reached, highlighting how the benchmark values proposed by this study could potentially be used as reference values to assess the performance of centers and individual surgeons, and act

	Primary localized retroperitoneal WD/DDLPS (N = 913)		Low-risk benchmarking cohort (N = 147)	
	N	Statistic	N	Statistic
Age (years)	913	62.7 ± 12.1	147	57.0 ± 11.0
Gender (% male)	913	507 (55.5%)	147	84 (57.1%)
Ethnicity (% white)	716	662 (92.5%)	106	102 (96.2%)
Body mass index (kg/m^2)	909	25.7 (22.9–29.0)	147	25.6 (23.2–28.4)
ECOG performance status	898		147	
0		635 (70.7%)		129 (87.8%)
1		212 (23.6%)		18 (12.2%)
2		41 (4.6%)		–
3		9 (1.0%)		–
4		1 (0.1%)		–
ASA score	874		147	
1		99 (11.3%)		28 (19.0%)
2		514 (58.8%)		119 (81.0%)
3		238 (27.2%)		–
4		22 (2.5%)		–
5		1 (0.1%)		–
COPD	903	43 (4.8%)	147	–
Coronary artery disease	873	90 (10.3%)	147	–
Diabetes mellitus	903	88 (9.7%)	147	–
Albumin (g/dl)	823	3.8 ± 0.7	147	4.1 ± 0.5
Creatinine (mg/dl)	888	0.80 (0.67–0.95)	147	0.78 (0.66–0.90)
Tumor histology[a]	913		147	
WDLPS		251 (27.5%)		45 (30.6%)
DDLPS		662 (72.5%)		102 (69.4%)
Max. tumor dimension (mm)[a]	911	271 ± 115	147	286 ± 104
Treated at low-volume center	913	164 (18.0%)	147	–
Non-comprehensive resection	906[b]	282 (31.1%)	147	–
Spleen and pancreas resected	906[b]	171 (18.9%)	147	–
Vascular resection	906[b]	50 (5.5%)	147	–
Pancreatoduodenectomy	906[b]	6 (0.7%)	147	–
Major liver resection	900[b,c]	5 (0.6%)	147	–

Data are reported as: "mean ± standard deviation", "median (interquartile range), or as "N (%)", as appropriate. ASA: American Society of Anesthesiologists, COPD: chronic obstructive pulmonary disease, DD(WD)LPS: dedifferentiated (well-differentiated) liposarcoma, ECOG: Eastern Cooperative Oncology Group, Max.: maximum. [a]On pathology. [b]Excludes patients where the organs resected were not reported. [c]Additionally excludes patients undergoing liver resection where it was unclear whether this was "major".

Table 1: **Cohort characteristics.**

as a target for improvement where these are not achieved.

Benchmarking is a quality improvement approach, which has gained its popularity in surgery, mainly due to the necessity of establishing reference values to enable comparisons against the "best in class".[21,22,30] Prior to the introduction of this methodology, the average values for the whole cohort of either patient- or center-level data were commonly used as reference thresholds. However, the use of quartiles of low-risk cohorts to define thresholds in the benchmarking methodology gives a more demanding, yet still achievable target, which can potentially drive improvements in practice. For benchmarking to be reliable and valid, it is important to use a well-defined, homogenous cohort. This can be challenging in RPS, which comprises several heterogenous tumor histologies which require a variety of surgical approaches, ranging from simple excision of the tumor for histologies with a very low risk of LR (e.g., solitary fibrous tumors), to a complex comprehensive resection for histologies with a high risk of LR, such as liposarcoma.[31] With this in mind, we used the largest prospective database for RPS sarcoma (RESAR) to select a homogenous, low-risk cohort, with a single histology (WD/DDLPS), and undergoing surgery at high-volume, academic centers using the current standard of care (comprehensive resection) for the benchmarking analysis.

To demonstrate how benchmarking could be utilized real-world practice, the benchmark values were applied to two additional cohorts. The first used exclusion criteria that were consistent with the benchmarking

Outcome	Median (IQR) of center-level data	Benchmark value
Intraoperative outcomes		
Duration of surgery (mins)	257 (234–278)	≤278
Blood loss (ml)	500 (300–588)	≤588
PRC transfusion required	25% (24–30%)	≤30%
Postoperative outcomes		
R0/R1 surgical margins[a]	95% (89–100%)	≥89%[a]
Length of hospital stay (days)	9 (9–15)	≤15
PRC transfusion required	25% (19–37%)	≤37%
Major complication (CDC grade ≥ 3)	18% (6–21%)	≤21%
Reoperation due to complication	12% (0–13%)	≤13%
Bowel anastomotic leak	0% (0–3%)	≤3%
Bleeding	0% (0–6%)	≤6%
Sepsis	0% (0–4%)	≤4%
90-day mortality	0% (0-0%)	0%
Failure-to-rescue	0% (0-0%)	0%
Survival outcomes		
Overall survival[a]		
Six month	100% (97–100%)	≥97%[a]
One year	98% (96–100%)	≥96%[a]
Three year	93% (88–97%)	≥88%[a]
Local recurrence		
Six month	0% (0–1%)	≤1%
One year	8% (0–14%)	≤14%
Three year	18% (11–22%)	≤22%
Distant metastases		
Six month	0% (0–8%)	≤8%
One year	10% (5–11%)	≤11%
Three year	13% (7–15%)	≤15%

Analysis is based on the center-level data for the N = 7 high-volume centers. Benchmark values are defined based on the 75th percentile, unless stated otherwise. CDC: Clavien-Dindo Classification, IQR: interquartile range, PRC: packed red cells, R0/R1: macroscopically complete. [a]Defined based on the 25th percentile.

Table 2: **Defining benchmark values for intra- and postoperative outcomes.**

cohort, but had ASA ≥ 3. As would be expected, this higher-risk cohort had inferior outcomes, and failed to achieve benchmark values for outcomes including intraoperative blood loss and major complication rates. The second cohort comprised low-risk patients undergoing surgery at low-volume specialist sarcoma centers. This cohort failed to achieve the benchmark values for 10 of the 22 outcomes, including those relating to major postoperative complications, resection margins and rates of LR and DM. Whilst this analysis was limited by the small sample size (N = 17), the findings were consistent with a recent English study, which found low-volume specialist sarcoma centers to have inferior outcomes, relative to those of high-volume centers.[24] Similarly, another recent study which updated the prognostic tool "Sarculator" found that case volume was associated with survival outcomes.[25]

The benchmark values proposed by this study may have important implications for clinical practice, particularly in the context of the rarity of RPS. Historically, specialist sarcoma centers have been defined by the presence of a sarcoma-specific multidisciplinary team (MDT). However, it has recently emerged that this alone is not sufficient to offer the best possible care. Specifically, low-volume centers may lack the experience to achieve optimal outcomes for patients[24] and, even for high-volume centers, there can be considerable variability in the treatment approaches proposed by MDTs.[32] This was observed in the present study, with patients treated at low-volume centers being significantly less likely to undergo comprehensive resections, compared to those treated a high-volume centers. Setting targets and demonstrating areas of improvement by

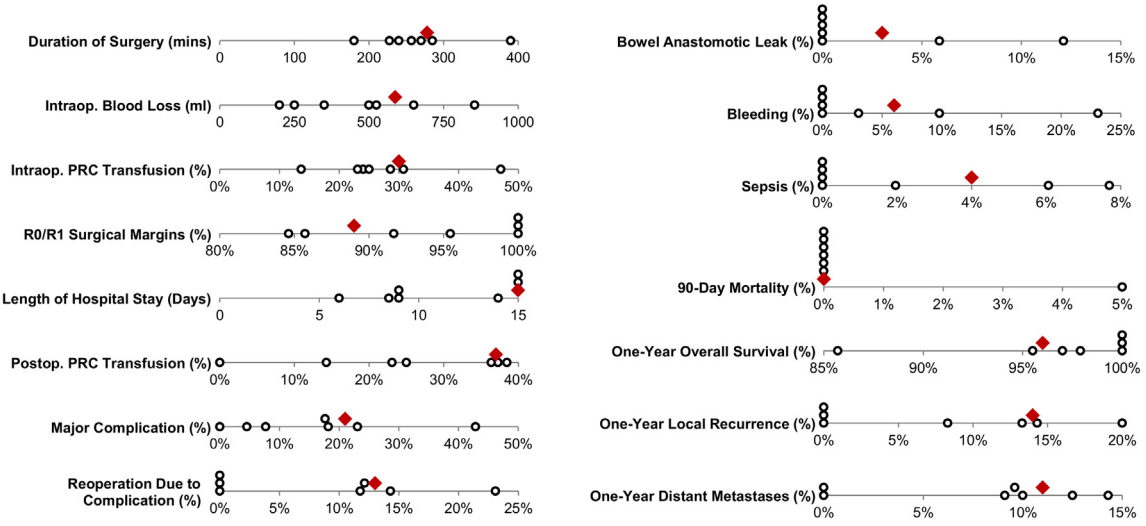

Fig. 3: **Center-level results and benchmark values for intra- and post-operative outcomes.** Black points represent data for the N = 7 individual centers in the benchmarking cohort, namely medians for continuous variables and rates for binary outcomes. Survival and recurrence rates were estimated from Kaplan–Meier and cumulative incidence function curves, respectively; only the one-year outcomes are reported, for brevity. Overlapping points are stacked for clarity. Red diamonds represent the benchmark value for each outcome. Abbreviations: Intraop: intraoperative, Postop: postoperative, PRC: packed red cells, R0/R1: macroscopically complete.

Outcome	Benchmark value	ASA ≥ 3 (N = 32)		Low-volume centers (N = 17)	
		N	Statistic	N	Statistic
Intraoperative outcomes					
Duration of surgery (mins)	≤278	32	248 (210–315)	17	**293 (247–360)**
Blood loss (ml)	≤588	32	**600 (300–838)**	17	500 (350–800)
PRC transfusion required	≤30%	30	**17 (57%)**	16	4 (25%)
Postoperative outcomes					
R0/R1 surgical margins	≥89%	32	32 (100%)	17	**14 (82%)**
Length of hospital stay (days)	≤15	32	13 (9–16)	17	11 (8–19)
PRC transfusion required	≤37%	32	10 (31%)	17	3 (18%)
Major complication (CDC Grade ≥ 3)	≤21%	32	**9 (28%)**	17	**6 (35%)**
Reoperation due to complication	≤13%	32	3 (9%)	17	**6 (35%)**
Bowel anastomotic leak	≤3%	32	**1 (3%)**	17	**2 (12%)**
Bleeding	≤6%	32	1 (3%)	17	**4 (24%)**
Sepsis	≤4%	32	**2 (6%)**	17	**4 (24%)**
90-day mortality	0%	28	0 (0%)	17	0 (0%)
Failure-to-rescue[a]	0%	9	0 (0%)	6	0 (0%)
Survival outcomes					
Overall survival		32		17	
Six month	≥97%		100%		100%
One year	≥96%		96%		**94%**
Three year	≥88%		91%		94%
Local recurrence		32		17	
Six month	≤1%		**4%**		**12%**
One year	≤14%		4%		12%
Three year	≤22%		**29%**		12%
Distant metastases		32		17	
Six month	≤8%		4%		0%
One year	≤11%		7%		6%
Three year	≤15%		11%		**19%**

Analyses are performed on patient-level data, which was summarized as "N (%)" for binary outcomes; "median (interquartile range)" for continuous outcomes; Kaplan–Meier estimated rates for overall survival; and cumulative incidence rates for recurrence-related outcomes. Bold values indicate outcomes where the benchmark value was not achieved for the cohort, i.e., where the median or rate was outside the threshold defined by the benchmark value. Abbreviations: ASA: American Society of Anesthesiologists score, CDC: Clavien-Dindo Classification, PRC: packed red cells, R0/R1: macroscopically complete. [a]In patients with major complications.

Table 3: **Comparison of outcomes in comparative cohorts to the benchmark values.**

establishing benchmark values is a useful and practical way to improve performance. However, it is important to highlight that the newly established benchmark values are not meant to judge or scrutinize individuals' or centers' performances, but to identify areas for potential improvement in the strive for excellence. Specifically, failure to achieve benchmarks should not necessarily be interpreted as being indicative of unacceptable performance, particularly where the difference between the achieved and benchmark values is small. Instead, the benchmarks should be used to identify areas to focus on in the ongoing process of continuous improvement in the treatment of RPS.

This study had several strengths, including the use of a large, prospectively maintained, international, multicenter database (RESAR). This allowed for strict exclusion criteria to be used, to ensure that the benchmarking cohort was homogenous and low risk. However, there are also several limitations that need to be considered when interpreting the findings. Primarily, whilst the RESAR database was extensive, the rarity and heterogeneity of RPS meant that the benchmarking analysis was based on a relatively small sample size and small number of centers. The resulting small within-center sample size meant center-level estimates of outcome rates would have been of low precision for the binary outcomes with low prevalence, such as R2 surgical margins. As such, there is a risk that some center-level estimates may have been overly influenced by outliers. A similar issue resulted from the small number of centers used when deriving the benchmark values. Therefore, if the small within-center sample size resulted in more than one center having an artificially high (or low) outcome rate, then these outliers would have influence on the final benchmark value, potentially leading to excessively lenient (or strict) benchmark values. Secondly, due to the sample size, it was only feasible to produce benchmark values for the whole

cohort, and not to consider subgroups of patients for whom outcomes may have differed; for example, complication rates in left-vs. right-sided disease, or oncological outcomes in different tumor grades. Finally, the overall postoperative complication burden was summarized based only on the CDC grade of the most severe complication, as data were not available for quantifications of the cumulative complication burden, such as the comprehensive complication index.[33,34]

In conclusion, this international multicenter study establishes novel benchmark values for outcomes of patients undergoing comprehensive resection for primary retroperitoneal WDLPS/DDLPS. These values can be used as a reference for quality improvement, enabling comparison between centers, surgeons, or geographical areas. Offering these new benchmark values should help centers to level up by targeting the best possible outcomes, rather than meeting in the middle.

Contributors

Fabio Tirotta (conceptualization, methodology, data curation, formal analysis, project administration, visualization, writing—original draft, writing—review and editing); Marco Fiore (data curation, writing—review and editing); Sylvie Bonvalot (data curation, writing—review and editing); Dirk Strauss (data curation, writing—review and editing); Piotr Rutkowski (data curation, writing—review and editing); David E. Gyorki (data curation, writing—review and editing); Winan J. van Houdt (data curation, writing—review and editing); Dario Callegaro (data curation, writing—review and editing); Markus Albertsmeier (data curation, writing—review and editing); Dimitri Tzanis (data curation, writing—review and editing); Ferdinando Cananzi (data curation, writing—review and editing); Jason K. Sicklick (data curation, writing—review and editing); John Mullinax (data curation, writing—review and editing); Michelle Wilkinson (data curation, writing—review and editing); Valerie P. Grignol (data curation, writing—review and editing); Kenneth Cardona (data curation, writing—review and editing); Toufik Bouhadiba (data curation, writing—review and editing); Marko Novak (data curation, writing—review and editing); Sergio Valeri (data curation, writing—review and editing); Mark Fairweather (data curation, writing—review and editing); Samuel J. Ford (data curation, writing—review and editing); Jacek Skoczylas (data curation, writing—review and editing); Hayden Snow (data curation, writing—review and editing); Andrew J. Hayes (data curation, writing—review and editing); James Hodson (formal analysis, writing—review and editing); Chandrajit P. Raut (data curation, writing—review and editing); Alessandro Gronchi (methodology, data curation, writing—review and editing).

Fabio Tirotta, Marco Fiore, James Hodson, and Alessandro Gronchi have accessed and verified the data. All authors read and approved the final version of the manuscript.

Data sharing statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of interests

Piotr Rutkowski received honoraria for lectures and Advisory Boards from Bristol-Myers Squibb, MSD, Novartis, Pierre Fabre, Philogen, Genesis, Medison Pharma outside of the scope of the manuscript. His institution received research funding from Novarits, Pfizer, Roche, Bristol-Myers Squibb.

Jason Sicklick serves as a consultant for CureMatch, Deciphera and Kura; received speakers' fees from Daiichi Sankyo, Deciphera, Foundation Medicine, La-Hoffman Roche, Merck, QED, and SpringWorks; and owns stock in CureMatch and Personalis.

Ferdinando Carlo Maria Cananzi received a lecture fee from Istituto Gentili.

All the other authors declare no conflict of interest.

Acknowledgements

We thank the Australia and New Zealand Sarcoma Association (ANZSA) as a sponsor of the RESAR project in Australia. This work was funded by 5xmille funds for healthcare research (Ministry of Health).

References

1 Tirotta F, Parente A, Hodson J, et al. Cumulative burden of postoperative complications in patients undergoing surgery for primary retroperitoneal sarcoma. *Ann Surg Oncol*. 2021;28(12):7939–7949.
2 Gronchi A, Strauss DC, Miceli R, et al. Variability in patterns of recurrence after resection of primary retroperitoneal sarcoma (RPS): a report on 1007 patients from the multi-institutional collaborative RPS working group. *Ann Surg*. 2016;263(5):1002–1009.
3 Bonvalot S, Rivoire M, Castaing M, et al. Primary retroperitoneal sarcomas: a multivariate analysis of surgical factors associated with local control. *J Clin Oncol*. 2009;27(1):31–37.
4 Gronchi A, Lo Vullo S, Fiore M, et al. Aggressive surgical policies in a retrospectively reviewed single-institution case series of retroperitoneal soft tissue sarcoma patients. *J Clin Oncol*. 2009;27(1):24–30.
5 Gronchi A, Miceli R, Colombo C, et al. Frontline extended surgery is associated with improved survival in retroperitoneal low- to intermediate-grade soft tissue sarcomas. *Ann Oncol*. 2012;23(4):1067–1073.
6 Swallow CJ, Strauss DC, Bonvalot S, et al. Management of primary retroperitoneal sarcoma (RPS) in the adult: an updated consensus approach from the transatlantic australasian RPS working group. *Ann Surg Oncol*. 2021;28(12):7873–7888.
7 Sánchez-Velázquez P, Muller X, Malleo G, et al. Benchmarks in pancreatic surgery: a novel tool for unbiased outcome comparisons. *Ann Surg*. 2019;270(2):211–218.
8 Muller X, Marcon F, Sapisochin G, et al. Defining benchmarks in liver transplantation: a multicenter outcome analysis determining best achievable results. *Ann Surg*. 2018;267(3):419–425.
9 Rössler F, Sapisochin G, Song G, et al. Defining benchmarks for major liver surgery: a multicenter analysis of 5202 living liver donors. *Ann Surg*. 2016;264(3):492–500.
10 Sousa Da Silva RX, Breuer E, Shankar S, et al. Novel benchmark values for open major anatomic liver resection in non-cirrhotic patients: a multicentric study of 44 international expert centers. *Ann Surg*. 2023;278(5):748–755.
11 Breuer E, Mueller M, Doyle MB, et al. Liver transplantation as a new standard of care in patients with perihilar cholangiocarcinoma? Results from an international benchmark study. *Ann Surg*. 2022;276(5):846–853.
12 Li Z, Rammohan A, Gunasekaran V, et al. Novel benchmark for adult-to-adult living-donor liver transplantation: integrating eastern and western experiences. *Ann Surg*. 2023;278(5):798–806.
13 Müller PC, Breuer E, Nickel F, et al. Robotic distal pancreatectomy: a novel standard of care? Benchmark values for surgical outcomes from 16 international expert centers. *Ann Surg*. 2023;278(2):253–259.
14 Gero D, Vannijvel M, Okkema S, et al. Defining global benchmarks in elective secondary bariatric surgery comprising conversional, revisional, and reversal procedures. *Ann Surg*. 2021;274(5):821–828.
15 Raptis DA, Sánchez-Velázquez P, Machairas N, et al. Defining benchmark outcomes for pancreatoduodenectomy with portomesenteric venous resection. *Ann Surg*. 2020;272(5):731–737.
16 Goh BKP, Han HS, Chen KH, et al. Defining global benchmarks for laparoscopic liver resections: an international multicenter study. *Ann Surg*. 2023;277(4):e839–e848.
17 Durin T, Marchese U, Sauvanet A, et al. Defining benchmark outcomes for distal pancreatectomy: results of a French multicentric study. *Ann Surg*. 2023;278(1):103–109.
18 Staiger RD, Rössler F, Kim MJ, et al. Benchmarks in colorectal surgery: multinational study to define quality thresholds in high and low anterior resection. *Br J Surg*. 2022;109(12):1274–1281.
19 Russolillo N, Aldrighetti L, Cillo U, et al. Risk-adjusted benchmarks in laparoscopic liver surgery in a national cohort. *Br J Surg*. 2020;107(7):845–853.

20 Schneider MA, Kim J, Berlth F, et al. Defining benchmarks for total and distal gastrectomy: global multicentre analysis. *Br J Surg*. 2024;111(2):znad379.
21 Staiger RD, Schwandt H, Puhan MA, et al. Improving surgical outcomes through benchmarking. *Br J Surg*. 2019;106(1):59–64.
22 Gero D, Muller X, Staiger RD, et al. How to establish benchmarks for surgical outcomes?: a checklist based on an international expert Delphi consensus. *Ann Surg*. 2022;275(1):115–120.
23 Villano AM, Zeymo A, Chan KS, et al. Identifying the minimum volume threshold for retroperitoneal soft tissue sarcoma resection: merging national data with consensus expert opinion. *J Am Coll Surg*. 2020;230(1):151–160.e2.
24 Tirotta F, Bacon A, Collins S, et al. Primary retroperitoneal sarcoma: a comparison of survival outcomes in specialist and non-specialist sarcoma centres. *Eur J Cancer*. 2023;188:20–28.
25 Callegaro D, Barretta F, Raut CP, et al. New sarculator prognostic nomograms for patients with primary retroperitoneal sarcoma: case volume does matter. *Ann Surg*. 2024;279(5):857–865.
26 MacNeill AJ, Gronchi A, Miceli R, et al. Postoperative morbidity after radical resection of primary retroperitoneal sarcoma. *Ann Surg*. 2018;267:959–964.
27 Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg*. 2004;240:205–213.
28 Clavien PA, Barkun J, de Oliverira M, et al. The Clavien–Dindo classification of surgical complications: five-year experience. *Ann Surg*. 2009;250:187–196.
29 Ghaferi AA, Birkmeyer JD, Dimick JB. Complications, failure to rescue, and mortality with major inpatient surgery in Medicare patients. *Ann Surg*. 2009;250:1029–1034.
30 Domenghino A, Walbert C, Birrer DL, et al. Consensus recommendations on how to assess the quality of surgical interventions. *Nat Med*. 2023;29(4):811–822.
31 Bonvalot S, Roland C, Raut C, et al. Histology-tailored multidisciplinary management of primary retroperitoneal sarcomas. *Eur J Surg Oncol*. 2022;49(6):1061–1067.
32 Tirotta F, Hodson J, Alcorn D, et al. Assessment of inter-centre agreement across multidisciplinary team meetings for patients with retroperitoneal sarcoma. *Br J Surg*. 2023;110(9):1189–1196.
33 Slankamenac K, Graf R, Barkun J, et al. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann Surg*. 2013;258:1–7.
34 Slankamenac K, Nederlof N, Pessaux P, et al. The comprehensive complication index: a novel and more sensitive endpoint for assessing outcome and reducing sample size in randomized controlled trials. *Ann Surg*. 2014;260:757–762.