Table 6 The relationships between the NmGPS (CRP cutoff; 0.5 mg/dL) and the clinicopathological features

Variable	NmGPS (0.5) 0 ($n = 121$)	NmGPS (0.5) 1 ($n = 26$)	NmGPS (0.5) 2 ($n = 21$)	p
Age, years				0.7471
≤65	56 (46.3 %)	10 (38.5 %)	10 (47.6 %)	
>65	65 (53.7 %)	16 (61.5 %)	11 (52.4 %)	
Sex				0.1744
Male	96 (79.3 %)	24 (92.3 %)	15 (71.4 %)	
Female	25 (20.7 %)	2 (7.7 %)	6 (28.6 %)	
Depth of tumor invasion				<0.0001
T0/1/2	78 (64.5 %)	5 (19.2 %)	3 (14.3 %)	
T3/4	43 (35.5 %)	21 (80.8 %)	18 (85.7 %)	
Lymph node metastasis				0.0008
N0	58 (47.9 %)	9 (34.6 %)	1 (4.8 %)	
N1/2/3	63 (52.1 %)	17 (65.4 %)	20 (95.2 %)	
Distant metastasis				0.0008
M0	121 (100 %)	26 (100 %)	19 (90.5 %)	
M1	0	0	2 (9.5 %)	
Histological type				0.1994
Differentiated	87 (71.9 %)	21 (80.8 %)	12 (57.1 %)	
Undifferentiated	34 (28.1 %)	5 (19.2 %)	9 (42.9 %)	
Lymphatic invasion				0.1399
0	53 (43.8 %)	8 (30.8 %)	5 (23.8 %)	
1, 2, 3	68 (56.2 %)	18 (69.2 %)	16 (76.2 %)	
Venous invasion				0.0015
0	72 (59.5 %)	6 (23.1 %)	8 (38.1 %)	
1, 2, 3	49 (40.5 %)	20 (76.9 %)	13 (61.9 %)	
Maximum tumor size, mm				<0.0001
≤40	79 (65.3 %)	6 (23.1 %)	4 (19.0 %)	
>40	42 (34.7 %)	20 (76.9 %)	17 (81.0 %)	
Lymph node dissection				0.4501
Two fields	44 (36.4 %)	12 (46.2 %)	6 (28.6 %)	
Three fields	77 (63.6 %)	14 (53.8 %)	15 (71.4 %)	
Neoadjuvant therapy				0.0007
Yes	11 (9.1 %)	2 (7.7 %)	8 (38.1 %)	
No	110 (90.9 %)	24 (92.3 %)	13 (61.9 %)	
Adjuvant therapy				0.0166
Yes	52 (43.0 %)	11 (42.3 %)	16 (76.2 %)	
No	69 (57.0 %)	15 (57.7 %)	5 (23.8 %)	
TNM stage				<0.0001
0, I, II	68 (56.2 %)	8 (30.8 %)	0	
III, IV	53 (43.8 %)	18 (69.2 %)	21 (100 %)	
Residual tumor				<0.0001
R0	121 (100 %)	23 (88.5 %)	15 (71.4 %)	
R1/2	0 (0 %)	3 (11.5 %)	6 (28.6 %)	

NmGPS new modified Glasgow Prognostic Score, *T* tumor, *N* node, *M* metastasis, *R* residual tumor

groups separated by one point (NmGPS [CRP cutoff; 0.5] 0 vs 1: $p < 0.0001$, NmGPS [CRP cutoff; 0.5 mg/dL] 1 vs 2: $p = 0.0099$). In the multivariate analysis of cancer-specific survival, a NmGPS (CRP cutoff; 0.5 mg/dL) of 2 was found to be a more independent prognostic indicator of a worse prognosis than a mGPS of 2. Moreover, we evaluated the quality of the prognostic scoring system using the AIC. The results of that evaluation suggested that the system using the NmGPS (CRP cutoff; 0.5 mg/dL) had higher quality than the mGPS and other NmGPS. These findings demonstrate that the NmGPS (CRP cutoff; 0.5 mg/dL) is more sensitive than the mGPS in patients with ESCC.

It is known that Asian countries, especially Japan, Korea and China, have the highest rates of esophageal cancer in the world [3], and people in Japan, Korea and China may be closely related in terms of genetics [19]. Shah et al. [20] reported that the mean CRP value of East Asians was less

Table 7 The univariate prognostic factors for esophageal cancer (including the NmGPS [CRP cutoff; 0.5 mg/dL])

Variable	*p*	HR	95 % CI
Sex (male)	0.4597	1.312	0.639–2.694
Age (>65 years)	0.1343	1.550	0.873–2.751
NmGPS (CRP cutoff; 0.5 mg/dL) (2)	<0.0001	7.807	4.215–14.461
Depth of tumor invasion (T3, 4)	<0.0001	3.801	2.057–7.025
Lymph node metastasis			
N1	0.0215	3.064	1.179–7.962
N2	0.0165	3.476	1.256–9.625
N3	<0.0001	10.200	4.575–22.739
Distant metastasis (M1)	<0.0001	3.745	2.165–6.476
Histological type (undifferentiated)	0.4363	0.765	0.390–1.501
Lymphatic invasion (1, 2, 3)	0.0047	2.541	1.332–4.850
Venous invasion (1, 2, 3)	0.0021	2.453	1.383–4.351
Tumor size (>40 mm)	0.0013	2.750	1.484–5.098
Lymph node dissection (three fields)	0.6601	1.142	0.632–2.064
Neoadjuvant therapy	0.5469	1.302	0.552–3.069
Adjuvant therapy	0.0020	2.468	1.392–4.374
Residual tumor (R1, 2)	<0.0001	17.248	7.826–38.016

NmGPS new modified Glasgow Prognostic Score, *HR* hazard ratio, *CI* confidence interval, *T* tumor, *N* node, *M* metastasis

Table 8 The multivariate prognostic factors for esophageal cancer (including the NmGPS [CRP cutoff; 0.5 mg/dL])

Variable	*p*	HR	95 % CI
NmGPS (CRP cutoff; 0.5 mg/dL) (2)	0.0002	4.437	2.000–9.844
Depth of tumor invasion (T3, 4)	0.2794	1.583	0.688–3.642
Lymph node metastasis			
N1	0.1379	2.236	0.772–6.476
N2	0.1119	2.553	0.804–8.108
N3	0.0201	3.731	1.229–11.323
Distant metastasis (M1)	0.3020	1.470	0.707–3.054
Lymphatic invasion (1, 2, 3)	0.3928	0.710	0.323–1.558
Venous invasion (1, 2, 3)	0.0190	2.286	1.145–4.563
Tumor size (>40 mm)	0.5685	1.251	0.579–2.702
Adjuvant therapy	0.0949	1.807	0.902–3.618
Residual tumor (R1, 2)	0.0209	3.230	1.194–8.737

NmGPS new modified Glasgow Prognostic Score, *HR* hazard ratio, *CI* confidence interval, *T* tumor, *N* node, *M* metastasis

than half the mean CRP value of people in other countries. Regarding the cause of the low CRP in East Asia, it was suggested that the haplotype map (HapMap) frequencies of CRP polymorphisms known to be associated with the CRP concentration might differ by ancestry; but for the most part, the difference in CRP is still unexplained [20]. Therefore, it could be that the low CRP cutoff value, we identified reflects the low mean CRP value of East Asians.

The mechanism responsible for the association between a systemic inflammatory response (SIR) and a poor outcome in patients with advanced cancer is not well understood. However, there is increasing evidence that there is a relationship between SIR and cancer survival. Cancer cells might influence the tumor microenvironment through the upregulation of inflammatory pathways by producing proinflammatory mediators, such as cytokines, chemokines, cyclooxygenase-2 (COX-2), prostaglandins, inducible nitric oxide synthase and nitric oxide [21]. These proinflammatory mediators markedly promote tumor progression, invasion and metastasis [21]. Interleukin-6 (IL-6) is a proinflammatory cytokine associated with angiogenesis, and it induces both the development and progression of cancer [22]. CRP is produced by hepatocytes in response to inflammatory cytokines, particularly interleukin-6, which is present in the tumor microenvironment [23]. Because the SIR is also associated with lymphocytopenia and an impaired T-lymphocytic response within the tumor microenvironment, it reflects compromised cell-mediated immunity [23].

Hypoalbuminemia often develops secondary to an ongoing SIR [24]. In addition, the occurrence of a SIR and the associated nutritional decline may influence the

tolerance of and compliance with active treatment [11]. Therefore, the combination of an elevated serum CRP level and hypoalbuminemia reflects both SIR and the progressive nutritional decline of the patient with advanced cancer, and can predict the malignant potential of the tumor and a worse prognosis of cancer patients.

CRP and albumin are routinely evaluated parameters. The GPS is simpler and cheaper than other techniques such as computed tomography, magnetic resonance imaging and positron emission tomography. Therefore, because we can easily predict the prognosis of cancer patients using the NmGPS (CRP cutoff; 0.5 mg/dL), we can take appropriate measures to care for postoperative patients to improve their survival.

In conclusion, we developed a simple and sensitive prognostic scoring system for patients with esophageal squamous cell carcinoma based on the GPS. This scoring system may also be useful for predicting the prognosis of patients with other carcinomas.

Conflict of interest M. Nakamura and the co-authors have no conflict of interest to declare.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010;127:2893–917.
2. Yan W, Wistuba II, Emmert-Buck MR, Erickson HS. Squamous cell carcinoma-similarities and differences among anatomical sites. Am J Cancer Res. 2011;1:275–300.
3. Tran GD, Sun XD, Abnet CC, Fan JH, Dawsey SM, Dong ZW, et al. Prospective study of risk factors for esophageal and gastric cancers in the Linxian general population trial cohort in China. Int J Cancer. 2005;113:456–63.
4. Vizcaino AP, Moreno V, Lambert R, Parkin DM. Time trends incidence of both major histologic types of esophageal carcinomas in selected countries, 1973–1995. Int J Cancer. 2002;99: 860–8.
5. Crumley AB, McMillan DC, McKernan M, Going JJ, Shearer CJ, Stuart RC. An elevated C-reactive protein concentration, prior to surgery, predicts poor cancer-specific survival in patients undergoing resection for gastro-oesophageal cancer. Br J Cancer. 2006;94:1568–71.
6. Nozoe T, Mori E, Takahashi I, Ezaki T. Preoperative elevation of serum C-reactive protein as an independent prognostic indicator of colorectal carcinoma. Surg Today. 2008;38:597–602.
7. Forrest LM, McMillan DC, McArdle CS, Angerson WJ, Dunlop DJ. Evaluation of cumulative prognostic scores based on the systemic inflammatory response in patients with inoperable non-small-cell lung cancer. Br J Cancer. 2003;89:1028–30.
8. Ishizuka M, Nagata H, Takagi K, Horie T, Kubota K. Inflammation-based prognostic score is a novel predictor of postoperative outcome in patients with colorectal cancer. Ann Surg. 2007;246:1047–51.
9. Nozoe T, Iguchi T, Egashira A, Adachi E, Matsukuma A, Ezaki T. Significance of modified Glasgow prognostic score as a useful indicator for prognosis of patients with gastric carcinoma. Am J Surg. 2011;201:186–91.
10. Glen P, Jamieson NB, McMillan DC, Carter R, Imrie CW, McKay CJ. Evaluation of an inflammation-based prognostic score in patients with inoperable pancreatic cancer. Pancreatology. 2006;6:450–3.
11. Forrest LM, McMillan DC, McArdle CS, Angerson WJ, Dunlop DJ. Comparison of an inflammation-based prognostic score (GPS) with performance status (ECOG) in patients receiving platinum-based chemotherapy for inoperable non small-cell lung cancer. Br J Cancer. 2004;90:1704–6.
12. Leitch EF, Chakrabarti M, Crozier JE, McKee RF, Anderson JH, Horgan PG, et al. Comparison of the prognostic value of selected markers of the systemic inflammatory response in patients with colorectal cancer. Br J Cancer. 2007;97:1266–70.
13. La Torre M, Nigri G, Cavallini M, Mercantini P, Ziparo V, Ramacciato G. The glasgow prognostic score as a predictor of survival in patients with potentially resectable pancreatic adenocarcinoma. Ann Surg Oncol. 2012;19:2917–23.
14. Dutta S, Crumley AB, Fullarton GM, Horgan PG, McMillan DC. Comparison of the prognostic value of tumour and patient related factors in patients undergoing potentially curative resection of gastric cancer. Am J Surg. 2012;204:294–9.
15. Ishizuka M, Nagata H, Takagi K, Kubota K. Influence of inflammation-based prognostic score on mortality of patients undergoing chemotherapy for far advanced or recurrent unresectable colorectal cancer. Ann Surg. 2009;250:268–72.
16. Hwang JE, Kim HN, Kim DE, Choi HJ, Jung SH, Shim HJ, et al. Prognostic significance of a systemic inflammatory response in patients receiving first-line palliative chemotherapy for recurred or metastatic gastric cancer. BMC Cancer. 2011;11:489.
17. Sobin LH, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumors. 7th ed. Oxford: Wiley-Blackwell; 2010.
18. Matsubara T, Kaise T, Ishiguro M, Nakajima T. Better grading systems for evaluating the degree of lymph node invasion in cancer of the thoracic esophagus. Surg Today. 1994;24:500–5.
19. Oota H, Saitou N, Matsushita T, Ueda S. Molecular genetic analysis of remains of a 2,000-year-old human population in China-and its relevance for the origin of the modern Japanese population. Am J Hum Genet. 1999;64:250–8.
20. Shah T, Newcombe P, Smeeth L, Addo J, Casas JP, Whittaker J, et al. Ancestry as a determinant of mean population C-reactive protein values: implications for cardiovascular risk prediction. Circ Cardiovasc Genet. 2010;3:436–44.
21. Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002;420:860–7.
22. Cohen T, Nahari D, Cerem LW, Neufeld G, Levi BZ. Interleukin 6 induces the expression of vascular endothelial growth factor. J Biol Chem. 1996;271:736–41.
23. McArdle PA, McMillan DC, Sattar N, Wallace AM, Underwood MA. The relationship between interleukin-6 and C-reactive protein in patients with benign and malignant prostate disease. Br J Cancer. 2004;91:1755–7.
24. Al-Shaiba R, McMillan DC, Angerson WJ, Leen E, McArdle CS, Horgan P. The relationship between hypoalbuminaemia, tumour volume and the systemic inflammatory response in patients with colorectal liver metastases. Br J Cancer. 2004;91:205–7.

Surg Today (2014) 44:1708–1715
DOI 10.1007/s00595-013-0782-3

ORIGINAL ARTICLE

Clinical benefits of thoracoscopic esophagectomy in the prone position for esophageal cancer

Makoto Iwahashi · Mikihito Nakamori · Masaki Nakamura · Toshiyasu Ojima · Masahiro Katsuda · Takeshi Iida · Keiji Hayata · Hiroki Yamaue

Received: 11 August 2013 / Accepted: 11 October 2013 / Published online: 20 November 2013

Abstract

Purposes The clinical benefits of thoracoscopic radical esophagectomy in the prone position compared to conventional open esophagectomy have not been fully documented.

Methods Forty-six patients with esophageal cancer who underwent MIE in the prone position (MIE-P group) were enrolled, and 46 case-matched controls that underwent open esophagectomy (OE group) were identified using propensity score methods to achieve a valid comparison of outcomes between MIE and open esophagectomy.

Results The duration of systemic inflammatory response syndrome was shorter in the MIE-P group than in OE group ($P = 0.005$). The time to first walking was earlier in the MIE-P group ($P < 0.001$). Although the vital capacity ratio (%VC) declined after the operation in both groups, the change ratio of the %VC was 85.3 % in the MIE-P group and 69.6 % in the OE group ($P < 0.001$). No mortality occurred in either group. The postoperative morbidity rate was lower in the MIE-P group (13 %) than in the OE group (30.4 %) ($P = 0.020$). Two patients (4.3 %) in the OE group and one patient in the MIE-P group (2.2 %) had pneumonia.

Conclusions MIE in the prone position was associated with less impairment of the pulmonary function, earlier recovery of activity and lower subsequent morbidity compared to open esophagectomy. Further investigation of the long-term outcomes is, therefore, needed.

M. Iwahashi · M. Nakamori · M. Nakamura · T. Ojima · M. Katsuda · T. Iida · K. Hayata · H. Yamaue (✉)
Second Department of Surgery, School of Medicine, Wakayama Medical University, 811-1 Kimiidera, Wakayama 641-8510, Japan
e-mail: yamaue-h@wakayama-med.ac.jp

Keywords Minimally invasive esophagectomy · Thoracoscopic esophagectomy · Esophageal cancer · Prone position · Postoperative pulmonary function · Postoperative morbidity

Introduction

A number of studies have demonstrated the safety and possible advantages of minimally invasive esophagectomy (MIE) in selected cohorts of patients [1–6]. MIE is, therefore, being performed with increasing frequency [7], and evidence of the short-term benefits of MIE over traditional open procedures with a similar oncological outcome is accumulating. Most comparative studies have demonstrated clinical advantages of MIE, such as less blood loss, a shorter intensive care unit (ICU) stay and similar survival. Nevertheless, systemic reviews of studies involving MIE have been equivocal and have failed to draw definitive conclusions [8]. A population-based national study in England has shown that there were no significant benefits demonstrated in the mortality and overall morbidity [9]. Most recently, a randomized controlled trial has shown the benefits of MIE in terms of a lower incidence of pulmonary infection and better quality of life compared to open esophagectomy [10].

Various types of MIE for patients with esophageal cancer have been described, and the most generally performed technique involves thoracoscopic mobilization of the esophagus in the left lateral decubitus position [1, 2, 4, 11]. Recently, the advantages of thoracoscopic esophageal mobilization in the prone position have also been reported [12, 13]. Compared to the left lateral decubitus position, the prone position allows better operative exposure and improved surgeon ergonomics, resulting in reduced

pulmonary complications, a shorter operation and less blood loss [12, 13]. The advantages in terms of surgeon ergonomics and operative exposure were apparent even in an aggressive esophagectomy with a three-field lymphadenectomy [14]. Therefore, thoracoscopic esophagectomy in the prone position would be expected to have potential benefits in radical esophagectomy with extended lymph node dissection for patients with esophageal cancer. However, the efficacy of this approach as a minimally invasive surgery compared to conventional open esophagectomy has not yet been fully documented [15].

The aim of the present study was to examine the clinical benefits of the thoracoscopic radical esophagectomy with extensive lymphadenectomy in the prone position as a minimally invasive surgery compared to open esophagectomy for patients with esophageal cancer, using a propensity score-matching analysis to evaluate the outcomes without selection bias.

Patients and methods

Patient populations

From January 2004 to December 2011, 288 patients with esophageal cancer underwent esophagectomy at Wakayama Medical University Hospital (WMUH). Until 2009, our traditional standard surgical procedure for patients with esophageal cancer was transthoracic open esophagectomy. Thoracoscopic esophagectomy with extensive lymphadenectomy in the prone position was adopted in 2009 and was performed in 51 patients with thoracic or abdominal esophageal cancer at WMUH. The selection criteria for this minimally invasive procedure were as follows: no previous thoracic surgery, no possibility of severe pleural adhesion and no previous radiation therapy to the esophagus. During the first year, this procedure was only performed in patients with clinical T1 tumors, but from the second year, it was also performed in patients with clinical T2 or T3 tumors. From among these 51 patients, 46 patients who underwent curative resection with gastric conduit reconstruction were enrolled in our study (MIE-P group). A case-matched control group (OE group) was identified from patients who underwent open transthoracic esophagectomy between January 2004 to December 2011 using propensity score methods to achieve a valid comparison of outcomes between MIE and open esophagectomy.

The patients were staged according to the TNM classification (7th edition) of the American Joint Committee on Cancer and the International Union Against Cancer.

This study was approved by the Ethics Committee on Human Research of WMUH, and informed consent was obtained from all patients.

Perioperative management

The perioperative care and anesthesia of esophageal cancer surgery are standardized at WMUH, as reported previously [16]. Respiratory physiotherapy and oral care were performed in all patients in our study groups. Intravenous methylprednisolone (125 mg) was administered twice to each patient, at the beginning of the thoracic procedure and at the end of the surgery. The patients were usually admitted to the ICU immediately after the operation. The patients were discharged from the ICU after extubation if their conditions remained stable. Epidural analgesia was used routinely for postoperative pain management for 1 week. Postoperative physical rehabilitation strategies, such as deep breathing with huffing and coughing and postural drainage to assist breathing and expectoration, were performed in all patients. In addition, the patients were instructed to stand up and walk as early as possible.

Surgical procedures

The patients underwent a radical esophagectomy with a total mediastinal lymphadenectomy (extended two-field) or three-field lymph node dissection via a cervicothoracoabdominal approach. The operation was conducted in three stages. In the first stage, the intrathoracic manipulation was performed as described in detail in the following sections. The patient was placed in the prone position for the MIE procedure or in the left lateral decubitus position for the open esophagectomy. In the second stage, the patient was rotated to a supine position, and gastric mobilization and abdominal lymphadenectomy (around the left gastric pedicle and celiac axis) were performed. The entire isolated thoracic esophageal specimen with dissected lymph nodes (LNs) was removed through the esophageal hiatus, and a gastric conduit was constructed. In the MIE-P group, a hand-assisted laparoscopic approach was used for these abdominal procedures. However, eight of 46 patients were converted to open surgery because of obvious LN metastases around the celiac axis or due to abdominal adhesion from previous upper abdominal surgery. In the third stage, the cervical esophagus was mobilized, and a cervical lymphadenectomy was performed. Finally, a gastric conduit was delivered up through the retrosternal rout or through the posterior mediastinum to reconstruct the anastomosis with the cervical esophagus in the cervical field.

Thoracoscopic esophagectomy in the prone position

The patient was placed in the prone position while under epidural and general anesthesia, which used a single-lumen flexible endotracheal tube with a blocking balloon

inserted into the right main bronchus for single-lung ventilation. The port positions were as follows: A 12-mm blunt port was first carefully inserted into the seventh intercostal space (ICS) behind the posterior axillary line, and CO_2 was then insufflated at a pressure of 6 mmHg. Four other ports were inserted under thoracoscopic control: a 12-mm port in the ninth ICS on the scapular angle line for the thoracoscope, a 5-mm port in the fifth ICS on the midaxillary line, a 12-mm port in the third ICS on the midaxillary line and a 5-mm port in the seventh ICS on the scapular angle line. Intrathoracic procedures began with mobilizing the middle and lower esophagus with regional LNs along the layer that exposed the pericardium, the descending aorta and the left mediastinal pleura. The azygos vein was then divided and the right bronchial artery was also divided. The entire thoracic duct was carefully preserved. Then, the upper thoracic esophagus, the right main branch of the vagal nerve and the right subclavian artery were exposed. The LNs around the right recurrent nerve were dissected up to the thyroid gland. Then, the upper thoracic esophagus was circumferentially mobilized, and tape was placed around the esophagus to facilitate retraction. The upper thoracic esophagus was retracted to the dorsal side by pulling the tape via thread outside the thorax, and the trachea was rolled back to the right and ventrally by a grasper holding small gauze, as described previously [14]. The tissue, including the LNs around the left recurrent nerve, was precisely dissected. Next, the esophagus was divided, the LNs below the aortic arch were dissected and the bilateral esophageal branches of the vagal nerve were divided, while pulmonary branches were preserved. The subcarinal LNs were dissected. The thoracic esophagus was completely mobilized circumferentially, and the paraesophageal LNs were dissected and maintained *en bloc* with the surgical specimen. A single 28-F chest tube was placed through the 12-mm port site for postoperative drainage.

Open transthoracic esophagectomy

The patient was placed in the left lateral decubitus position. A right vertical incision (12 cm) was made, and thoracotomy was performed through the fourth ICS. A 12-mm port was placed in the sixth ICS on the midaxillary line for the thoracoscope. The operator looked directly at the surgical field, and assistant surgeons viewed the area through a video monitor. The mobilization of the esophagus and lymphadenectomy was performed almost the same way as in the thoracoscopic esophagectomy. However, the definite difference from thoracoscopic procedures was that the retraction of the lung, the trachea, the right bronchus and the heart by assistants was necessary to expose the surgical field.

Patient monitoring and data collection

The data were recorded prospectively for all patients who underwent thoracoscopic esophagectomy in the prone position. However, the data were retrospectively collected from medical records for patients in the control OE group. Patients with disorders such as angina pectoris or previous myocardial infarction were defined as having cardiovascular disease. Patients with abnormal pulmonary function on spirograms (vital capacity ratio [%VC] <70 % or forced expiratory volume in 1 s [FEV_1]/forced vital capacity [FVC] <60 %) were defined as having comorbid pulmonary disease. Diabetes mellitus was noted if the patient had a fasting blood glucose concentration >126 mg/dL or was receiving antidiabetic therapy. Patients with chronic hepatitis or liver cirrhosis that required treatment were defined as having liver disease. Patients with renal disease that required treatment were defined as having renal disease.

Evaluation of the postoperative clinical course and respiratory function

The duration of systemic inflammatory response syndrome (SIRS) and ICU stay, the time to first independent sitting, time to first standing and time to first walking outside of the room were evaluated by comparing the two groups. The white blood cell (WBC) count and C-reactive protein (CRP) level were compared before the operation, and on postoperative days (PODs) 1, 3 and 5. Postoperative complications were analyzed according to the Clavien–Dindo classification [17], and postoperative complications greater than grade II were regarded as being clinically significant. The surgical mortality (Clavien–Dindo classification; grade V) included in-hospital deaths (by POD 90). Pulmonary function was evaluated by the %VC and the ratio of FEV1 to FVC (FEV1%) in spirograms before and 3 to 4 weeks after the operation. The change ratio of the %VC (postoperative %VC/preoperative %VC × 100) and the change ratio of the FEV1% (postoperative FEV1%/preoperative FEV1% × 100) were also compared between the two groups.

Statistical analysis

The case-matched control group was identified using a propensity score matching method. The propensity score was calculated for each patient by a logistic regression analysis based on the following variables: age, sex, tumor location, depth of tumor invasion, the degree of LN involvement, pathological stage, histological type of tumor, preoperative chemotherapy and concomitant diseases. The quantitative results were expressed as the mean ± standard

Table 1 Patient and tumor demographics

	OE ($n = 46$)	MIE-P ($n = 46$)	P value[§]
Age[a]	65.9 ± 8.9	65.0 ± 10.2	0.689
Sex			0.797
Male/female	37/9	36/10	
Location of tumor			0.663
Upper	4	2	
Middle	30	30	
Lower	12	14	
Depth of tumor invasion			1.00
T1	35	35	
T2	4	4	
T3	7	7	
Lymph node metastases			0.717
N0	27	28	
N1	8	8	
N2	5	7	
N3	6	3	
Stage			0.947
IA + IB	25	26	
IIA + IIB	13	14	
IIIA + IIIB + IIIC	5	4	
IV	3	2	
Histology			1.00
Squamous cell carcinoma	44	44	
Adenocarcinoma	1	1	
Others	1	1	
Neoadjuvant chemotherapy			0.216
Yes/no	4/42	8/38	
Lymph node dissection			
Extended 2-field[b]/3-field	35/11	38/8	0.440
Comorbidity			
Cardiac Yes/no	2/44	1/45	0.500
Respiratory Yes/no	8/38	9/37	0.788
Diabetes Yes/no	6/40	2/44	0.133
Liver Yes/no	3/43	2/44	0.500

OE open esophagectomy, *MIE-P* minimally invasive esophagectomy in the prone position

§ P value between the OE group and the MIE-P group

[a] Data are expressed as the mean ± SD

[b] Extended 2-field: total mediastinal lymphadenectomy

deviation or the medians (range). The statistical analysis was performed by Student's t test, the Mann–Whitney U test or Fisher's exact test, as appropriate. Values of $P < 0.05$ were considered significant. All statistical analyses were carried out using the SPSS software package (v. 19.0; SPSS, Chicago, IL).

Table 2 Procedure-related data

	OE ($n = 46$)	MIE-P ($n = 46$)	P value[‡]
Length of operation (min)[a]			
Total	488 ± 59	609 ± 54	<0.001
Chest	234 ± 44	362 ± 40	<0.001
Estimated blood loss (mL)[b]	255 (72–925)	125 (30–420)	<0.001
Number of dissected lymph nodes[b]			
Chest	22 (9–54)	23 (9–36)	0.774
Along the recurrent nerve	6 (1–14)	7 (1–17)	0.253
Length of ICU stay (d)[c]	1.07 (1–3)	1.07 (1–2)	0.669
Duration of SIRS (d)[c]	0.82 (0–4)	0.22 (0–3)	0.005
Days until independent sitting (d)[c]	1.74 (1–5)	1.02 (1–2)	<0.001
Days until first standing (d)[c]	1.82 (1–5)	1.02 (1–2)	<0.001
Days until first walking (d)[c]	3.47 (1–21)	1.02 (1–2)	<0.001

OE open esophagectomy, *MIE-P* minimally invasive esophagectomy in the prone position, *SIRS* systemic inflammatory response syndrome

‡ P value between the OE group and the MIE-P group

[a] Data are expressed as the mean ± SD

[b] Data are expressed as the medians (range)

[c] Data are expressed as the means (range)

Results

Patients and tumor demographics

Forty-six patients who underwent thoracoscopic esophagectomy in the prone position were enrolled in this study and were matched 1:1 with 46 patients who underwent open esophagectomy. The patient and tumor demographics and the associated preoperative comorbidities for the matched groups are summarized in Table 1. There were no significant differences between the groups.

Surgical outcomes

The surgical outcomes after esophagectomy are shown in Table 2. The mean total length of the operation was significantly longer in the MIE-P group (609 ± 54 min) than in the OE group (488 ± 59 min), which was due to the significantly longer duration of the thoracoscopic procedure ($P < 0.001$). The median estimated blood loss in the MIE-P group was 125 mL (range 30–420 mL) and that in the OE group was 255 mL (range 72–925 mL); the difference was statistically significant between the two groups ($P \leq 0.001$). The number of dissected LNs in the chest and along the recurrent nerve in the OE group was similar to that in the MIE-P group. There were no significant differences in the length of ICU stay between

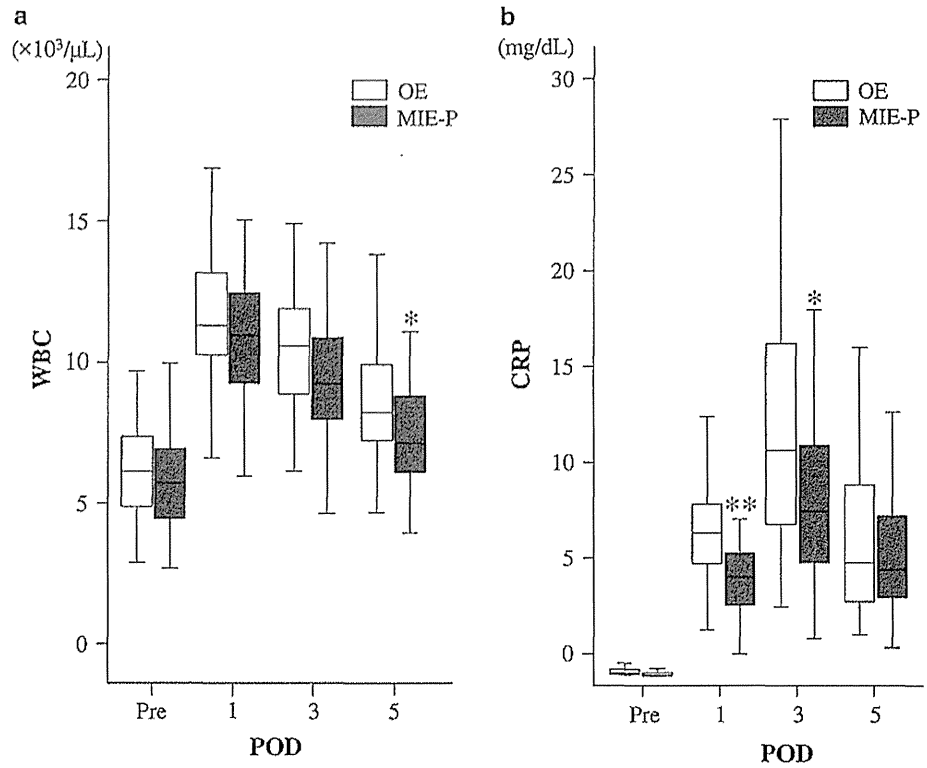

Fig. 1 The pre- and postoperative changes in the white blood cell (WBC) count (a) and serum C-reactive protein (CRP) level (b). *The horizontal bars* indicate the median values. *The vertical bars* indicate the range, and the horizontal boundaries of the *boxes* represent the first and third quartiles. *OE* open esophagectomy, *MIE-P* minimally invasive esophagectomy in the prone position. $^{**}P < 0.01$, $^{*}P < 0.05$ significantly different between the OE group and the MIE-P group

the two groups, while the duration of SIRS was shorter in the MIE-P group than in the OE group ($P = 0.005$). In terms of the activity of daily living (ADL), the time to first independent sitting, time to first standing and time to first walking outside of the room were all significantly earlier in the MIE-P group compared to the OE group ($P < 0.001$).

The perioperative changes in the WBC count and serum CRP level are shown in Fig. 1. The WBC count peaked on POD 1, and the CRP level peaked on POD 3. The peak WBC counts were similar between the OE and MIE-P groups; however, the WBC count in the MIE-P group was significantly lower than that in the OE group on POD 5 ($P < 0.05$). The serum CRP levels in the MIE-P group were significantly lower than those in the OE group on POD 1 and 3 ($P < 0.01$, $P < 0.05$).

Perioperative pulmonary function

The perioperative changes in the respiratory function were evaluated by spirography (Table 3). The %VC was decreased after the operation in both the OE and MIE-P groups. Importantly, the %VC was significantly higher in the MIE-P group than in the OE group after the operation ($P = 0.002$), although it was similar between the two groups before the operation. The change ratio of the %VC was significantly different between the two groups ($P < 0.001$): 85.3 % in the MIE-P group and 69.6 % in the OE group. The FEV1% was not influenced by the operation in either group. There were no differences in the FEV1% between the two groups before or after the operation.

Table 3 Perioperative changes in the pulmonary functions

	OE ($n = 46$)	MIE-P ($n = 46$)	P value‡
%VC			
Preoperative	112.8 ± 17.6	108.9 ± 15.7	0.412
Postoperative	78.0 ± 12.9	92.6 ± 16.3	0.002
Change ratio[a]	69.6 ± 8.8	85.3 ± 10.7	<0.001
FEV1%			
Preoperative	74.8 ± 10.4	76.4 ± 9.7	0.558
Postoperative	77.9 ± 9.5	79.0 ± 9.5	0.683
Change ratio[b]	104.8 ± 9.5	103.9 ± 9.5	0.687

Data are expressed as the mean ± SD

OE open esophagectomy, *MIE-P* minimally invasive esophagectomy in the prone position, *%VC* %vital capacity, *FEV1%* the ratio of the forced expiratory volume in one second (FEV_1) to the forced vital capacity (FVC)

‡ P value between the OE group and the MIE-P group

[a] Change ratio = postoperative %VC/preoperative %VC × 100

[b] Change ratio = postoperative FEV1%/preoperative FEV1% × 100

Table 4 Postoperative morbidity and mortality

	Number (%)		
	OE ($n = 46$)	MIE-P ($n = 46$)	P value‡
Mortality	0	0	1.0
Complications[a]	14 (30.4)	6[b] (13.0)	0.020
Pneumonia	2 (4.3)	1 (2.2)	0.500
Grade II/IIIa/IIIb, IV	1/1/0	0/1/0	
Chylothorax	1 (2.2)	0	0.315
Grade II/IIIa/IIIb, IV	1/0/0	–	
Anastomotic leakage	4 (8.7)	1 (2.2)	0.181
Grade II/IIIa/IIIb, IV	3/1/0	0/1/0	
Surgical site infection	2 (4.3)	0	0.153
Grade II/IIIa/IIIb, IV	2/0/0	–	
Recurrent nerve palsy	5 (10.9)	5 (10.9)	1.0
Grade II/IIIa/IIIb, IV	4/1/0	4/1/0	

‡ P value between the OE group and the MIE-P group

OE open esophagectomy, *MIE-P* minimally invasive esophagectomy in the prone position

[a] Clavien–Dindo classification

[b] One of six patients developed both pneumonia and anastomotic leakage

Postoperative morbidity and mortality

There was no operative mortality associated with either approach in this study. The incidence of postoperative complications was significantly lower in the MIE-P group (13 %) than in the OE group (30.4 %) ($P = 0.020$). The specific complications of grades II, III and IV according to the Clavien–Dindo classification are summarized in Table 4. None of the patients developed severe complications of grade IIIb or IV. Two patients (4.3 %) in the OE group and one patient in the MIE-P group (2.2 %) had pneumonia. Four patients (8.7 %) in the OE group and one (2.2 %) in the MIE-P group developed anastomotic leakage. One patient (2.2 %) developed chylothorax and two patients (4.3 %) developed a surgical site infection in the OE group, while no patients in the MIE-P group developed chylothorax or a surgical site infection. Five patients (10.9 %) in each group, respectively, developed recurrent nerve palsy. There were no statistically significant differences in the incidence of each specific complication between the two groups. The statistical data are shown in Table 4.

Discussion

The present study shows that MIE in the prone position has several advantages over conventional open esophagectomy in the radical surgical treatment with extensive lymphadenectomy. The patients who underwent MIE had an earlier recovery of the ADL, lower levels of the inflammatory response, less impairment of pulmonary function and lower morbidity. In addition, the propensity score matching technique was used to minimize any selection bias, and the patient and tumor demographics showed good matching.

The minimally invasive procedure in itself is expected to reduce the surgical stress after esophagectomy. However, there have been few studies that have evaluated the postoperative inflammatory response in patients who underwent MIE. The most recent study demonstrated that the serum levels of inflammatory cytokines, such as interleukin-6, immediately after the operation are significantly lower in patients who underwent MIE in the prone position than in those who underwent conventional open esophagectomy; further, the incidence of SIRS was lower in patients who underwent MIE [18]. In the present study, the peak serum levels of CRP in the MIE-P group were significantly lower than those in the OE group, and the duration of SIRS was significantly shorter in the MIE-P group than in the OE group. This is consistent with the study by Tsujimoto et al. [18], and it suggests that MIE in the prone position in itself is a less invasive procedure.

Less invasive surgical procedures may allow patients to achieve comparatively earlier recovery. A thoracoscopic esophagectomy minimizes the injury to the chest wall and would, therefore, contribute to an early physical recovery. In this context, there have not been enough reports that have examined the postoperative course of recovery. In the present study, the time to first independent sitting, time to first standing and time to first walking outside of the room were significantly earlier in the MIE-P group than in the OE group. All patients except one in the MIE-P group could walk outside of their room on POD 1. However, there is a possibility of some bias in the background of the patients because the early recovery program was launched in 2007 at our hospital.

The postoperative pulmonary function may affect not only the postoperative morbidity, but also the postoperative quality of life. In particular, the vital capacity is usually impaired after open transthoracic surgery due to the injury to the chest wall, including respiratory muscles. Preservation of the pulmonary function is one of the most important issues after esophagectomy. However, this outcome has not been fully investigated in previous studies, especially during the early postoperative period. In the present study, there were no significant differences in the FEV1% between the OE and MIE-P groups before or after the operation. This is likely because the airway resistance is not influenced by esophagectomy, even when the open approach is used. In contrast, the %VC was diminished 3 to 4 weeks after the operation in both groups. Importantly, however, the impairment of the %VC was significantly lower in the MIE-P group than in the OE group after the

operation. MIE can minimize the injury to the chest wall, and moreover, especially when performed in the prone position, it may minimize the direct damage to the lung. This is because the physical retraction of the lung is not needed in the prone position. Therefore, the vital capacity could be preserved even in the early postoperative period following MIE. These results suggest that MIE in the prone position has an obvious benefit in terms of the preservation of pulmonary function.

The postoperative morbidity and mortality remain important issues in patients who undergo esophagectomy for esophageal cancer. Considering the various advantages of minimally invasive surgical procedures, MIE has been expected to reduce the morbidity and mortality. Although MIE is reported to be an independent factor predicting a lower frequency of postoperative respiratory failure [19], many previous reports have failed to demonstrate that MIE actually reduces the morbidity and mortality [1, 2, 4, 8]. In addition, a recent population-based national study from England showed that there were no significant benefits demonstrated in the mortality and mobility [9]. With regard to MIE in the prone position, there have been no studies that have demonstrated obvious advantages over conventional open esophagectomy in terms of postoperative mortality and morbidity [5, 6, 12]. This lack of evidence may be due to a selection bias of patients, as well as the considerable variations in the definition of postoperative complications. In the present study, we used the propensity score matching technique to diminish the selection bias, and as a matter of course, we analyzed the postoperative complications according to the Clavien–Dindo classification [17].

In the present study there was no mortality due to either approach, and the overall morbidity was generally low compared to previous reports [1, 3–5, 12, 19]. The morbidity rate was significantly lower in the MIE-P group (13.0 %) than in the OE group (30.4 %). These results suggest that thoracoscopic esophagectomy with extensive lymphadenectomy in the prone position is less invasive compared to conventional open esophagectomy. The incidence of pneumonia was 2.2 % in the ME-P group and 4.3 % in the OE group. Both rates were low compared with other reports [1, 3–5, 9, 12], and no statistically significant difference could be detected between the groups in terms of the development of pneumonia.

In conclusion, we examined the clinical benefits of thoracoscopic esophagectomy in the prone position for patients with esophageal cancer compared to propensity score-matched control patients who underwent open esophagectomy for esophageal cancer. MIE with extensive lymphadenectomy in the prone position had obvious advantages over conventional open esophagectomy in terms of the lower levels of inflammatory response, less impairment of the pulmonary function, earlier recovery of the ADL and lower subsequent morbidity. However, the present study has some limitations with regard to showing the definitive benefits of MIE in the prone position because it was a nonrandomized study. Further investigations are also needed to determine the long-term outcomes of this procedure, including the survival.

Conflict of interest Makoto Iwahashi and the co-authors have no conflicts of interest to declare.

References

1. Luketich JD, Alvelo-Rivera M, Buenaventura PO, Christie NA, McCaughan JS, Litle VR, et al. Minimally invasive esophagectomy: outcomes in 222 patients. Ann Surg. 2003;238:486–94.
2. Nguyen NT, Follette DM, Wolfe BM, Schneider PD, Roberts P, Goodnight JE Jr. Comparison of minimally invasive esophagectomy with transthoracic and transhiatal esophagectomy. Arch Surg. 2000;135:920–5.
3. Nguyen NT, Hinojosa MW, Smith BR, Chang KJ, Gray J, Hoyt D. Minimally invasive esophagectomy: lessons learned from 104 operations. Ann Surg. 2008;248:1081–91.
4. Osugi H, Takemura M, Higashino M, Takada N, Lee S, Kinoshita H. A comparison of video-assisted thoracoscopic oesophagectomy and radical lymph node dissection for squamous cell cancer of the oesophagus with open operation. Br J Surg. 2003;90:108–13.
5. Smithers BM, Gotley DC, Martin I, Thomas JM. Comparison of the outcomes between open and minimally invasive esophagectomy. Ann Surg. 2007;245:232–40.
6. Zingg U, McQuinn A, DiValentino D, Esterman AJ, Bessell JR, Thompson SK, et al. Minimally invasive versus open esophagectomy for patients with esophageal cancer. Ann Thorac Surg. 2009;87:911–9.
7. Lazzarino AI, Nagpal K, Bottle A, Faiz O, Moorthy K, Aylin P. Open versus minimally invasive esophagectomy: trends of utilization and associated outcomes in England. Ann Surg. 2010;252:292–8.
8. Decker G, Coosemans W, De Leyn P, Decaluwe H, Nafteux P, Van Raemdonck D, et al. Minimally invasive esophagectomy for cancer. Eur J Cardiothorac Surg. 2009;35:13–20.
9. Mamidanna R, Bottle A, Aylin P, Faiz O, Hanna GB. Short-term outcomes following open versus minimally invasive esophagectomy for cancer in England: a population-based national study. Ann Surg. 2012;255:197–203.
10. Biere SSAY, van Berge Henegouwen MI, Maas KW, Bonavina L, Rosman C, Garcia JR, et al. Minimally invasive versus open oesophagectomy for patients with oesophageal cancer: a multicentre, open-label, randomised controlled trial. Lancet. 2012;379:1887–92.
11. Akaishi T, Kaneda I, Higuchi N, Kuriya Y, Kuramoto J, Toyoda T, et al. Thoracoscopic en bloc total esophagectomy with radical mediastinal lymphadenectomy. J Thorac Cardiovasc Surg. 1996;112:1533–40.
12. Fabian T, Martin J, Katigbak M, McKelvey AA, Federico JA. Thoracoscopic esophageal mobilization during minimally invasive esophagectomy: a head-to-head comparison of prone versus decubitus positions. Surg Endosc. 2008;22:2485–91.
13. Palanivelu C, Prakash A, Senthilkumar R, Senthilnathan P, Parthasarathi R, Rajan PS, et al. Minimally invasive esophagectomy: thoracoscopic mobilization of the esophagus and mediastinal lymphadenectomy in prone position—experience of 130 patients. J Am Coll Surg. 2006;203:7–16.

14. Noshiro H, Iwasaki H, Kobayashi K, Uchiyama A, Miyasaka Y, Masatsugu T, et al. Lymphadenectomy along the left recurrent laryngeal nerve by a minimally invasive esophagectomy in the prone position for thoracic esophageal cancer. Surg Endosc. 2010;24:2965–73.
15. Watanabe M, Baba Y, Nagai Y, Baba H. Minimally invasive esophagectomy for esophageal cancer: an updated review. Surg Today. 2013;43:237–44.
16. Nakamura M, Iwahashi M, Nakamori M, Ishida K, Naka T, Iida T, et al. An analysis of the factors contributing to a reduction in the incidence of pulmonary complications following an esophagectomy for esophageal cancer. Langenbecks Arch Surg. 2008;393:127–33.
17. 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;240:205–13.
18. Tsujimoto H, Takahata R, Nomura S, Yaguchi Y, Kumano I, Matsumoto Y, et al. Video-assisted thoracoscopic surgery for esophageal cancer attenuates postoperative systemic responses and pulmonary complications. Surgery. 2012;151:667–73.
19. Zingg U, Smithers BM, Gotley DC, Smith G, Aly A, Clough A, et al. Factors associated with postoperative pulmonary morbidity after esophagectomy for cancer. Ann Surg Oncol. 2011;18:1460–8.

Research Article

Statistics in Medicine

Received 9 May 2013, Accepted 2 March 2014 Published online 27 March 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6154

Sample size determination in group-sequential clinical trials with two co-primary endpoints

Koko Asakura,[a,b] Toshimitsu Hamasaki,[a,b*†] Tomoyuki Sugimoto,[c] Kenichi Hayashi,[a] Scott R. Evans[d] and Takashi Sozu[e]

We discuss sample size determination in group-sequential designs with two endpoints as co-primary. We derive the power and sample size within two decision-making frameworks. One is to claim the test intervention's benefit relative to control when superiority is achieved for the two endpoints *at the same interim timepoint* of the trial. The other is when superiority is achieved for the two endpoints *at any interim timepoint, not necessarily simultaneously*. We evaluate the behaviors of sample size and power with varying design elements and provide a real example to illustrate the proposed sample size methods. In addition, we discuss sample size recalculation based on observed data and evaluate the impact on the power and Type I error rate. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: average sample number; conditional power; Cui–Hung–Wang statistics; co-primary endpoints; group-sequential methods; maximum sample size; sample size recalculation; Type I error

1. Introduction

Traditionally, in clinical trials, a single outcome is selected as a primary endpoint. This endpoint is then used as the basis for the trial design including sample size determination, interim monitoring, and final analyses. However, many recent clinical trials, especially in medical product development, have utilized more than one endpoint as *co-primary*. 'Co-primary' in this setting means that the trial is designed to evaluate if the new intervention is superior to the control on *all* endpoints, thus evaluating the intervention's multidimensional effects. Regulators have issued guidelines recommending co-primary endpoints in some disease areas. For example, the Committee for Medicinal Products for Human Use issued a guideline [1] recommending the use of cognitive, functional, and global endpoints to evaluate symptomatic improvement of dementia associated with Alzheimer's disease, indicating that primary endpoints should be stipulated reflecting the cognitive and functional disease aspects. Offen *et al.* [2] provides other examples with co-primary endpoints for regulatory purposes.

The resulting need for new approaches to the design and analysis of clinical trials with co-primary endpoints has been noted [2–4]. Utilizing multiple endpoints may provide the opportunity for characterizing intervention's multidimensional effects and also create challenges. Specifically controlling the Type I and Type II error rates when the multiple co-primary endpoints are potentially correlated is non-trivial. When designing the trial to evaluate the joint effects on *all* of the endpoints, no adjustment is needed to

[a]*Department of Biomedical Statistics, Osaka University Graduate School of Medicine, Osaka, Japan*
[b]*Office of Biostatistics and Data Management, Research and Development Initiative Center, National Cerebral and Cardiovascular Center, Osaka, Japan*
[c]*Department of Mathematical Sciences, Graduate School of Science and Technology, Hirosaki University, Aomori, Japan*
[d]*Department of Biostatistics and the Center for Biostatistics in AIDS Research, Harvard School of Public Health, Boston, MA, U.S.A.*
[e]*Department of Biostatistics, Kyoto University School of Public Health, Kyoto, Japan*
**Correspondence to: Toshimitsu Hamasaki, Office of Biostatistics and Data Management, Research & Development Initiative Center, National Cerebral and Cardiovascular Center, 5-7 Fujishiro-dai, Suita, Osaka 565-8565, Japan.*
[†]*E-mail: toshi.hamasaki@ncvc.go.jp*

control the Type I error rate. However, the Type II error rate increases as the number of endpoints to be evaluated increases. Thus, adjustments in design (i.e., sample size) are needed to maintain the overall power. Methods for clinical trials with co-primary endpoints have been discussed in fixed sample size designs by many authors [5–16]. Even if the correlation among the endpoints is incorporated into the sample size calculation, existing methods often result in large and impractical sample sizes as the testing procedure for co-primary endpoints is conservative. Chuang-Stein *et al.* [7] and Kordzakhia *et al.* [10] discuss the methods to adjust the significance levels that depend on the correlation among the endpoints in the fixed sample size designs. The methods may provide smaller sample sizes and also introduce the other challenges. For example, the sample size calculated to detect the joint effect may be smaller than the sample size calculated for each individual endpoint. The prespecified correlation incorporated into the significance level adjustment is usually unknown and may be incorrect. This calls into question whether or not the significance level should be updated based on the observed correlation.

In this paper, we extend previous work for the fixed sample size designs, considering sample size evaluation in the group-sequential setting with co-primary endpoints. As suggested by Hung and Wang [3], a group-sequential design may be a remedial but practical approach because it offers the possibility to stop a trial early when evidence is overwhelming and thus offers efficiency (i.e., potentially fewer patients than the fixed sample size designs). We discuss the case of two positively correlated continuous outcomes. We consider a two-arm parallel-group trial designed to evaluate if an experimental intervention is superior to a control. The paper is structured as follows. In Section 2, we describe the statistical setting, decision-making frameworks for rejecting the null hypothesis, and definitions of power. In Section 3, we evaluate the behaviors of sample size and power with varying design elements and then provide a real example to illustrate the methods. In Section 4, we describe sample size recalculation and the resulting effect on power and Type I error rate. In Section 5, we summarize the findings and discuss the further developments.

2. Group-sequential designs with two co-primary endpoints

2.1. Statistical setting

Consider a randomized, group-sequential clinical trial of comparing the test intervention (T) with the control intervention (C). Two continuous outcomes are to be evaluated as co-primary endpoints. Suppose that a maximum of L analyses are planned, where the same number of analyses with the same information space are selected for both endpoints. Let n_l and $r_l n_l$ be the cumulative number of participants on the test and the control intervention groups at the lth analysis ($l = 1, \ldots, L$), respectively, where r_l is the sampling ratio. Hence, up to n_L and $r_L n_L$ participants are recruited and randomly assigned to the test and the control intervention groups, respectively. Then, there are n_L paired outcomes $(Y_{\mathrm{T}1i}, Y_{\mathrm{T}2i})$ $(i = 1, \ldots, n_L)$ for the test intervention group and $r_L n_L$ paired outcomes $(Y_{\mathrm{C}1j}, Y_{\mathrm{C}2j})$ $(j = 1, \ldots, r_L n_L)$ for the control intervention group. Assume that $(Y_{\mathrm{T}1i}, Y_{\mathrm{T}2i})$ and $(Y_{\mathrm{C}1j}, Y_{\mathrm{C}2j})$ are independently bivariate-normally distributed as $(Y_{\mathrm{T}1i}, Y_{\mathrm{T}2i}) \sim \mathrm{N}_2(\mu_{\mathrm{T}1}, \mu_{\mathrm{T}2}, \sigma^2_{\mathrm{T}1}, \sigma^2_{\mathrm{T}2}, \rho_{\mathrm{T}})$ and $(Y_{\mathrm{C}1j}, Y_{\mathrm{C}2j}) \sim \mathrm{N}_2(\mu_{\mathrm{C}1}, \mu_{\mathrm{C}2}, \sigma^2_{\mathrm{C}1}, \sigma^2_{\mathrm{C}2}, \rho_{\mathrm{C}})$, respectively. For simplicity, the variances are assumed to be known and common, that is, $\sigma^2_{\mathrm{T}1} = \sigma^2_{\mathrm{C}1} = \sigma^2_1$ and $\sigma^2_{\mathrm{T}2} = \sigma^2_{\mathrm{C}2} = \sigma^2_2$. Note that the method can be applied to the case of unknown variances. For the fixed sample size designs, Sozu *et al.* [12] discuss a method for the unknown variance case and show that the calculated sample size is nearly equivalent to that for the known variance in the setting of a one-sided significance level $\alpha = 0.025$ and power $1 - \beta = 0.8$ or 0.9. By analogy from the fixed sample designs, there is no practical difference in the group-sequential setting, and the methodology for a known variance provides a reasonable approximation for the unknown variances case.

Let (δ_1, δ_2) denote the differences in the means for the test and the control intervention groups, respectively, where $\delta_k = \mu_{\mathrm{T}k} - \mu_{\mathrm{C}k} (k = 1, 2)$. Suppose that positive values of (δ_1, δ_2) represent the test intervention's benefit. We are interested in conducting a hypothesis test to evaluate if the intervention is superior to the control intervention, that is, the null hypothesis $\mathrm{H}_0 : \delta_1 \leqslant 0$ or $\delta_2 \leqslant 0$ versus the alternative hypothesis $\mathrm{H}_1 : \delta_1 > 0$ and $\delta_2 > 0$. Let (Z_{1l}, Z_{2l}) be the statistics for testing the hypotheses at the lth analysis, given by $Z_{kl} = (\bar{Y}_{\mathrm{T}kl} - \bar{Y}_{\mathrm{C}kl}) / (\sigma_k \sqrt{\kappa_l / n_l})$, where $\kappa_l = (1 + r_l)/r_l$, and $\bar{Y}_{\mathrm{T}kl}$ and $\bar{Y}_{\mathrm{C}kl}$ are the sample means given by $\bar{Y}_{\mathrm{T}kl} = n_l^{-1} \sum_{i=1}^{n_l} Y_{\mathrm{T}ki}$ and $\bar{Y}_{\mathrm{C}kl} = (r_l n_l)^{-1} \sum_{j=1}^{r_l n_l} Y_{\mathrm{C}kj}$. Z_{1l} and Z_{2l} are normally distributed as $\mathrm{N}(\sqrt{n_l/\kappa_l}\delta_1/\sigma_1, 1^2)$ and $\mathrm{N}(\sqrt{n_l/\kappa_l}\delta_2/\sigma_2, 1^2)$, respectively. Thus, (Z_{1l}, Z_{2l}) is bivariate-normally distributed with the correlation $(r_l \rho_{\mathrm{T}} + \rho_{\mathrm{C}})/(1 + r_l)$. Furthermore, the

joint distribution of $(Z_{11}, Z_{21}, \ldots, Z_{1L}, Z_{2L})$ is $2L$ multivariate normal with their correlations given by $\mathrm{corr}[Z_{kl}, Z_{k'l'}] = \sqrt{\kappa_l n_{l'} / \kappa_{l'} n_l}$ if $k = k'$; $\sqrt{\kappa_l n_{l'}}(r_l \rho_T + \rho_C) / \{\sqrt{\kappa_{l'} n_l}(1 + r_l)\}$ if $k \neq k'$.

2.2. *Decision-making framework, stopping rules, and power*

When evaluating the joint effects on both of the endpoints within the context of group-sequential designs, there are the two decision-making frameworks associated with hypothesis testing. One is to reject H_0 if and only if superiority is achieved for the two endpoints simultaneously (i.e., at the same interim timepoint of the trial) (DF-1). The other is to reject H_0 if superiority is achieved for the two endpoints at any interim timepoint (i.e., not necessarily simultaneously) (DF-2). We will discuss the two decision-making frameworks separately as the corresponding stopping rules and power definitions are unique.

DF-1: The DF-1 is relatively simple: if superiority is demonstrated on only one endpoint at an interim, then the trial continues and the hypothesis testing is repeated for both endpoints until the joint significance for the two endpoints is established simultaneously. The stopping rule for DF-1 is formally given as follows:

At the lth analysis $(l = 1, \ldots, L-1)$

If $Z_{1l} > c_{1l}$ and $Z_{2l} > c_{2l}$,	then reject H_0 and stop the trial,
otherwise,	continue to the $(l+1)$th analysis,

at the Lth analysis

if $Z_{1L} > c_{1L}$ and $Z_{2L} > c_{2L}$,	then reject H_0,
otherwise,	do not reject H_0,

where c_{1l} and c_{2l} are the critical values, which are constant and selected separately, using any group-sequential method such as the Lan–DeMets (LD) alpha-spending method [17] to control the overall Type I error rate of α, as if they were a single primary endpoint, ignoring the other co-primary endpoint. The testing procedure for co-primary endpoints is conservative. For example, if a zero correlation between the two endpoints is assumed and each endpoint is tested at the one-sided significance level of 2.5%, then the Type I error rate is 0.0625%. As shown in Section 4, the maximum Type I error rate associated with the rejection region of the null hypothesis increases as the correlation goes toward one, but it is not greater than the targeted significance level.

The power corresponding to DF-1 is

$$1 - \beta = \Pr\left[\bigcup_{l=1}^{L} \{A_{1l} \cap A_{2l}\} \middle| H_1\right], \tag{1}$$

where $A_{kl} = \{Z_{kl} > c_{kl}\} (k = 1, 2;\ l = 1, \ldots, L)$. The power (1) can be numerically assessed by using multivariate normal integrals. A detailed calculation is provided in Appendix A.1.

DF-2: DF-2 is more flexible than DF-1. If superiority is demonstrated on one endpoint at the interim, then the trial will continue, but subsequent hypothesis testing is repeatedly conducted only for the previously non-significant endpoint until superiority is demonstrated. The stopping rule for DF-2 is formally given as follows:

At the lth analysis $(l = 1, \ldots, L-1)$

If $Z_{1l} > c_{1l}$ and $Z_{2l'} > c_{2l'}$ for some $1 \leqslant l' \leqslant l$,	then reject H_0 and stop the trial,
if $Z_{2l} > c_{2l}$ and $Z_{1l'} > c_{1l'}$ for some $1 \leqslant l' \leqslant l$,	then reject H_0 and stop the trial,
otherwise,	continue to the $(l+1)$th analysis,

at the Lth analysis

if $Z_{1L} > c_{1L}$ and $Z_{2l'} > c_{2l'}$ for some $1 \leqslant l' \leqslant L$,	then reject H_0,
if $Z_{2L} > c_{2L}$ and $Z_{1l'} > c_{1l'}$ for some $1 \leqslant l' \leqslant L$,	then reject H_0,
otherwise,	do not reject H_0.

Therefore, following DF-2, the power is

$$1 - \beta = \Pr\left[\left\{\bigcup_{l=1}^{L} A_{1l}\right\} \cap \left\{\bigcup_{l=1}^{L} A_{2l}\right\} \middle| H_1\right]. \tag{2}$$

Similarly as in the power (1), the power (2) can be calculated by using multivariate normal integrals. For the details, please refer to Appendix A.1.

For simplicity, consider a two-stage group-sequential design with one interim and one final analysis. The probability of rejecting the null hypothesis at the interim analysis is the same for DF-1 and DF-2. The difference in power between DF-1 and DF-2 is due to whether or not the null hypothesis is rejected at the final analysis. The difference in decision making for DF-1 and DF-2 comes from the following two situations where the interim analysis result is inconsistent with the final analysis result even the alternative hypothesis is true, that is, (i) endpoint 1 is statistically significant at the interim, but not at the final analysis and similarly, and (ii) endpoint 2 is statistically significant at the interim, but not at the final analysis. Thus, DF-1 fails to reject the null hypothesis in both situations even if the alternative hypothesis is true, but DF-2 is able to reject the null hypothesis in both situations. However, the likelihood of this scenario occurring is quite low. Thus, there is little practical difference in the power and sample size determinations for DF-1 and DF-2. However, DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. Stopping measurement may be desirable if the endpoint is very invasive or expensive although stopping measurement may also introduce an operational difficulty into the trial. This will be illustrated in Section 3.

2.3. Maximum sample size and average sample number

We discuss two sample size concepts, that is, the maximum sample size (MSS) and the average sample number (ASN) based on DF-1 and DF-2, and the corresponding powers (1) and (2) discussed in the previous section.

The MSS is the sample size required for the final analysis to achieve the desired power $1-\beta$. The MSS is given by the smallest integer not less than n_L satisfying the power (1) or (2) for a group-sequential design at the prespecified δ_1, δ_2, ρ_T and ρ_C, with Fisher's information time for the interim analyses, n_l/n_L, $l = 1, \ldots, L$. To find a value of n_L, an iterative procedure is required to numerically solve for the power (1) or (2). This can be accomplished by using a grid search to gradually increase n_L until the power under n_L exceeds the desired power, although this often requires considerable computing resources. To reduce the computational resources, the Newton–Raphson algorithm in [14] or the basic linear interpolation algorithm in [15] may be utilized.

The ASN is the expected sample size under a specific hypothetical reference. Given these prespecifications, the ASN per intervention group for DF-1 is given by

$$\mathrm{ASN} = n_L\left(1 + \sum_{l=1}^{L-1} \Pr\left[\{\bar{A}_{11} \cup \bar{A}_{21}\} \cap \cdots \cap \{\bar{A}_{1l} \cup \bar{A}_{2l}\}\right]\right) \Big/ L. \tag{3}$$

and for DF-2,

$$\mathrm{ASN} = n_L\left(1 + \sum_{l=1}^{L-1} \Pr\left[\{\bar{A}_{11} \cap \cdots \cap \bar{A}_{1l}\} \cup \{\bar{A}_{21} \cap \cdots \cap \bar{A}_{2l}\}\right]\right) \Big/ L, \tag{4}$$

where $r_l = 1$ and $n_l = ln_1$, $l = 1, \ldots, L$. The representations for calculating ASN (3) and (4) are described in Appendix A.2.

The powers, MSS, and ASN will depend on the design parameters including differences between means, the correlation structure between the endpoints, the testing procedure (e.g., O'Brien–Fleming (OF) boundary [18], Pocock (PC) boundary [19]), the number of analyses, and the information time.

3. Evaluation of the sample size

3.1. Behavior of the sample size

In this section, we evaluate the behavior of the power, MSS, and ASN as the design parameters vary. Here, without loss of generality, $\sigma_1^2 = \sigma_2^2 = 1^2$ is chosen for simplicity, so that δ_1 and δ_2 are interpreted as (standardized) effect sizes.

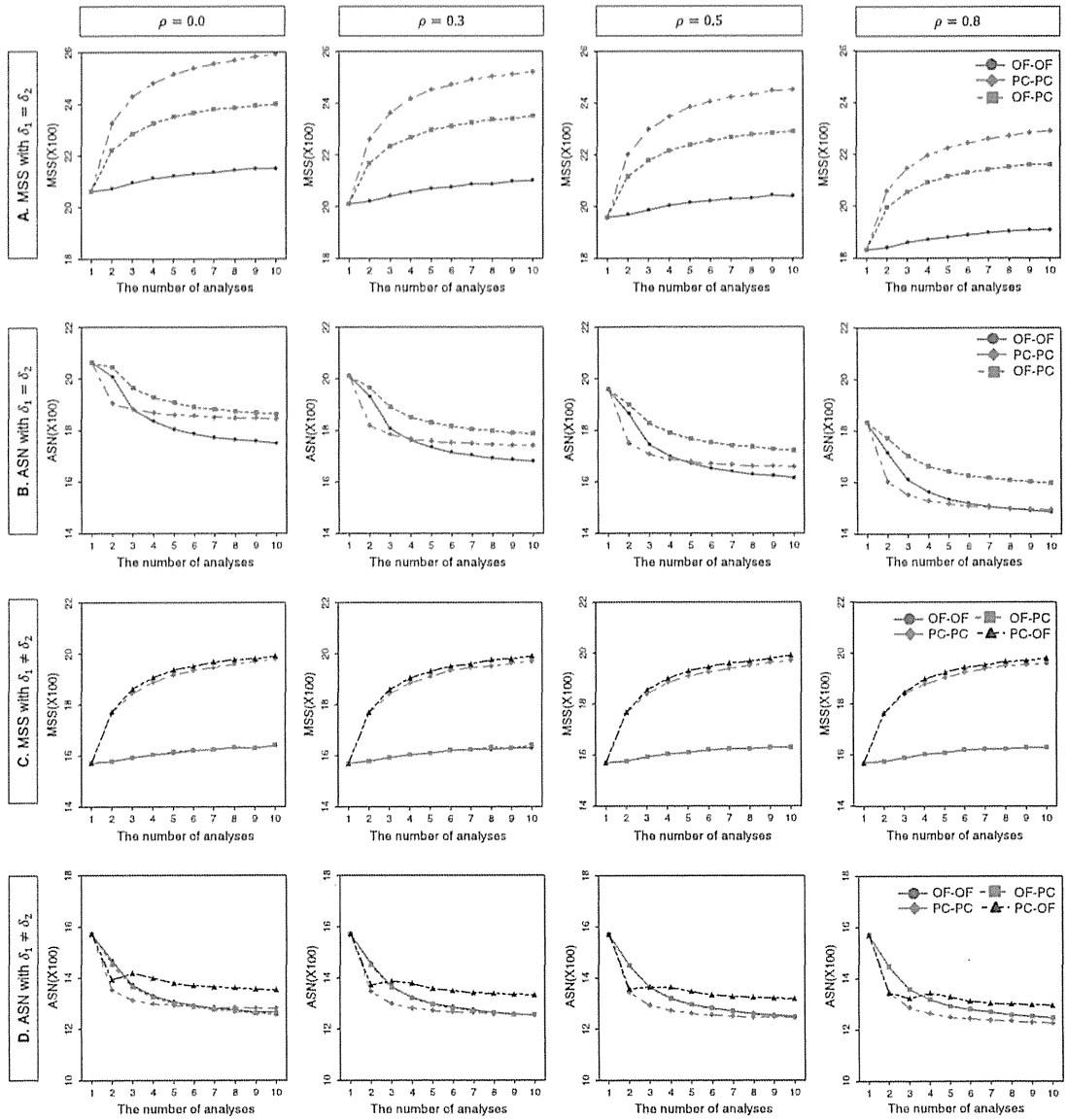

Figure 1. Behavior of MSS and ASN for DF-1 as the number of analyses and boundaries vary. The MSS and ASN per intervention group (equally-sized groups: $r_I = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for A and B, and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for C and D; $\sigma_1^2 = \sigma_2^2 = 1^2$. When differences between means are equal, the critical values are determined by the three boundary combinations, that is, (i) the OF for both endpoints, (ii) the PC for both endpoints, and (iii) the OF for δ_1 and the PC for δ_2, with the LD alpha-spending method with equal information space. When differences between means are unequal, in addition to the three combinations, (iv) the PC for δ_1 and the OF for δ_2 is considered.

Figure 1 illustrates how the MSS and ASN per intervention group for DF-1 behave as a function of the number of analyses and the boundaries when effect sizes are equal and unequal, that is, $\delta_1 = \delta_2$ and $\delta_1 \neq \delta_2$ between the two endpoints. The MSS and ASN for DF-1 and DF-2 (equally-sized groups: $r_I = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for equal effect sizes and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for unequal effect sizes; $\sigma_1^2 = \sigma_2^2 = 1^2$; and $\rho_T = \rho_C = \rho = 0.0, 0.3, 0.5$ and 0.8. The critical values are determined by the three boundary combinations, that is, (i) the OF for both endpoints (OF–OF), (ii) the PC for both endpoints (PC–PC), and (iii) the OF for δ_1 and the PC for δ_2 (OF–PC), with the LD alpha-spending method with equal information space.

When effect sizes are equal, the MSS for the three boundary combinations increases as the number of analyses increases and the correlation is smaller. In all of $\rho = 0, 0.3, 0.5$ and 0.8, the largest MSS is given by PC–PC and the smallest MSS by OF–OF. On the other hand, the ASN for the three boundary combinations decreases as the number of analyses increases and the correlation is larger. In all of $\rho = 0.0, 0.3, 0.5$ and 0.8, the largest ASN is given by OF–PC.

When effect sizes are unequal $\delta_1 < \delta_2$, in addition to the three boundary combinations, one more combination of (iv) the PC for δ_1 and the OF for δ_2 (PC–OF) is considered, $\delta_1 = 0.1$ and $\delta_2 = 0.2$. Similarly as seen with equal effect sizes, the MSS for the four boundary combinations increases as the number of analyses increases, but it does not change as with the correlation varies. The largest MSS is given by PC–PC and PC–OF and the smallest MSS by OF–OF and OF–PC. On the other hand, the ASN for the four boundary combinations decreases as the number of analyses increases independently of the correlation. The largest ASN is given by OF–OF and OF–PC and the smallest ASN by PC–PC and PC–OF. When one effect size is smaller (or larger) than the other, the MSS and ASN will be driven by the smaller effect size. In this illustration, as the OF is selected for the smaller effect size and the PC for the larger, the MSS and ASN by OF–PC are approximately equal to those by OF–OF.

Figure 2 illustrates how the MSS and ASN per intervention group for DF-2 behave as a function of the number of analyses and the boundaries when effect sizes are equal $\delta_1 = \delta_2$ and unequal $\delta_1 \neq \delta_2$ between the two endpoints with the same parameter settings as in Figure 1. The MSS and ASN behaviors are similar to those observed for DF-1. The major difference between DF-1 and DF-2 is that the MSS and ASN for DF-2 are smaller than those for DF-1. They are notably smaller as the number of analyses increases, especially when the correlation is low.

If the trial was designed to detect effects on *at least one* endpoint with a prespecified ordering of endpoints, a choice of different boundaries for each endpoint (i.e., the OF for the primary endpoint and the PC for the secondary endpoint) can provide a higher power than using the same boundary for both endpoints [20, 21]. However, as shown in Figures 1 and 2, the selection of a different boundary has a minimal effect on the power.

3.2. *Example*

We provide an example to illustrate the sample size methods discussed in the previous sections. Consider the clinical trial, 'Effect of Tarenflurbil on Cognitive Decline and Activities of Daily Living in Patients With Mild Alzheimer Disease', a multicenter, randomized, double-blind, placebo-controlled trial in patients with mild Alzheimer disease (AD) [22]. Co-primary endpoints were cognition as assessed by the Alzheimer Disease Assessment Scale Cognitive Subscale (ADAS-Cog; 80-point scale) and functional ability as assessed by the Alzheimer Disease Cooperative Study activities of daily living (ADCS-ADL; 78-point scale). A negative change score from baseline on the ADAS-Cog indicates improvement while a positive change score on the ADCS-ADL indicates improvement. The original sample size per intervention group of 800 patients provided an overall power of 96% to detect the joint difference in the two primary endpoints between the tarenflurbil and placebo groups, by using a one-sided test at 2.5% significance level, with the standardized effect size of 0.2 for both endpoints. In addition, the correlation between the two endpoints was assumed to be zero in the calculation of the sample size although the two endpoints were expected to be correlated (for example, see Doraiswamy *et al.* [23]).

Table I displays the MSS and ASN per intervention group (equally-sized groups: $r_l = 1$) for the DF-1 and DF-2. The sample size was with an alternative hypothesis of a difference for both ADAS-Cog ($\delta_1 = 0.2$) and ADCS-ADL ($\delta_2 = 0.2$), with the overall power of 96% at the one-sided significance level of 2.5%, where $\rho = \rho_T = \rho_C = 0.0, 0.3, 0.5$, and 0.8 and $L = 1, 2, 3, 5, 8$, and 10. The critical values are determined by the three boundary combinations, that is, the OF for both endpoints (OF–OF), the PC for both endpoints (PC–PC), and OF for ADAS-Cog and the PC for ADCS-ADL (OF–PC).

Based on the selected parameters described in [22], that is, $L = 1$ and $\rho = 0.0$, the sample size per intervention group is calculated as 804. If four interims and one final analysis are planned (i.e., $L = 5$) with DF-1, and conservatively assuming a zero correlation between the endpoints, then the MSS is 825 for OF–OF, 945 for PC–PC and 895 for OF–PC, and the ASN is 604 for OF–OF, 548 for PC–PC, and 608 for OF–PC. If the correlation is incorporated into the calculation when $\rho = 0.3, 0.5$, and 0.8, then the MSS are 820, 810, and 785 for OF–OF; 940, 930, and 900 for PC–PC; and 890, 885, and 860 for OF–PC. The ASN are 589, 574, and 543 for OF–OF; 525, 506, and 469 for PC–PC; and 593, 582, and 556 for OF–PC. When comparing DF-2 to DF-1, there are no major differences in MSS and ASN for all of the boundary combinations, although DF-2 provides a slightly smaller MSS and ASN than DF-1, for

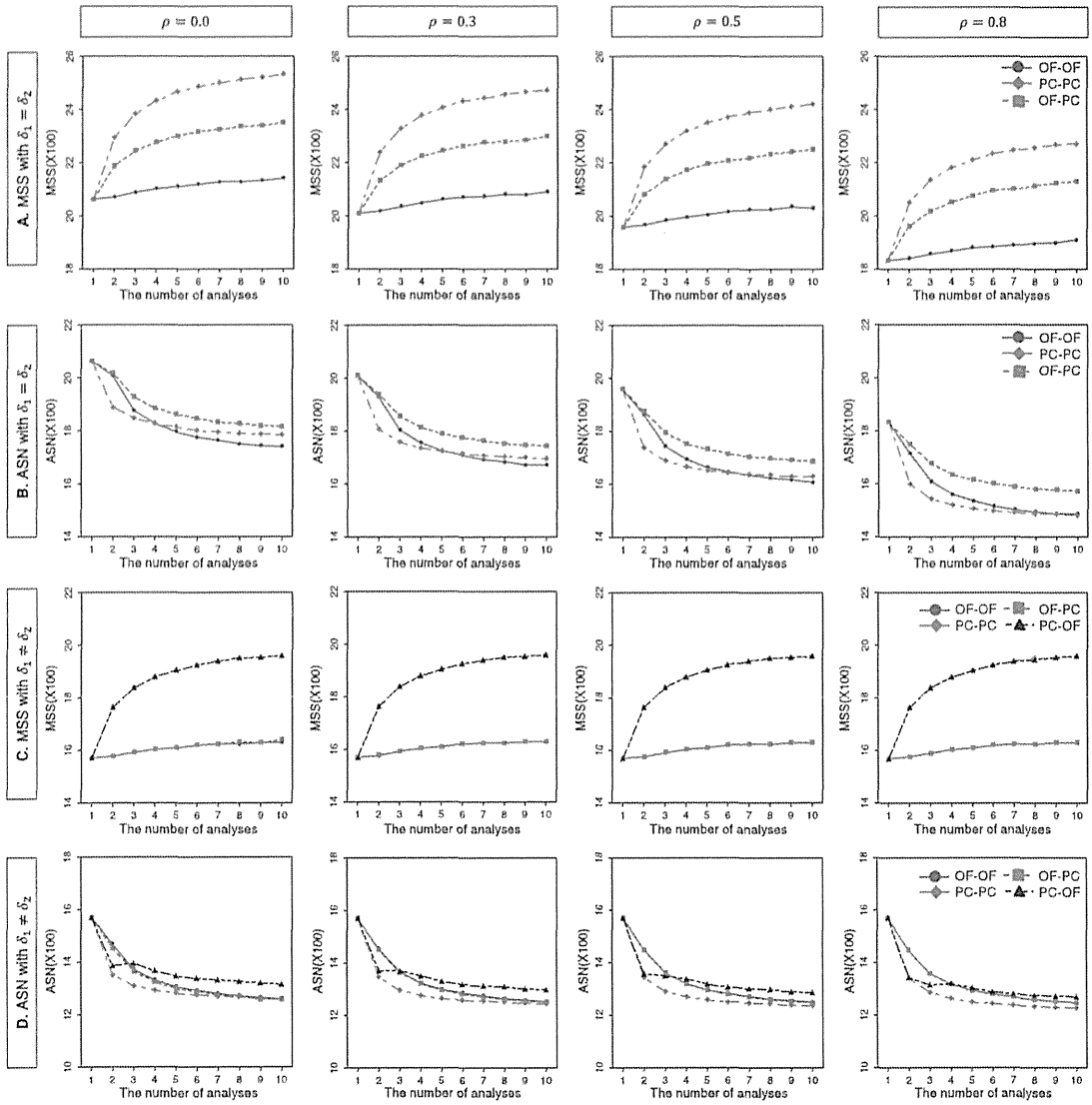

Figure 2. Behavior of MSS and ASN for DF-2 as the number of analyses and boundaries vary. The MSS and ASN per intervention group (equally-sized groups: $r_l = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for A and B and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for C and D; $\sigma_1^2 = \sigma_2^2 = 1^2$. When differences between means are equal, the critical values are determined by the three boundary combinations, that is, (i) the OF for both endpoints, (ii) the PC for both endpoints, and (iii) the OF for δ_1 and the PC for δ_2, with the LD alpha-spending method with equal information space. When differences between means are unequal, in addition to the three combinations, (iv) the PC for δ_1 and the OF for δ_2 is considered.

PC–PC and OF–PC. However, if the endpoint is very invasive and thus stopping measurement may be ethically desirable, there is a benefit of using DF-2 as DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. For example, when four interims and one final analysis with DF-2 are planned (i.e., $L = 5$), the average total number of measurements for each intervention group are 1052, 1045, 1041, and 1021 for OF–OF; 846, 845, 841, and 831 for PC–PC; and 966, 961, 958, and 944 for OF–PC, corresponding to $\rho = 0.0, 0.3, 0.5$, and 0.8. They are smaller than those for DF-1 as the average total number of measurements for DF-1 are 1208, 1178, 1148, and 1086 for OF–OF; 1096, 1050, 1012, and 938 for PC–PC; and 1216, 1186, 1164, and 1112 for OF–PC.

Table I. MSS and ASN per intervention group (equally-sized groups) for detecting the joint difference for ADAS-Cog (0.2) and ADCS-ADL (0.2), with DF-1 and DF-2 and the overall power of 96% at the one-sided significance level of 2.5%.

			(i) OF–OF		(ii) PC–PC		(iii) OF–PC	
Decision-making framework	Correlation	Number of analyses	MSS	ASN	MSS	ASN	MSS	ASN
DF-1	0.0	1	804	804	804	804	804	804
		2	808	725	886	607	854	693
		3	816	647	918	572	876	652
		5	825	604	945	548	895	608
		8	832	579	968	535	912	587
		10	840	573	970	530	920	581
	0.3	1	799	799	799	799	799	799
		2	802	702	880	593	850	676
		3	810	633	912	552	870	638
		5	820	589	940	525	890	593
		8	824	563	960	511	904	571
		10	830	556	970	507	910	564
	0.5	1	791	791	791	791	791	791
		2	794	684	872	580	842	662
		3	801	620	903	536	864	627
		5	810	574	930	506	885	582
		8	816	549	952	492	896	558
		10	820	542	960	488	900	551
	0.8	1	764	764	764	764	764	764
		2	768	644	842	549	818	635
		3	774	588	873	501	840	603
		5	785	543	900	469	860	556
		8	792	520	920	453	872	533
		10	800	514	920	447	880	527
DF-2	0.0	1	804	804	804	804	804	804
		2	808	725	882	605	848	690
		3	813	645	912	569	867	646
		5	825	603	940	540	890	602
		8	832	578	960	524	904	579
		10	830	568	960	518	910	572
	0.3	1	799	799	799	799	799	799
		2	802	702	876	591	842	672
		3	807	632	906	549	861	632
		5	815	586	935	520	880	586
		8	824	562	952	503	896	564
		10	830	555	960	498	900	556
	0.5	1	791	791	791	791	791	791
		2	794	684	868	579	834	658
		3	801	620	897	533	855	621
		5	810	574	925	502	875	575
		8	816	549	944	486	888	552
		10	820	541	950	481	890	544
	0.8	1	764	764	764	764	764	764
		2	768	644	840	549	810	631
		3	774	588	870	499	831	597
		5	785	543	895	467	850	550
		8	792	520	912	450	864	528
		10	790	510	920	445	870	521

4. Sample size recalculation

Clinical trials are designed based on assumptions often constructed based on prior data. However, prior data may be limited or an inaccurate indication of future data, resulting in trials that are over/under-powered. Interim analyses provide an opportunity to evaluate the accuracy of the design assumptions and potentially make design adjustments (i.e., to the sample size) if the assumptions were markedly inaccurate. The tarenflurbil trial mentioned in the previous section, failed to demonstrate a beneficial effect of tarenflurbil on both ADAS-Cog and ADCS-ADL. The observed treatment effects were smaller than the assumed effects. Group-sequential designs allow for early stopping when there is sufficient statistical evidence that the two treatments are different. However, more modern adaptive designs may also allow for increases in the sample size if effects are smaller than assumed. Such adjustments must be conducted carefully for several reasons. Challenges include the following: (i) maintaining control of statistical error rates, (ii) developing a plan to make sure that treatment effects cannot be inferred via back-calculation of a resulting change in the sample size, (iii) consideration of the clinical relevance of the treatment effects, and (iv) practical concerns such as an increase in cost and the challenge of accruing more trial participants. In this section, we discuss sample size recalculation based on the observed intervention's effects at an interim analysis with a focus on control of statistical error rates.

4.1. Test statistics and conditional power

Consider that the maximum sample size is recalculated to n'_L based on the interim data at the Rth analysis. Suppose that n'_L is subject to $n_R < n'_L \leqslant \lambda n_L$, where λ is a prespecified constant for the maximum allowable sample size. For simplicity, assume a common correlation between the treatment groups, that is, $\rho_T = \rho_C = \rho$. Let $(\tilde{\delta}_1, \tilde{\delta}_2)$ and let (δ_1^*, δ_2^*) be the mean differences used for planned sample size and for recalculated sample size, respectively.

Here, we consider the Cui–Hung–Wang (CHW) statistics [24] for sample size recalculation in group-sequential designs with two co-primary endpoints to preserve the overall Type I error rate at a prespecified alpha level even when the sample size is increased and conventional test statistics are used. The CHW statistics are

$$Z'_{km} = \sqrt{\frac{n_R}{n_m}} Z_{kR} + \sqrt{\frac{n_m - n_R}{n_m}} \frac{\sum_{i=n_R+1}^{n'_m} Y_{Tki} - \sum_{j=n_R+1}^{n'_m} Y_{Ckj}}{\sqrt{2\left(n'_m - n_R\right)}},$$

where $n'_m = (n_m - n_R)(n'_L - n_R)/(n_L - n_R) + n_R$ and $r_R = r_m = 1 (k = 1, 2; \quad R = 1, \ldots, L-1; m = R+1, \ldots, L)$. The same critical values utilized for the case without sample size recalculation are used.

The sample size is increased or decreased when the conditional power evaluated at the Rth analysis is lower or higher than the desired power $1 - \beta$. Under the planned maximum sample size and a given observed value of (Z_{1R}, Z_{2R}), for DF-1, the conditional power is defined by

$$CP = \Pr\left[\bigcup_{m=R+1}^{L} \{A_{1m} \cap A_{2m}\} \,|a_{1R}, a_{2R}\right] \tag{5}$$

if $Z_{1l} \leqslant c_{1l}$ or $Z_{2l} \leqslant c_{2l}$ for all $l = 1, \ldots, R$, where (a_{1R}, a_{2R}) is a given observed value of (Z_{1R}, Z_{2R}). On the other hand, the conditional power for DF-2 is given by

$$CP = \begin{cases} \Pr\left[\bigcup_{m=R+1}^{L} A_{1m} \,|a_{1R}, a_{2l'}\right] \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{2l'} > c_{2l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[\bigcup_{m=R+1}^{L} A_{2m} \,|a_{2R}, a_{1l'}\right] \\ \quad \text{if } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R \text{ and } Z_{1l'} > c_{1l'} \text{ for some } l' = 1, \ldots, R, \\ \Pr\left[\left\{\bigcup_{m=R+1}^{L} A_{1m}\right\} \cap \left\{\bigcup_{m=R+1}^{L} A_{2m}\right\} \,|a_{1R}, a_{2R}\right] \\ \quad \text{if } Z_{1l} \leqslant c_{1l} \text{ and } Z_{2l} \leqslant c_{2l} \text{ for all } l = 1, \ldots, R. \end{cases} \tag{6}$$

The detailed calculation of the conditional powers for DF-1 and DF-2 are provided in Appendix A.3. Because (δ_1, δ_2) is unknown, it is customary to substitute (δ_1^*, δ_2^*), the estimated mean differences at