The Spine Journal 19 (2019) 65–70

Clinical Study

Are patient-reported outcome measures biased by method of follow-up? Evaluating paper-based and digital follow-up after lumbar fusion surgery

Marc L. Schröder, MD, PhD[a], Marlies P. de Wispelaere, MSc[b], Victor E. Staartjes, BMed[a,c,*]

[a]*Department of Neurosurgery, Bergman Clinics, Rijksweg 69, 1411 GE Naarden, The Netherlands*
[b]*Department of Clinical Informatics, Bergman Clinics, Gooimeer 11, 1411 GE Naarden, The Netherlands*
[c]*Faculty of Medicine, University of Zurich, Pestalozzistrasse 3, 8091, Zurich, Switzerland*

Received 6 February 2018; revised 1 May 2018; accepted 1 May 2018

Abstract

BACKGROUND CONTEXT: Long-term follow-up of patient-reported outcome measures (PROM) is essential in both modern spinal care and research. Lack of time and staff are commonly reported barriers to implementing long-term follow-up of PROM. Automated and digital follow-up systems for PROM collection are seeing widespread use, yet their validity and comparative effectiveness have never been evaluated.
PURPOSE: The present study aimed to assess the validity of digital follow-up systems in comparison with the conventional paper-based follow-up (PB-FU).
STUDY DESIGN: This is a retrospective analysis of prospectively collected double follow-up data.
PATIENT SAMPLE: Patients who underwent lumbar spinal fusion for spondylolisthesis or degenerative disc disease between 2013 and 2016 were included in the study.
OUTCOME MEASURES: The study determined the Oswestry Disability Index (ODI) and Numeric Rating Scale (NRS) for back and leg pain severity at baseline, 6 weeks, 12 months, and 24 months.
MATERIALS AND METHODS: After lumbar spinal fusion surgery, a double follow-up of PROM was carried out by conventional PB-FU during clinical visits, while simultaneously completing an automatically dispatched digital follow-up questionnaire. As the primary end point, we assessed the intraindividual discrepancy in PROM between PB-FU and automated digital follow-up (AD-FU).
RESULTS: Forty patients completed all parts of the dual follow-up trajectory and were analyzed. We detected no discrepancy in ODI or NRS for back and leg pain severity at any of the baseline, 6-week, 12-month, or 24 month follow-ups (all $p>.05$). This was confirmed in a sensitivity analysis.
CONCLUSIONS: In an analysis of dual paper-based and digital follow-up after lumbar fusion surgery, patients report highly similar values using either method of follow-up. It appears that AD-FU without incentives produces lower response rates. To reassess the validity of these systems for data collection in spinal patient care, a prospective validation with higher statistical power is warranted.

Keywords: Digital; Follow-up; Outcome measurement; Patient-reported outcome measure; Spinal fusion; Spine

FDA device/drug status: Not applicable.

Author disclosures: ***MLS***: Speaking and/or Teaching Arrangements: Mazor Technologies, Ltd (A), outside the submitted work. ***MPdW:*** Nothing to disclose. ***VES:*** Nothing to disclose.

The disclosure key can be found on the Table of Contents and at www.TheSpineJournalOnline.com.

This research has never previously been submitted for review or presented at any conferences.

MLS has received travel compensation for presentations from Mazor Robotics, Ltd, in the past. The other authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author. Bergman Clinics, Rijksweg 69, Naarden 1411 GE, The Netherlands. Tel.: 0031 88 900 0500; fax: 0031 88 900 0568.

E-mail address: victor.staartjes@gmail.com (V.E. Staartjes)

https://doi.org/10.1016/j.spinee.2018.05.002
1529-9430/

Introduction

Repeated follow-up is an essential part in the modern-day treatment of degenerative spinal diseases. This applies to both clinical practice and research, which depend on the continued completion of standardized measurement tools that assess pain severity, functional impairment, and health-related quality of life. For this reason, a large number of national and institutional databases have been initiated, all of which heavily rely on self-reported outcome measures [1].

Lumbar fusion surgery, particularly for degenerative diseases, requires the short- and long-term monitoring of patients, and it has even been called into question if 12 months of follow-up is adequate in this patient population [2]. With longer follow-up time spans and repeated clinical visits required, it is becoming increasingly difficult to monitor an increasing number of patients in a time- and cost-effective, controlled way [3]. One way of tackling these difficulties would be to implement an automated digital follow-up (AD-FU) system. Although such systems have already become standard practice in some centers, there are no studies that examine and validate their methodological accuracy. It is conceivable that patients may report different outcomes in an anonymous digital questionnaire as opposed to a clinical setting. Moreover, to the best of the authors' knowledge, as of yet, no study has performed a dual follow-up that compares traditional and digital follow-up.

We aim to determine if there are any short- and long-term differences between conventional paper-based follow-up (PB-FU) and AD-FU of measures of pain severity and functional impairment in patients undergoing lumbar fusion surgery.

Materials and methods

Study design

From a prospectively collected cohort of 429 lumbar fusion procedures, all patients who had completed a full baseline, 6-week, 12-month, and 24-month assessment by both PB-FU and AD-FU were retrospectively analyzed. Standardized patient-reported outcome measures (PROM) were employed. We compare the data collected using PB-FU and AD-FU. The present study was approved by the local institutional review board (Medical Research Ethics Committees United, Registration Number: W16.065) and was conducted according to the Declaration of Helsinki. Informed consent was obtained from all individual participants.

Patient population

Indications for surgery were degenerative spondylolisthesis and degenerative disc disease (DDD). All patients were operated on at a single specialized spine surgery center using minimally invasive or mini-open techniques between 2013 and 2016. Depending on their clinical history and demographics, patients were treated with minimally invasive robot-guided transforaminal or mini-open posterior lumbar interbody fusion, anterior lumbar interbody fusion (ALIF), or minimally invasive transaxial lumbar interbody fusion (AxiaLIF) as described in detail before [4].

Paper-based follow-up

During clinical baseline and follow-up visits at all time points, patients filled in a standardized questionnaire that contained social and demographic queries, as well as the Dutch version of the Oswestry Disability Index (ODI) and Numeric Rating Scales (NRS) ranging from 0 to 10, separately for back and leg pain [5]. Patients completed the questionnaires in a room separate from the clinician to avoid a potential Hawthorne effect, and any questions regarding the questionnaire were answered [6].

Automated digital follow-up

Before the first visit, patients received an invitation to complete an online baseline questionnaire on a specific web-based application designed by the Department of Clinical Informatics of our institution. At 6 weeks, 12 months, and 24 months after the day of surgery, scheduled follow-up questionnaires were automatically sent out to patients by email and completed in the same fashion. The AD-FU questionnaires contained the same questions as on paper. Both PB-FU and AD-FU were usually completed within weeks of each other.

Statistics

Continuous data are reported as mean±standard deviation, and categorical data as numbers (percentages). Intrasubject differences in PROM were assessed using Wilcoxon signed-rank test. A Benjamini-Hochberg correction for multiple testing was applied to control the false discovery rate while retaining power [7]. All analyses were carried out using version 3.4.1 of R (The R Foundation for Statistical Computing, Vienna, Austria) [8]. A two-tailed $p<.05$ was considered significant.

Results

Patients

The flow of patients throughout this analysis is reported in Fig. 1. Of 429 patients in the database, 40 (9%) completed a full dual PB-FU and AD-FU at all four time points. Detailed patient characteristics are given in Table 1. Most patients (58%) were operated for DDD, and L5-S1 was the index level in the majority (77%) of surgical procedures. Surgical data, complications, and reoperations during the follow-up period are reported in Table 2. Most patients underwent MI-PLIF (38%), followed by AxiaLIF (28%), minimally invasive robot-guided transforaminal (20%), and ALIF (15%). The only

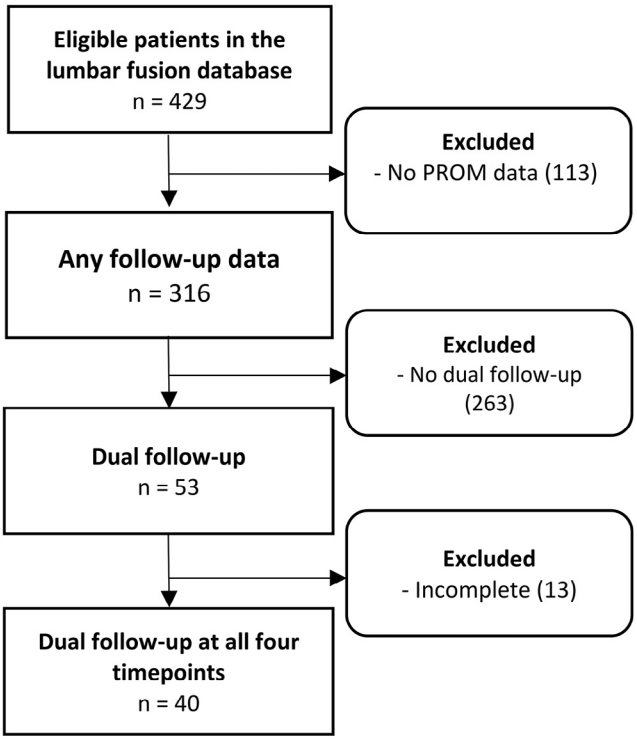

Fig. 1. Flowchart demonstrating the selection of patients for this analysis. PROM, patient-reported outcome measure.

complications encountered were 2 (5%) cases of iatrogenic extensor paresis that both recovered spontaneously throughout the first year of follow-up and 1 (3%) case of excessive intraoperative blood loss. Two (5%) patients had to undergo additional posterior fixation at the index level for DDD that developed due to non-fusion after an AxiaLIF procedure (Table 3).

Table 1
Summary of demographic and clinical data

Characteristic	Value
Male gender	17 (43)
Age (y)	48.7±10.6
Height (cm)	175.7±9.7
Weight (kg)	76.2±14.2
BMI (kg/m^2)	24.5±2.9
History of pain (mo)	20.3±27.2
Fully able to work	11 (28)
Smoking status	
Active	13 (32)
Ceased smoking	15 (38)
Never smoked	12 (30)
Indication	
Degenerative disc disease	23 (58)
Spondylolisthesis, Grade 1	14 (35)
Spondylolisthesis, Grade 2	3 (8)
Index level	
L4-L5	9 (23)
L5-S1	31 (77)

BMI, body mass index.

Table 2
Surgical data, complications, and reoperations during the follow-up period

Characteristic	Value
Surgical time (min)	120.8±64.1
Length of stay (d)	1.9±0.8
Estimated blood loss (mL)	258.8±325.0
Surgical technique	
MI-TLIF	8 (20)
MI-PLIF	15 (38)
ALIF	6 (15)
AxiaLIF	11 (28)
Complications	
Excessive blood loss	1 (3)
Transient extensor paresis	2 (5)
Reoperations	
DDD at index level after AxiaLIF	2 (5)

MI-TLIF, minimally invasive transforaminal lumbar interbody fusion; MI-PLIF, minimally invasive posterior lumbar interbody fusion; ALIF, anterior lumbar interbody fusion; AxiaLIF, transaxial lumbar interbody fusion; DDD, degenerative disc disease.

Patient-reported outcome measures

No significant discrepancy between PB-FU and AD-FU was detected at baseline, 6 weeks, 12 months, or 24 months (all p>.05). This was true for ODI, NRS for back pain severity, and NRS for leg pain severity. Although modest, the greatest discrepancies were seen at the 6-week FU. A sensitivity analysis without correction of p-values for the false discovery rate was carried out. Even without correction, the intraindividual differences between PB-FU and AD-FU remained p>.2 in all analyses. Fig. 2 demonstrates the evolution of PROM throughout the 2-year follow-up period, as well as

Table 3
Paper-based and automated digital baseline and follow-up values of patient-reported outcome measures

Data	Paper-based follow-up	Automated digital follow-up	p
Baseline			
ODI	44.65±15.98	49.70±18.94	.633
NRS-BP	7.38±2.00	7.00±2.34	.879
NRS-LP	5.23±3.33	6.15±2.64	.420
6 wk			
ODI	24.72±15.20	17.05±13.85	.420
NRS-BP	3.94±2.58	2.67±1.65	.202
NRS-LP	2.51±2.86	1.67±1.59	.632
12 mo			
ODI	14.77±14.92	16.48±16.82	.954
NRS-BP	2.93±2.65	3.06±2.87	.999
NRS-LP	2.03±2.61	1.64±2.32	.879
24 mo			
ODI	16.82±19.09	17.48±15.07	.420
NRS-BP	3.72±3.08	2.78±2.28	.999
NRS-LP	2.90±3.12	2.65±3.01	.785

ODI, Oswestry Disability Index; NRS-BP, Numeric Rating Scale for back pain severity; NRS-LP, Numeric Rating Scale for leg pain severity.

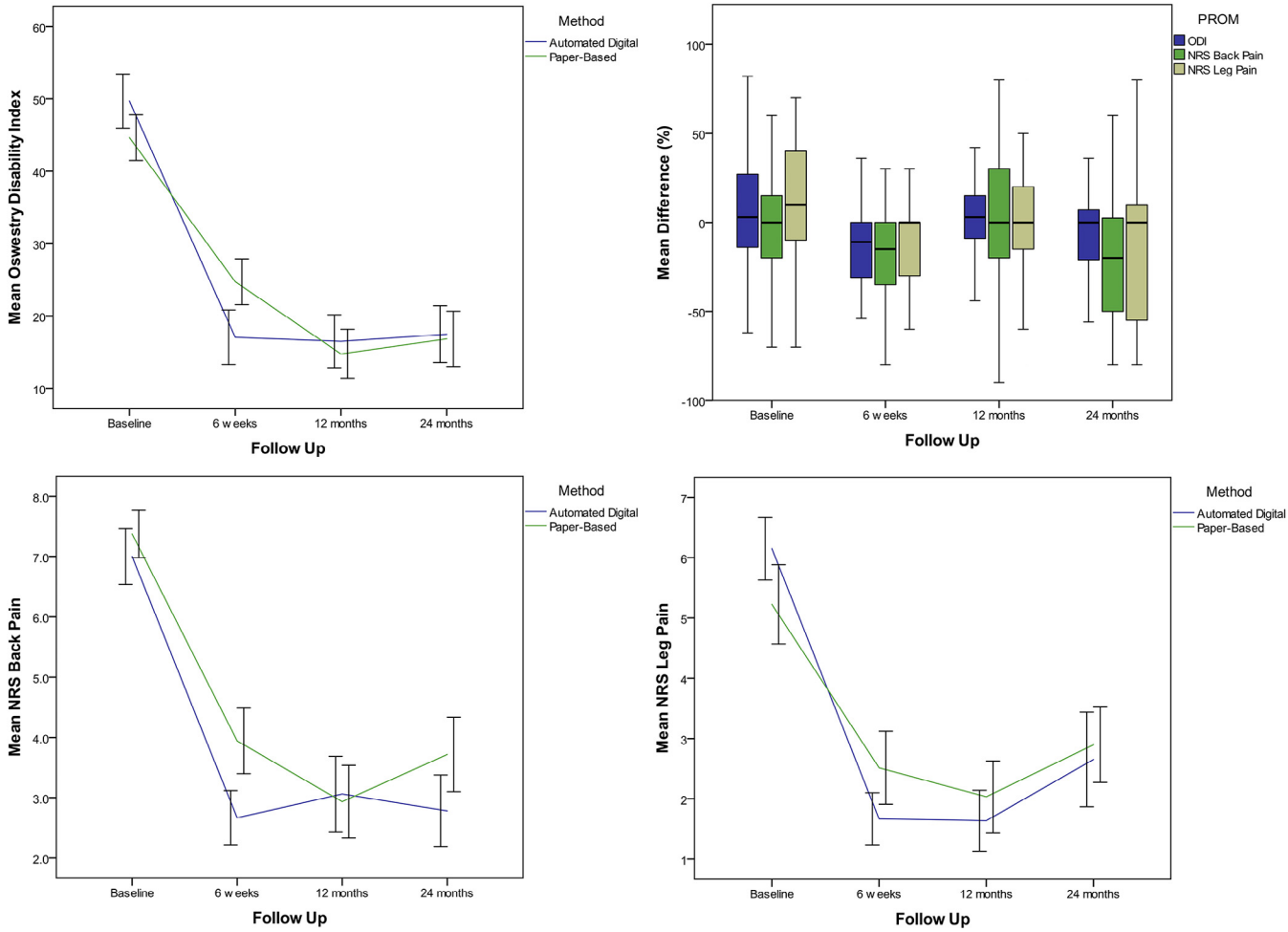

Fig. 2. Graphic representation of patient-reported outcome measures during the follow-up period. Error bars represent the standard error of the mean. The boxplots demonstrate the percentagewise differences between paper-based and automated digital follow-up. PROM, patient-reported outcome measure; ODI, Oswestry Disability Index; NRS, Numeric Rating Scale.

the mean percentagewise differences between PB-FU and AD-FU.

Discussion

Forty patients were followed up for 24 months after lumbar spinal fusion surgery. A double follow-up of PROM was carried out by conventional PB-FU during clinical visits, while simultaneously completing an automatically dispatched digital follow-up questionnaire. We detected no discrepancy in ODI and NRS for back and leg pain severity at any of the baseline, 6-week, 12-month, or 24-month follow-ups, suggesting that an AD-FU system represents a valid alternative to conventional PB-FU.

Although today's PROM are certainly not perfect measures of treatment success, they narrow the gap between the subjective perception of symptoms of patients and clinicians [3,9]. A recent worldwide survey reports that 32% of spine surgeons do not routinely use PROM to monitor their patients' outcome and identifies lack of time to administer the questionnaires, the time to complete them, and lack of staff as the main barriers to implementing PROM [3]. Short global assessments have unfortunately not proven reliable for clinical follow-up [10,11]. An AD-FU system, if properly validated and implemented, could mitigate most of these barriers and potentially lead to a streamlined and effective follow-up system.

Our data support the use of digital follow-up as an adjunct to normal clinical follow-up visits. Overall, there was only a slight, but not statistically or clinically relevant tendency toward lower scores in the digital follow-up. This refutes the hypothesis that patients may be more or less lenient, or in some other way biased, when filling in questionnaires during clinical follow-up visits as opposed to a domestic setting [6]. Generally, measures of pain and functional disability as measured by AD-FU appeared to be marginally higher for baseline, whereas they appeared marginally lower during follow-up in comparison with PB-FU, as Fig. 2 demonstrates.

There are obvious drawbacks to an automated follow-up system. Most importantly, patients are less likely to complete

all questionnaires than they would be if called in for a clinical follow-up visit. Overall, depending on the indication for surgery, we achieve an overall response rate of 40% at our institution using the automated follow-up, whereas 100% of patients who are handed a paper-based questionnaire during a clinical visit complete the entire form. Although the differences among paper-based and digital follow-up may play a role, the most important explanation is that we did not use mailed paper-based questionnaires, but instead had patients fill in paper-based questionnaires as part of the clinical routine. Nonetheless, a large analysis by Ebert et al. [12] also found that digital response rates are generally lower than for paper-based mailed questionnaires, but that digital follow-up led to a lower number of missing values and was more cost-effective. They conclude that, because non-respondents did not differ in socioeconomic variables, the lower response rate in the digital group does not imply an increased level of selection bias. It may be conceivable that the response rate would be higher if patients were explicitly asked by the treating physician to complete the digital follow-up, or if they were enrolled in a clinical trial. An extensive meta-analysis by Edwards et al. [13] demonstrated that contacting patients about the questionnaires before sending them out led to an increased response rate. The same was true for monetary incentives, and for renewed contact regarding mailed follow-up questionnaires. It must also be considered that patients highly value the opportunity to return to their physician after surgery, and to report their current situation in person as opposed to an anonymous automated survey. Overall, it appears that AD-FU without incentives produces lower response rates.

There is some evidence that loss of follow-up influences health-care data [14]. Nonetheless, recent data from the DaneSpine Registry suggest that loss of follow-up does not necessarily lead to a bias in PROM. In a methodologically sound analysis of 506 patients, Højmark et al. [15] found that the discrepancy between responders and non-responders in PROM is negligible. Solberg et al. [16] came to the same conclusion. However, their registry sported an unusually low rate of loss of follow-up of 12%, which limits the generalizability of their findings to other cohorts with greater amounts of non-responders. For example, an analysis of 13 large prospective spine registries revealed loss of follow-up rates ranging from 78% to 21% [1].

The power of our analysis lies within the study design. We were able to assess PB-FU and AD-FU in the same individuals by carrying out a double follow-up using both methods at the same time. Consequently, we were able to apply paired statistical tests. This eliminates multiple sources of bias, and in theory prevents the risk of "comparing apples and pears." Other retrospective designs must rely on large sample sizes or exact matching to minimize these sources of bias.

The methodological evidence on the various methods of follow-up in the peer-reviewed literature is still scarce. With the growing importance of PROM in clinical practice and research, it is crucial that we know exactly what the potential sources of bias are when implementing different methods of follow-up. The findings of the present study warrant prospective examination in a larger cohort. Although our analysis can be seen as a proof of concept, validation in larger prospective studies must be carried out before AD-FU systems can reliably be applied to streamline long-term data collection in prospective registries. Digital follow-up must also be validated in conservatively treated patients and healthy individuals, as long-term observational studies could profit from implementing automated online data collection systems.

Limitations

Although we employed secure statistical methods, this analysis is limited first and foremost by sample size. We identified all patients with a complete dual follow-up from a large prospective registry, which still resulted in a comparatively low follow-up rate of 9%. Consequently, the sample size for our statistical analysis was low, which only allows us to make limited claims as to the similarity of outcomes between the two follow-up methods. The high dropout rate could have altered the effect size of surgical treatment itself caused by selection bias. However, this bias is mitigated by the fact that the present study does not look at the outcomes proper, but instead is concerned with comparing two methods of collecting outcomes. Nonetheless, this limitation does limit the generalizability of our findings. This registry included lumbar fusion procedures that were carried out using four different surgical techniques, which might have further biased our findings. Furthermore, all data stem from a single center, limiting the generalizability of our findings to other centers and countries with varying demographics. Lastly, it is conceivable that, for various reasons, patients may have tried to match the two values that they specified on PB-FU and AD-FU, which would further confound the findings.

Conclusions

In an analysis of dual paper-based and digital follow-up after lumbar fusion surgery, patients report highly similar values using either method of follow-up. It appears that AD-FU without incentives produces lower response rates. To reassess the validity of these systems for data collection in spinal patient care, a prospective validation with higher statistical power is warranted.

Acknowledgments

We cordially thank Femke Beusekamp, BSc, and Nathalie Schouman for their assistance in maintaining the database.

References

[1] McGirt MJ, Parker SL, Asher AL, Norvell D, Sherry N, Devin CJ. Role of prospective registries in defining the value and effectiveness of spine care. Spine 2014;39:S117. doi:10.1097/BRS.0000000000000552.

[2] Adogwa O, Elsamadicy AA, Han JL, Cheng J, Karikari I, Bagley CA. Do measures of surgical effectiveness at 1 year after lumbar spine surgery accurately predict 2-year outcomes? J Neurosurg Spine 2016;25:689–96. doi:10.3171/2015.8.SPINE15476.

[3] Falavigna A, Dozza DC, Teles AR, Wong CC, Barbagallo G, Brodke D, et al. Current status of worldwide use of patient-reported outcome measures (PROMs) in spine care. World Neurosurg 2017;108:328–35. doi:10.1016/j.wneu.2017.09.002.

[4] Staartjes VE, Vergroesen P-PA, Zeilstra DJ, Schröder ML. Identifying subsets of patients with single-level degenerative disc disease for lumbar fusion: the value of prognostic tests in surgical decision making. Spine J 2018;18:558–566. doi:10.1016/j.spinee.2017.08.242.

[5] van Hooff ML, Spruit M, Fairbank JCT, van Limbeek J, Jacobs WCH. The Oswestry Disability Index (version 2.1a): validation of a Dutch language version. Spine 2015;40:E83–90. doi:10.1097/BRS.0000000000000683.

[6] Adair JG. The Hawthorne effect: a reconsideration of the methodological artifact. J Appl Psychol 1984;69:334–45. doi:10.1037/0021-9010.69.2.334.

[7] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57:289–300. doi:10.2307/2346101.

[8] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

[9] Scheer JK, Keefe M, Lafage V, Kelly MP, Bess S, Burton DC, et al. Importance of patient-reported individualized goals when assessing outcomes for adult spinal deformity (ASD): initial experience with a Patient Generated Index (PGI). Spine J 2017;17:1397–405. doi:10.1016/j.spinee.2017.04.013.

[10] Godil SS, Parker SL, Zuckerman SL, Mendenhall SK, Devin CJ, Asher AL, et al. Determining the quality and effectiveness of surgical spine care: patient satisfaction is not a valid proxy. Spine J 2013;13:1006–12. doi:10.1016/j.spinee.2013.04.008.

[11] Parai C, Hägg O, Lind B, Brisby H. The value of patient global assessment in lumbar spine surgery: an evaluation based on more than 90,000 patients. Eur Spine J 2017;1–10. doi:10.1007/s00586-017-5331-0.

[12] Ebert JF, Huibers L, Christensen B, Christensen MB. Paper- or web-based questionnaire invitations as a method for data collection: cross-sectional comparative study of differences in response rate, completeness of data, and financial cost. J Med Internet Res 2018;20:e24. doi:10.2196/jmir.8353.

[13] Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, et al. Increasing response rates to postal questionnaires: systematic review. BMJ 2002;324:1183.

[14] Etter JF, Perneger TV. Analysis of non-response bias in a mailed health survey. J Clin Epidemiol 1997;50:1123–8.

[15] Højmark K, Støttrup C, Carreon L, Andersen MO. Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. Eur Spine J 2016;25:282–6. doi:10.1007/s00586-015-4127-3.

[16] Solberg TK, Sørlie A, Sjaavik K, Nygaard ØP, Ingebrigtsen T. Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? Acta Orthop 2011;82:56–63. doi:10.3109/17453674.2010.548024.