ORIGINAL RESEARCH

Get the most from your data: a propensity score model comparison on real-life data

Dennis Ferdinand[1]
Mirko Otto[2]
Christel Weiss[1]

[1]Department of Biomathematics and Medical Statistics, [2]Department of Surgery, University Medical Center Mannheim (UMM), University of Heidelberg, Mannheim, Germany

Purpose: In the past, the propensity score has been in the middle of several discussions in terms of its abilities and limitations. With a comprehensive review and a practical example, this study examines the effect of propensity score analysis of real-life data and introduces a simple and effective clinical approach.

Materials and methods: After the authors reviewed current publications, they applied their insights to the data of a nonrandomized clinical trial in bariatric surgery. This study examined weight loss in 173 patients where 127 patients received Roux-en-Y gastric bypass surgery and 46 patients sleeve gastrectomy. Both groups underwent analysis in terms of their covariate distribution using Mann–Whitney U and χ^2 testing. Mean differences within excess weight loss in native data were examined with Student's t-test. Three propensity score models were defined and matching was performed. Covariate distribution and mean differences in excess weight loss were checked with Mann–Whitney U and χ^2 testing.

Results: Native data implied a significant difference in excess weight loss. The propensity score models did not confirm this difference. All models proved that both surgical procedures were equal, due to their weight-loss induction. Covariate distribution improved after the matching procedure in terms of an equal distribution.

Conclusion: It seemed that a practical clinical approach with outcome-related covariates as a propensity score base is the ideal midpoint between an equal distribution in covariates and an acceptable loss of data. Nevertheless, propensity score models designed with clinical intent seemed to be absolutely suitable for overcoming heterogeneity in covariate distribution.

Keywords: nonrandomized clinical trial, statistics, logistic regression, study design, balancing score

Introduction

Propensity score models are quite popular in modern clinical research practice. While the randomized clinical trial (RCT) is the method of choice when it comes to clinical studies, it might occur that they are not suitable or applicable for various reasons, eg, financing or ethics. In this case, one might perform a non-RCT to test a hypothesis.[1] Unfortunately, this study design can lead to heterogeneity in covariate distribution between two therapy groups, which makes it difficult to compare between the outcomes of therapy A versus B. Propensity score models are a simple tool to solve this problem and generate unbiased data.[2]

In 1983, Rosenbaum and Rubin introduced the propensity score as a new balancing score.[2] The propensity score unifies all known covariates from a patient in one score

Correspondence: Dennis Ferdinand
Department of Biomathematics and Medical Statistics, Universitätsmedizin Mannheim (UMM), 13–17 Ludolf-Krehl-Strasse, Mannheim 68167, Germany
Email d.ferdinand@stud.uni-heidelberg.de

value between 0 and 1. Calculated by logistic regression, the propensity score is a patient's likelihood of receiving a specific therapy, considering his or her covariates. Covariates in this context are actually all the characteristics a patient has, like blood type, weight, height, age, and sex, and also nonmedical aspects, such as education and leisure activities.[2,3] With the knowledge of all patients' propensity scores, there are various ways to use them in further analysis. Common applications are nearest-neighbor matching and stratification.[4] Nearest-neighbor matching, for instance, compares two individuals from separate groups with the most similar propensity score values in terms of their outcome. Patients without a match are dropped.

Since its introduction, there has been great discussion on which covariates should be considered in propensity score calculation, the calculation methods, the abilities and quality of propensity score results, and many other aspects of this tool. Stürmer et al[5] argued that there is no factual evidence for superiority of the propensity score compared to traditional multivariate-outcome models. Hahn[6] even claimed that the propensity score reduces the efficiency of outcome examinations. Bryson et al[7] reduced the role of the propensity score to a simple average treatment-effect analysis method without further advantages. On the other hand, Cattaneo[8] posited that propensity score models are superior to common multivariate-adjustment methods, especially when there are fewer than eight covariates included. However, in stating that there are legitimate reasons for using the propensity score, one might have a closer look at studies that examine more specialized and detailed applications in clinical research settings. Regularly used for binary-treatment comparisons, there are now approaches for multiple-treatment settings, as they are quite closer to reality, eg, when it comes to dose adaptation within a medical treatment group.[9] Unfortunately, most propensity score evaluations are performed with Monte Carlo studies instead of real-life data.

One of the most critical issues in implementing propensity score analysis is the question of which covariates should be part of the model. While Rubin[10] stated that propensity score models should be exhaustive and contain all available covariates, there are various voices against this procedure. Augurzky and Schmidt[11] supported Rubin's thesis, claiming that this would lead to the true average treatment effect. As shown by Austin et al,[12] propensity score models based on outcome-relevant covariates might be a better tool. While models based on all covariates lead to a holistic propensity score, there is a significant reduction in matching pairs, because it becomes less probable to find a fitting match for

a patient from group A in group B. Using just covariates that are significant within the logistic regression does not seem to be a suitable approach. Models based on this calculation are far too vague, even though one will find the lowest patient dropouts here. D'Agostino and D'Agostino[13] also stated that there is clinical evidence to include these insignificant covariates.

Another aspect to consider is the choice of the best-fitting matching method. While Frölich[14] could not find a significant difference in validity, Augurzky and Schmidt[11] showed that there is a significant loss of data due to patient dropout within nearest-neighbor matching, especially compared to stratification with quartiles. Furthermore, stratification would also lead to a reduced standard error. There are many more possibilities of matching methods, some of which were mentioned by Rosenbaum and Rubin,[2] like Mahalanobis or caliper matching. Those very special calculation methods are not part of this examination.

In respect of all those contrary arguments, the purpose of this study was to examine what kind of influence different propensity score models have on real non-RCT data and how they distinguish from one another in terms of outcome and data structure. Furthermore, we wanted to compare the outcome of this analysis to our native data and to findings in similar RCTs.

Materials and methods
Data
For our research purposes, we used a non-RCT by Otto et al.[15] A total of 173 patients were recruited to compare the weight-loss effects of sleeve gastrectomy versus Roux-en-Y gastric bypass; 127 patients decided to receive bypass surgery, and 46 patients received sleeve gastrectomy. To become part of this study, patients had to be older than 18 years and fulfill the criteria of body mass index (BMI) >40 kg/m^2 or BMI >35 kg/m^2 in combination with comorbidity.

In general, patients could choose between both therapy options, but there was a clear recommendation for sleeve gastrectomy with patients who suffered from gastroesophageal reflux disease, psychiatric diseases, who needed dialysis, or had a BMI >60 kg/m^2.

The surgical procedure was performed with a standard technique and laparoscopically. Roux-en-Y gastric bypass was applied with an antecolic Roux limb of 150 cm and a 50 cm biliopancreatic limb. For the sleeve gastrectomy, the calibration of the gastric sleeve was done with a 42-French bougie, inserted along the lesser gastric curvature. Linear stapling went from 5 cm proximal to the pylorus to the angle of

Hiss. In cases of bleeding or inadequate staples, the stapling line got oversewn.

Follow-up was performed in an outpatient clinic after 6 weeks, 18 weeks, 6 months, 9 months, and 12 months. For comparison, presurgical baselines were measured 1 day before the intervention. Measured data included body composition using the bioelectrical impedance-analysis tool Nutriguard-MS (Data Input GmbH, Darmstadt, Germany) and Nutri Plus software. The relevant criterion for this study was excess weight loss (EWL). EWL is defined as the percentage loss of overweight. We performed native statistic analysis with SAS version 9.3 (SAS Institute Inc, Cary, NC, USA).

Propensity score analysis

In a first step in order to perform propensity score analysis, we checked both therapy groups for heterogeneity within the distribution of their covariates, specifically their sex, height, overweight, BMI, percentage body fat, absolute body water, dry mass, extracellular mass, body cell mass, basal metabolism, phase angle, albumin, glucose, cholesterol, triglycerides, creatinine, leukocytes, erythrocytes, and Quick's value. For this purpose, we used Mann–Whitney U tests and the χ^2 test. Given the native data, we also performed two-sample t-tests with all relevant EWL percentages.

Afterward, we defined three different propensity score models that we used for statistical analysis. These three models were based on the most popular ways to choose included covariates. The first model contained just outcome-related covariates, and the second model was based on all covariates in order to calculate an exhaustive score. The third model contained just significant covariates in logistic regression. Within this analysis, we defined significance as $P<0.05$. We identified outcome-related covariates by reviewing current scientific publications and discussion with bariatric surgeons. This model included sex, overweight, BMI, percentage body fat, body cell mass, basal metabolism, phase angle, glucose, and creatinine. Calculation of the exhaustive propensity score includes sex, height, overweight, BMI, percentage body fat, absolute body water, dry mass, extracellular mass, body cell mass, basal metabolism, phase angle, albumin, glucose, cholesterol, triglycerides, creatinine, leukocytes, erythrocytes, and Quick's value.[16–22] To identify significant covariates, we used logistic regression with backward elimination. Therefore, we used BMI and creatinine as the covariates of choice.

Using the defined propensity score models, we performed a logistic regression to calculate propensity score values for each patient and each model. Knowing these values, we started the matching procedure. In respect of the recent discussion on the best method of data adjustment with propensity scores, we decided to perform nearest-neighbor matching with a maximum difference of 5% in propensity scores between both matching partners and stratification by quartiles.

After propensity score adjustment, we checked both therapy groups again for heterogeneity in covariates with Mann–Whitney U tests and performed two-sample t-tests for every model and adjustment procedure to examine the difference in mean EWL. We performed all statistical analysis with SPSS version 22 for Mac OS X (IBM, Armonk, NY, USA).

Results
Native data

Student's t-test (Table 1) proved that there was a significant difference in EWL between the therapy groups after 18 weeks and 1 year. As emphasized earlier, one has to consider heterogeneity within the groups' covariate distribution. Mann–Whitney U tests and χ^2 testing indicated that the groups contained an unequal distribution for the following covariates, as shown in Table 2: overweight, BMI, percentage body fat, absolute body water, dry mass, extracellular mass, body cell mass, basal metabolism, phase angle, and triglycerides. The P-value of the χ^2 test for the distribution of sex was 0.174.

Propensity score analysis with significant covariates

This model led to an outcome comparison between 54 patients: 27 matching pairs. The mean propensity score for sleeve gastrectomy was 0.4442 and for Roux-en-Y gastric bypass was 0.8391. Values in parentheses are minima and maxima. After the matching procedure had been performed, mean values were 0.5691 (0.029–0.9571) for sleeve gastrectomy and 0.5802 (0.0191-0.9668) for Roux-en-Y gastric bypass (Table 3). Table 4 shows that heterogeneous distribution in covariates was reduced to leukocytes. Student's t-tests for mean differences in EWL showed no significant difference between the groups after matching.

Propensity score analysis with outcome-related covariates

Containing 48 patients, or 24 matching pairs, this model led to a mean propensity score of 0.3976 for sleeve gastrectomy and 0.8539 for Roux-en-Y gastric bypass

Table 1 Student's *t*-test for mean differences in native data

Statistical model	Outcome	Levene's test for equal variance (upper line: equal variance)	Student's *t*-test for mean differences		95% confidence interval	
		P-value	P-value	Mean difference	Lower	Upper
Native data	EWL 6 weeks	0.179	0.171	−0.0962	−0.2345	0.0421
			0.107	−0.0962	−0.2136	0.0212
	EWL 18 weeks	0.093	0.011	−0.1515	−0.2684	−0.0346
			0.002	−0.1515	−0.2441	−0.059
	EWL 6 months	0.192	0.058	−0.1652	−0.336	0.0055
			0.034	−0.1652	−0.3177	−0.0128
	EWL 9 months	0.264	0.06	−0.2071	−0.4229	0.0088
			0.049	−0.2071	−0.4128	−0.0014
	EWL 12 months	<0.0001	0.007	−0.2882	−0.4963	−0.0801
			0.001	−0.2882	−0.4614	−0.115
Propensity score with significant covariates (BMI, creatinine)	EWL 6 weeks	0.255	0.382	−0.0855	−0.1091	0.2801
			0.382	−0.0855	−0.1096	0.2806
	EWL 18 weeks	0.342	0.311	−0.0917	−0.0884	0.2719
			0.312	−0.0917	−0.0887	0.2722
	EWL 6 months	0.528	0.633	−0.0618	−0.1964	0.3199
			0.633	−0.0618	−0.1967	0.3202
	EWL 9 months	0.258	0.942	−0.0107	−0.2805	0.3019
			0.942	−0.0107	−0.2809	0.3023
	EWL 12 months	0.014	0.159	−0.2165	−0.0877	0.5208
			0.16	−0.2165	−0.0883	0.5214
Propensity score with outcome-related covariates	EWL 6 weeks	0.0436	0.2529	−0.1426	−0.3906	0.1053
			0.2542	−0.1426	−0.392	0.1068
	EWL 18 weeks	0.4016	0.3209	−0.0613	−0.1843	0.0617
			0.3244	−0.0613	−0.1866	0.064
	EWL 6 months	0.6461	0.7301	0.0517	−0.2483	0.3518
			0.7301	0.0517	−0.2484	0.3519
	EWL 9 months	0.7753	0.5968	−0.0986	−0.4712	0.274
			0.5968	−0.0986	−0.4713	0.274
	EWL 12 months	0.111	0.5037	−0.108	−0.4305	0.2145
			0.5039	−0.108	−0.4312	0.2152
Propensity score with all covariates	EWL 6 weeks	0.0043	0.1326	−0.3342	−0.7788	0.1104
			0.1474	−0.3342	−0.808	0.1396
	EWL 18 weeks	0.9795	0.7601	−0.0568	−0.4394	0.3259
			0.7601	−0.0568	−0.4394	0.3259
	EWL 6 months	0.9393	0.8611	−0.0383	−0.4895	0.4128
			0.8611	−0.0383	−0.4896	0.4129
	EWL 9 months	0.506	0.8158	0.074	−0.5799	0.7279
			0.8158	0.074	−0.58	0.728
	EWL 12 months	0.029	0.2816	−0.2988	−0.8621	0.2645
			0.2833	−0.2988	−0.8668	0.2692

Abbreviations: EWL, excess weight loss; BMI, body mass index.

before we performed matching. Afterward, these values were 0.5608 (0.1479–0.9593) and 0.572 (0.1277–0.9925) (Table 3). Unequally distributed covariates remained within the number of leukocytes and erythrocytes. Again, *t*-tests for mean EWL between the therapy groups showed no significant difference.

Propensity score analysis with all covariates

With the highest loss of data, this model contained 22 patients in eleven matching pairs. Mean propensity score values before

matching were 0.2991 for sleeve gastrectomy and 0.8526 for Roux-en-Y gastric bypass. Compared to our other models, this was the biggest difference between mean propensity scores, which also indicated relevant heterogeneity in covariate distribution between the therapy groups. After our matching procedure, these values adapted to 0.7196 (0.2476–0.9729) for sleeve gastrectomy and 0.725 (0.267–0.9936) for Roux-en-Y gastric bypass (Table 3). The exhaustive propensity score model was the only one without any residual observed heterogeneously distributed covariate. As seen earlier, the *t*-test for mean propensity score values showed no significant difference for EWL.

Table 2 Comparison between both therapy groups

Treatment	Covariate	Patients (n)	Minimum	Maximum	Mean	Standard deviation
Sleeve gastrectomy	Height	46	1.49	1.88	1.7087	0.09588
	Overweight	46	46.4722	161.6755	98.8788	25.7022
	BMI	46	38.5488	76.2308	55.9492	7.8175
	Percentage body fat	46	30	62	49.28	7.735
	Absolute body water	46	42.1	89	61.878	14.0704
	Dry mass	46	35.6	121.6	83.554	20.5338
	Extracellular mass	46	24.5	65.5	40.804	10.4248
	Body cell mass	46	27.6	66.5	43.733	10.6046
	Basal metabolism	45	1,490	2,940	2,023.33	347.968
	Phase angle	46	3.9	9.2	6.067	1.084
	Albumin	39	23.8	136	40.241	16.104
	Glucose	44	68	235	105.07	31.462
	Cholesterol	37	98	257	173.95	41.831
	Triglycerides	37	51	441	175.86	93.558
	Creatinine	46	0.56	6.64	1.1241	1.00326
	Leukocytes	46	4.4	14.8	8.286	2.2288
	Erythrocytes	46	3.34	5.9	4.6917	0.51346
	Quick's value	45	79	122	97.31	10.083
Roux-en-Y gastric bypass	Height	129	1.48	1.95	1.6828	0.09665
	Overweight	129	25.5746	126.2864	66.6147	18.2266
	BMI	129	34.5599	69.8046	45.7421	5.7738
	Percentage body fat	126	12	67	46.07	7.545
	Absolute body water	127	16.3	95.5	52.127	12.0184
	Dry mass	127	44.3	130.5	71.425	15.9515
	Extracellular mass	127	17.5	63	32.977	8.1355
	Body cell mass	127	24.8	83.1	38.455	9.7078
	Basal metabolism	123	1,400	3,060	1,824.55	289.885
	Phase angle	127	3.6	14.8	6.485	1.1767
	Albumin	86	30.6	49.2	39.156	3.1825
	Glucose	123	66	207	101.42	25.749
	Cholesterol	83	96	295	182.63	42.451
	Triglycerides	83	50	435	142.28	76.901
	Creatinine	127	0.46	4.56	0.8921	0.38614
	Leukocytes	123	3.8	16.4	7.928	2.0977
	Erythrocytes	123	3.68	6.42	4.7654	0.41175
	Quick's value	124	48	125	96.14	10.893

Abbreviation: BMI, body mass index.

Table 3 Propensity score comparison after matching

Propensity score model	Treatment	Patients (n)	Minimum	Maximum	Mean	Standard deviation
Significant covariates	Sleeve gastrectomy	27	0.029	0.9571	0.5691	0.2525
	Roux-en-Y gastric bypass	27	0.0191	0.9668	0.5802	0.2545
Outcome-related covariates	Sleeve gastrectomy	24	0.1479	0.9597	0.5608	0.2498
	Roux-en-Y gastric bypass	24	0.1277	0.9925	0.572	0.2571
All covariates	Sleeve gastrectomy	11	0.2746	0.9729	0.7196	0.2239
	Roux-en-Y gastric bypass	11	0.267	0.9936	0.725	0.2197

Summary

In contrast to our native data, which contained significant heterogeneity in covariate distribution, all propensity score models led to the result that there was actually no significant difference in EWL between sleeve gastrectomy and Roux-en-Y gastric bypass. This conclusion is similar to the findings of Peterli et al,[23] who performed an RCT to examine the weight loss between these therapy options. The outcome within all propensity score models based on non-RCT was equal to the results of an actual RCT, while the native non-RCT analysis showed a significant difference.

Table 4 Mann–Whitney *U* tests

Covariate	Native data set	Significant covariates	Outcome-related covariates	All covariates
	P-value	P-value	P-value	P-value
Height	0.112	0.499	0.650	0.797
Overweight	<0.0001	0.411	0.918	0.898
BMI	<0.0001	0.659	0.711	0.949
Percentage body fat	0.006	0.612	0.959	0.478
Absolute body water	<0.0001	0.81	0.975	0.519
Dry mass	<0.0001	0.929	0.975	0.562
ECM	<0.0001	0.551	0.695	0.438
BCM	0.001	0.866	0.635	0.519
Basal metabolism	<0.0001	0.707	0.741	0.748
Phase angle	0.011	0.373	0.432	0.606
Albumin	0.075	0.879	0.988	0.27
Glucose	0.715	0.887	0.288	0.332
Cholesterol	0.331	1	0.954	0.652
Triglycerides	0.045	0.309	0.665	1
Creatinine	0.16	0.917	0.302	0.847
Leukocytes	0.358	0.042	0.027	0.847
Erythrocytes	0.509	0.258	0.049	0.847
Quick's value	0.876	0.173	0.628	0.797

Abbreviations: BMI, body mass index; ECM, extracellular mass; BCM, body cell mass.

Differences within our models were seen in their ability to eliminate heterogeneity in covariate distribution and the loss of data. Residual heterogeneous distribution in covariate distribution might still be a reason for confounding, but is very unlikely, especially when using propensity score calculation with outcome-related covariates. Within other models, it is necessary to discuss their effect on the outcome. The more covariates we included in the models, the fewer the residual heterogeneous covariates, but the larger the loss of data in terms of matching dropouts. It seems adequate to balance both effects by including just covariates that might have an effect on the observed outcome. Therefore, it is necessary to emphasize preclinical preparation as an important consideration in planning a non-RCT if one wants to include propensity score analysis. Reviewing literature and discussing pathophysiologic pathways, which might affect the outcome beyond medical aspects, are the main points in forming the best-fitting propensity score model.

This scheme for organizing a non-RCT with propensity score analysis is shown in Figure 1. We marked sections that have to receive special attention when it comes to propensity score matching. To calculate adequate propensity scores, one needs to identify all outcome-related covariates. Reviewing actual research literature and discussing extramedical aspects

must be emphasized. Also, data collection during the first patient contact in a prestudy setting and during the clinical study needs to be focused. Only with a reliable and holistic data set is there a chance to compare both therapy groups as objectively as possible.

In conclusion, we showed that there is no need to apply highly specialized statistical methods for reliable calculations, as recent literature often might imply. The key advantage of propensity score analysis is its simplicity if adequate preparation is done.

Discussion

Our work proved that propensity score analysis is actually able to eliminate heterogeneity within covariate distribution in patients of non-RCTs. Nevertheless, one needs to consider the impact of dropouts on the *P*-value. With rising standard deviation and decreasing sample size, the *t*-test is more likely to be significant. Due to this possible bias, it is necessary to interpret these findings in a holistic statistical analysis. Therefore, it might be useful to plan future studies with an increased number of patients. In our case, we accepted this because of our predetermined limited resources and an additional clinical consideration of our data afterward. We did not observe critical issues in terms of propensity score usage. Some former studies implied that these might occur within choosing the right covariates or comparative tests. It just seems to be important to prepare statistical analysis with meticulous data collection and perform a holistic selection for outcome-related covariates. As seen in our work, propensity score models based on all available covariates led to absolute homogeneous covariate distribution but were accompanied by a tremendous loss of data. A useful trade-off we chose was propensity score calculation based on outcome-related covariates. Residual heterogeneous distributions should be critically examined for their influence on the study. Our data showed a remaining imbalance between the therapy groups in respect to their amounts leukocytes and erythrocytes. There is actually no reason to see an impact of these factors on EWL. Observed patient dropouts in the outcome-related propensity score model stood in a rational connection to bias reduction. Concerning patient dropouts, one might argue that patients without matches referring to their extreme propensity score values also do not reflect the regular potential target audience for the specific therapy, which narrows the influence of this loss of data.

A definite advantage of this work lies within the holistic comparison of various propensity score models on a real-life
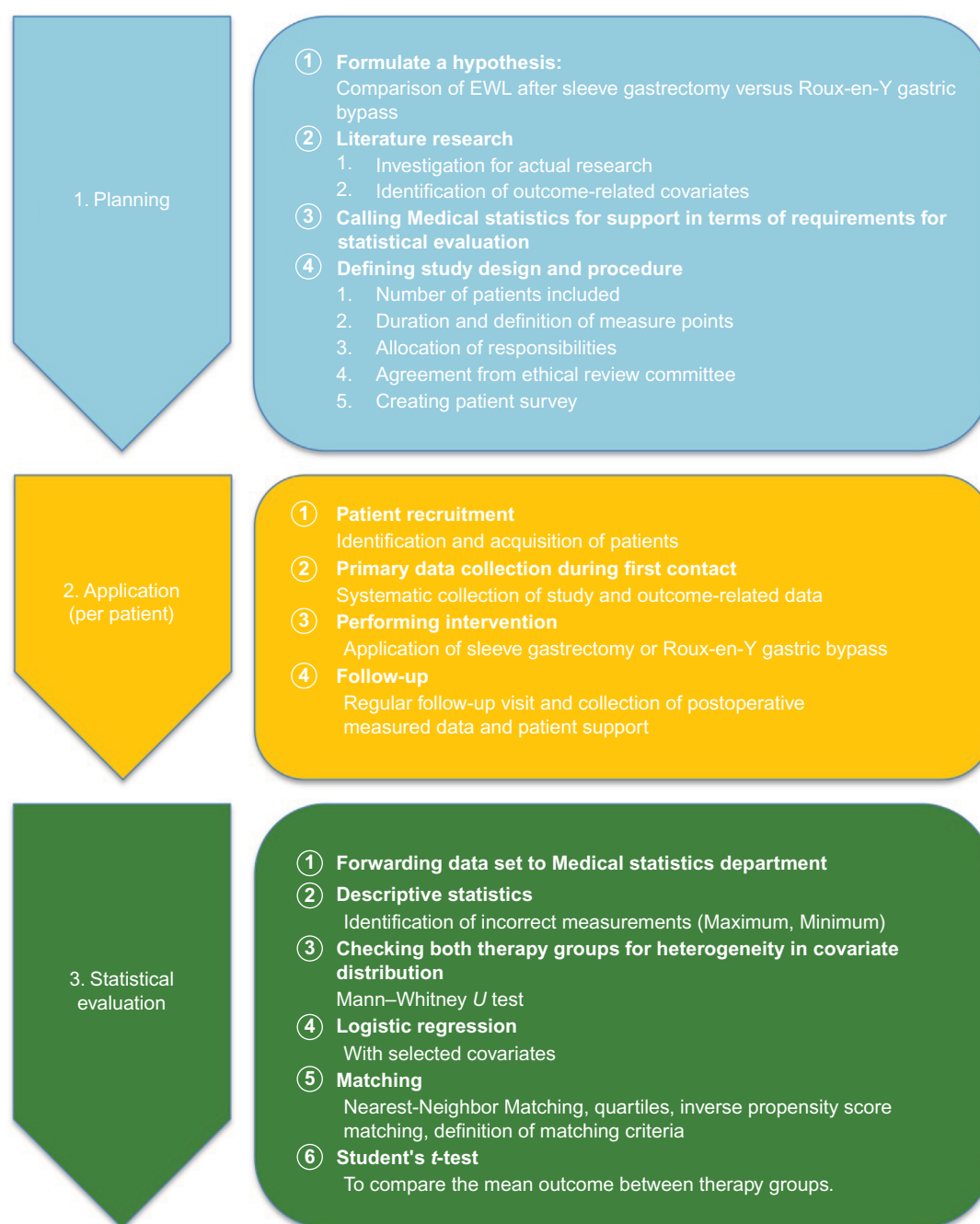
Figure 1 Suggestion for an optimized clinical trial procedure including propensity score matching.
Abbreviation: EWL, excess weight loss.

database. Earlier works often focused on specific and very complex statistical procedures while using Monte Carlo data sets. Furthermore, this work emphasized the central role of data collection, which is the key aspect when it comes to reliable propensity score analysis.

Also, with regard to keeping data as native as possible, we did not use oversampling for our calculations. Oversampling uses a patient various times as matching partner for different members of the other therapy group. We decided to stay as near to a natural study environment as possible.

Methodologically, we were restricted, as we had to work with a preexisting non-RCT. Therefore, we had no influence on data collection. Also, we analyzed mean differences for EWL with *t*-tests and did not use tools like inverse probability of treatment weighting using the propensity scores or caliper analysis. These advanced statistical methods seemed not suitable for our goal of proving the efficiency and simplicity of propensity score models.

Unfortunately, we only had the capacity to perform these calculations with one study. For further examina-

tion, it would be very interesting to analyze these simple propensity score implementations for a wide range of non-RCTs. By doing so, we might get a deeper understanding of chances going along with propensity score matching, and it might appear that these simple models especially are the perfect match for non-RCT analysis. Talking about validating and establishing propensity scores in statistical examinations, we also have to ask whether there is a necessity to perform RCTs. From an ethical and economical point of view, there are a lot of arguments against RCTs, like their higher organizational effort and difficult recruitment of patients. Of course, there is no way to simulate natural conditions as perfectly as appears in a RCT, but as in every other aspect of science, we should consider tools like propensity scores a valuable compromise. By doing so, it seems suitable to set a standardized framework for propensity score analysis in terms of covariate selection and matching procedures in regard to comparability and transparency.

Acknowledgments

Disclosure

The authors report no conflicts of interest in this work.

References

1. Dannehl K. Experimental study versus non-experimental study: the non-experimental (non-randomized) study as a methodological compromise. In: Abel U, Koch A, editors. *Nonrandomized Comparative Clinical Studies*. Düsseldorf: Symposion; 1998:41–52.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
3. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–1156.
4. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996;52(1):249–264.
5. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59(5):437–447.
6. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*. 1998;66(2):315–331.
7. Bryson A, Dorsett R, Purdon S. *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies*. London: Department for Work and Pensions; 2002.
8. Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *J Econom*. 2010;155(2):138–154.
9. Feng P, Zhou XH, Zou QM, Fan MY, Li XS. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med*. 2012;31(7):681–697.
10. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8 Pt 2):757–763.
11. Augurzky B, Schmidt CM. The propensity score: a means to an end. 2001. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=270919. Accessed April 5, 2016.
12. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26:734–753.
13. D'Agostino RB Jr, D'Agostino RB Sr. Estimating treatment effects using observational data. *JAMA*. 2007;297(3):314–316.
14. Frölich M. Finite-sample properties of propensity score matching and weighting estimators. *Rev Econ Stat*. 2004;86(1):77–90.
15. Otto M, Elrefai M, Krammer J, Weiss C, Kienle P, Hasenberg T. Sleeve gastrectomy and Roux-en-Y gastric bypass lead to comparable changes in body composition after adjustment for initial body mass index. *Obes Surg*. 2016;26(3):479–485.
16. Chagnac A, Weinstein T, Herman M, Hirsh J, Gafter U, Ori Y. The effects of weight loss on renal function in patients with severe obesity. *J Am Soc Nephrol*. 2003;14(6):1480–1486.
17. Cottam DR, Schaefer PA, Shaftan GW, Velcu L, Angus LD. Effect of surgically-induced weight loss on leukocyte indicators of chronic inflammation in morbid obesity. *Obes Surg*. 2002;12(3):335–342.
18. Folsom AR, Qamhieh HT, Wing RR, et al. Impact of weight loss on plasminogen activator inhibitor (PAI-1), factor VII, and other hemostatic factors in moderately overweight adults. *Arterioscler Thromb*. 1993;13(2):162–169.
19. Kyle UG, Genton L, Pichard C. Low phase angle determined by bioelectrical impedance analysis is associated with malnutrition and nutritional risk at hospital admission. *Clin Nutr*. 2013;32(2):294–299.
20. Layman DK, Boileau RA, Erickson DJ, et al. A reduced ratio of dietary carbohydrate to protein improves body composition and blood lipid profiles during weight loss in adult women. *J Nutr*. 2003;133(2):411–417.
21. Norman K, Stobäus N, Pirlich M, Bosy-Westphal A. Bioelectrical phase angle and impedance vector analysis – clinical relevance and applicability of impedance parameters. *Clin Nutr*. 2012;31(6):854–861.
22. Pelleymounter MA, Cullen MJ, Baker MB, et al. Effects of the obese gene product on body weight regulation in ob/ob mice. *Science*. 1995;269(5223):540–543.
23. Peterli R, Borbély Y, Kern B, et al. Early results of the Swiss Multicentre Bypass or Sleeve Study (SM-BOSS): a prospective randomized trial comparing laparoscopic sleeve gastrectomy and Roux-en-Y gastric bypass. *Ann Surg*. 2013;258(5):690–695.