SUPPLEMENTARY APPENDIX

Assessment of data

The following parameters were collected in the CRFs: age and gender, autoantibody profile, disease duration (defined as time since onset of first non-Raynaud symptoms), disease subtype (diffuse vs. limited disease according to Leroy). The modified Rodnan Skin Score (mRSS) was used to evaluate skin fibrosis.(1) EUSTAR investigators are regularly trained to perform the mRSS.(2-4) Pulmonary function was primarily evaluated by measurement of forced vital capacity (FVC) % predicted, diffusing capacity of the lung for carbon monoxide (DLCO) % predicted and high-resolution CT (HRCT). Moreover, safety of TCZ was assessed by various other clinical and laboratory parameters. Data were collected at baseline as well as 3-, 6- and 12-months follow-up.

Supplementary tables

Supplementary Table S1. Overlap syndrome with RA	
	TCZ treated patients (n=93)
Anti CCP positive	13/55 (23.6%)
RF positive	22/60 (36.7%)
RF and anti- CCP both positive	11/60 (18.3%)
Overlap syndrome RA as reported by principal investigator	20/72 (27.8%)
No data was available for the control patients. Abbreviations: anti-CCP- anti cyclic citrullinated peptide antibody, RA-rheumatoid arthritis; RF-rheumatoid factor; TCZ-tocilizumab	

Supplementary Table S2. Pooled SMD between groups after multiple imputation and propensity score matching (nearest neighbour matching algorithm)	
	SMD mean across imputations
age	0.077
sex	0.080
Scleroderma subtype	0.073
Baseline mRSS	0.061
Baseline FVC % predicted	0.093
Baseline DLCO % predicted	0.085
Immunosuppressive co-therapy	0.061
Rituximab within 6 months	0.076
Disease duration	0.053
Year of treatment	0.063
Abbreviations: SMD- standardised mean difference; mRSS- modified Rodnan skin score; FVC- forced vital capacity; DLCO-diffusing capacity for carbon monoxide;	

Supplementary Table S3. Progressive and regressive patients		
Secondary outcome	group	Percentage (%)
Progressive lung fibrosis	TCZ treated	14.1
Progressive lung fibrosis	control	17.4
Progressive skin fibrosis	TCZ treated	4.0
Progressive skin fibrosis	control	3.9
Regressive lung fibrosis	TCZ treated	22.3
Regressive lung fibrosis	control	16.1
Regressive skin fibrosis	TCZ treated	20.1
Regressive skin fibrosis	control	18.9
Progression/regression of the skin fibrosis was defined as: an increase/decrease in mRSS >5 units AND ≥25% from baseline to 12±3 months follow up. Progression/regression of lung fibrosis is defined as decrease from baseline to12±3 months follow-up in FVC ≥10%, or FVC decrease 5%–9% combined with diffusing capacity for carbon monoxide (DLCO) ≥15%. After multiple imputation the **most.recent** selection of patients and the **nearest** neighbor matching was applied. Abbreviations: TCZ- tocilizumab		

Supplementary Table S4. Baseline demographics of random drawn dataset after multiple imputation and propensity score matching

	Tocilizumab	Control group	p-values	SMD
	N=93	N=93		
Age (mean±SD; years)	50.9±13.5	48.4±15.1	0.23	0.18
Sex				
Female (n, %)	73 (78.5)	73 (78.5)	1.00	<0.001
Scleroderma subtype				
Diffuse (n, %)	54 (58.1)	55 (59.1)	1.00	0.02
Immunosuppressive co-therapy				
Yes (n, %)	75 (80.6.)	75 (80.6)	1.00	<0.001
Prednisone (≥10 mg/day; n; %))	44 (47.3)	16 (17.2)	<0.001	0.68
Cyclophosphamide (n, %)	0	6 (7.2)	NaN	NaN
Methotrexate (n, %)	41 (44.1)	31 (33.3)	0.18	0.22
Azathioprine (n, %)	6 (6.5)	20 (21.5)	0.01	0.45
Mycophenolate mofetil (n, %)	6 (6.5)	18 (19.4)	0.02	0.39
D-penicillamine (n, %)	0	0	NaN	NaN
Rituximab within 6 months (n, %)	1 (1.1)	1 (1.1)	1.00	<0.001
Imatinib (n, %)	0	0	NaN	NaN
TNF-alpha antagonist (n, %)	0	0	NaN	NaN
Abatacept (n, %)	1 (1.1)	0	NaN	NaN
Disease duration (mean±SD; years)	6.4±5.4	6.2±4.9	0.79	0.04
Autoantibodies positive				
ANA (n, %)	73 (92.4)	78 (96.3	0.47	0.17
ACA (n, %)	12 (16.7)	10 (13.5)	0.76	0.09
Anti-Scl-70 (n, %)	54 (65.1)	42 (53.8)	0.20	0.23
CRP ≥ 5mg/l (n, %)	38 (43.7)	80 (87.9)	<0.001	1.05
ESR elevation (>25mm/h)	38 (54.3)	29 (34.9)	0.03	0.40
mRSS (median, IQR)	13.0 (6.0, 22.0)	11.0 (6.0, 21.0)	0.67	0.07
FVC % predicted (mean±SD)	87.7±21.4	88.0±22.8	0.93	0.01
DLCO % predicted (mean±SD)	62.6±21.9	65.1±19.0	0.41	0.12
HRCT positive or X-ray positive	49 (73.1)	37 (47.4)	0.003	0.54
Digital ulcers (n, %)	16 (17.8)	12 (18.5)	1.00	0.02
Joint synovitis (n, %)	44 (62.0)	14 (15.4)	<0.001	1.09
Tendon friction rubs (n, %)	25 (31.2)	11 (12.1)	0.004	0.48

Demographics and clinical Characteristics are defined according to EUSTAR criteria.[5] Above shown data is before multiple imputation and propensity score matching. Abbreviations: SD-standard deviation; mRSS- modified Rodnan skin score; FVC- forced vital capacity; DLCO-diffusing capacity for carbon monoxide;NaN- not available number;

Supplementary Table S5. Subgroup analysis (nearest neighbour matching and selection of most recent observation)		
	outcome	p-value for interaction test
mRSS ≥10 versus mRSS <10	mRSS	0.77
	FVC	0.47
diffuse versus limited	mRSS	0.81
	FVC	0.69
HRCT positive vs HRCT negative	mRSS	0.63
	FVC	0.90
FVC <80% versus FVC ≥80%	mRSS	0.68
	FVC	0.84
FVC <80% & HRCT positive	mRSS	0.76
	FVC	0.41
Disease duration (in years) ≤3 versus disease duration >3	mRSS	0.23
	FVC	0.63
C-reactive protein (CRP) ≤ 5 mg/l versus CRP> 5 mg/l	mRSS	0.95
	FVC	0.55
Abbreviations: mRSS- modified Rodnan skin score; FVC- forced vital capacity % predicted; HRCT- high resolution computed tomography		

Supplementary Table S6. mRSS at follow-up (12±3months), sensitivity analyses					
		nearest neighbour matching		exact matching	
		most recent	random	most recent	random
Tocilizumab	mean estimate (95% CI)	11.2 (9.1 to 13.3)	11.1 (9.1 to 13.2)	11.2 (9.1 to 13.3)	11.1 (9.1 to 13.2)
Controls	mean estimate (95% CI)	12.2 (9.7 to 14.6)	12.0 (9.6 to 14.4)	12.5 (10.3 to 14.7)	12.1 (9.8 to 14.4)
between group difference	mean estimate (95% CI)	-1.0 (-3.7 to 1.8)	-0.8 (-3.8 to 2.1)	-1.3 (-3.5 to 0.9)	-1.0 (-3.4 to 1.5)
p-value		0.48	0.58	0.25	0.44
Abbreviations: mRSS- modified Rodnan skin score; CI-confidence interval					

Supplementary Table S7. FVC at follow-up (12±3months), sensitivity analyses					
		nearest neighbour matching		exact matching	
		most recent	random	most recent	random
Tocilizumab	mean estimate (95% CI)	88.7 (83.7 to 93.7)	88.7 (83.6 to 93.8)	88.7 (83.7 to 93.7)	88.7 (83.7 to 93.8)
Controls	mean estimate (95% CI)	87.2 (80.8 to 93.6)	87.2 (80.6 to 93.9)	87.9 (82.2 to 93.6)	87.6 (81.2 to 93.9)
between group difference	mean estimate (95% CI)	1.5 (-6.1 to 9.07)	1.5 (-6.4 to 9.3)	0.8 (-6.1 to 7.7)	1.2 (-6.6 to 8.9)
p-value		0.70	0.72	0.82	0.77
Abbreviations: FVC- forced vital capacity % predicted; CI-confidence interval					

Supplementary Table S8. Progression and regression of mRSS, sensitivity analyses				
	nearest neighbour matching		exact matching	
	most recent	random	most recent	random
Estimated treatment effect of TCZ for progression OR (95% CI)	0.73 (0.11 to 4.83)	0.66 (0.10 to 4.39)	0.83 (0.12 to 5.76)	0.76 (0.10 to 5.51)
p-values	0.74	0.66	0.85	0.78
Estimated treatment effect of TCZ for regression OR (95% CI)	1.09 (0.44 to 2.70)	1.01 (0.39 to 2.64)	1.11 (0.39 to 3.13)	1.10 (0.39 to 3.09)
p-value	0.86	0.99	0.84	0.85
Abbreviations: mRSS- modified Rodnan skin score; CI-confidence interval; TCZ-tocilizumab; OR-odds ratio				

Supplementary Table S9. Progression and regression of FVC, sensitivity analyses				
	nearest neighbour matching		exact matching	
	most recent	random	most recent	random
Estimated treatment effect of TCZ on FVC decline OR (95% CI)	0.77 (0.27 to 2.24)	0.68 (0.24 to 1.95)	0.80 (0.27 to 2.37)	0.68 (0.23 to 2.00)
p-values	0.63	0.48	0.69	0.48
Estimated treatment effect of TCZ for on FVC improvement OR (95% CI)	1.49 (0.58 to 3.84)	1.46 (0.53 to 4.06)	1.29 (0.52 to 3.18)	1.35 (0.53 to 3.40)
p-value	0.41	0.47	0.58	0.53
Abbreviations: FVC- Forced vital capacity % predicted; CI-confidence interval; TCZ-tocilizumab; OR-odds ratio				

References:

1. Clements P, Lachenbruch P, Siebold J, White B, Weiner S, Martin R, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol. 1995;22(7):1281-5.
2. Czirják L, Foeldvari I, Müller-Ladner U. Skin involvement in systemic sclerosis. Rheumatology (Oxford). 2008;47 Suppl 5:v44-5.
3. Ionescu R, Rednic S, Damjanov N, Varjú C, Nagy Z, Minier T, et al. Repeated teaching courses of the modified Rodnan skin score in systemic sclerosis. Clin Exp Rheumatol. 2010;28(2 Suppl 58):S37-41.
4. Khanna D, Furst DE, Clements PJ, Allanore Y, Baron M, Czirjak L, et al. Standardization of the modified Rodnan skin score for use in clinical trials of systemic sclerosis. J Scleroderma Relat Disord. 2017;2(1):11-8.
5. Walker UA, Tyndall A, Czirják L, et al. Clinical risk assessment of organ manifestations in systemic sclerosis: a report from the EULAR Scleroderma Trials And Research group database. *Ann Rheum Dis. 2007 Jun; 66(6):754-763.*

Statistical analysis plan

University of Zurich[UZH]

Epidemiology, Biostatistics
and Prevention Institute,
Biostatistics Department
+41 44 634 46 41
www.biostat.uzh.ch

Statistical Analysis Plan: Treatment effect of tocilizumab in systemic sclerosis patients estimated from the EUSTAR database with propensity-score matching.

Analysis for Dr. Suzana Jordan, Prof. Oliver Distler

PD Dr. Ulrike Held (*ulrike.held@uzh.ch*)
BSc. Klaus Steigmiller (*klaus.steigmiller@uzh.ch*)

Version 1.5 of September 6, 2018

Contents

1 Introduction

Systemic sclerosis (SSc) is a rare autoimmune disease, associated with a high risk of mortality (Elhai et al., 2012). Although treatments have become available for some manifestations such as pulmonary arterial hypertension, until now, there is no effective therapy to counteract the fibrotic process of this

disease. Increasing evidence supports a contribution of the immune system and in particular interleukin 6 (IL-6) in the pathogenesis of SSc (O'Reilly et al., 2013). In SSc-patients

preliminary data have suggested that tocilizumab, a humanised anti-IL-6 receptor antibody, might improve dermal fibrosis and joint involvement in refractory polyarthritis associated with SSc (Shima et al., 2010) (Elhai et al., 2013). Following these results, a phase II randomized controlled trial was performed enrolling 87 diffuse cutaneous SSc with progressive skin disease (Khanna et al., 2016). The primary endpoint (difference in mean change from baseline in modified Rodnan skin score at 24 weeks) was not met, but Rodnan skin score had a greater decreasing trend in patients treated with Tocilizumab than in those receiving placebo (p=0.06). Furthermore, one exploratory analysis revealed that fewer patients in the Tocilizumab group had a decline in forced vital capacity at 48 weeks (p=0.04). An ongoing phase III trial might help to better analyse safety and efficacy of tocilizumab in SSc. However, limited follow-up periods and strict criteria for recruitment in clinical trials prevent from documenting effects in a broader population and the long-term outcomes of the patients and might lead to an underestimation of adverse events. Therefore, data from large "real- life" registries could be used to better determine the effects of Tocilizumab on different involvements of SSc, to identify who, among SSc-patients might benefit from this treatment, and to analyse long- term efficacy and safety of tocilizumab in SSc.

2 Data source

We will use data from the EUSTAR ("European Scleroderma Treatment And Research") database. There are currently more than 200 scleroderma centres in the EUSTAR cohort, treating more than 15,000 patients with SSc. To document the course and treatment of scleroderma, the EUSTAR cohort maintains an online patient database. The data is stored centrally in Switzerland (New Win AG). Patient data were entered starting from January 2009. Further information is available on www.EUSTAR.org.

3 Aims

The aim of this project is to evaluate the outcomes of SSc-patients receiving tocilizumab in routine care in comparison with SSc-patients not receiving tocilizumab.

4 Analysis sets and populations

Definition of treated and controls: Patients reported from the EUSTAR database receiving tocilizumab for at least 3 months are considered as TREATED. Patients from the database not receiving tocilizumab are considered as potential CONTROLS.

Inclusion criteria:

- Age ≥ 17 years
- Observations of patients entering the database after 1.1.2010
- Classification criteria for SSc fulfilled (ACR 1980 criteria or 2013 ACR/EULAR criteria)
- Potential control patients with disease duration ≤ 35 years

5 Endpoints, covariates, and subgroups

Primary endpoints: Primary endpoints are mRSS (modified Rodnan Skin Score) and FVC, both at 12 months follow-up time. The time window for the 12 months assessment will be a 9-15 months window. The two different endpoints will be addressed separately, but a single matched set of patient treated-control pairs will be used for analysis.

Exploratory secondary endpoints: The percentage of progressive patients for skin under therapy, defined as an increase in (mRSS of 5 points AND of 25% compared to baseline). The percentage of progressive patients for lung under therapy, defined as a decrease in either (FVC > 10%) or (FVC > 5% AND DLCO > 15%).

The percentage of regressive patients for skin under therapy, defined as a (decrease in mRSS of 5 points AND of 25% compared to baseline). The percentage of regressive patients for lung under therapy, defined as an increase in either (FVC > 10%) or (FVC > 5% AND DLCO > 15%).

Covariates identified to impact outcomes are

- Age (years)
- Gender
- Subtype (limited / diffuse)
- Baseline mRSS, baseline FVC, baseline DLCO
- Co-therapy immunosuppressive DMARDs (prednisone > 10mg/d, Methotrexate, azathioprine and mycophenolate mofetil), either one (coded 1) or none (coded 0)
- Rituximab (biologic) within 6 months before baseline
- Disease duration (years)
- Year of treatment

The above defined covariates (confounders) will be used for matching.

Pre-specified exploratory subgroup analysis for primary endpoints, and secondary endpoints where applicable:

- Subgroup analysis for baseline mRSS: mRSS $\geq$ 10 versus $<$ 10.
- Subgroup analysis for subtype analysis: diffuse SSc vs limited SSc.
- Subgroup analysis for baseline FVC: FVC < 80% versus $\geq$ 80%.
- Subgroup analysis for baseline FVC: FVC < 80% and HRCT-diagnosed lung fibrosis versus complementary group.

A test for interaction will precede the subgroup analyses. Only if $p < 0.05$, we assume that there is evidence for differential treatment effect between subgroups, and subgroup results will be reported.

6 Handling of missing values and other data conventions

Missing values Missing values will be present in the database and the case report forms.

- Potential control patients: Missing data in each of the two outcomes (FVC and mRSS) of control patients will be excluded listwise.
- Treated patients: within patients receiving tocilizumab, none will be excluded due to missing outcomes prior to multiple imputation.
- If the percentage of missing values in a covariate is above 50%, the covariate will be excluded from the analysis. Otherwise, missingness patterns will be assessed. If it can be assumed that data are missing completely at random (MCAR) or missing at random (MAR), multiple imputation using chained equations (MICE) will be applied, leading to the analysis of > 50 completed datasets. The MI will be used for the defined covariates and outcomes, at baseline and follow-up.

Preselection of potential control observations across the database The follow-up time is±12 ±3months, leaving a time window from 9 - 15 months follow-up duration. A preselection of observations in control patients will extract those observations for each control patient from the database that have exactly 2 observations within this time window. The first of these will be considered as baseline, the second as follow-up. In case that still multiple control observations exist for single patients, we will use the most recent one, but this approach will be revisited in the sensitivity analysis. Eventually, one single observation (line in dataset) for each potential CONTROL patient, and all patients considered as TREATED will be used for estimating the propensity score and matching.

7 Statistical methodology

Estimation of the propensity score and matching The propensity score will be estimated with logistic regression, including the covariates described above. We plan to use an optimal matching algorithm (Ho et al., 2011), attempting to minimize a global measure of distance between treated and controls. Attention needs to be paid to matched sets in the subgroup analysis. 1:1 matching will be performed.
Balancing of baseline covariates will be assessed with descriptive statistics, exploratory p-values, and the standardized mean difference (SMD), resulting in a Table 1 before and after matching. If SMD < 0.1, the covariate will be assumed balanced between matched samples, if SMD ≥ 0.1 the covariate will be adjusted for in a regression model.

Analysis after matching If all baseline covariates are balanced after matching, the treatment effect will be estimated as mean difference between treatment groups, including 95% confidence interval for continuous outcomes and as odds ratios with 95% confidence intervals for binary outcomes. If unbalanced baseline confounders remain after matching, the analysis will be using a linear random effects model for continuous outcomes, including treatment group and potentially unbalanced covariates after matching as independent variables. Random effects will be used to account for each matched pair of TREATED-CONTROL patients. Binary outcomes will be addressed with logistic regression models under the same specifications. The results will include confirmatory p-values for primary outcomes, as well as exploratory p-values for all other hypotheses. The significance level for confirmatory p-values is set to 0.05.

8 Sensitivity analysis

Preselection of observations in control patients, not the most recent observation per each patient will be used, but randomly an observation will be chosen.

Exact matching for the variables Gender, Subtype, Co-therapy immunosuppressive DMARDs and Rituximab (biologic) within 6 months before baseline will be applied as sensitivity analysis.

A common critique of the matching approach is that unmeasured baseline covariates may still affect the estimated treatment effect and cannot be accounted for. Robustness of the results, in case of unmeasured confounders, will be addressed with Rosenbaum bounds for p-values and Hodges-Lehmann point estimates (Rosenbaum, 1993) (Rosenbaum, 2013). The results will be presented in combination with the calculated level of robustness, if the significance level of the first analysis was met.

9 Rationale for any deviation from pre-specified analysis plan

- Percentage of missing values too large for some of the defined confounders.
- Missing values in the outcomes.
- Missingness generating mechanism not at random.

10 References

Elhai, M., Meune, C., Avouac, J., Kahan, A. and Allanore, Y. (2012). Trends in mortality in patients with systemic sclerosis over 40 years: a systematic review and meta-analysis of cohort studies. *Rheumatology* **51** 1017–1026.
URL http://dx.doi.org/10.1093/rheumatology/ker269

Elhai, M., Meunier, M., Matucci-Cerinic, M., Maurer, B., Riemekasten, G., Leturcq, T., Pel- lerito, R., Von Mühlen, C. A., Vacca, A., Airo, P., Bartoli, F., Fiori, G., Bokarewa, M., Riccieri, V., Becker, M., Avouac, J., Müller-Ladner, U., Distler, O. and Allanore, Y. (2013). Outcomes of patients with systemic sclerosis-associated polyarthritis and myopathy treated with tocilizumab or abatacept: a eustar observational study. *Annals of the Rheumatic Diseases* **72** 1217– 1220.
URL https://ard.bmj.com/content/72/7/1217

Ho, D., Imai, K., King, G. and Stuart, E. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, Articles* **42** 1–28.
URL https://www.jstatsoft.org/v042/i08

Khanna, D., Denton, C. P., Jahreis, A., van Laar, J. M., Frech, T. M., Anderson, M. E., Baron, M., Chung, L., Fierlbeck, G., Lakshminarayanan, S., Allanore, Y., Pope, J. E., Riemekasten, G., Steen, V., MÃijller-Ladner, U., Lafyatis, R., Stifano, G., Spotswood, H., Chen-Harris, H., Dziadek, S., Morimoto, A., Sornasse, T., Siegel, J. and Furst, D. E. (2016). Safety and efficacy of subcutaneous tocilizumab in adults with systemic sclerosis (fasscinate): a phase 2, randomised, controlled trial. *The Lancet* **387** 2630 – 2640.
URL http://www.sciencedirect.com/science/article/pii/S0140673616002324

O'Reilly, S., Cant, R., Ciechomska, M. and van Laar, J. M. (2013). Interleukin-6: a new therapeutic target in systemic sclerosis? *Clinical & Translational Immunology* **2** e4.
URL https://onlinelibrary.wiley.com/doi/abs/10.1038/cti.2013.2

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Rosenbaum, P. (2013). *Observational Studies*. Springer Series in Statistics, Springer New York.
URL https://books.google.ch/books?id=a7whBQAAQBAJ

Rosenbaum, P. R. (1993). Hodges-lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association* **88** 1250–1253.
URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1993.10476405

Shima, Y., Kuwahara, Y., Murota, H., Kitaba, S., Kawai, M., Hirano, T., Arimitsu, J., Narazaki, M., Hagihara, K., Ogata, A., Katayama, I., Kawase, I., Kishimoto, T. and Tanaka, T. (2010). The skin of patients with systemic sclerosis softened during the treatment with anti-il-6 receptor antibody tocilizumab. *Rheumatology* **49** 2408–2412.
URL http://dx.doi.org/10.1093/rheumatology/keq275