Predicting Response to Tocilizumab Monotherapy in Rheumatoid Arthritis: A Real-World Data Analysis Using Machine Learning

Authors:
Fredrik D Johansson, PhD (1), Jamie E Collins, PhD (2), Vincent Yau, PhD (3), Hongshu Guan, MSc (4), Seoyoung C Kim, MD, ScD (4,5), Elena Losina, PhD (2), David Sontag, PhD (6), Jacklyn Stratton, BA (4), Huong Trinh, PhD (3), Jeffrey Greenberg, MD, MPH (7,8), Daniel H Solomon MD, MPH(4,5)

Key indexing terms: Rheumatoid arthritis, disease-modifying anti-rheumatic drug, remission, prediction model, machine learning

Departments & institutions:
(1) Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden
(2) Department of Orthopedic Surgery, Brigham and Women's Hospital, Boston MA
(3) Genentech, South San Francisco, CA
(4) Division of Rheumatology, Brigham and Women's Hospital, Boston MA
(5) Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston MA
(6) Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA
(7) NYU School of Medicine, New York, NY
(8) Corrona, Waltham MA

Support: This work was supported by a research grant from Roche/Genentech to Brigham and Women's Hospital.

Potential Conflicts of interest: DHS receives salary support from research grants to Brigham and Women's Hospital from Abbvie, Amgen, Corrona, Janssen, Pfizer, and Roche/Genentech. He also serves on the editorial board of Arthritis & Rheumatology and on the FDA Arthritis Advisory Committee. FDJ receives salary support from research grants to Chalmers University of Technology from AstraZeneca. SCK received research grants to Brigham and Women's Hospital from AbbVie, Pfizer, Roche, and Bristol-Myers Squibb for unrelated studies. JEC is a consultant to Boston Imaging Core Labs and serves as Associate Editor for Statistics at Osteoarthritis and Cartilage.

Appointments:
(1) FD Johansson, Assistant Professor, PhD
(2) J Collins, Assistant Professor, PhD
(3) V Yau, Principal Data Scientist, PhD
(4) H Guan, Sr Statistical Programmer, Biostatistician, MS
(5) SC Kim, Associate Professor of Medicine, MD, ScD

(6) D Sontag, Associate Professor, PhD
(7) J Greenberg, Chief Medical Officer, MD, MPH
(8) DH Solomon, Professor of Medicine, MD, MPH

Correspondence: Fredrik D Johansson, E-mail: fredrik.johansson@chalmers.se

Running head: Predicting response to Tocilizumab

Words: 3498
Tables: 3
Figures: 1
References: 19
Appendix: Yes

ABSTRACT

Objectives

Tocilizumab (TCZ) had similar efficacy when used as monotherapy or in combination with other treatments for rheumatoid arthritis (RA) in randomized controlled trials (RCT). We derived a remission prediction score for TCZ monotherapy (TCZm) using RCT data and now performed an external validation of the prediction score using "real world data" (RWD).

Methods

We identified patients in Corrona-RA who used TCZm (n=453), matching the design and patients from four RCTs used in previous work (n=853). Patients were followed to determine remission status at 24 weeks. We compared the performance of remission prediction models in RWD, first based on variables determined in our prior work in RCTs, and then using an extended variable set, comparing logistic regression and random forest models. We included patients on other biologic DMARD monotherapies (bDMARDm) to improve prediction.

Results

The fraction of patients observed reaching remission on TCZm by their follow-up visit was 12% (n=53) in RWD vs 15% (n=127) in RCTs. Discrimination was good in RWD for the risk score developed in RCTS with AUROC of 0.69 (95% CI 0.62, 0.75). Fitting the same logistic regression model to all bDMARDm patients in the RWD improved the AUROC on held-out TCZm patients to 0.72 (95% CI 0.63, 0.81). Extending the variable set and adding regularization further increased it to 0.76 (95% CI 0.67, 0.84).

Conclusion

The remission prediction scores, derived in RCTs, discriminated patients in RWD about as well as in RCTs. Discrimination was further improved by retraining models on RWD.

INTRODUCTION

An expanding treatment armamentarium means more treatment options for patients with rheumatoid arthritis (RA), however clinicians face difficult decisions attempting to make evidence-based recommendations regarding which disease-modifying antirheumatic drug (DMARD) treatment will be most effective in a given patient. While the majority of patients with RA will find an effective treatment, not all do; many spend months trying medications that may not work for them.[1] Prior investigations have attempted to find biomarkers which can help personalize treatments, but most efforts have not produced useful results.[2] While an exhaustive comparison between all treatment options is a desirable goal, a natural first step is to identify and understand predictors of a single drug's success.

We recently examined clinical data from several randomized controlled trials (RCTs) and were able to derive and validate a prediction score for remission among patients using tocilizumab monotherapy (TCZm).[3] Monotherapy with TCZ has been found more effective than monotherapy with some targeted therapies for RA.[4] However, RCT data may not always replicate in typical practice with real world data.[5] These differences may derive from different patient populations, different treatment patterns, or other more subtle differences.[5]

Real-world data (RWD) offer important advantages over RCTs in that patients are more heterogeneous, with a greater variety of clinical characteristics as well as experience with other biologic and targeted DMARD treatments. We examined the performance of our original prediction score[3] for remission among patients using TCZm among patients in Corrona, a large

RWD set from the US.[6] We employed various machine learning algorithms to take advantage of the copious data contained within Corrona.

MATERIALS & METHODS

Study Design

No research occurred before patients gave written informed consent for these analyses. The research was reviewed and approved by the New England IRB (Corrona-RA-100, IRB Tracking # 120160610), a centralized human ethics committee.

Our study sought to answer the following questions regarding remission in patients with RA using TCZm: A) to what extent do the findings of Collins et al.[3] replicate in RWD? B.1) will expanding the set of remission predictors improve model fit? B.2) can data from patients on other therapies improve a predictive model for TCZm patients? C) what gains are there from applying non-parametric estimators of remission probability?

Question A. Evaluating baseline model in real-world data

As a baseline model, we used the remission model due to Collins et al.[3] derived from patients on TCZm from two RCTs.[7,8] In the most successful model from the Collins study, twelve variables representing demographics, basic RA characteristics and treatment history were included in a logistic regression (LR) model based on two criteria: an a priori baseline set of covariates and model odds ratio (OR). We refer to this model as LR-OR.

In the analysis by Collins et al.,[3] two RCTs were used for derivation and two for validation.[4,7-9] In the current analyses, our focus is predictive discrimination in RWD. Hence, all four trials were used for derivation of models validated in the RWD. Access to the RCT and RWD data was granted following de-identification and IRB approval from the Partners Healthcare Human Studies. Variables from the four RCTs were harmonized by Collins et al. The less than 5% of subjects with missing values were removed from the study.

We evaluated the baseline LR-OR model using patients from the RWD dataset. This was pursued using the parameter values fit in the four RCTs, and also by refitting the parameters of the LR-OR model to RWD. Using both methods serves to estimate the extent to which model fit is affected by the cohort discrepancies between RCT and RWD.

Question B. Expanding the variable set and derivation population

The original variable set used in LR-OR was limited to the covariates collected in the four RCTs underlying its derivation.[4,7-9] The Corrona RWD used in the current analyses contains a greatly expanded feature set not available in the RCTs, see below and in the **Appendix**. All models fit with this expanded set were trained and evaluated using only RWD. Adding covariates comes at a "statistical power price" however, since they contribute to increased variance if the cohort remains of fixed size. To address this, we used a regularized logistic regression model (LR-Reg) which penalizes models with many large coefficients.

To further reduce the variance of our model parameters, we fit models also to an expanded population including patients on monotherapy with *any* biologic DMARD, including TCZm, (bDMARDm). Variables that predict remission in RA are likely to be predictive for patients on different therapies. By including an indicator for TCZm therapy, this design choice greatly increased the cohort size while enabling the model to remain predictive for our cohort of interest. Validation was then pursued for the cohort including only patients using TCZm.

Question C. Applying ML algorithms in prediction of remission

A limitation of using logistic regression for predicting remission is that it models the log-odds of an event as a *linear* function of the covariates. As an alternative, we used non-parametric *random forest* estimators—ensembles of tree-structured decision rules.[10] Random forests are capable of *discovering* meaningful interactions and transformations of variables while mitigating increased variance through the use of bootstrapping. A drawback is that it is often difficult to describe ensembles of learned decision rules concisely.[11] For this reason, we use random forests primarily to get an indication for how limited linear models are in this task.

Study Populations

Following Collins et al.,[3] we used four RCTs—ACT-RAY, FUNCTION, ADACTA and AMBITION—for derivation of our baseline model.[4,7-9] Our RWD cohort was extracted from the Corrona RA registry—the largest prospective cohort study of RA in the world.[6] The registry comprises

medical history including conditions, diagnoses, labs and treatments as well as demographic and lifestyle data. Records are collected through at regular visits through two questionnaires filled in by the patient and their physician, respectively. We used a version of the registry exported on February 4, 2018, containing visits from 54,646 patients recorded from October 2001 to December 2017.

Patients in the Corrona RWD were eligible for inclusion if they were older than 18 years and on bDMARDm for a minimum of three months with at least one follow-up visit no later than nine months after initiation. Patients starting a bDMARD in combination with other DMARDs were included if the monotherapy with the target drug was started at most three months after target drug (not necessarily monotherapy) initiation; this occurred when the non bDMARD was stopped resulting in bDMARDm. See **Appendix** for a list of bDMARDs. Subjects had to be on monotherapy for at least 3 months and have a follow-up at most 9 months after initiation of monotherapy. If patients were eligible at multiple time-points, only the first instance was included. For models fit to all bDMARD subjects, an indicator for TCZ treatment was added. Finally, the most striking difference between the RWD and RCT cohorts was CDAI at baseline. For this reason, we evaluated models both for all RWD TCZm patients and for the subset of patients with baseline CDAI > 20, see tables 1,2.

Study Outcome (RA Remission)

The primary outcome of interest was disease remission at 24 weeks following initiation of monotherapy, to match the outcome of the RCTs used by Collins et al.[3] Remission was defined

by a Clinical Disease Activity Index (CDAI) < 2.8.[12] A benefit of the CDAI remission criteria is that it does not require access to a laboratory measurement and is thus widely available in RWD. In the Corrona RWD, remission status was evaluated at the follow-up visit closest to 24 weeks after start of monotherapy, but no sooner than 3 months or later than 9 months after initiation.

Potential Predictors

The variables used in the models derived from the RCTs were identical to the "OR" set used by Collins et al.[3] These included demographic variables, RA characteristics and previous use of DMARDs (biologic or non-biologic).The baseline variable set is listed in its entirety in **Table 1**.

The extended variable set derived from the RWD included additional disease activity scores; history of cancer, hypertension, rheumatoid factor, joint erosions & deformity; additional comorbidities; prescriptions of non-steroidal anti-inflammatory drugs and glucocorticoids; work status; education; general medical problems; physical disability; current and number of previous DMARDs. A full description of this set (EXT) is provided in the **Appendix**.

Most of the RWD were collected through questionnaires filled out at each Corrona patient visit (typically every six months) for each patient. Baseline features for the Corrona subjects were defined as the last recorded measurements taken *prior* to initiation of the target drug (TCZm or bDMARDm). In particular, if the target drug was prescribed for the first time between patient visits, the data of the last visit before prescription were used. Both baseline (OR) and extended

(EXT) variable sets were extracted from the registry data. Missing values were imputed with the R package MICE, see **Appendix.**

Statistical Analyses

The envisioned clinical use-case for the developed risk scores is to aid in treatment of *new* patients. Therefore, out-of-sample and out-of-distribution generalization are primary concerns. To address the former, we use sample splitting when deriving and validating models on RWD. For a single experiment, the full RWD cohort was first split at random into a derivation set and a validation set, the former used for fitting model parameters and the latter only for evaluation. The overall quality of each method was then computed as an average over a large number of repeated experiments. To assess out-of-distribution generalization, we evaluate the baseline model fit to the RCT cohorts on the RWD. The reverse (RWD to RCT) was not considered here as the extended variable set is not available in the RCTs. The primary quality metric was the area under the receiver-operating curve (AUROC) which measures the extent to which models successfully rank subjects' probability of remission. Standard errors were computed using the classical model of Hanley & McNeil.[14] Calibration of estimated probabilities, following Platt scaling fit to held-out data[15], was evaluated in the standard way.

Models from three different families were fit to multiply imputed data, pooled and evaluated: logistic regression (LR), L1-regularized logistic regression (LR-Reg) and random forests, see **Appendix for details**. For logistic models, fit to standardized variables, we use the magnitudes of regression coefficients as proxies for the variables' importance. For random forests, as is

common, we measure variable importance by a variable's ability to discriminate between subjects with different outcomes when used in splitting nodes in the trees, captured by the *mean decrease in impurity* (MDI).

Weighting of Subjects in the Extended Cohort

Extending the cohort to include patients on bDMARDs other than TCZ (see Question B) induces a shift between the derivation (all BDMARDm) and evaluation cohorts (only TCZm). In particular, TCZm patients make up a small proportion of the overall RWD cohort and would have limited impact on an unweighted model. To mitigate this, we made use of inverse propensity weighting, as is standard practice for handling distributional shift between treatment groups. A logistic regression model was fit to estimate the propensity of patients in the RWD to receive TCZm treatment compared to receiving any bDMARDm therapy (including TCZm). This was used to weight samples to emphasize those who had higher propensity to be put on TCZm. The weights were then used to fit weighted logistic regression and weighted random forest models tailored to TCZm patients.

RESULTS

Patient Sample Characteristics

From the RCTs, a total of 853 subjects were enrolled in the TCZm arms and had complete data. Among these, 80% were female and 80% were white. At baseline, the mean CDAI was 40.1 and 52% of subjects had been treated previously with both MTX and another DMARD. In the RWD,

out of 54,646 subjects, 3204 subjects were identified fitting our criteria for bDMARDm. 76% of these subjects were female and 93% white. The mean baseline CDAI was 17.4 and 83% were previously treated with both MTX and another DMARD. In the bDMARDm cohort, 452 were treated with TCZm.

Missingness at baseline in the RWD, before imputation, was low (< 2%) for variables in the original feature set, with the exception of disease duration (11%), ESR (30%) and HAQ-DI (25%). For the extended feature set, indicator variables representing certain past comorbidities, joint erosion, rheumatoid factor, smoking and previous pregnancy had high (> 30%) missingness. For evaluation of models fit to the RWD, the RWD was repeatedly split into a validation set, containing 50% (n=226) of TCZm subjects and 20% (n=550) of other bDMARDm subjects, and a derivation set containing remaining subjects.

Derivation and Validation of the Prediction Model

The full results of our evaluation of different sets of predictors (original/extended), models (LR, LR-Reg, random forest) and derivation sets (RCT, RWD TCZm, RWD bDMARDm) are presented in **Table 2**. Each combination was evaluated within two cohorts, each containing only TCZm subjects: the full RWD TCZm population and, for a closer comparison with RCTs, the subset of patients among these with baseline CDAI > 20. The cohorts were comparable on demographic characteristics (see **Table 1**). The calibration of the LR and random forest models, following Platt scaling, is illustrated in **Figure 1**.

All LR-Reg and random forest models trained on the extended feature set demonstrated larger AUROCs than models using the original feature set. For example, when trained on all bDMARDm subjects, LR-Reg (Extended) achieved 0.76 (95% CI 0.67, 0.84) AUROC compared to 0.72 (95% CI 0.63, 0.81) for LR (Original), with numbers in parenthesis indicating the 95% CIs. This suggests that there are gains to be made in predictive performance from including additional comorbidity, lifestyle and treatment variables in the risk score. Note, however, that the CIs overlap. We saw no advantage of using the random forest model over the regularized logistic regression model, in either setting, indicating that the remaining variance in the outcome is unlikely to be due to underutilized interactions or nonlinearities.

We found that expanding the derivation cohort to include non-TCZ bDMARDm patients improved the AUROC for both the TCZm and the TCZm high-CDAI cohorts. Compare, for example 0.75 (95% CI 0.66, 0.83) AUROC for LR (Extended) trained on bDMARDm to 0.65 (95% CI 0.54, 0.75) for the same model fit to TCZm patients only. For LR-Reg, the AUROCs were 0.76 (95% CI 0.67, 0.84) and 0.74 (95% CI 0.65, 0.82) when fitting to bDMARDm and TCZm, respectively. Comparable gains were seen for other models in the full group and for all models in the CDAI>20 group, but due to its small validation set, the variance in the latter results is high. The estimated probability of reaching remission was low (around 10%) for most subjects. Subjects who are more likely to reach remission than others may be identified by thresholding estimated probabilities. For thresholds at 10%, 12.5% and 15% the sensitivity/specificity of the best performing model, LR (Extended), were (0.75/0.54), (0.64/0.64), and (0.58/0.72), respectively. The original model, derived in the RCTs performed substantially worse in the high-

CDAI cohort than in the full TCZm cohort, even though the criterion CDAI>20 was meant to increase the similarity to the RCTs. However, as we can see in **Table 1**, significant differences in the cohorts remained. The results may be partially explained by higher variance in outcome for the high-CDAI cohort, after controlling for baseline disease severity.

For all models and derivation sets, different measures of disease severity (e.g., DAS28 and CDAI) at baseline were consistently highly predictive of remission. In **Table 3**, we list the features with highest estimated importance in the LR-Reg and random forest models trained on the extended feature set of the full bDMARDm cohort, ordered by feature importance (random forests) or coefficient magnitude (LR-Reg). The highest-ranked features are mostly unsurprising: the majority pertain to measures of disease severity either explicitly (CDAI, MDAS, global assessment) or implicitly (larger number of previous DMARDs). For the LR-Reg model, disability and unusual fatigue were associated with lower chances of remission.

DISCUSSION

Machine learning applied to real-world data may offer new opportunities to better define the course of disease and to identify better treatment strategies. In RA, the expanded treatment options present a challenge for clinicians and patience, as predictors for response to specific treatments are lacking. In prior work, we used RCT data to derive and validate predictors of remission among patients initiating TCZm.[3] In the current study, we tested this prediction rule using RWD and attempted to refine the prediction rule using machine learning. We found that

the original prediction rule held up well in RWD from Corrona, despite notable differences between the RCT and RWD populations. This and the fact that the original rule contains only commonly available variables points toward the feasibility of implementing these rules in clinical practice.

The implications of this work are several. First, we examined the validity of prediction models derived and validated in RCTs in RWD. RCTs, while appropriate for estimating average treatment effects on the selected cohort, are limited in their generalizability to a broader population. RWD offer an insight into how patients are treated in the healthcare system, and what their outcomes are in the absence of strict inclusion criteria and potential experiment effects. Thus, RWD provide a "laboratory" for testing findings from RCTs. Indeed, we found that the RCT and RWD cohorts differ substantially in terms of disease duration and severity, as well as treatment history; the FUNCTION trial enrolled subjects that were MTX naïve with short disease duration[4] while the other RCTs enrolled subjects that showed inadequate response to MTX.[7-9] Despite this, the discrimination between patients was good for the transferred model. This indicates that a) good predictors in the RCTs are good predictors in the RWD, and b) the variables observed for at baseline appropriately controlled for the cohort differences. However, the overall performance characteristics (sensitivity and specificity) of our models was modest and gains could be expected with additional data. Thus, the current prediction rules should not be used in daily practice without further improvements.

Second, we developed and validated in RWD several prediction models for remission with TCZm. As noted above, TCZ when given as monotherapy seems to be more effective than other bDMARDs as monotherapy.[4] The variables identified in our prediction rule were disabled working status, prior DMARDs, and baseline disease severity. These variables are not surprising and may be obvious to some clinicians. A prediction rule attempts to put together variables (some new and some old) that have never been put together in one prediction rule that may have utility in the clinic. They also point out the value of several non-clinical variables. We anticipate further refining this rule in other datasets with more variables and less missing data. After improving the rule, we may test it among other bDMARDs; it may be that future iterations of this rule could be programmed in an electronic medical record and help clinicians and patients identify therapy likely to be effective in patients with a given set of characteristics.

Finally, this set of analyses used a robust RWD dataset, Corrona, with an expanded set of variables. Because of the presence of many potentially correlated variables, we used machine learning algorithms to analyze these data. We recognize that disabled working status is contained in the HAQ-DI, so that there is the potential for overlap between potential predictor variables. In high-dimensional settings such as these, machine learning may be used together with sample splitting to discover models with reduced predictive variance at the cost of a small increase in bias.[16] This was confirmed in this work, in particular for the regularized logistic regression models. Such an approach is appropriate particularly in applications where out-of-sample prediction is the goal, rather than parameter identification. The benefits of machine learning are smaller when domain knowledge is strong enough to identify a successful model

without the need to search over a large set of variables. In some cases, a model with slightly lower predictive accuracy may be preferred if it is easier to interpret, explain or communicate.[19]

Strengths of the current analyses include the validation of a previously derived and validated algorithm using RWD as an external validation dataset. However, several limitations should be noted. The work needs to be expanded to consider the prediction rule across other bDMARD; this is planned future work. We had significant rates of missing values for some variables in the RWD; this is typical but likely has some impact on model fit. Our strategy of pooling model estimates on multiply imputed data is standard practice, but not immune to bias. Corrona encompasses patients from North America only; while more generalizable and much larger than many RCTs, this is a limitation.

In conclusion, we were able to test a prediction rule for remission with TCZm among patients with RA in RWD and found that it worked well. Additional variables enhanced the prediction rule further. Moreover, using data from other bDMARDs allowed us to improve the model fit. We encourage other investigators to derive and validate prediction models for RA treatment across RCTs and RWD. Machine learning algorithms may play important roles in optimizing prediction rules.

REFERENCES

1. Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis: A Review. JAMA 2018;320:1360-1372.
2. Boire G, Allard-Chamard H. The 4-H of Biomarkers in Arthritis: A Lot of Help, Occasional Harm, Some Hype, Increasing Hope. J Rheumatol 2019;46:758-763.
3. Collins JE, Johansson FD, Gale S, Kim SC, Shrestha S, Sontag D, et al. Predicting Remission Among Patients With Rheumatoid Arthritis Starting Tocilizumab Monotherapy: Model Derivation and Remission Score Development. ACR Open Rheumatol 2020;2:65-73.
4. Gabay C, Emery P, van Vollenhoven R, Dikranian A, Alten R, Pavelka K, et al. Tocilizumab monotherapy versus adalimumab monotherapy for treatment of rheumatoid arthritis (ADACTA): a randomised, double-blind, controlled phase 4 trial. Lancet 2013;381:1541-1550.
5. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017;102:924-933.
6. Kremer J. The CORRONA database. Ann Rheum Dis 2005;64 Suppl 4:37-41.
7. Dougados M, Kissel K, Conaghan PG, et al. Clinical, radiographic and immunogenic effects after 1 year of tocilizumab-based treatment strategies in rheumatoid arthritis: the ACT-RAY study. Ann Rheum Dis 2014;73:803-809.
8. Burmester GR, Rigby WF, van Vollenhoven RF, Kay J, Rubbert-Roth A, Blanco R, et al. Tocilizumab in early progressive rheumatoid arthritis: FUNCTION, a randomised controlled trial. Ann Rheum Dis 2016;75:1081-1091.
9. Jones G, Sebba A, Gu J, Lowenstein MB, Calvo A, Gomez-Reino JJ, et al. Comparison of tocilizumab monotherapy versus methotrexate monotherapy in patients with moderate to severe rheumatoid arthritis: the AMBITION study. Ann Rheum Dis 2010;69:88-96.
10. Breiman L. Random forests. Machine Learning 2001;45:5-32.
11. Friedman JH, Popescu BE. Predictive learning via rule ensembles. The Annals of Applied Statistics 2008;2:916-954.
12. Aletaha D, Smolen J. The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. Clin Exp Rheumatol 2005;23:S100-108.
13. Greenberg JD, Kremer JM, Curtis JR, Hochberg MC, Reed G, Tsao P, et al. Tumour necrosis factor antagonist use and associated risk reduction of cardiovascular events among patients with rheumatoid arthritis. Ann Rheum Dis 2011;70:576-582.
14. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-843.
15. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 1999;10:61-74.
16. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research 2011;12:2825-2830.
18. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Paper presented at: Proceedings of the 9th Python in Science Conference. 2010.

19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 2019;1:206-215.

Table 1: Baseline characteristics. N (%) or median (IQR). Variables in the original set across RWD (All, TCZm), RCTs, following imputation. RWD is also stratified by baseline CDAI > 20 to achieve a closer comparison to the RCTs. The extended variable set used in our analysis included variables representing: additional disease activity scores; history of cancer, hypertension, rheumatoid factor, joint erosions & deformity; comorbidities; prescriptions of NSAIDs and steroids; work status; education; general medical problems; physical disability; current and number of previous DMARDs.

Characteristic:	**All RWD** (n=3204)	**TCZm** (n=452, RWD)	**TCZm, CDAI>20** (n=240, RWD)	**RCTs** (n=853)
Age, years	57.0 (48.0, 66.0)	59.0 (49.0, 67.0)	59.0 (49.8, 67.0)	53.0 (44.0, 61.0)
Sex: Female	2439 (76.4%)	370 (82.0%)	194 (80.8%)	680 (79.7%)
Race: White	2932 (93.0%)	420 (94.4%)	227 (96.2%)	680 (79.7%)
BMI, kg/m^2	28.6 (24.9, 33.7)	28.7 (24.7, 33.7)	28.3 (24.0, 34.1)	26.5 (23.5, 30.5)
HAQ-DI	1.0 (0.4, 1.5)	1.2 (0.6, 1.6)	1.4 (1.0, 1.9)	1.6 (1.1, 2.0)
ESR, mm/hr	17.0 (8.0, 33.0)	16.0 (7.0, 36.0)	15.0 (6.5, 38.0)	40.0 (30.0, 60.0)
Hematocrit, %	40.0 (37.5, 43.0)	40.0 (37.4, 42.8)	39.8 (37.9, 42.9)	40.3 (40.1, 40.5)
Disease duration, years	8.0 (3.0, 16.0)	10.0 (5.0, 17.0)	10.0 (5.0, 18.0)	1.8 (0.5, 7.3)
Past DMARD/MTX:				
Both No	215 (6.7%)	14 (3.1%)	5 (2.1%)	291 (34.1%)
Both Yes	2669 (83.3%)	400 (88.5%)	216 (90.0%)	441 (51.7%)
DMARD Yes / MTX No	320 (10.0%)	38 (8.4%)	19 (7.9%)	121 (14.2%)
DMARD: TCZ	452 (14.1%)	452 (100.0%)	240 (100.0%)	853 (100.0%)
Baseline CDAI	17.4 (8.8, 28.0)	21.5 (12.0, 32.5)	32.0 (25.5, 40.3)	40.1 (30.7, 49.4)
Follow-up duration, weeks	32.7 (25.1, 52.7)	30.9 (25.0, 51.2)	28.1 (24.8, 42.1)	24.0 (24.0, 24.0)
Remission	563 (17.6%)	53 (11.7%)	13 (5.4%)	127 (14.9%)

Notes. RWD (Real-world data)

Table 2. Discrimination in prediction of remission, measured by the area under the ROC curve (AUROC), for different models evaluated in the TCZm and TCZm (high CDAI) cohorts. Parentheses indicate 95% CIs. We compare models trained using the original variable set from Collins et al., (2019) and the extended feature set described in the Methods section. Additionally, we compare learning from only TCZm patients and learning from patients on any biologic DMARD monotherapy (bDMARDm). Cohort sizes (n=X) refer to the size of the respective validation set. AUROC near 0.5 is no better than random selection.

Model	**Variables**	**AUROC, TCZm (n=226)**	**AUROC, TCZm, CDAI>20 (n=120)**
Model from Collins et al., (2019), trained on all RCT patients			
LR	Original	0.70 (0.64, 0.77)	0.56 (0.41-0.72)
Training on only RWD TCZm patients in derivation set			
LR	Original	0.68 (0.58, 0.78)	0.47 (0.23, 0.71)
LR	Extended	0.61 (0.51, 0.72)	0.52 (0.30, 0.75)
Random forest	Extended	0.74 (0.65, 0.83)	0.63 (0.45, 0.84)
LR-Reg	Extended	0.73 (0.64, 0.82)	0.67 (0.48, 0.85)
Training on all RWD bDMARDm patients in derivation set, evaluating on TCZm cohort			
LR	Original	0.73 (0.64, 0.82)	0.56 (0.34, 0.78)
LR	Extended	0.74 (0.65, 0.82)	0.67 (0.48, 0.86)
Random forest	Extended	0.76 (0.68, 0.84)	0.68 (0.50, 0.86)
LR-Reg	Extended	0.77 (0.68, 0.85)	0.72 (0.55, 0.89)

Notes. AUROC (Area under the receiver-operating characteristic curve) TCZ (Tocilizumab), TCZm (TCZ monotherapy), CDAI (Clinical Disease Activity Index), LR (Logistic Regression), LR-Reg (Regularized Logistic Regression), bDMARD (Biologic DMARD), RWD (Real-world data)

Table 3. Features with highest estimated importance measured by the mean decrease in impurity (MDI) for random forests and the magnitude of coefficients $|\beta|$ for LR-Reg.

Random Forest (Extended)	**MDI**	**LR-Reg (Extended)**	$\boldsymbol{\beta}$ **(95% CI)**
CDAI	0.10	In remission at baseline	0.30 (0.14, 0.39)
DAS	0.09	DAS	-0.30 (-0.38, -0.14)
MD Global Assessment	0.07	Work status: Disabled	-0.29 (-0.46, -0.25)
HAQ-DI	0.06	Past steroid prescription	-0.27 (-0.36, -0.07)
Remission at baseline	0.05	Education: High-school or less	-0.25 (-0.39, -0.15)
Trouble dressing self	0.04	Sex: Female	-0.25 (-0.36, -0.11)
Num. past DMARDs	0.04	Past DMARD/MTX: Both Yes	-0.25 (-0.36, -0.14)

Notes. LR-Reg (Regularized Logistic Regression), Extended (Extended variable set), CDAI (Clinical Disease Activity Index), DAS (Disease Activity Score), HS (High-school), HAQ-DI (Health Assessment Questionnaire without Disability Index). Confidence intervals for LR-Reg coefficients were computed using the empirical bootstrap over the derivation set. This method was chosen due to the inclusion of regularization and sample weighting in the procedure. Subjects were resampled with replacement and the propensity and outcome models were fit to each bootstrap sample. The stated results are for the best tuning parameters from the experiment presented in Table 2.

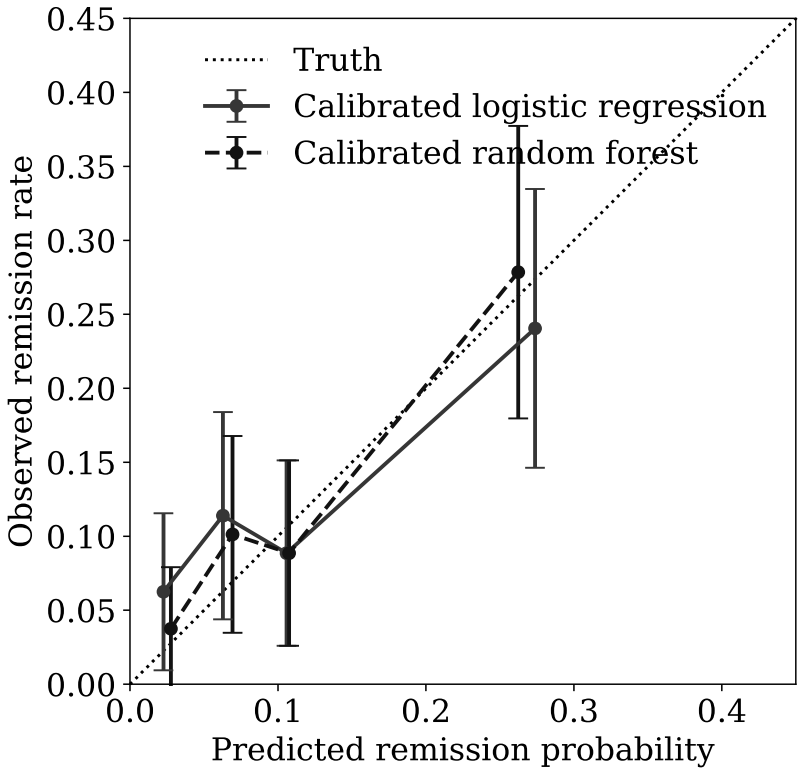

Figure 1. Calibration of logistic regression (LR) and random forest models trained on all bDMARDs in RWD using the extended feature set and evaluated on held-out TCZm patients from the RWD. The predictions of each model have been adjusted using Platt scaling. Calibration is assessed in the four quartiles of predicted remission probability. We note that the majority of patients (75%) have a predicted probability of remission around or below 0.1.