Title: Development and validation of a prognostic model for leflunomide discontinuation with abnormal blood-tests during long-term treatment: cohort study using data from Clinical Practice Research Datalink Gold and Aurum.

Authors: Georgina Nakafero[1], Matthew J. Grainge[2], Tim Card[2,3], Maarten W. Taal[4], Guruprasad P. Aithal[3,5] Weiya Zhang[1], Michael Doherty[1], Christopher P. Fox[6], Christian D. Mallen[7] and Abhishek Abhishek[1,5].

Affiliation: [1]Academic Rheumatology, University of Nottingham, Nottingham, UK. [2]Population and Lifespan Sciences, University of Nottingham, Nottingham, UK. [3]Nottingham Digestive Diseases Centre, School of Medicine, University of Nottingham, Nottingham, UK. [4]Centre for Kidney Research and Innovation, University on Nottingham, Nottingham, UK. [5]NIHR Nottingham BRC, Nottingham University Hospitals NHS Trust and the University of Nottingham, Nottingham, UK. [6]Department of Haematology, Nottingham University Hospital NHS Trust, Nottingham, UK. [7]Primary Care Centre Versus Arthritis, School of Medicine, Keele University, Keele, UK.

Corresponding author: Georgina Nakafero

Address for correspondence:

G Nakafero

A23, Academic Rheumatology,

Clinical Sciences Building,

School of Medicine,

The University of Nottingham, Nottingham, UK

Email: georgina.nakafero@nottingham.ac.uk

Phone: 01158231655

ORCiD iD: https://orcid.org/0000-0002-3859-7354

Abstract:

Objective: To develop and validate a prognostic model for leflunomide discontinuation with abnormal blood-test results.

Methods: Data from CPRD Gold and Aurum were used for model development and external validation respectively. Participants prescribed leflunomide between 01/01/2007 and 31/12/2019 were followed-up from six-months after first GP-prescription to the earliest of date of outcome, death, 5-year follow-up or 31/12/2019. Candidate prognostic factors were ascertained using theory and data driven approaches. Penalised Cox regression was performed to develop the risk equation, followed by internal validation using 500-bootstraps to correct for optimism. Multiple imputation was applied to handle missing data. Model performance was assessed in terms of calibration and discrimination.

Results: Data for 1,487 and 2,329 participants contributing 3,140 and 5,246 person-years follow-up were included in the development and validation cohorts, respectively. Thirteen candidate predictors were included in the model. Epilepsy, and either cytopenia or elevated liver enzymes during first six months of shared-care leflunomide prescription were strong predictors of drug discontinuation with hazard ratio (95%CI) 4.39 (1.74 -11.06) and 3.06 (2.15 - 4.35), respectively. The unadjusted and optimism adjusted calibration slope in development data was 1.00 (95% CI 0.75-1.25) and 0.72 (95% CI 0.47-0.97), respectively. The calibration slope in validation data was 0.91 (95% CI 0.74-1.07). The model showed prognostic separation with optimism adjusted Royston D statistic of 0.73 (95% CI 0.44-1.02).

Conclusion: We have developed and externally validated an easy-to-use prognostic model that may be used to risk-stratify monitoring for leflunomide toxicity and to make informed choices about risks when choosing treatments.

Keywords:

leflunomide, rheumatoid arthritis, psoriatic arthritis, drug toxicity, monitoring

Rheumatology key messages

- One in five patients established on long-term leflunomide discontinue treatment with abnormal monitoring blood-tests.
- This is the first prognostic model to discriminate patients at varying risk of leflunomide toxicity.
- The developed tool may be used to risk-stratify monitoring after successful stabilisation on leflunomide.

Introduction

Leflunomide is used in the treatment of inflammatory arthritis when low-dose weekly methotrexate is either contraindicated, ineffective, or causes side-effects (1). Although head-to-head trials suggest comparable efficacy to methotrexate ≤15 mg/week, leflunomide is less well tolerated, with a higher risk of treatment discontinuation, mainly due to cytopenia and elevated liver enzymes (2-5). For instance, up to 7.1% patients commenced on leflunomide discontinued it by 12 months due to elevated liver enzymes in clinical trials (2, 4). Real world data indicates that 9.3% and 20.5% patients initiated on long-term leflunomide discontinue treatment with abnormal blood test results by 1-year and 5-years, respectively (5).

The risk factors for target-organ damage from leflunomide are not well understood. In the absence of this information, those prescribed long-term leflunomide undergo monitoring blood tests every three months (6, 7). This strategy of routine periodic testing may not be necessary for those at low risk. Additionally, better understanding of predictors for target organ damage will aid patients and rheumatologists when choosing disease modifying anti-rheumatic drugs (DMARDs). Thus, the aim of this study was to develop and externally validate a prognostic model for leflunomide discontinuation due to abnormal monitoring blood-tests at 5-years.

Methods

Data source

Data from Clinical Practice Research Datalink (CPRD) Gold and Aurum were used for model development and external validation respectively (8, 9).

CPRD is an anonymised longitudinal database of electronic health records, and its' participants are representative of the UK population in terms of age, sex, and ethnicity (8). It includes information on demographic details, lifestyle factors (e.g., smoking, alcohol intake), diagnoses, results of investigations including blood tests, and details of general practitioner (GP) prescriptions during clinical care.

CPRD Gold and Aurum complement each other in terms of nationwide coverage of general practice surgeries. The former uses Vision software while the latter uses EMIS. Some general practice surgeries have contributed data to both CPRD Gold and Aurum databases. Data from such surgeries were excluded from the validation cohort using a bridging file provided by the CPRD to allow for true independent external validation.

Approvals

Ethical approval was obtained from the Independent Scientific Advisory Committee (ISAC) of the Medicines and Healthcare Products Regulatory Agency (Reference: 19_275R).

Study design

This was a cohort study. Study period was 1st January 2007 to 31st December 2019.

Study population comprised those who received first shared-care leflunomide prescription from GP in study period.

In the UK, DMARDs are initiated in hospital rheumatology clinic and prescriptions are initially issued by the rheumatologist until a stable, effective, and well tolerated dose

is reached. During this period, the rheumatology team oversees monitoring blood-tests. Once the patient is established on treatment, the responsibility for prescribing and monitoring is handed to the patients' GP under a shared-care protocol endorsed by the British Society for Rheumatology (BSR) and the Royal College of General Practitioners (6). The GP consults with the rheumatologist if there are abnormal blood-test results or any side-effects, and changes in treatments are directed by the rheumatologist.

Inclusion and exclusion criteria

Participants with autoimmune rheumatic disease (AIRD, e.g. rheumatoid arthritis, axial spondyloarthritis etc.), age ≥ 18 years, with ≥ 12-month follow-up in CPRD Gold (Aurum for validation) prior to first ever prescription of leflunomide were eligible (5). Exclusion criteria comprised of chronic liver disease, haematological malignancy, myelodysplasia, haemolytic anaemia, neutropenia, idiopathic thrombocytopenic purpura, or chronic kidney disease (CKD) stage ≥4 as detailed previously (5).

Outcome

Drug discontinuation with abnormal blood-test result, defined as a prescription gap of ≥ 90 days, with abnormal blood-test result (or diagnostic code indicating abnormal blood-test result) within ±60 days of the last prescription (5). See the Supplementary Methods (available at *Rheumatology* online) for thresholds used to define abnormal blood-test results.

Start of follow-up: Participants were followed-up from 180 days after the first leflunomide prescription issued by the GP until the earliest of date of outcome, death, transfer out of the practice, date of last data collection from the practice, 5-years or 31/12/2019.

Predictors

Predictors were ascertained using theory and data driven approaches.

(A) Theory driven: Clinical members of the team comprising a hepatologist, nephrologist, haematologist, rheumatologist, gastroenterologist, and GP suggested potential predictors. These were supplemented with drugs that increase the risk of leflunomide toxicity according to the British National Formulary (BNF).

(a) Demographic or lifestyle factors. Age, sex, body mass index (BMI), and alcohol intake were included as they increase the risk of drug induced liver injury (DILI) and smoking was included as it increases the clearance of leflunomide (10, 11).

(b) Drugs that increase the risk of leflunomide toxicity as per BNF, specifically statins, paracetamol, methotrexate, 5-acetyl salicylates, carbamazepine, and sodium valproate.

(c) Comorbidities. Diabetes was included as it increases the risk of DILI (10).

(d) Cytopenia (neutrophil count <2 x 10^9/l, total leucocyte count <4 x 10^9/l, or platelet count <150 10^9/l,) or liver enzyme elevation (ALT/AST levels >35 IU/l) during the first six months of shared-care leflunomide prescription were included. This is because blood-test abnormalities predict cytopenia and/or transaminitis due to other DMARDs (12, 13).

The latest record of demographic and lifestyle factors prior to start of follow-up, diagnostic code for comorbidities in the 2-years prior to start of follow-up, and prescription and blood-test results in the six-month prior to start of follow-up were used to define the prognostic factors. A longer look-back was used to capture data on comorbidities as GPs usually review patients with chronic illnesses annually.

(B) Data driven: All diagnoses for study participants within 2-years of start of follow-up were extracted and classified into chronic disease categories. Hypothesis-free logistic regression adjusted for age and gender was undertaken to identify potential prognostic

factors that associate with outcome of interest. Potential risk factors associated with outcome with $p < 0.10$ and present in >=1% of the derivation cohort were included in the prognostic model. Uncommon prognostic factors were excluded to avoid model imbalance.

Sample size

To minimise model overfitting and ensure precise estimation of overall risk, the minimum sample size required for new model development is 1398 participants (189 events) based on a maximum of 20 parameters, Cox-Snell R^2 value of 0.12, estimated event rate of 0.057/person-year, a 5-year time horizon, and a mean follow-up period of 2.36 years using the findings from our earlier work (5). (See Supplementary Methods, available at *Rheumatology* online, for Stata syntax)

Statistical analysis

Mean (standard deviation (SD)) and n (%) were used for descriptive purposes. We applied multiple imputation to handle missing values using chained equations. We carried out 10 imputations in the development dataset as there tends to be no additional benefit for using more than 5-10 imputations (14). We used five imputations for the validation data - a pragmatic approach considering the large size of CPRD Aurum. The imputation model included all candidate predictors, Nelson-Aalen cumulative hazard function and outcome variables.

Model development

All candidate predictors were included in the Cox model and coefficients of each predictor estimated and combined using Rubin's rule across the imputed datasets. We formed the risk equation for predicting an individual's risk of leflunomide discontinuation due to abnormal blood-test results at 5-years follow-up, using the developed model's baseline survival function at *t=5 years*, a non-parametric estimate

of survival function when all predictor values are set to zero, which is equivalent to the Kaplan-Meier product-limit estimate, along with the estimated regression coefficients (β) and the individual's predictor values (X). This process ultimately led to an equation for the predicted absolute risk over time (15):

Predicted event risk at 5-years $=1 - S_0(t_{=5})^{\exp(X\beta)}$ where $S_0(t_{=5})$ is the baseline survival function at 5-years of follow-up and βX is the linear predictor, $\beta_1x_1+ \beta_2x_2+ \ldots + \beta_px_p$. Regression coefficients (β) are estimated from the developed model.

Model validation

We assessed the performance of the model in terms of calibration (where 1.00 is the ideal) by plotting agreement between predicted and observed events. We performed internal validation to correct calibration for optimism (overfitting) by bootstrapping with replacement 500 samples of the development data in each imputed dataset. We fitted the full model in each bootstrap sample to quantify performance in bootstrap sample (apparent performance) and applied the same model to the original sample to test model performance and optimism (difference in test performance and apparent performance). Uniform shrinkage factor was then estimated as the average of calibration slopes from each of the bootstrap samples. This process was repeated in each imputed dataset, and the final uniform shrinkage calculated by averaging across the estimated shrinkage estimates from all imputations. To account for overfitting during model development process, the original β coefficients were penalised by the final uniform shrinkage factor and the baseline hazard re-estimated on the basis of the shrunken β coefficients to ensure that overall calibration was maintained, producing a final model. We calculated the D statistic, a measure of discrimination, interpreted as a log hazard ratio (HR), the exponential of which gives the HR comparing two groups defined by above/below the median of the linear predictor, and plotted Kaplan-Meier

curves in risk groups to visually assess separation. The cut-points are the 16th, 50th and 84th centiles of the linear predictor (mean +/- 1 SD) as determined by Cox's method (16, 17).

External validation of the model

Independent external validation of the final model was performed using data from CPRD Aurum within the same start and end of follow-up periods. General practice surgeries that also contributed data to CRPD Gold were excluded from the validation cohort. The final developed model equation was applied to each individual in the validation dataset, and then we examined calibration and discrimination as described above. In addition, we examined calibration at 5 years by plotting agreement between predicted risk and observed event rate.

We used Stata-MP version 16 for all statistical analyses. This study was reported in line with the transparent reporting of a multivariate prediction model for individual prediction or diagnosis (TRIPOD) guidelines (18).

Results

Study participants

Data for 1,487 and 2,329 participants contributing 3,140 and 5,246 person-years follow-up were included in the development and validation cohorts, respectively (Table 1; Supplementary Figures S2 and S3, available at *Rheumatology* online). The majority of participants in both cohorts had rheumatoid arthritis, were female and the cohorts had similar prevalence of lifestyle factors, comorbidities and drug treatments.

On data-driven analyses in the derivation cohort, epilepsy, CKD and nutritional intolerances were associated with the outcome of interest with $p < 0.10$ (Supplementary Table S1, available at *Rheumatology* online). As nutritional intolerances were only present in 0.15% of the derivation cohort, it was not taken forward as a candidate predictor. A diagnosis of epilepsy and prescription of sodium valproate or carbamazepine was merged together to create a single candidate predictor epilepsy to avoid multicollinearity. We used fraction polynomials to model non-linear risk relationships with continuous predictors (BMI and age) but these were found not to be better than the linear terms, hence BMI and age were not transformed (data not shown). Thirteen candidate predictors (17 predictor parameters) were selected to be included in the model (Table 2).

Model development and identification of candidate predictors

In the development dataset, 136 outcome events occurred during the follow-up period at a rate (95% CI) of 43.32 (36.62 - 51.25) per 1,000 person-years. Epilepsy, and presence of cytopenia or elevated liver enzymes during the first six months of shared-care leflunomide prescription were strong predictors of leflunomide discontinuation with adjusted hazard ratio (95%CI) 4.39 (1.74 -11.06) and 3.06 (2.15 - 4.35) respectively (Table 2).

Apparent and internal validation performance statistics

As expected, the calibration slope in the development data was 1.00 (95% CI 0.75-1.25). From the bootstrap a uniform shrinkage factor of 0.73 was obtained and used to shrink predictor coefficients in the final model for optimism (Table 3), and after re-estimation the final model's $S_0(5)$ was 0.914.

Royston *D* statistic was 1.06 (95% CI 0.77 – 1.35), corresponding to HR (95% CI) 2.89 (2.16-3.86) comparing the risk group above the median of linear predictor to that below the median. The optimism adjusted Royston *D* statistic was 0.73 (95% CI 0.44-1.02) corresponding to HR (95% CI) 2.08 (1.55-2.77).

External validation

In the CPRD Aurum cohort, there were 260 outcome events at a rate (95% CI) of 49.94 (44.25-56.37) per 1000 person-years. Application of our final prognostic model to the independent population from CPRD Aurum yielded excellent calibration, with a calibration slope (95% CI) of 0.91 (0.74-1.07) (Figure 1). The Royston D statistic in the validation data was 0.97 (95% CI 0.89 -1.05), corresponding to HR (95% CI) 2.64 (2.44 -2.86) which suggests that our prediction model provided similar prognostic separation to the development dataset. Model discrimination in the derivation and validation data was broadly similar but the model seemed less able to distinguish between the lowest two risk groups, particularly in the validation data (Figures 2). The observed (and predicted) 5-year survival probabilities in validation data in these four risk groups were similar: 0.87 (0.90), 0.84 (0.87), 0.73 (0.79), 0.56 (0.59) respectively.

Worked examples

A prognostic score to predict the absolute risk of leflunomide discontinuation after six months of primary care prescription and within the next 5-years may be calculated using the risk-equation (Figure 3, Supplementary Figure S1, available at

Rheumatology online). Participants with 16th centile and median linear predictor scores had 10.8% and 15.7% absolute risk of outcome event respectively over the 5-year follow-up period in the development datasets. The corresponding values were 10.9% and 15.3% in the validation dataset.

Discussion

This is the first study to develop and validate a prognostic model that predicts leflunomide discontinuation due to target organ damage. It includes routinely collected data and provides a readily applicable means of risk stratification. It has excellent calibration and good discrimination between higher and lower risk groups. It focussed on patients successfully initiated on leflunomide and treated for >6 months as this includes majority of burden of monitoring. Current guidelines recommend three-monthly blood-test monitoring during long-term leflunomide treatment and more frequent monitoring in context of polypharmacy or comorbidities (6, 7). However, with the exception of concurrent methotrexate prescription, these factors are poorly understood (19). Utilising the results from this study, patients at high-risk of leflunomide toxicity may be offered more careful monitoring or alternate treatments, while those at very-low risk may undergo less frequent monitoring e.g. six-monthly testing. Additionally, this study reports that cytopenia and elevated liver enzymes including those not sufficiently severe to withdraw treatment within first six-months of shared-care GP prescription strongly predict target-organ drug-toxicity. This is a novel finding for leflunomide and is consistent with previous observations regarding methotrexate (12, 13). Similarly, epilepsy and/or treatment with carbamazepine and sodium valproate was a strong predictors of target-organ drug toxicity. These data may help inform drug choice in these patients. Statins and paracetamol were also strong prognostic factors while other DMARDs such as methotrexate and 5-ASA were weak prognostic factors. We did not observe a statistically significant association between demographic and lifestyle factors including alcohol excess, and AIRD type and outcomes of interest. There is week evidence that alcohol consumption may be a risk factors for DILI due to specific drugs such as methotrexate, but not with other

drugs (20). Alcohol use in the preceding 12 months was a negative predictor of severe DILI in general (OR (95%CI) 0.33 (0.15–0.76) in a previous study(20). These findings should be interpreted with caution as our study was not powered to detect these associations.

Overall, the prognostic model performed well in the external validation dataset with excellent calibration. It had low discriminant ability for those at very-low and low predicted risk. This is unsurprising as the absolute difference in risk over a 5-year horizon between these two groups was only 5%. Reassuringly, our model discriminated between low and high-risk subsets which it could be argued is important for clinical application. In future, discrimination may be improved by including variants associated with leflunomide transaminitis (e.g. C163A in CYP1A2 gene; and rs4244285 and rs12248560 in CYP2C19 gene); reduced leflunomide metabolism (e.g. rs3213422 in dihydroorotate dehydrogenase gene) and excretion (rs2231137 in ABCG2 gene, also linked with gout) (21-26).

Not all prognostic models change practice. To facilitate this, evidence from this study will be disseminated to the BSR DMARD monitoring guideline writing group and the monitoring strategy will be changed if the BSR recommendations are modified in light of the findings. The risk calculators will be available online and included in the in-practice software used by GPs.

Strengths of this study include adequate power, use of time to event methods, external validation in an independent dataset, and the inclusion of prognostic factors that are simple to obtain during routine care, and at no additional cost. We followed TRIPOD guidelines and used robust statistical methodology to develop and evaluate the prognostic model. The study included internal correction for optimism and missing data was estimated by multiple imputations. Generalisability of the model was enhanced by

the use of a database with nationwide coverage. We used an exhaustive list of potential predictors using data driven and theory driven approaches.

However, there are several limitations of this study. Firstly, dose reduction due to abnormal blood test results was not used to define the outcome as 30% of data on leflunomide dose is missing in the CPRD making it difficult to ascertain dose reductions (5). Some outcomes may have been related to toxicity to other drugs. These two factors may have reduced our model's performance due to misclassification bias. Secondly, it is possible that some outcome events may actually be due to a combination of lack of efficacy of leflunomide and a concurrent illness resulting in blood test abnormality. However, our validation exercise revealed that 95% of outcome events were not explained by a concurrent illness (5). Patients prescribed leflunomide from rheumatology clinic were excluded from the study. However, this is unlikely to affect the generalisability of our findings as vast majority of long-term prescriptions in the UK are issued from primary-care under shared-care prescribing and monitoring agreement. Our development dataset had a high shrinkage factor indicating a degree of overfitting.

In conclusion, we have developed and validated a risk prediction equation to quantify the absolute risk of leflunomide discontinuation due to abnormal monitoring blood-test results over 5-years. We ascertained several strong risk factors that may be useful when choosing between DMARDs. Further research is warranted to validate the model in other populations and to evaluate the clinical outcomes using this model.

Funding: This work was funded by the National Institute for Health Research (NIHR) under its Research for Patient Benefit Programme (Grant Reference Number PB-PG-1217–20030)

Disclosure statement: This article presents independent research funded by the National Institute for Health Research (NIHR) under its Research for Patient Benefit Programme (Grant Reference Number PB-PG-1217–20030). The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR or the Department of Health and Social Care. C.D.M. is funded by the NIHR Applied Research Collaboration West Midlands, the NIHR School for Primary Care Research and a NIHR Research Professorship in General Practice (NIHR-RP-2014–04-026) for this research project. The study sponsor did not have any role in the conduct or reporting of this study.

A.A. has received departmental research grants from AstraZeneca and Oxford Immunotec, speaker bureau fees from Menarini, scientific meeting support from Pfizer, consulting fees from Inflazome and author royalties from UpToDate and Springer, unrelated to this work. M.D. has received honoraria for attending Ad hoc advisory boards on gout and osteoarthritis for Grunenthal, Mallinckrodt and Pfizer, and author royalties from UpToDate, and was an investigator in an AstraZeneca-funded, investigator-led, non-drug study (the 'Sons of Gout' study), unrelated to this work. W.Z. has received honoraria from Regeneron and Eli Lilly for advice on treatment of OA. G.P.A. reports consulting fees from Astrazenca, Amryt Pharma, FRACTYL, Median technologies, Bergen Bio ASA; advisory fees from Kandy therapeutics, GSK, Owlstone, Inventiva Pharma; research grant support from Preglem, Pfiezer inc; and meeting support from Roche Diagnostics unrelated to this work. The other authors have no conflict of interest to declare.

Data availability: This study used data from the Clinical Practice Research Datalink (CPRD). Due to the CPRD data sharing policy, we unable to share this study's data. However, access to CPRD data can be directly requested from the CPRD.

Patient and Public Involvement (PPI): The study question was discussed at a PPI meeting in Nottingham and received support from all present. Study results were reported to PPI group and modes of dissemination of study findings were also discussed and agreed with them.

Table 1: Baseline characteristics of study population

Variable[1]	Development cohort (CPRD Gold) n=1,487	Validation cohort (CPRD Aurum) n=2,329
Age, mean (SD) year	57 (13)	57 (13)
Female sex	979 (65.8)	1,580 (67.8)
BMI		
<18.5 kg/m^2	28 (1.9)	28(1.2)
18.5-24.9 kg/m^2	426 (28.7)	651 (28.0)
25.0-29.9 kg/m^2	470 (31.6)	728 (31.3)
≥30 kg/m^2	495 (33.3)	821(35.3)
Missing	68 (4.6)	101(4.3)
Current smoker		
No	1,168 (78.6)	1,878 (80.6)
Yes	319 (21.5)	451 (19.4)
Alcohol use		
Non-user	329 (22.1)	519 (22.3)
Low (1-14 units/week)	805 (54.1)	931 (40.0)
Moderate (15-21 units/week)	43 (2.9)	109 (4.7)
Hazardous (>21 units/week)	76 (5.1)	112 (4.8)
Ex-user	88 (5.9)	354 (15.2)
Missing	146 (9.8)	304 (13.1)
Autoimmune rheumatic disease		
Rheumatoid Arthritis	970 (65.2)	1,518 (65.2)
Polymyalgia rheumatica/giant cell arteritis	91 (6.1)	201 (8.6)
Spondyloarthritis	426 (28.7)	610 (26.2)
Comorbidities		
Epilepsy or prescribed carbamazepine or valproate	19 (1.3)	26 (1.1)
Diabetes	149 (10.2)	278 (11.9)
Chronic kidney disease	74 (5.0)	57 (2.5)
Other DMARDs		
Methotrexate or 5-aminosalicylates	467 (31.4)	758 (32.6)
Other drugs		
Statins	341 (22.9)	531 (22.8)
Paracetamol	287 (19.3)	464 (19.92)
Blood-test abnormalities		
Mild cytopenia or liver enzyme elevation in six-months preceding start of follow-up	325 (21.9)	514 (22.1)

[1]Values are numbers (percentage) unless stated otherwise. DMARDS: Disease modifying anti-rheumatic drugs, SD: Standard deviation, CPRD: Clinical Practice Research Datalink.

Table 2: Final model hazard ratios and β-coefficients

Predictors	Adjusted HR (95% CI)	Coefficients
Age	1.01 (0.99 to 1.03)	0.0094981
Female sex	1.24 (0.83 to 1.83)	0.2128283
Body mass index (kg/m^2)	0.98 (0.95 to 1.01)	-0.0171081
Smoking status		
Non-smoker/not recorded/ex-smoker	Reference	-
Current smoker	0.90 (0.57 to 1.42)	-0.1056694
Alcohol consumption		
Non-drinker	Reference	-
Low (1-14 units/week)	0.96 (0.63 to 1.46)	-0.0400223
Moderate (15-21 units/week)	0.86 (0.26 to 2.86)	-0.1474903
Hazardous (>21 units/week)	1.12 (0.47 to 2.69)	0.1171966
Ex-drinker	0.84 (0.37 to 1.87)	-0.1774794
AIRD type		
Rheumatoid arthritis	Reference	-
PMR or GCA	1.03 (0.46 to 2.30)	0.026971
Spondyloarthritis	1.14 (0.76 to 1.70)	0.1266522
Comorbidities		
Epilepsy[1]	4.39 (1.74 to 11.06)	1.479007
Diabetes	0.88 (0.48 to 1.60)	-0.1311263
Chronic Kidney Disease	1.72 (0.96 to 3.06)	0.5400153
Other DMARDs		
Methotrexate or 5-aminosalicylates	0.93 (0.64 to 1.35)	-0.0733462
Other drugs		
Statins	1.44 (0.94 to 2.22)	0.3666838
Paracetamol	1.45 (0.98 to 2.16)	0.3747208
Blood-test abnormalities		
Mild cytopenia or liver enzyme elevation in six-months preceding start of follow-up	3.06 (2.15 to 4.35)	1.117226

[1]includes participants prescribed carbamazepine or valproate without a Read code for epilepsy.
HR: hazard ratio, CI: confidence interval, PMR: polymyalgia rheumatica, GCA : Giant cell arteritis

Table 3: Model diagnostics¶

Measure	Apparent performance*	Test performance§	Average optimism¥	Optimism corrected performance†	External validation (CPRD Aurum)
Overall calibration slope	1.00 (0.75 to 1.25)	0.72 (0.50 to 0.94)	0.28	0.72 (0.47 to 0.97)	0.91 (0.74 to 1.07)
Royston D statistic	1.06 (0.77 to1.35	0.90 (0.63 to 1.17)	0.33	0.73 (0.44 to 1.02)	0.97 (0.89 to 1.05)
R-squared	0.21 (0.12 to 0.30)	0.16 (0.08 to 0.24)	0.10	0.11 (0.02 to 0.20)	0.18 (0.16 to 0.21)

¶ Results from a single imputed dataset but similar across the other imputations (data not shown).
*Refers to performance (95% CI) estimated directly from the data that was used to develop the model.
§ Determined by executing full model in each bootstrap sample (500 samples with replacement), calculating bootstrap performance, and applying same model in original sample.
¥ Average difference between model performance in bootstrap data and test performance in original dataset
†subtracting average optimism from apparent performance.
CPRD: Clinical Practice Research Datalink

Risk score = 1 – $0.918^{e(X\beta)}$, where Xβ= 0.0094981*Age in years at first primary-care prescription + 0.2128283 *female-sex - 0.0171081 *BMI - 0.0400223*low alcohol intake - 0.1474903*moderate alcohol intake + 0.1171966*hazardous alcohol intake - 0.1774794*ex-alcohol intake - 0.1056694*current smoker - 0.1311263 *diabetes + 0.5400153*CKD + 0.026971*GCA/PMR + 0.1266522* Axial spondyloarthritis - 0.0733462*other-DMARDs + 0.3666838*statins + 0.3747208*paracetamol + 1.479007*epilepsy or carbamazepine or valproate + 1.117226 *mild cytopenia or liver enzyme elevation within six months of primary care LEF prescription.

All variables are code 0, and 1 if absent or present respectively, except for BMI and age that were continuous variables. 0.914 is the baseline survival function at 5-years and the other numbers are the estimated regression coefficients for the predictors, which indicate their mutually adjusted relative contribution to the outcome risk.

Figure 3: Equation to predict the risk of leflunomide discontinuation after 6 months of primary care prescription and within the next 5-years.

Figure legends

Figure 1: Calibration plot in the validation dataset. C-slope of 0.91 (0.74-1.07)

Figure 2: Kaplan-Meier survival estimates in the model development and validation datasets.

Figure 3: Equation to predict the risk of leflunomide discontinuation after 6 months of primary care prescription and within the next 5-years.

References

1. Smolen JS, Landewé RBM, Bijlsma JWJ, Burmester GR, Dougados M, Kerschbaumer A, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. Annals of the Rheumatic Diseases. 2020;79(6):685-99.
2. Strand V, Cohen S, Schiff M, Weaver A, Fleischmann R, Cannon G, et al. Treatment of active rheumatoid arthritis with leflunomide compared with placebo and methotrexate. Leflunomide Rheumatoid Arthritis Investigators Group. Archives of internal medicine. 1999;159(21):2542-50.
3. Cohen S, Cannon GW, Schiff M, Weaver A, Fox R, Olsen N, et al. Two-year, blinded, randomized, controlled trial of treatment of active rheumatoid arthritis with leflunomide compared with methotrexate. Utilization of Leflunomide in the Treatment of Rheumatoid Arthritis Trial Investigator Group. Arthritis and rheumatism. 2001;44(9):1984-92.
4. Emery P, Breedveld FC, Lemmel EM, Kaltwasser JP, Dawes PT, Gömör B, et al. A comparison of the efficacy and safety of leflunomide and methotrexate for the treatment of rheumatoid arthritis. Rheumatology (Oxford, England). 2000;39(6):655-65.
5. Nakafero G, Grainge MJ, Card T, Mallen CD, Zhang W, Doherty M, et al. What is the incidence of methotrexate or leflunomide discontinuation related to cytopenia, liver enzyme elevation or kidney function decline? Rheumatology (Oxford, England). 2021.
6. Ledingham J, Gullick N, Irving K, Gorodkin R, Aris M, Burke J, et al. BSR and BHPR guideline for the prescription and monitoring of non-biologic disease-modifying anti-rheumatic drugs. Rheumatology (Oxford, England). 2017;56(12):2257.
7. Singh JA, Saag KG, Bridges SL, Jr., Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis & rheumatology (Hoboken, NJ). 2016;68(1):1-26.
8. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.
9. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019;48(6):1740-g.
10. Chalasani N, Björnsson E. Risk factors for idiosyncratic drug-induced liver injury. Gastroenterology. 2010;138(7):2246-59.
11. Prakash A, Jarvis B. Leflunomide: a review of its use in active rheumatoid arthritis. Drugs. 1999;58(6):1137-64.
12. Meijer B, Wilhelm AJ, Mulder CJJ, Bouma G, van Bodegraven AA, de Boer NKH. Pharmacology of Thiopurine Therapy in Inflammatory Bowel Disease and Complete Blood Cell Count Outcomes: A 5-Year Database Study. Ther Drug Monit. 2017;39(4):399-405.
13. Dirven L, Klarenbeek NB, van den Broek M, van Groenendael JH, de Sonnaville PB, Kerstens PJ, et al. Risk of alanine transferase (ALT) elevation in patients with rheumatoid arthritis treated with methotrexate in a DAS-steered strategy. Clin Rheumatol. 2013;32(5):585-90.

14. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3-15.
15. Steyerberg EW. Clinical Prediction Models : A Practical Approach to Development, Validation, and Updating. Cham, SWITZERLAND: Springer International Publishing AG; 2019.
16. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology. 2013;13(1):33.
17. Cox DR. Note on Grouping. Journal of the American Statistical Association. 1957;52(280):543-7.
18. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
19. Curtis JR, Beukelman T, Onofrei A, Cassell S, Greenberg JD, Kavanaugh A, et al. Elevated liver enzyme tests among patients with rheumatoid arthritis or psoriatic arthritis treated with methotrexate and/or leflunomide. Annals of the rheumatic diseases. 2010;69(1):43-7.
20. EASL Clinical Practice Guidelines: Drug-induced liver injury. J Hepatol. 2019;70(6):1222-61.
21. Grabar PB, Rozman B, Logar D, Praprotnik S, Dolžan V. Dihydroorotate dehydrogenase polymorphism influences the toxicity of leflunomide treatment in patients with rheumatoid arthritis. Annals of the Rheumatic Diseases. 2009;68(8):1367-8.
22. Sandoval-Plata G, Morgan K, Abhishek A. Variants in urate transporters, ADH1B, GCKR and MEPE genes associate with transition from asymptomatic hyperuricaemia to gout: results of the first gout versus asymptomatic hyperuricaemia GWAS in Caucasians using data from the UK Biobank. Ann Rheum Dis. 2021.
23. Bohanec Grabar P, Rozman B, Tomsic M, Suput D, Logar D, Dolzan V. Genetic polymorphism of CYP1A2 and the toxicity of leflunomide treatment in rheumatoid arthritis patients. European journal of clinical pharmacology. 2008;64(9):871-6.
24. Hopkins AM, Wiese MD, Proudman SM, O'Doherty CE, Upton RN, Foster DJ. Genetic polymorphism of CYP1A2 but not total or free teriflunomide concentrations is associated with leflunomide cessation in rheumatoid arthritis. British journal of clinical pharmacology. 2016;81(1):113-23.
25. Wiese MD, Schnabl M, O'Doherty C, Spargo LD, Sorich MJ, Cleland LG, et al. Polymorphisms in cytochrome P450 2C19 enzyme and cessation of leflunomide in patients with rheumatoid arthritis. Arthritis Research & Therapy. 2012;14(4):R163.
26. Kim KA, Joo HJ, Park JY. Effect of ABCG2 genotypes on the pharmacokinetics of A771726, an active metabolite of prodrug leflunomide, and association of A771726 exposure with serum uric acid level. European journal of clinical pharmacology. 2011;67(2):129-34.

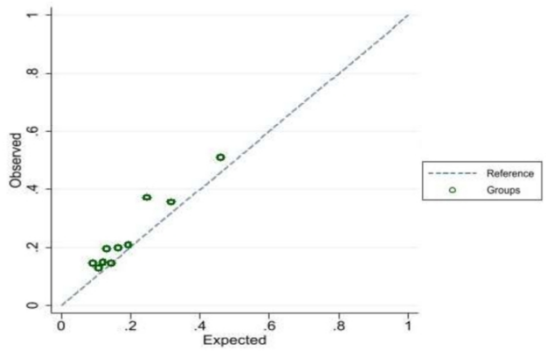

Figure 1: Calibration plot in the validation dataset. C-slope of 0.91 (0.74-1.07)

Figure 1: Calibration plot in the validation dataset. C-slope of 0.91 (0.74-1.07)

54x30mm (600 x 600 DPI)

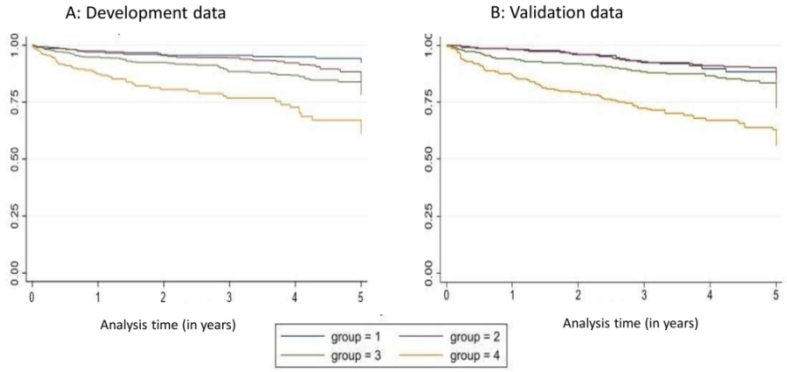

Figure 2: Kaplan-Meier survival estimates in the model development and validation datasets. Groups 1,2,3 and 4 were defined using the cut-offs for the 16th, 50th, 84th centile of the linear predictor. The range of linear predictor in each of the four groups was -0.67 to 0.24, 0.24 to 0.64, 0.64 to 1.45 and 1.45 to 3.83 in the development dataset. The corresponding values were -0.40 to 0.25, 0.25 to 0.61, 0.61 to 1.40 and 1.40 to 3.92 in the validation dataset.

Figure 2: Kaplan-Meier survival estimates in the model development and validation datasets.

54x30mm (600 x 600 DPI)