Artificial Intelligence

A Machine Learning Approach to Identify Predictors of Severe COVID-19 Outcome in Patients With Rheumatoid Arthritis

Sara M. Burns, MS[1,2], Tiffany S. Woodworth, MPH[1], Zeynep Icten, PhD[1], Trenton Honda, PhD[3], and Justin Manjourides, PhD[2]

From: [1]Panalgo LLC, Boston, MA; [2]Department of Health Sciences, Bouvé College of Health Sciences, Northeastern University, Boston, MA; [3]School of Clinical and Rehabilitation Sciences, Bouvé College of Health Sciences, Northeastern University, Boston, MA

Address Correspondence:
Sara M. Burns, MS
Panalgo LLC
265 Franklin Street, Suite 1101
Boston, MA 02110
E-mail:
saramariaburns@gmail.com

Disclaimer: There was no external funding in the preparation of this manuscript.

Conflict of interest: Each author certifies that he or she, or a member of his or her immediate family, has no commercial association (i.e., consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted manuscript.

Manuscript received: 04-15-2022
Revised manuscript received: 07-15-2022
Accepted for publication: 08-29-2022

Free full manuscript:
www.painphysicianjournal.com

Background: Rheumatoid arthritis (RA) patients have a lowered immune response to infection, potentially due to the use of corticosteroids and immunosuppressive drugs. Predictors of severe COVID-19 outcomes within the RA population have not yet been explored in a real-world setting.

Objectives: To identify the most influential predictors of severe COVID-19 within the RA population.

Study Design: Retrospective cohort study.

Setting: Research was conducted using Optum's de-identified Clinformatics® Data Mart Database (2000-2021Q1), a US commercial claims database.

Methods: We identified adult patients with index COVID-19 (ICD-10-CM diagnosis code U07.1) between March 1, 2020, and December 31, 2020. Patients were required to have continuous enrollment and have evidence of one inpatient or 2 outpatient diagnoses of RA in the 365 days prior to index. RA patients with COVID-19 were stratified by outcome (mild vs severe), with severe cases defined as having one of the following within 60 days of COVID-19 diagnosis: death, treatment in the intensive care unit (ICU), or mechanical ventilation. Baseline demographics and clinical characteristics were extracted during the 365 days prior to index COVID-19 diagnosis. To control for improving treatment options, the month of index date was included as a potential independent variable in all models. Data were partitioned (80% train and 20% test), and a variety of machine learning algorithms (logistic regression, random forest, support vector machine [SVM], and XGBoost) were constructed to predict severe COVID-19, with model covariates ranked according to importance.

Results: Of 4,295 RA patients with COVID-19 included in the study, 990 (23.1%) were classified as severe. RA patients with severe COVID-19 had a higher mean age (mean [SD] = 71.6 [10.3] vs 63.4 [13.7] years, $P < 0.001$) and Charlson Comorbidity Index (CCI) (3.8 [2.4] vs 2.4 [1.8], $P < 0.001$) than those with mild cases. Males were more likely to be a severe case than mild (29.1% vs 18.5%, $P < 0.001$). The top 15 predictors from the best performing model (XGBoost, AUC = 75.64) were identified. While female gender, commercial insurance, and physical therapy were inversely associated with severe COVID-19 outcomes, top predictors included a March index date, older age, more inpatient visits at baseline, corticosteroid or gamma-aminobutyric acid analog (GABA) use at baseline or the need for durable medical equipment (i.e., wheelchairs), as well as comorbidities such as congestive heart failure, hypertension, fluid and electrolyte disorders, lower respiratory disease, chronic pulmonary disease, and diabetes with complication.

Limitations: The cohort meeting our eligibility criteria is a relatively small sample in the context of machine learning. Additionally, diagnoses definitions rely solely on ICD-10-CM codes, and there may be unmeasured variables (such as labs and vitals) due to the nature of the data. These limitations were carefully considered when interpreting the results.

Conclusions: Predictive baseline comorbidities and risk factors can be leveraged for early detection of RA patients at risk of severe COVID-19 outcomes. Further research should be conducted on modifiable factors in the RA population, such as physical therapy.

Key words: COVID-19, rheumatoid arthritis, machine learning, real-world data, predictive modeling, RA, SARS-CoV-2, real-world evidence, physical therapy, corticosteroid use

Pain Physician 2022: 25:593-602

Rheumatoid arthritis (RA) patients have long been known to have a higher risk of infection than the general population (1,2). This increased risk may be due to a number of factors, including immunosuppressive medication use, chronic inflammation, and immunosuppressive comorbid conditions (3,4). As such, RA patients may be particularly vulnerable to the impacts of COVID-19 and its sequelae. There is a lack of literature in this area, and few studies utilizing large electronic health records or claims databases. In the context of this public health emergency, understanding the relationship between COVID-19 infection and potentially vulnerable RA patients is increasingly critical. The development of an approach for using these data to identify risk factors for severe disease is a critical need for future pandemic responses.

The numerous options available to treat RA add complexity to understanding the relationship between COVID-19 severity and RA-specific treatment and characteristics. For example, some studies show that certain RA medications, such as corticosteroids and some biologic disease-modifying anti-rheumatic (bDMAR) drugs (5), may be associated with a higher risk of severe COVID-19 outcomes (3). However, certain drugs that are commonly used to treat RA, such as hydroxychloroquine and tocilizumab, have been studied as treatments for COVID-19 (2,6). In addition to medication use, factors specific to RA patients, such as specific comorbidities, socioeconomic disparities, and mental health, may influence the severity of disease among patients with COVID-19 (3).

Understanding risk factors that are predictive of COVID-19 severity in RA patients could facilitate early detection of patients at high risk of a severe COVID-19 outcome and allow doctors to better advise their RA patients about modifiable risk factors. In this study, we use a data-driven approach to address this gap in knowledge by leveraging real-world data in combination with several machine learning algorithms to identify risk factors within the RA population that are most predictive of severe COVID-19 outcomes.

Methods

Data Source

All models used Optum de-identified Clinformatics® Data Mart Database (Optum), a US administrative claims database containing approximately 88 million patient lives. Optum data consists of a large, privately insured population with both medical and pharmacy benefits (primarily from the UnitedHealth Group) and data elements include demographics, diagnosis and procedure codes, and prescribed medications. Data from March 1, 2018, through March 31, 2021, were utilized for this study. This timeframe reflects the first wave of the pandemic, prior to vaccines becoming widely available.

Study Population

The study population consisted of RA patients who received a COVID-19 diagnosis during the COVID-19 risk window, March 1, 2020, to the end of the year, December 31, 2020. The index date was defined as the first occurrence of a COVID-19 International Classification of Disease (ICD-10-CM) diagnosis code (U07.1) and included patients who were required to have an inpatient RA ICD-10-CM diagnosis code (M05.*, M06.*) or 2 outpatient RA diagnosis codes, at least 7 days apart during the 365 days preceding the index date. In addition to having one inpatient or 2 outpatient diagnoses for RA, patients were required to have a least one prescription for an RA medication in the year prior to RA diagnosis. RA medications included hydroxychloroquine, leflunomide, methotrexate, sulfasalazine, tumor necrosis factor inhibitor (TNFi) biologic, non-TNFi biologic, and Janus kinase inhibitors (JAKi) prescription (7,8).

Patients were excluded if they had a diagnosis of juvenile idiopathic arthritis, psoriasis, psoriatic arthritis, ankylosing spondylitis, Crohn's disease, or ulcerative colitis in the 2 years prior to the index date, as patients with these indications may be treated with RA medications. Study participants were required to be continuously enrolled during the baseline time period, defined as 365 days prior to the index date minus 1 day, to ensure that baseline characteristics and risk factors were captured in the data. Lastly, we required patients to be at least 18 years of age at the index date.

Outcome Definition

The outcome of interest was COVID-19 severity, categorized as mild or severe. With the index date being the first occurrence of a COVID-19 diagnosis code, the definition of a severe case was composed of any of the following outcomes within 60 days of index: 1) all-cause death, 2) required mechanical ventilation, or 3) treatment in the intensive care unit (ICU) or critical care unit (CCU). COVID-19 cases which were not categorized as severe were assigned as mild.

Feature Definitions

Candidate features were selected based on literature review, clinical input, and data-driven decision making. Demographics, healthcare utilization metrics, RA-specific measures, and data-driven diagnoses, procedures, and medications were extracted from the baseline time period (index date minus 365 days to index date minus 1 day) and included as candidate features in each model.

Baseline demographics included age, gender, region, payor, and Charlson Comorbidity Index (CCI) categories (9). Healthcare utilization metrics included the number of emergency room (ER) visits, number of inpatient visits, and number of office/clinic visits during baseline. RA-specific characteristics included biologic or JAKi use (defined as a TNFi biologic, non-TNFi biologic, or JAKi prescription claim during baseline) (10), corticosteroid prescription, methotrexate prescription, last c-reactive protein (CPR) lab result and last erythrocyte sedimentation rate (ESR) lab result prior to index. In a data-driven approach, the top 30 ICD-10-CM diagnoses, top 20 Current Procedure Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) procedures, and top 20 National Drug Code (NDC) and HCPCS medications were also included as candidate features. To adjust for improving treatment options of COVID-19, the month of index was also included as a candidate feature in all models (Table 1).

Statistical Analysis

The target outcome of interest was COVID-19 severity (mild or severe). Descriptive statistics were obtained on all candidate features, stratified by COVID-19 severity, and comparisons were made using independent t-tests for continuous variables and chi-squared tests for categorical variables with $\alpha = 0.05$.

Mean imputation was applied to missing values in continuous variables, and a "missing" category was added for missing values in categorical and binary variables. Data were partitioned into training and testing sets, with 80% of the data allocated to training and 20% of the data allocated to testing. Feature scaling was applied when appropriate, with consideration to the corresponding model type. For example, standardization, which transforms features to be centered and scaled to a mean of zero and a standard deviation equal to one, was used for the logistic regression model. To identify the top predictors of severe COVID-19 in the RA population, 4 machine learning models were fit using 5-fold cross-validation: logistic regression, random forest (RF), support vector machine (SVM), and XGBoost. Model discrimination was assessed using the area under the ROC curve (AUC), accuracy (defined as the ratio of number of correctly classified predictions divided by the total number of samples), negative predictive value (defined as the number of true negatives divided by the number of observed negatives) and F1 score (defined as the harmonic mean of precision and recall) and model calibration was assessed using the Brier score. The best performing model was identified based on AUC.

For each of the 4 models, selected features were ranked using appropriate importance metrics corresponding to the model type. The logistic regression model used the absolute value of the standardized coefficient estimate, the random forest model used Gini importance, the SVM model used the weights assigned to the features, and the XGBoost model used feature gain to rank importance. To determine the direction of association between each predictor and COVID-19 severity, the Shapley Additive exPlanations (SHAP) methodology was utilized (11). Most influential predictors of severe COVID-19 were identified using the best performing model. A sensitivity analysis was conducted to capture top predictors of severe COVID-19 across all models in a robust manner, using the average rank importance.

Statistical analyses were performed using Instant Health Data (IHD) software (Panalgo, Boston, MA).

Results

Baseline Characteristics

There were 5,503 patients in the database that had COVID-19 and met our RA definition. After removing patients with an exclusion diagnosis, patients that did not meet the continuous enrollment in baseline, and patients younger than 18 on the index date, 4,295 patients remained in the study population (Table 2).

Of the 4,295 RA patients with COVID-19 included in the study, 990 (23.1%) were classified as "severe." RA patients with severe COVID-19 were older (mean age = 71.6 [SD = 10.3] vs 63.4 [13.7] years, $P < 0.001$) and had a higher Charlson Comorbidity Index (CCI) (3.8 [2.4] vs 2.4 [1.8], $P < 0.001$) than those with mild cases. Males were more likely to be a severe case than mild (29.1% vs 18.5%, $P < 0.001$), and higher instances of all measured comorbidity groups were observed in patients with severe cases (Table 1).

Table 1. *Candidate features from baseline.*

		COVID-19 Outcome	
		Mild (n = 3,305)	**Severe (n = 990)**
Adjustments			
Month of Index	03 (March)	14 (0.42%)	29 (2.93%)
	04 (April)	136 (4.11%)	72 (7.27%)
	05 (May)	154 (4.66%)	46 (4.65%)
	06 (June)	178 (5.39%)	54 (5.45%)
	07 (July)	327 (9.89%)	111 (11.21%)
	08 (August)	229 (6.93%)	73 (7.37%)
	09 (September)	220 (6.66%)	55 (5.56%)
	10 (October)	359 (10.86%)	114 (11.52%)
	11 (November)	732 (22.15%)	207 (20.91%)
	12 (December)	956 (28.93%)	229 (23.13%)
Demographics			
Age (years) at Index	Mean (SD)	63.43 (13.67)	71.62 (10.33)
	Median (IQR)	65 (55-73)	73 (66-79)
	Min-Max	18-90	36-90
Gender	Female	2,694 (81.51%)	702 (70.91%)
	Male	611 (18.49%)	288 (29.09%)
Region	Midwest	748 (22.63%)	212 (21.41%)
	Northeast	341 (10.32%)	84 (8.48%)
	South	1,679 (50.80%)	534 (53.94%)
	West	532 (16.10%)	160 (16.16%)
	Missing	5 (0.15%)	0 (0%)
Payor	Commercial	1,278 (38.67%)	120 (12.12%)
	Medicare Advantage	2,027 (61.33%)	870 (87.88%)
Charlson Comorbidity Index (CCI)	Mean (SD)	2.39 (1.82)	3.75 (2.36)
	Median (IQR)	2 (1-3)	3 (2-5)
	Min-Max	1-14	1-14
CCI Category	Rheumatologic Disease	3,305 (100%)	990 (100%)
	Chronic Pulmonary Disease	1,001 (30.29%)	488 (49.29%)
	Renal Disease	556 (16.82%)	331 (33.43%)
	Diabetes w/ Complications or Comorbidity	493 (14.92%)	329 (33.23%)
	Congestive Heart Failure	407 (12.31%)	325 (32.83%)
	Mild Liver Disease	311 (9.41%)	114 (11.52%)
	Any Primary Malignancy	262 (7.93%)	139 (14.04%)
	Dementia	158 (4.78%)	97 (9.80%)
	Hemiplegia or Paraplegia	44 (1.33%)	33 (3.33%)
	Metastatic Solid Tumor	36 (1.09%)	32 (3.23%)
	Moderate or Severe Liver Disease	10 (0.30%)	11 (1.11%)
	AIDS	3 (0.09%)	1 (0.10%)
RA Specific			
Biologic or JAKi Use	n (%)	1,253 (37.91%)	281 (28.38%)
Corticosteroid Use	n (%)	2,310 (69.89%)	768 (77.58%)

Table 1 cont. *Candidate features from baseline.*

		COVID-19 Outcome	
		Mild (n = 3,305)	**Severe (n = 990)**
Methotrexate Use	n (%)	1,590 (48.11%)	490 (49.49%)
Last CRP	n	1,195	322
	Mean (SD)	7.14 (14.35)	11.60 (23.31)
	Median (IQR)	3 (1-7)	5 (1.90-13)
	Min-Max	0-168	0-272.40
Last ESR	n	1,157	291
	Mean (SD)	21.94 (21.98)	30.57 (25.83)
	Median (IQR)	15 (6-31)	25 (10-44)
	Min-Max	0-121	0-120
Healthcare Utilization			
Number of ER Visits	Mean (SD)	0.69 (1.49)	1.10 (1.97)
	Median (IQR)	0 (0-1)	0 (0-2)
	Min-Max	0-21	0-28
Number of Inpatient Visits	Mean (SD)	0.40 (1.29)	0.85 (1.43)
	Median (IQR)	0 (0-0)	0 (0-1)
	Min-Max	0-38	0-14
Number of Office or Clinic Visits	Mean (SD)	16.70 (13.06)	18.20 (12.81)
	Median (IQR)	13 (8-22)	16 (9-25)
	Min-Max	0-153	0-102
Data Driven			
Top 30 ICD-10 Diagnoses			
Top 20 CPT and HCPCS Procedures			
Top 20 NDC and HCPCS Medications			

Model Performance

Among the 4 models assessed (logistic regression, RF, SVM, and XGBoost), the XGBoost model performed best based on AUC and F1 scores. XGBoost selected 60 of the 107 candidate features yielding an AUC = 75.6, accuracy = 67.6, NPV = 88.1, F1 = 0.50, and test AUC = 74.4. The XGBoost model is parsimonious in the sense that it performs slightly better while using a much smaller number of features than the other models (Table 3).

Table 2. *Cohort attrition.*

	Count	% Dropped
Total Database	89,711,387	
COVID-19	25,588	99.97%
1 Inpatient or 2 Outpatient RA Dx	9,622	62.4%
RA Medication	5,503	42.8%
Excluded Dx	5,000	9.1%
Continuous Enrollment in Baseline	4,295	14.1%
Age ≥ 18 at Index	4,295	0.0%

Top Predictors of Severe COVID-19 Outcome

As no clear-cut point of diminishing return of feature importance was identified, the top 15 features based on importance rank from the best performing model were selected, including: 1) congestive heart failure (CHF), 2) payor, 3) hypertension, 4) lower respiratory disease, 5) chronic pulmonary disease, 6) diabetes with complications or comorbidity, 7) fluid and electrolyte disorders, 8) gender, 9) number of inpatient visits, 10) gamma-aminobutyric acid analogs (GABA), 11) age, 12) March index date, 13) corticosteroid use, 14) physical therapy and 15) need for durable medical equipment (e.g., wheelchair) (Table 4).

Female gender, commercial insurance, and physical

Table 3. *Model performance metrics.*

Model	# of Features (a)	Binary Threshold	AUC	Average Precision	Brier Score	Accuracy	Recall	Precision	Youden Index	F1 Score	Negative Predictive Value	Specificity
XGBoost	60	0.23	75.64%	47.56%	0.151	67.55%	69.59%	38.72%	36.53%	0.497	88.10%	66.94%
RF	105	0.23	74.88%	46.29%	0.154	65.19%	72.49%	37.08%	35.50%	0.490	88.43%	63.01%
Logistic	104	0.23	74.58%	45.88%	0.156	67.57%	65.80%	38.31%	33.91%	0.484	86.92%	68.11%
SVM	107	0.23	74.42%	45.95%	0.154	67.05%	69.33%	38.24%	35.71%	0.493	87.85%	66.38%

(a) There were 107 candidate features

therapy were inversely associated with severe COVID-19 outcomes. Top predictors included March index date, older age, more inpatient visits in baseline, corticosteroid or gamma-aminobutyric acid analog (GABA) use in baseline, or the need for durable medical equipment, as well as comorbidities such as congestive heart failure, hypertension, fluid and electrolyte disorders, lower respiratory disease, chronic pulmonary disease, and diabetes with complication. A full list of features included in the XGBoost model can be found in the supplemental information (Table S1).

Top Shared Predictors of Severe COVID-19 Outcome Across All 4 Models

As an alternative approach to identifying the most important features, the average rank importance was calculated across all 4 models. Results revealed that 12 features with the highest importance rank in the XGBoost model were also among the 15 features with the highest average importance rank across all 4 models: 1) age, 2) payor, 3) more inpatient visits, 4) March index date, 5) diabetes with complications or comorbidity, 6) congestive heart failure (CHF), 7) gender, 8) lower respiratory disease, 9) hypertension, 10) physical therapy, 11) corticosteroid use, and 12) chronic pulmonary disease. The last ESR at baseline, month of index (April), and region (South) were also found to be robustly important across models (Table 5).

Discussion

In this study, we identify several risk factors for severe COVID-19 outcomes within a population of individuals with RA during the first wave of the pandemic, prior to available vaccination. Risk factors identified as predictive of severe COVID-19 outcomes included a March index date, older age, more inpatient visits at baseline, corticosteroid or gamma-aminobutyric acid analog (GABA) use at baseline or the need for durable medical equipment as well as comorbidities such as congestive heart failure, hypertension, fluid and electrolyte disorders, lower respiratory disease, chronic pulmonary disease, and diabetes with complication. Additionally, we identified several factors associated with a lower risk of COVID-19 severity, such as female gender, physical therapy, and commercial insurance. Together, these findings can help to guide the early detection of patients at high risk of a severe COVID-19 outcome and could inform future research to allow doctors to better advise their RA patients about modifiable risk factors. It is important to identify such predictors during the pre-vaccine stage of the pandemic as they would likely be similar in future covid variations or covid-like illnesses that escape vaccine-induced immunity. Additionally, the methodology outlined in this manuscript could be adapted to include data from later phases of the pandemic.

The nonmodifiable risk factors we identified within the RA community are largely similar to those in the general population and consistent with published research on this topic (3,12-16). For example, in a study early in the pandemic among a hospitalized Chinese population, male gender, and older age were associated with severe COVID-19 outcomes (17). Consistent with this, among those with severe COVID-19, these same risk factors were strongly associated with subsequent death. Many of the modifiable risk factors we identified for severe COVID-19 have also been extensively described previously in studies from across the world, with perhaps the greatest evidence for cardiovascular disease, hyperglycemia/diabetes mellitus, and chronic respiratory disease. The consistency of our findings with most prior research increases our confidence in the validity of our results.

Importantly, to our knowledge, we are the first to identify that commercial

Table 4. *Top 15 predictors of severe COVID-19 outcome from XGBoost model.*

Feature	Rank	SHAP Direction
CCI Category: Congestive Heart Failure	1	+
Payor: Commercial Insurance	2	-
ICD10 Dx: 7.1-Hypertension	3	+
ICD10 Dx: 8.8-Other lower respiratory disease [133.]	4	+
CCI Category: Chronic Pulmonary Disease	5	+
CCI Category: Diabetes with Complications or Comorbidity	6	+
ICD10 Dx: 3.8-Fluid and electrolyte disorders [55.]	7	+
Gender: Female	8	-
Number of Inpatient Visits	9	+
NDC and HCPCS Rx: gamma-aminobutyric acid analogs	10	+
Age at Index	11	+
Month of Index: 03 (March)	12	+
Corticosteroid Use	13	+
CPT and HCPCS Proc: Physical therapy exercises, manipulation, and other procedures	14	-
CPT and HCPCS Proc: Durable Medical Equipment and supplies	15	+

Table 5. *Top 15 shared predictors of severe COVID-19 outcome across all 4 models.*

		XGBoost		RF		Logistic		SVM	
Feature	Average Rank Importance	Rank	SHAP Direction	Rank	SHAP Direction	Rank	OR	Rank	Coefficient
Age at Index	5	11	+	1	+	7	1.04*	1	2.62
Payor: Commercial Insurance	7.25	2	-	3	-	9	0.59*	15	-0.46
Number of Inpatient Visits	7.75	9	+	4	+	15	1.17*	3	1.32
Month of Index: 03 (March)	9	12	+	14	+	8	12.76*	2	1.75
CCI Category: Diabetes with Complications or Comorbidity	9.25	6	+	6	+	13	1.63*	12	0.51
CCI Category: Congestive Heart Failure	9.75	1	+	2	+	18	1.41*	18	0.39
Gender: Female	13.5	8	-	19	-	11	0.59*	16	-0.45
Last ESR	13.5	22	+	7	+	21	1.01*	4	1.01
ICD10 Dx: 8.8-Other lower respiratory disease [133.]	14	4	+	12	+	14	1.44*	26	0.25
ICD10 Dx: 7.1-Hypertension	14.5	3	+	13	+	17	1.36*	25	0.25
CPT and HCPCS Proc: Physical therapy exercises, manipulation, and other procedures	16.75	14	-	26	-	10	0.58*	17	-0.44
Month of Index: 04 (April)	22	21	+	34	+	12	2.56*	21	0.30
Corticosteroid Use	22.25	13	+	32	+	20	1.30*	24	0.26
CCI Category: Chronic Pulmonary Disease	22.5	5	+	16	+	34	1.17	35	0.19
Region: South	22.75	17	+	45	+	1	1751.33	28	0.24

insurance and physical therapy are negatively associated with severe COVID-19 outcomes among RA patients. Although it is known that un- or underinsured populations have an increased risk of severe COVID-19 outcomes (18), patients in this cohort were either commercially insured or on Medicare Advantage. There-

fore, insurance may be a proxy for age, as we expect younger patients to be on commercial insurance plans. Physical therapy has been used as a treatment option for patients with RA (19,20), although it is unclear how physical therapy might be biologically related to a decreased likelihood of severe COVID-19. One previous study found that moderate to high disease activity (OR 1.4, 95% CI, 1.1, 1.8), social support (OR 2.1, 95% CI: 1.3, 3.5), and being on disability (OR 2.4, 95% CI: 1.3, 4.6) were associated with increased likelihood of physical therapy, while RA patients with less education were significantly less likely to receive physical therapy (OR 0.5, 95% CI: 0.2, 0.8). As we were unable to examine the impact of measures of socioeconomic status (SES) in our cohort, it is possible that access to physical therapy is acting as a proxy of this unmeasured variable (21). However, it is also possible that physical therapy has direct, biological impacts that may impact the trajectory of COVID-19. For example, physical therapy in RA patients has been previously shown to have a significant impact on symptomatology and physical activity (20). As prescribed physical activity regimens in RA patients may decrease chronic inflammation, vascular dysfunction and have positive effects on glucose metabolism (22), it is possible that physical therapy can impact several known risk factors for severe COVID-19, which may explain the findings we observe. As physical therapy is readily modifiable, the potential biological basis for the associations we observe should be further explored as it could be used by physicians when advising RA patients about preventative measures in reducing their risk of a severe COVID-19 outcome.

We identified several medications associated with increased COVID-19 risk, although the results were somewhat variable across models. Our study identified corticosteroid use as a top predictor of severe COVID-19 outcomes among RA patients. While several recent clinical trials have shown that corticosteroids administered acutely during severe COVID-19 infection likely afford beneficial effects by decreasing the magnitude of cytokine production (23), the prolonged use of these agents in RA patients can lead to significant immunosuppression. However, the degree to which this leads to adverse outcomes is a matter of debate, with several studies about chronic corticosteroids and COVID-19 severity finding detrimental outcomes among those on these agents (24); other studies have found no clinically important differences (25). Given the conflicting evidence around corticosteroid use and COVID-19 severity both within RA patients and the general population, additional research is warranted to guide clinical applications (3,26).

While we investigated predictors of COVID-19 severity in RA specifically, our findings may have implications for several similar disorders which can be avenues for future research. For example, our findings of inverse associations for physical therapy and positive associations for baseline corticosteroid use may warrant similar investigation among other autoimmune disorders for which these interventions are common, such as the HLA-B27 associated seronegative spondyloarthropathies (i.e., psoriatic arthritis, ankylosing spondylitis, and reactive arthritis). While these diseases are relatively rare compared to RA, which might limit the impact of the findings, other common connective tissue disorders or rheumatologic diseases for which joint manifestations and immunosuppressive medications are common, such as systemic lupus erythematosus and inflammatory bowel disease, may be high-yield and help guide future risk stratification in these patient populations.

The following limitations were carefully considered when drawing conclusions from this study. Although the Optum claims database is quite large, the cohort meeting our eligibility criteria is a relatively small sample in the context of machine learning. While our definition of RA was rigorous, the definition of COVID-19 relied solely on the ICD-10-CM diagnosis U07.1, which is not guaranteed to be a reliable form of classification. Additionally, there were unmeasured variables due to the nature of the data. For example, while select lab values were observed (and included), other labs and vitals (such as BMI) or measures of socioeconomic status were not available. When creating features from lab data with missing values, mean imputation was utilized, which has the potential to bias estimates as the variance is reduced. Lastly, this algorithm was validated to the best of its ability utilizing the data source selected, but validation on an external data source would strengthen the reliability of the findings.

These limitations are counterbalanced by a number of significant strengths. First, this study bridges a knowledge gap by identifying potential risk factors of severe COVID-19 outcomes specific to the RA population, who may be at high risk. Second, speed is important in the face of a public health emergency, and this study demonstrates how rapid analytics can be used to analyze routinely collected data quickly. This could be crucial to identifying risk factors for severe disease in future COVID-19 waves or pandemics for other populations.

Lastly, the predictors identified in the study can inform future research and contribute to the compounding body of evidence required to draw reliable conclusions.

Conclusion

This study exemplifies how predictive baseline comorbidities and risk factors can be leveraged for early detection of RA patients at high risk of a severe COVID-19 outcome. After further external validation, we can draw conclusions about modifiable factors in RA patients at risk of severe COVID-19 outcomes. Although COVID-19 treatment options have hastily progressed, it is important to understand preventative measures for severe COVID-19 within the RA population during the height of the pandemic. Using this data-driven method, risk factors have been identified that can inform more targeted future studies on the prevention of severe COVID-19 outcomes in RA patients.

Data Availability

The data underlying this article were provided by Optum under license / by permission. Data will be shared upon request to the corresponding author with the permission of Optum.

Supplementary material available at www.painphysicianjournal.com

References

1. Doran MF, Crowson CS, Pond GR, O'Fallon WM, Gabriel SE. Frequency of infection in patients with rheumatoid arthritis compared with controls: A population-based study. *Arthritis Rheum* 2002; 46:2287-2293.
2. Favalli EG, Ingegnoli F, De Lucia O, Cincinelli G, Cimaz R, Caporali R. COVID-19 infection and rheumatoid arthritis: Faraway, so close! *Autoimmun Rev* 2020; 19:102523.
3. D'Silva KM, Wallace ZS. COVID-19 and rheumatoid arthritis. Curr Opin Rheumatol 2021; 33:255-261.
4. Lahiri M, Dixon WG. Risk of infection with biologic antirheumatic therapies in patients with rheumatoid arthritis. *Best Pract Res Clin Rheumatol* 2015; 29:290-305.
5. Sparks JA, Wallace ZS, Seet AM, et al. Associations of baseline use of biologic or targeted synthetic DMARDs with COVID-19 severity in rheumatoid arthritis: Results from the COVID-19 Global Rheumatology Alliance physician registry. *Ann Rheum Dis* 2021; 80:1137-1146.
6. Roongta R, Ghosh A. Managing rheumatoid arthritis during COVID-19. *Clin Rheumatol* 2020; 39:3237-3244.
7. Chung CP, Rohan P, Krishnaswami S, McPheeters ML. A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine* 2013; 31:K41-K61.
8. Chandran U, Reps J, Stang PE, Ryan PB. Inferring disease severity in rheumatoid arthritis using predictive modeling in administrative claims databases. *PloS One* 2019; 14:e0226255.
9. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* 2011; 17:676-682.
10. Singh JA, Saag KG, Bridges SL Jr, et al. 2015 American College of Rheumatology guideline for the treatment of rheumatoid arthritis. *Arthritis Rheumatol* 2016; 68:1-26.
11. Lundberg S, Lee S. A unified approach to interpreting model predictions. *Adv Neural Inf Process Sys* 2017; 30:1-10.
12. Gallo Marin B, Aghagoli G, Lavine K, et al. Predictors of COVID-19 severity: A literature review. *Rev Med Virol* 2021; 13:1-10.
13. Palaiodimos L, Kokkinidis DG, Li W, et al. Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. *Metabolism* 2020; 108:154262.
14. Kragholm K, Andersen MP, Gerds TA, et al. Association between male sex and outcomes of Coronavirus Disease 2019 (Covid-19)-A Danish nationwide, register-based study. *Clin Infect Dis* 2021; 73:e4025-e4030.
15. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol* 2020; 146:110-118.
16. Ejaz H, Alsrhani A, Zafar A, et al. COVID-19 and comorbidities: Deleterious impact on infected patients. *J Infect Public Health* 2020; 13:1833-1839.
17. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol* 2020; 146:110-118.
18. Gaffney AW, Hawks L, Bor DH, Woolhandler S, Himmelstein DU, McCormick D. 18.2 million individuals at increased risk of severe COVID-19 illness are un-or underinsured. *J Gen Intern Med* 2020; 35:2487-2489.
19. Bell MJ, Lineker SC, Wilkins AL, Goldsmith CH, Badley EM. A randomized controlled trial to evaluate the efficacy of community based physical therapy in the treatment of people with rheumatoid arthritis. *J Rheumatol* 1998; 25:231-237.
20. Feldthusen C, Dean E, Forsblad-d'Elia H, Mannerkorpi K. Effects of person-centered physical therapy on fatigue-related variables in persons with rheumatoid arthritis: A randomized controlled trial. *Arch Phys Med Rehabil* 2016; 97:26-36.
21. Iversen MD, Chhabriya RK, Shadick N. Predictors of the use of physical therapy services among patients with rheumatoid arthritis. *Phys Ther* 2011; 91:65-76.
22. Metsios GS, Stavropoulos-Kalinoglou A, Sandoo A, et al. Vascular function and inflammation in rheumatoid arthritis: The role of physical activity. *Open Cardiovasc Med J* 2010; 4:89-96.
23. Jamilloux Y, Henry T, Belot A, et al. Should we stimulate or suppress immune responses in COVID-19? Cytokine and anti-cytokine interventions. *Autoimmun Rev* 2020; 19:102567.
24. Brenner EJ, Ungaro RC, Gearry RB,

et al. Corticosteroids, but not TNF antagonists, are associated with adverse COVID-19 outcomes in patients with inflammatory bowel diseases: Results from an international registry. *Gastroenterology* 2020; 159:481-491.

25. Andersen KM, Mehta HB, Palamuttam N, et al. Association between chronic use of immunosuppresive drugs and clinical outcomes from coronavirus disease 2019 (COVID-19) hospitalization: A retrospective cohort study in a large US health system. *Clin Infect Dis* 2021; 73:e4124-e4130.
26. Bhargava A, Fukushima EA, Levine M, et al. Predictors for severe COVID-19 infection. *Clin Infect Dis* 2020; 71:1962-1968.

Supplemental Table 1. *All predictors of severe COVID-19 outcome from XGBoost model.*

Feature	Rank	Weight	Gain	Cover
CCI Category: Congestive Heart Failure	1	137	1	859
Payor: Commercial Insurance	2	64	2	680
ICD10 Dx: 7.1-Hypertension	3	31	1	580
ICD10 Dx: 8.8-Other lower respiratory disease [133.]	4	30	2	578
CCI Category: Chronic Pulmonary Disease	5	25	1	574
CCI Category: Diabetes with CC	6	17	4	557
ICD10 Dx: 3.8-Fluid and electrolyte disorders [55.]	7	14	1	556
Gender: Female	8	11	4	543
Number of Inpatient Visits	9	10	6	536
NDC and HCPCS Rx: gamma-aminobutyric acid analogs	10	10	1	551
Age at Index	11	10	20	533
Month of Index: March	12	8	6	532
Corticosteroid Use	13	8	2	537
CPT and HCPCS Proc: Physical therapy exercises, manipulation, and other procedures	14	6	5	528
CPT and HCPCS Proc: DME and supplies	15	4	1	529
CCI Category: Any Primary Malignancy	16	4	2	520
Region: SOUTH	17	4	2	522
CPT and HCPCS Proc: Mammography	18	3	2	519
ICD10 Dx: 3.11-Other nutritional; endocrine; and metabolic disorders [58.]	19	3	2	519
Month of Index: December	20	3	3	518
Month of Index: April	21	3	5	517
Last ESR	22	3	13	513
CPT and HCPCS Proc: Laboratory - Chemistry and Hematology	23	2	3	516
Region: Northeast	24	2	4	516
NDC and HCPCS Rx: TNF alpha inhibitors	25	2	3	513
CPT and HCPCS Proc: Other therapeutic procedures	26	2	1	515
ICD10 Dx: 13.3-Spondylosis; intervertebral disc disorders; other back problems [205.]	27	2	1	515
CCI Category: Metastatic Solid Tumor	28	2	2	516
CPT and HCPCS Proc: Magnetic resonance imaging	29	2	3	512
CPT and HCPCS Proc: Ophthalmologic and otologic diagnosis and treatment	30	2	4	513
Last CRP	31	2	15	507
ICD10 Dx: 17.2-Factors influencing health care	32	2	2	513
NDC and HCPCS Rx: loop diuretics	33	2	1	515
NDC and HCPCS Rx: adrenergic bronchodilators	34	2	2	511
NDC and HCPCS Rx: viral vaccines	35	2	3	511
NDC and HCPCS Rx: skeletal muscle relaxants	36	2	2	509
ICD10 Dx: 3.5-Nutritional deficiencies [52.]	37	1	2	509
Month of Index: September	38	1	3	509
Number of ER Visits	39	1	12	507
Methotrexate Use	40	1	2	509
NDC and HCPCS Rx: antimalarial quinolines	41	1	1	510
NDC and HCPCS Rx: narcotic analgesic combinations	42	1	1	508

Supplemental Table 1 (continued). ***All predictors of severe COVID-19 outcome from XGBoost model.***

Feature	Rank	Weight	Gain	Cover
Number of Office or Clinic Visits	43	1	23	505
ICD10 Dx: 4.1-Anemia	44	1	4	505
Month of Index: July	45	1	2	505
CCI Category: Renal Disease	46	1	1	506
Month of Index: May	47	1	1	505
ICD10 Dx: 3.6-Disorders of lipid metabolism [53.]	48	1	2	505
CCI Category: Mild Liver Disease	49	1	2	505
ICD10 Dx: 6.8-Ear conditions	50	1	2	504
CCI Category: Dementia	51	1	1	503
ICD10 Dx: 2.16-Benign neoplasms	52	1	1	504
NDC and HCPCS Rx: omeprazole	53	1	1	502
CPT and HCPCS Proc: Arthrocentesis	54	1	1	502
ICD10 Dx: 7.4-Diseases of arteries; arterioles; and capillaries	55	1	1	502
ICD10 Dx: 13.2-Non-traumatic joint disorders	56	1	1	502
CPT and HCPCS Proc: Electrocardiogram	57	1	1	500
ICD10 Dx: 13.9-Other bone disease and musculoskeletal deformities [212.]	58	1	1	500
ICD10 Dx: 9.12-Other gastrointestinal disorders [155.]	59	1	1	498
Region: West	60	0	1	499